STAT 311: Homework 1
Due: Jul 1

Name:

The material covered include chapter 2 and the first part of chapter 3 in Mind on Statistics and Lectures 2 and 3. Questions 1-4 focus on lecture 2, Questions 5-7 focus on lecture 3. In general, rounding to 2 digits is sufficient.

1 Statistical Terminology

A 2016 study in the Journal of the American Medical Association Oncology examined 2400 women between the ages of 27 and 70 with early stage cancer. In their study, the found that the average women fasted 12.5 hours each night. They also found that "nighttime fasting period of less than 13 hours was linked to a 36% higher risk for a cancer recurrence."[1]

1. What are the observational units in this study?

   Each of the individual women

2. What might be the population of interest?

   Could be all women or all people; most likely women ages 27-70. In general, the population of interest depends on who exactly the researcher is interested in generalizing the results to. If we only have data on women, is it reasonable to generalize to men as well? Whenever you read a study, be careful to consider who was in the sample, and who it might be reasonable to generalize.

3. What is a parameter of interest? State in words. There may be multiple parameters, but you only need to list one.

   These are the quantities of interest in the population. We don't know what the exact values are, but we can describe them in words.

   - *Average number of hours a night an individual in the population fasts*
   - *Average change in cancer risk in the population when fasting less than 13 hours a night*

4. What is the sample?

   The sample are the observational units on which we have data. In this case, the 2400 women between 27 and 70 which were examined

5. What is the statistic? State in words and also give the numeric answer. Again, there may be multiple statistics, but only list the one which corresponds to the parameter you listed above.

   This is a quantity that describes our sample. In this case, we know that the sample

   - *Fasted an average of 12.5 hrs a night*

[1] More about the study can be found at `http://www.seattletimes.com/nation-world/study-eating-late-at-night-may-increase-cancer-risk/`

- *Had a 36% increase in cancer risk when fasting less than 13 hrs a night*

2 Descriptive Statistics

Below is a table of the 7 tight ends currently on the Seahawks roster and their official weight.

Name	weight (lb)	weight (kg)
Cottom, Brandon	262	119
Graham, Jimmy	265	120
Helfet, Cooper	239	109
Shields, Ronnie	245	111
Vannett, Nick	257	117
Williams, Brandon	247	112
Willson, Luke	252	115

Table 1: Data from www.seahawks.com/team/roster

1. Calculate the five number summary

   First, sort the weights:

   Min: 239

   Q1: 245

   Median: 252

   Q3: 262

   Max: 265 *239, 245, 247, 252, 257, 262, 265*

   Min: 239

   Q1: 245

   Median: 252

   Q3: 262

   Max: 265

2. *Calculate the mean*

   252.43

3. *Suppose tomorrow the Seahawks sign Baylor undrafted rookie LaQuan McGowan. LaQuan is listed at 410 lbs. What would the new mean be? What would the new median be?*

   median- 254.5

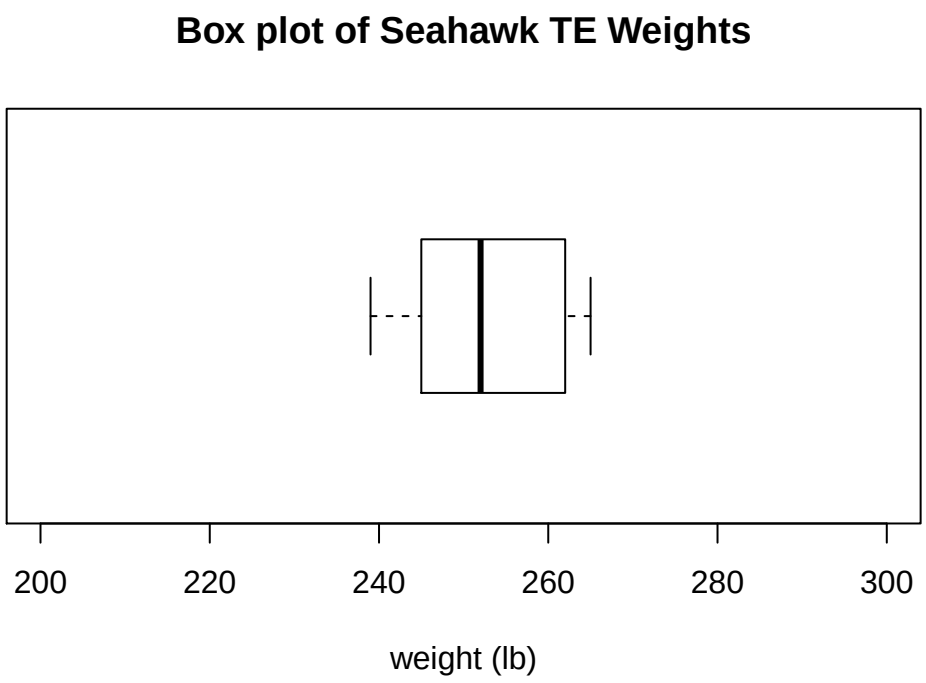
   In particular, note that the median changed much more than the mean. As we discussed in class, the median is less affected by outliers than the mean. mean- 272.13

   median- 254.5

   In particular, note that the median changed much more than the mean. As we discussed in class, the median is less affected by outliers than the mean.

3 Visualizing Data

1. Draw a boxplot for the weights of the Seahawk TE's (do not include LaQuan McGowan).



4 Changing the units of measure

(a) The standard deviation of the Seahawk Tight End weights (lb) is: 9.45 and the variance (which is just standard deviation squared) is 89.29. Now, let's see what changes as we work in kg instead of lb. The mean weight in kg is 114.71. Now calculate the standard deviation of the weights in kg. Remember that the standard deviation is calculated as

$$s_x = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2} \tag{1}$$

where n is the number of observations and $\bar{x}$ is the mean. You may find using the table below helpful, but it is not necessary.

Name	pounds	kg	x - x.bar	(x - x.bar)^2
Cottom, Brandon	262	119	4.29	18.37
Graham, Jimmy	265	120	5.29	27.94
Helfet, Cooper	239	109	-5.71	32.65
Shields, Ronnie	245	111	-3.71	13.80
Vannett, Nick	257	117	2.29	5.22
Williams, Brandon	247	112	-2.71	7.37
Willson, Luke	252	115	0.29	0.08

Summing the squared deviations (the right most column), gives us 105.43, so to calculate the standard deviation, we get

$$\sqrt{\frac{1}{7-1}105.43} = 4.19$$

(b) Compare the new mean and standard deviation (of kg) to the old mean and standard deviation (of lb). Considering that a 1 kg = 2.2 lb, how did the quantities change? Given the standard deviation that you calculated above, if we compare the variances instead (just square the standard deviation), how do those compare?

If we divide the standard deviation of the lbs by 2.2, that gives us the standard deviation in kg. If we divide the variance of the lbs by 2.2^2, that gives us the variance in kg. In general, multiply a data set by some constant results in the standard deviation being multiplied by the same constant.

5 Review on lines

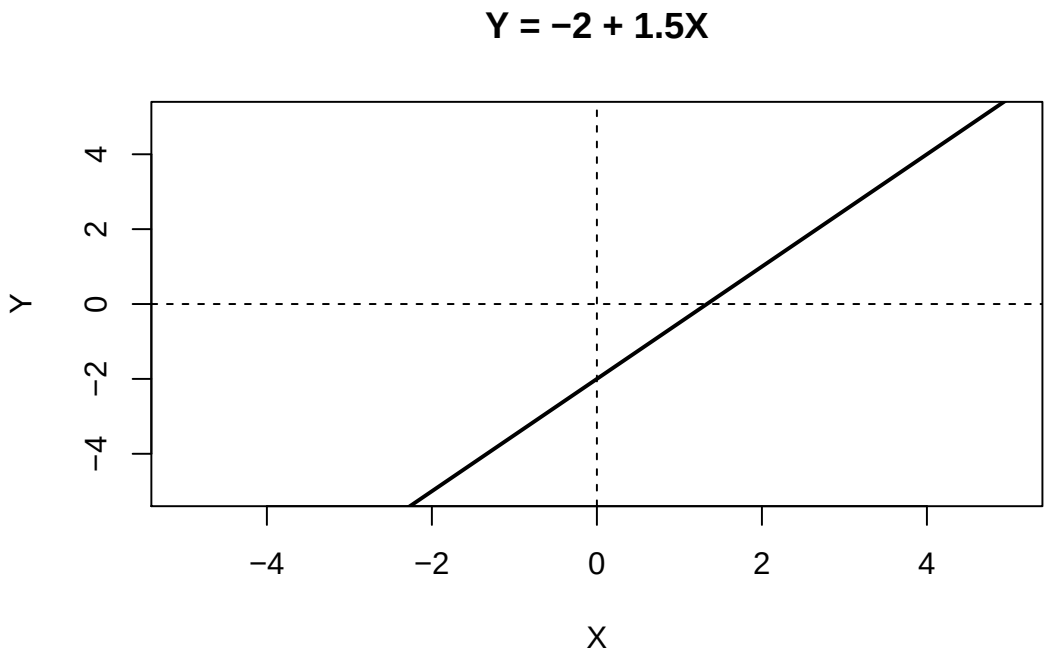
Let's review a bit of geometry first. Recall that lines can be defined by an equation in the form

$$Y = \mathrm{a} + \mathrm{b} \times X \tag{2}$$

where a represents the y-intercept and b represents the slope. The line crosses the y-axis at the y-intercept, and the slope represents the "rise-over-run."

(a) Plot the line

$$Y = -2 + 1.5 \times X \tag{3}$$



(b) What is the value of Y on the line when $X = 4$?

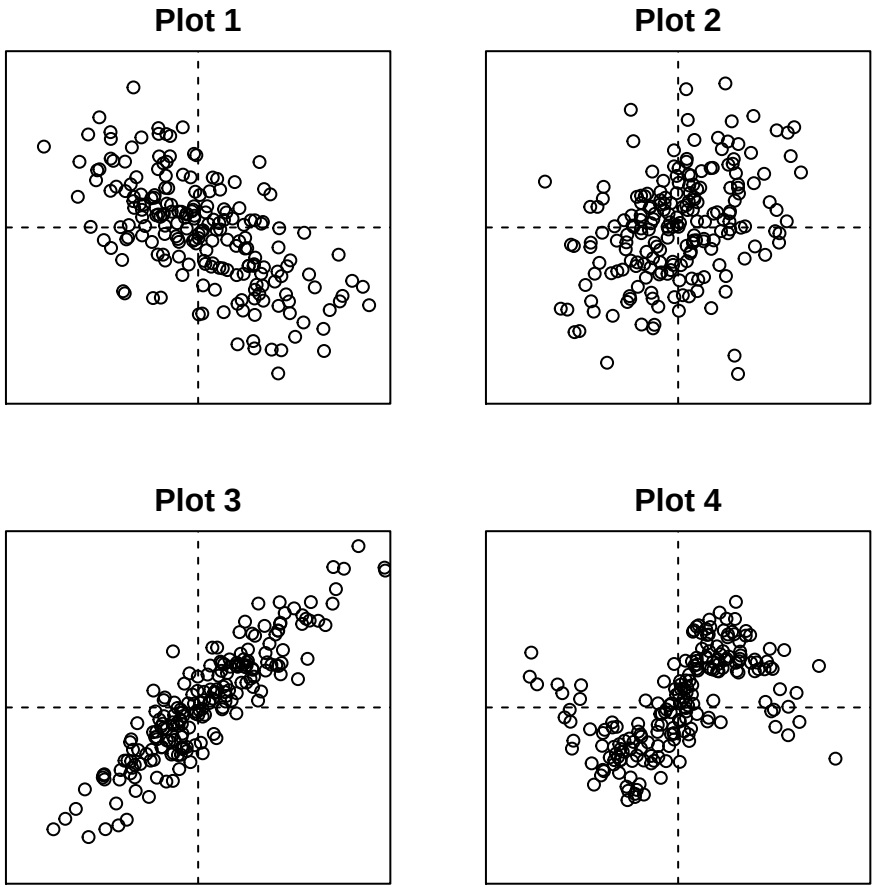
$$-2 + 1.5 \times 4 = 4$$

(c) What is the X value on the line when $Y = 1$?

$$-2 + 1.5 \times X = 1 \Rightarrow 1.5 \times X = 3 \Rightarrow X = 2 \tag{4}$$

6 Thinking about Correlation

Consider the plots below



- Which plot has a correlation of .9?

  Plot 3. It has the clearest linear pattern

- Which plot has a correlation of 0?

  Plot 4. It has a clear pattern, but it is non-linear and is increasing in parts and decreasing in other parts

- which plot has a correlation of -.6?

  Plot 1. It is the only plot with a clear negative association

- which plot has a correaltion of .4?

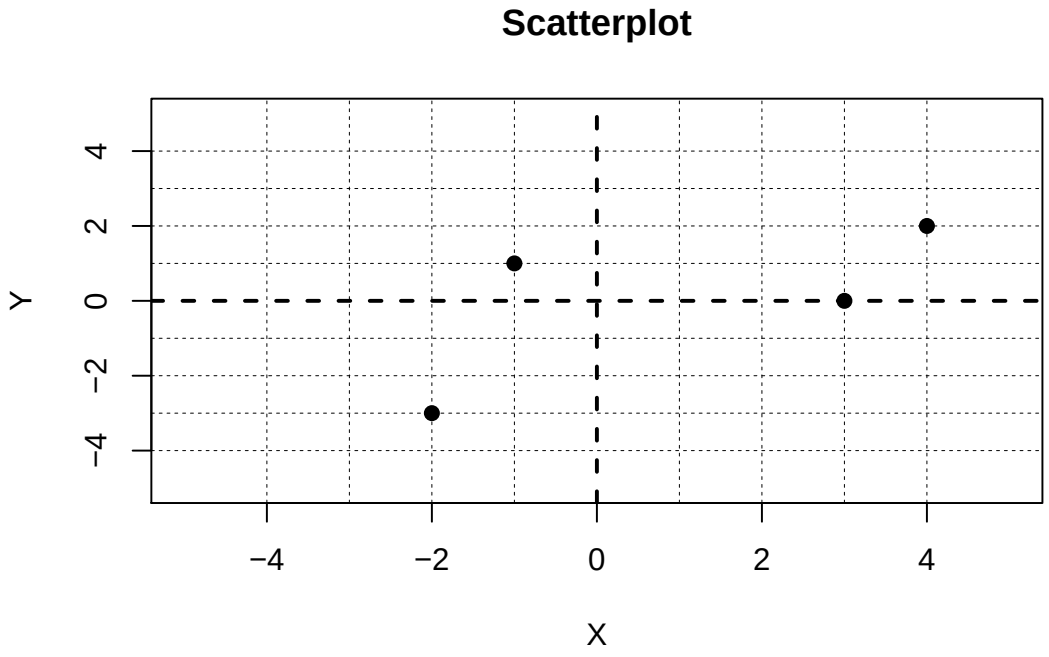
  Plot 2. It has a positive association, but is not as strong as plot 3

7 Calculating covariance

Consider the data below-

X	Y
-2	-3
4	2
3	0
-1	1

(a) Plot a scatterplot corresponding to the data



(b) Calculate the covariance of the data set by filling out the table below

	X	Y	(X - X.bar)	(Y - Y.bar)	(X - X.bar) x (Y - Y.bar)
Obs 1	-2.00	-3.00	-3.00	-3.00	9.00
Obs 2	4.00	2.00	3.00	2.00	6.00
Obs 3	3.00	0.00	2.00	0.00	0.00
Obs 4	-1.00	1.00	-2.00	1.00	-2.00
Sum	4.00	0.00	0.00	0.00	13.00

The covariance is

$$\frac{1}{4-1}13 = 4.33$$