# MST 109 - Midterm 1 - Fall 2017

Elementary Probability and Statistics

September 28, 2017

Instruction:

- You have 50 minutes to finish this exam.
- No calculators, phones or laptops are allowed during this exam.
- It is expected that each student during this exam will conduct himself, herself or themselves within the guidelines of the WFU Honor Code. All academic work should be done with the high level of honesty and integrity that the university demands.

**Name:** _____

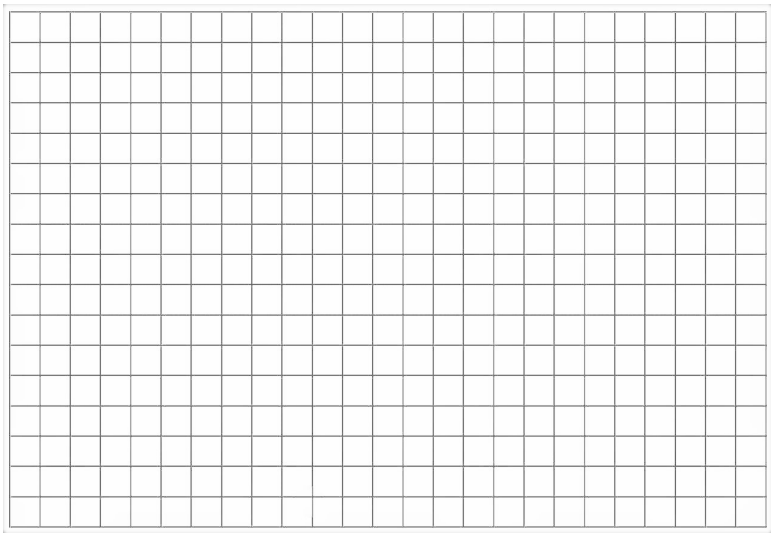
**Section:** Section G (1:00 pm- 1:50 pm) Section H ( 2:00pm - 2:50 pm)

I) The Aleppo pine and the Torrey pine are widely planted as ornamental trees in Southern California. Here are the length (in centimeters) of 15 Aleppo pine needles.

10.2 7.2 7.6 9.3 12.1 10.5 9.4 11.3 8.5 8.5 12.8 8.7 9.0 9.0 9.4

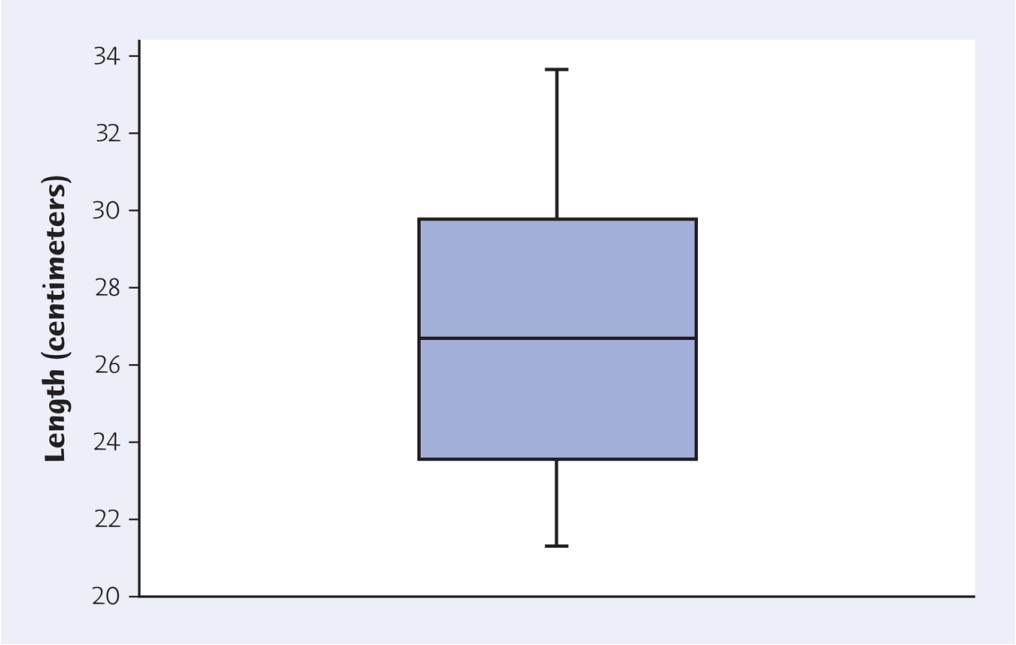
1. Find the five number summary.

2. Draw a boxplot of the representing the data. (Hint see Fig 1 for an example of a boxplot)



3. Use the IQR method to find if this data set has any outliers.

Fig 1 below represents a boxplot for the distribution of the length (in centimeters) of 18 Torrey pine needles. Use this information to help answer the remainder of this question.




**Box plot for the distribution of the length of 18 Torrey pine needles**
**Fig 1, The Basic Practice of Statistics, © 2015 W. H. Freeman**

4. The median of the distribution of Torrey pine needles is closet to which of the following values?

   24 25 27 30

5. Twenty five percent of the Torrey pine needles exceed what value?

6. Given only the length of a needle, do you think you could say which pine species it comes from? Explain briefly.

II) The given data represents the waiting times in minutes at your doctor's office.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 10.4 | 10.9 | 12.5 | 17.3 | 12.0 | 12.4 | 14.3 | 11.4 |
| 18.7 | 11.4 | 10.4 | 11.2 | 15.9 | 10.7 | 16.4 | 20.3 |
| 11.8 | 10.2 | 11.4 | 19.9 | 12.0 | 13.9 | 10.6 | 20.0 |
| 17.5 | 13.0 | 13.9 | 14.2 | 11.3 | 12.7 | 11.2 | 11.6 |

1. Fill out the following table by the number of occurrences (frequency) that falls in each of the given intervals.

| | frequency |
|---|---|
| $10 \le t \le 10.9$ | |
| $11 \le t \le 11.9$ | |
| $12 \le t \le 12.9$ | |
| $13 \le t \le 13.9$ | |
| $14 \le t \le 14.9$ | |
| $15 \le t \le 15.9$ | |
| $16 \le t \le 16.9$ | |
| $17 \le t \le 17.9$ | |
| $18 \le t \le 18.9$ | |
| $19 \le t \le 19.9$ | |
| $20 \le t \le 20.9$ | |

2. What does the standard deviation represent?

3. The mean of the given data is 13.48. By looking at the data and knowing the mean, choose the standard deviation s.d. from the following. (You don't need to calculate the s.d. using the formula)

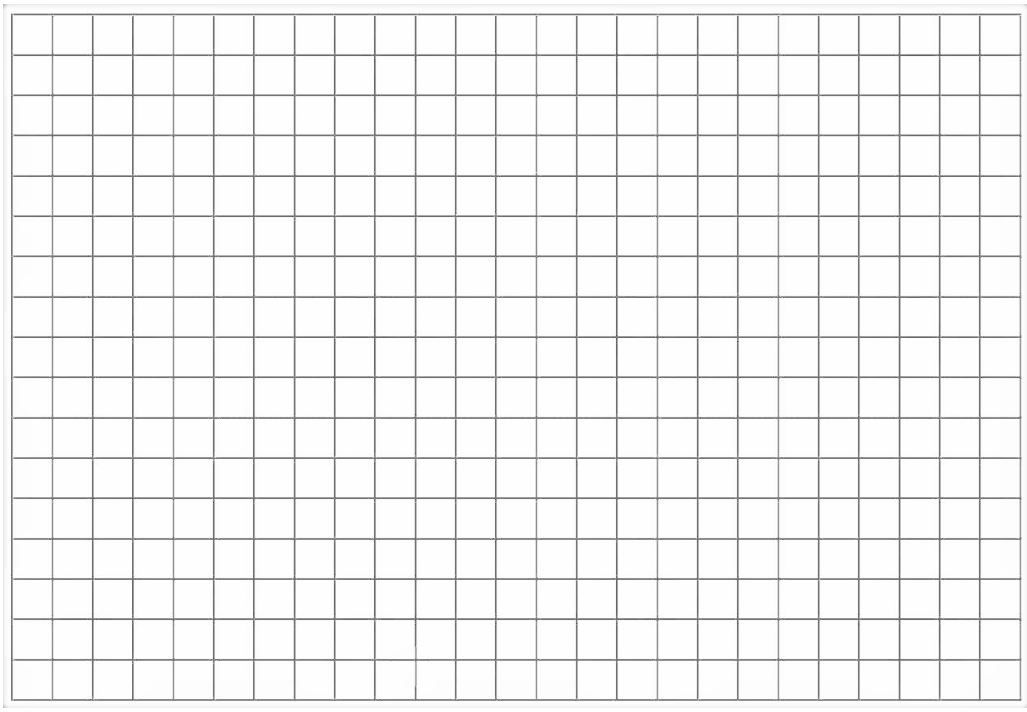
   (i) -3.5

   (ii) 9

   (iii) 0.6

   (iv) 3.05

4. Draw a histogram representing the data. (Use the intervals from the table)



III) Suppose you take 100 measurements on the speed of cars on Interstate, and that these measurements follow roughly a Normal distribution with a mean $\mu = 65\,mph$ (miles per hour) and standard deviation $\sigma = 10\,mph$

1. Using the z-table find the percentage of cars that is driving at a speed under $70\,mph$.

2. What is the median of this distribution?

3. If a car had a standardized speed of 2. What does this score tell us about the speed of the car in relation to the other cars on interstate. Draw a graph of the distribution, mark and shade the appropriate parts.

IV) Consider the following two way table

| | Sport Utility Vehicle (SUV) | Sport Car | Totals |
|---|---|---|---|
| Male | 44 | 56 | 100 |
| Female | 80 | 20 | 100 |
| Total | 124 | 76 | 200 |

1. What is the percentage of females and the percentage of males in this survey?

2. What is the percentage of SUV's and the percentage of Sports Cars in this survey?

3. What percentage of females prefer to have an SUV over a Sports Car?

4. What percentage of males prefer to have a Sports Car over an SUV?

V) Regression lines and Correlation. In the study of correlation between two variables $x$ and $y$ for $n$ data points such that the mean of the $x-$data ( similarly the $y-$data) is given by $\overline{x}$ (similarly $\overline{y}$) and

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \overline{x}}{s_x} \right) \left( \frac{y_i - \overline{y}}{s_y} \right)$$

and the least-squares regression line is given by

$$\hat{y} = r \, \frac{s_y}{s_x} \, x + \overline{y} - r \, \overline{x} \, \frac{s_y}{s_x}.$$

Knowing this information please answer these independent questions.

1. What kind of variables do you need to generate a scatter plot?

2. If we are studying the effect of sleeping (in hours) on the amount of productive work you have during the day (in hours).

   (a) Which variable is the response variable and which is the explanatory variable?

   (b) On which axis does each go when drawing a scatter plot?

   (c) If the correlation between the two variables described in part (b) is 0.7, what would be the correlation if we did all our measurements in minutes instead of hours?

3. If the regression line had an equation $y = -3x + 12$, and $r^2 = 0.64$. What is the correlation $r$.

VI) A psychologist wants to know if the difficulty of a task influences our estimate of how long we spend working at it. She designs two sets of mazes that subjects can work through on a computer. One set has easy mazes, and the other had hard mazes.

Subjects work until told to stop (after six minutes, but subjects didn't know this). They are then asked to estimate how long they worked. The psychologist has 30 students available to serve as subjects.

1. Describe the design of a completely randomized experiment to learn the effect of the difficulty on estimated time.

2. Describe the matched pairs design experiment using the same 30 subjects.

VII) Choose the best answer.

1. Which one of the following is a FALSE statement about density curves?
   (a) Always on or above the x-axis.
   (b) Area under the curve within an interval is the proportion of values expected in that interval.
   (c) Total area under the curve depends on the shape of the curve.
   (d) The curve is an idealized depiction of the distribution of a variable.

2. Which one of the following is a FALSE statement about the normal distribution?
   (a) The mean is greater than the median.
   (b) It is symmetric.
   (c) It is bell-shaped.
   (d) It has one peak.

3. If y is the response variable and x is the explanatory variable, what does the regression line allow you to do that correlation does not?
   (a) calculate the exact value of x given a value for y
   (b) calculate the predicted value of x given a value for y
   (c) calculate the exact value of y given a value for x
   (d) calculate the predicted value of y given a value for x

4. When two variables are actually not related to each other, even though they may have a very high correlation because they both result from some other, possibly hidden factor, this is an example of
   (a) an outlier.
   (b) a lurking variable.
   (c) extrapolation.
   (d) None of the above

5. In the Salk vaccine trial of 1954, almost 400,000 students (grades 13) in 11 states participated. Students were randomly assigned to either a vaccine or placebo injection. All students were observed for evidence of polio during the school year.

i- What is the factor in the Salk vaccine experiment?

(a) type of injection
(b) vaccine
(c) placebo
(d) polio status

ii- What are the treatments in the Salk vaccine experiment?

(a) syringe, school nurse
(b) vaccine, placebo
(c) polio, vaccine
(d) polio status

iii- What is the response variable in the Salk vaccine experiment?

(a) type of inoculation
(b) polio, vaccine
(c) polio status
(d) vaccine, placebo