Exam I

STAT 305, Section D
FALL 2019

Instructions

- The exam is scheduled for 80 minutes, from 8:00 to 9:20 AM. At 9:20 AM the exam will end.
- A forumula sheet is attached to the end of the exam. Feel free to tear it off.
- You may use a calculator during this exam.
- Answer the questions in the space provided. If you run out of room, continue on the back of the page.
- If you have any questions about, or need clarification on the meaning of an item on this exam, please ask your instructor. No other form of external help is permitted attempting to receive help or provide help to others will be considered cheating.
- **Do not cheat on this exam.** Academic integrity demands an honest and fair testing environment. Cheating will not be tolerated and will result in an immediate score of 0 on the exam and an incident report will be submitted to the dean's office.

Name: _____

Student ID: _____

1. (2 points) Circle the **bold face** term that makes the following statement true:

   A measurement device that reports the true measurement of the item on which the device is being used is (**precise** or **accurate**).

2. A sample of size 5 was drawn from a population and the resulting observations are reported below.

   12, 15, 18, 19, 26

   Using these observed values, report the following:

   (a) (2 points) the mean

   Solution:

$$\begin{aligned}\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i &= \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) \\ &= \frac{1}{5}(12 + 15 + 18 + 19 + 26) \\ &= \frac{1}{5}(90) \\ &= 18\end{aligned}$$

   (b) (2 points) the median

   Solution: We will need to use the quantile function.

   In this case, $i = \lfloor np + 0.5 \rfloor = \lfloor 5 \cdot 0.25 + 0.5 \rfloor = \lfloor 1.75 \rfloor = 1$.

$$\begin{aligned}Q(.50) &= x_i + (np + 0.5 - i) \cdot (x_{i+1} - x_i) \\ &= x_3 + (5 \cdot 0.50 + 0.5 - 3) \cdot (x_4 - x_3) \\ &= 18 + (0) \cdot (19 - 18) \\ &= 18 + (0) \cdot (1) \\ &= 18 + 0 \\ &= 18\end{aligned}$$

   (c) (2 points) the variance

   Solution: Since this is a sample, we must s^2:

$$
\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \\
&= \frac{1}{5-1}\sum_{i=1}^{5}(x_i-\bar{x})^2 \\
&= \frac{1}{4}\left((x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+(x_4-\bar{x})^2+(x_5-\bar{x})^2\right) \\
&= \frac{1}{4}\left((12-18)^2+(15-18)^2+(18-18)^2+(19-18)^2+(26-18)^2\right) \\
&= \frac{1}{4}\left((-6)^2+(-3)^2+(0)^2+(1)^2+(8)^2\right) \\
&= \frac{1}{4}\left(36+9+0+1+64\right) \\
&= \frac{1}{4}\left(110\right) \\
&= 27.5
\end{aligned}
$$

(d) (2 points) the standard deviation

Solution: We must use the sample standard deviation, s:

$$s = \sqrt{s^2} = \sqrt{27.5} = 5.2440442$$

(e) (2 points) the value of $Q(.25)$

Solution: We will need to use the quantile function.

$$
\begin{aligned}
Q(.25) &= x_i + (np + 0.5 - i)\cdot(x_{i+1} - x_i) \\
&= x_1 + (5\cdot 0.25 + 0.5 - 1)\cdot(x_2 - x_1) \\
&= 12 + (0.75)\cdot(15-12) \\
&= 12 + (0.75)\cdot(3) \\
&= 12 + 2.25 \\
&= 14.25
\end{aligned}
$$

(f) (2 points) the interquartile range

Solution: This is just $Q(.75) - Q(.25)$.

$$\begin{aligned} Q(.75) &= x_i + (np + 0.5 - i) \cdot (x_{i+1} - x_i) \\ &= x_4 + (5 \cdot 0.25 + 0.5 - 4) \cdot (x_5 - x_4) \\ &= 19 + (0.25) \cdot (26 - 19) \\ &= 19 + (0.25) \cdot (7) \\ &= 19 + 1.75 \\ &= 20.75 \end{aligned}$$

So the IQR is 6.5

3. An environmental engineer is testing four methods for reducing the concentration of a certain lake pollutant found in Iowa lakes. To do this he first randomly selected 20 Iowa lakes from which he took water samples, then split each of the 20 samples into 4 portions, and randomly labeled the four portions 1, 2, 3, and 4. Finally, he attempted to reduce the concentration of each of the portions labeled 1 using the the first method, of each of the portions labeled 2 using the second method, of each of the portions labeled 3 using the third method, and of each of the portions labeled portion 4 using the fourth method. After the methods had been applied, he measured the change in concentration.

   (a) (2 points) Is this an experiment or an observational study? Explain.

   Solution: This is an experiment. The engineer is taking an active role in manipulation the system under study by intentionally changing the cleaning method used.

   (b) Identify the following (if there was not one, simply put "not used").

      i. (2 points) Response variable(s):

      Solution: The change in concentration is the only response.

      ii. (2 points) Experimental variable(s):

      Solution: The method used to clean the portion is the only experimental variable.

      iii. (2 points) Blocking variable(s):

      Solution: The lakes from which the samples are taken are acting as a blocking variable. We are not interested in studying the effect of the lake on the response, but we can reasonably believe that the portions from the same lake's sample will be similar. So we are treating the lake the sample came from as a smaller, homogenous environment in our experiment. We also use all the methods on each lake's sample which is another indication that it is working as a block.

   (c) (2 points) Was replication used in this experiment? If so, where was it applied? If not, how could we have applied it?

   Solution: No replication was used. While each cleaning method was used multiple times across the entire experiment, they were never used in the same block (meaning, for each of the 20 lakes we only used each method once). This means that we did not truly replicate.

4. Aisha recently discovered she has the opportunity to upgrade her smart phone. She narrowed her choices down to two phones (we will call them phone A and phone B) but had a hard time making her final decision. She decided to interviewed people she knew who had one of the phones to rate their satisfaction from 0% to 100%. She also asked them if they would prefer to have the other phone. In order to help put their feelings in perspective, she also made note of how negative she thought they were in general (since critical people might be harsher in their criticism in general), using three descriptions: the interviewee's personality was classified as overly critical, appropriately critical, or not critical enough.

   (a) (2 points) Is this an experiment or an observational study?

   Solution: It is an observational study. Aisha only records information she can observer (even if she may be a biased observer, or her sample of people she interviews may be biased).

   (b) (2 points) What is the population under study?

   Solution: The population under study is people Aisha knows who have one of the two phones.

   (c) (2 points) Identify the response variable(s).

   Solution: There are multiple response variables. (1) The interviewee's satisfaction, (2) whether or not the interviewee would have preferred the other phone, (3) Aisha's rating of their negativity, and (4) which of the phones the individual owns (since that is information we are collecting from each sampling unit and it is information that will change between sampling units). Note: it could be argued that since this is not an experiment, there is no response variable at all and all the variables are just "variables of interest" - tread lightly with that response though, since it seems a lot like word play instead of an honest attempt at the problem...

   (d) For each of the following variables,

   - Identify whether it is qualitative or quantitative variable, and
   - If it is qualitative, what are the possible values it can take? If it is quantitative, is it continuous or discrete?

   i. the individual's reported phone satisfaction percentage.

   Solution: This is quantitative and continuous.

   ii. Aisha's appraisal of the interviewee's negativity.

   Solution: This is qualitative with three levels: overly critical, appropriately critical, not critical enough.

   iii. whether or not the interviewee would prefer to have the other phone.

   Solution: This is qualitative with two levels: yes or no.

   iv. the type of phone the interviewee currently owns.

   Solution: This is qualitative with two levels: Phone A or Phone B.

5. The strength of an internet connection is often described in terms of its download speed, measured in megabits per second (or Mbps). A systems administrator is concerned that recent changes in her company's main server framework may be having a negative impact on the local network's download speed. Every 2 minutes for an hour, she recorded the network speed at that moment and collected her results into the following stem-and-leaf plot:

```
The decimal point is at the |

 0 | 9
 1 | 8
 2 | 7
 3 | 6
 4 | 134
 5 | 7
 6 | 1145677
 7 | 01338
 8 | 2346
 9 | 79
10 | 45
11 |
12 | 17
```

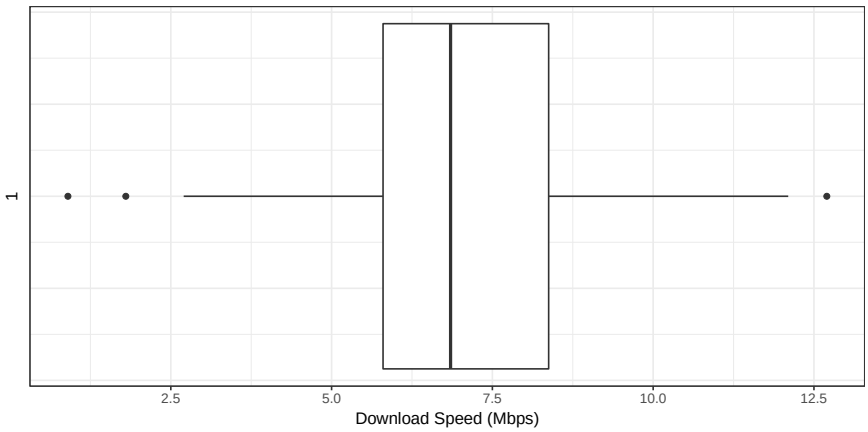
Note that `0 | 9` represents 0.9. In this case, the first quartile is $Q(.25) = 5.7$, the median is 6.85, and the IQR is 2.7.

(a) (10 points) Complete the following frequency table:

Value Range	**Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
0.00 - 2.00	2	0.07	0.07
2.01 - 4.00	2	0.07	0.14
4.01 - 6.00	4	0.13	0.27
6.01 - 8.00	12	0.4	0.67
8.01 - 10.00	6	0.2	0.87
10.01 - 12.00	2	0.07	0.94
12.01 - 14.00	2	0.07	1.01

(b) (10 points) Create a box plot to summarize the data. Carefully label the axes.

Solution: The boxplot is below:



Please note: the values on the y-axis are meaningless

(c) (4 points) Are there any unusually low or high observations? If so, what pressures caused those beams to fail?

Solution: Yes, there are two unusually low observations as indicated by the box plot. They had download speeds of at 0.9 Mbps and 1.8 Mbps. There is also one unusually high observations as indicated by the box plot. It had download speeds of 12.7 Mbps.

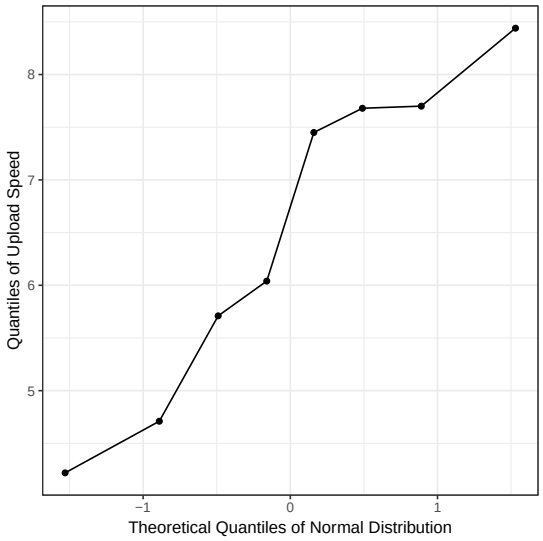
(d) (10 points) She also measured upload speed, obtaining the following 8 values.

7.45, 4.22, 7.7, 6.04, 7.68, 5.71, 4.71, 8.44

Create a theoretical Q-Q plot using the following quantiles from the normal distribution as the theoretical quantiles. Carefully label your axes. What does this graph tell us about the upload speeds?

	1	2	3	4	5	6
p	1/12	3/12	5/12	7/12	9/12	11/12
$Q(p)$	-1.53	-0.89	-0.49	-0.16	0.16	0.49

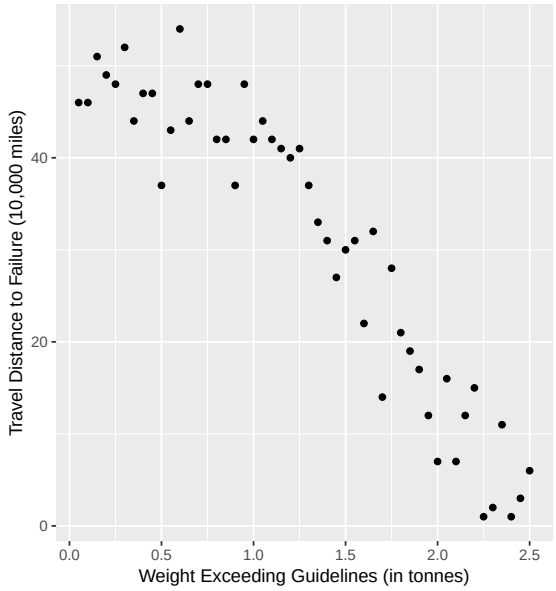
Solution: We get the QQ-plot by plotting the ordered values of our sample against the ordered quantiles from the normal distribution (as given above):



The points do seem to somewhat linear - an argument could be made that because of this the upload speed is normally distributed.

6. The major cause of axel failure in freight trucks is when shippers exceed the recommended weight limits that can be handled by the axels. Issues resulting from these failures have been becoming more frequent as shippers try to cut corners, leading members of the state's Department of Transportation to ask one of their civil engineers to look into the available data and better advise them on the relationship between excessive weight and axel failure.

   A company manufacturing axels provides the engineer with data gathered from conducting experiments loading axels with excessive weight and simulating traveling conditions. The data consists of two columns, **excessive weight (in tonnes)** is the amount of weight over the limit that was placed on the axel, and **distance to failure (in tens of thousands of miles)** is the simulated distance to the axel's failure.



   Here are some summaries of the data:

$$\sum_{i=1}^{50} x_i = 64 \qquad \sum_{i=1}^{50} x_i^2 = 107$$

$$\sum_{i=1}^{50} y_i = 1558 \qquad \sum_{i=1}^{50} y_i^2 = 61528$$

$$\sum_{i=1}^{50} x_i y_i = 1444$$

   (a) Using the summaries above, fit a linear relationship between **weight exceeding guidelines** (x) and **travel distance to failure** (y).

      i. (5 points) Write the equation of the fitted linear relationship.

Solution: The fitted line equation is

$$\hat{y} = b_0 + b_1 \cdot x$$

We can use the information above to get the value for b_1 and b_0:

$$\begin{aligned} b_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \\ &= \frac{(1444) - (50)(64/50)(1558/50)}{107 - 50(65/50)^2} \\ &= -24.4551111 \end{aligned}$$

and with b_1 we can find the value fo b_0:

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ &= (1558/50) - (-24.4551111)(64/50) \\ &= 62.4625422 \end{aligned}$$

Which gives us the fitted equation of

$$\hat{y} = 62.47 - 24.46 \cdot x$$

ii. (5 points) Find and interpret the value of R^2 for the fitted linear relationship.

Solution: Since we are using a linear relationship, we can get R^2 from r:

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)\,(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}} \\ &= \frac{(1444) - (50)(64/50)(1558/50)}{\sqrt{(107 - 50(64/50)^2)\,(61528 - (50)(1558/50)^2)}} \\ &= -0.9643596 \end{aligned}$$

So $R^2 = (r)^2 = 0.9299894$
This means that 93.00% of our the variablity in travel distance to failure can be explained by the linear relationship with weight exceeding guidelines.

iii. (5 points) Using the fitted line, provide a predicted value of travel distance to failure when the weight exceeding the guidelines is 3.4 tonnes.

Solution: $\hat{y} = 62.47 - 24.46(3.4) = -20.694$

iv. (5 points) Sketch what you believe the plot of residuals vs weight would look like. Why would this be a problem?

Solution: I'm not doing this all the way (sorry not sorry!). The residual plot should have a roughly parabolic shape, with negative residuals at the start, postive residuals through the peak of the arch, and negative residuals at the end again. This is a problem because the form of our fitted relationship does not actually match the real form of the relationship seen on our data.

(b) The JMP output below comes from fitting a quadratic model using x and x^2.

Response Distance to Failure

Summary of Fit

RSquare	REDACTED
RSquare Adj	REDACTED
Root Mean Square Error	5.281589
Mean of Response	0.16
Observations (or Sum Wgts)	50

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	13229.647	6614.82	237.1314
Error	47	1311.073	27.90	Prob > F
C. Total	49	14540.720		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	16.27602	2.333507	6.97	<.0001*
Weight Exceeding Limit	4.6604349	4.221593	1.10	0.2752
(Weight Exceeding Limit)^2	-10.2775	1.604983	-6.40	<.0001*

i. (5 points) Write the equation of the fitted quadratic relationship.

Solution:

$$\hat{y} = 16.27602 + 4.6604349x - 10.2775x^2$$

ii. (5 points) Find and interpret the value of R^2 for the fitted quadratic relationship.

Solution:

$$R^2 = 1 - SSE/SSTO = 1 - (1311.073/14540.720) = 0.9098344$$

In other words, 90.98% of the variability in travel distance to failure can be explained by the linear relationship with weight exceeding guidelines.

iii. (5 points) Using the fitted quadratic relationship, provide a predicted value of travel distance to failure when the weight exceeding the guidelines is 3.4 tonnes.

Solution:

$$\hat{y} = 16.27602 + 4.6604349(3.4) - 10.2775(3.4)^2 = -86.6864013$$