ST 411/511 - Methods of Data Analysis I

Midterm Exam

Chimdi Chikezie

Due: Wednesday, February 9th before 11:59PM

Directions:

Please read the directions below *very* carefully as failure to follow them will result in point deductions, receiving a zero on the exam, and/or having your name forwarded to the Office of Student Affairs.

- Submissions must be made as entirely typed PDFs to Gradescope prior to the deadline. Exams which are handwritten or contain images, scans, etc. of handwritten work will not be accepted/graded.
- No late submissions are allowed for any reason unless you have prior approval from the instructor. See the course syllabus for more details.
  - Failing to "knit" your document to PDF is not a valid reason to submit your exam after the deadline. Start the exam early and knit your document frequently.
  - Failing to upload your document to Gradescope is not a valid reason to submit your exam after the deadline. Verify your exam has been successfully submitted and allow yourself time to "troubleshoot" any potential issues.
  - Failing to upload the correct document to Gradescope is not a valid reason to submit your exam after the deadline. Verify that you've uploaded the correct document prior to the deadline.
- Exams should be "knitted" directly to PDF using the provided template.
  - Knitting your document to a Word Doc file and then converting it to PDF is not desirable but I'll live if you do this.
  - Do not knit your document to an HTML file and then convert it to PDF. The formatting is terrible and it makes them difficult to read. Exams where this occurs will be penalized.
- Indicate all pages containing solutions to the individual questions in Gradescope. Exams which are not properly indicated will be penalized on a per-question basis.
- Do not include extraneous code/outputs (i.e. "Scratch work"). Exams which include extraneous code/outputs will be penalized.
- Answer questions using complete sentences. Solutions which are not written as complete sentences or are difficult to read and/or understand will be penalized.
- Exams which are messy, hard to read, unclear, disorganized, etc. will be penalized.
- The exam is open book and open note. You can utilize any material found in the text or the course Canvas page but nothing else.
- Do not talk or seek assistance from anyone else (students, peers, other faculty, online message boards, etc.).
  - Exams where "copying" or student collaboration is suspected will receive grades of zero until students arrange to meet with the instructor to explain the similarities.
- Do not share code or solutions with anyone else.

Question 1 (10 points)

For each part below, indicate whether the statement is TRUE or FALSE and provide a 1-2 sentence explanation as to why.

a. (2 points) If a 95% confidence interval for an unknown population mean is (21.4, 78.7), this means there is a 95% *probability* that the true population mean is between 21.4 and 78.7.

- Answer: FALSE
- Reason: Confidence intervals is not probability. It gives us 95% confidence that the true population mean will fall between 21.4 and 78.7.

b. (2 points) After conducting a Z-test, the corresponding p-value of the test is 0.0456. This means, precisely, that the probability that the null hypothesis is true is 0.0456.

- Answer: FALSE
- Reason: P-value does not determine if the null hypothesis is true, however, it determines the probability of getting a result at least as extreme as the statistic observed..

c. (2 points) In a study on "handedness" in households with exactly two children, it was found that older siblings are much more likely to be left-handed than younger siblings (A statistically significant result). The researchers should conclude that being born first *causes* the older children (at least in some cases) to be left-handed.

- Answer: FALSE
- Reason: This is an observational study and we cannot make a causal conclusion. There might be other factors that make those children to be left handed and not only being a first born.

d. (2 points) The Los Angeles Dodgers, a major league baseball team, played 81 home games against 17 other teams. For these games, the Dodgers played the same team over the course of two to four consecutive nights. Since these were different games, i.e. they always started with a score of 0-0, the games should be considered to be *independent* of one another.

- Answer: TRUE
- Reason: Because the winning probability of all the games are the same for all games.

e. (2 points) A professional DK64 speedrunner, and current 101% record holder, wishes to purchase 14 different candies for their significant other on Valentine's Day. They visit two different large grocery stores and record the prices of the same 14 candies. If they want to see if the average price of the 14 candies are different between the two stores then they should conduct an equal variance or a Welch's two-sample t-test.

- Answer: Welch's two-sample t-test
- Reason: Welch's two-sample t-test is used to test that means are the same when their variances are significantly different.

Question 2 (13 points)

Dairy scientists are interested in researching the ascorbic acid content of dairy cows raised in Benton County. Specifically, they would like to know if the population mean ascorbic acid content (measured in mg. per cc), is different than 220 mg/cc. To conduct their study, they gather a random sample from six different cows raised in Benton County, the values or which are below (You can *knit* the exam to see the printed table):

255	190	55	137	138	174

a. (2 points) Based on the researcher's question of interest, write out the appropriate null and alternative hypotheses for a statistical hypothesis test using correct statistical notation.

$H_0 : \mu_p = 220mg/cc \backslash\ H_A : \mu_P \neq 220mg/cc$

b. (4 points) Conduct an appropriate statistical hypothesis test, by hand (i.e. you can use R code or mathematical notation, but do not use any built-in test functions like `t.test()`, `wilcox.test()`, `binom.test()`, etc.), to answer the researcher's question of interest at the 5% significance level. Specifically, write out the value of the test statistic, the p-value, and a conclusion in the spaces provided below. For the conclusion, be sure to state the statistical outcome of the test, why you arrived at that conclusion, and what this conclusion means for the researchers using the context of the question. Please show all relevant work in the space indicated/provided below.

- Test statistic: (xbar - mu) / sqrt(s/n) = (158.167 - 220) / sqrt(4427.767/6) = -2.276164
- p-value: 2 * (1 - pt(abs(150.0685), 6 - 1)) = 0.071882
- Conclusion: I fail to reject the null hypothesis because the P-value is greater than alpha. We do have have enough evidence to show that the population mean ascorbic acid content (measured in mg. per cc), is equal to 220 mg/cc.

Show your work for this question below. If you write R code, please make sure the computed test statistic and p-value are shown as outputs (And please don't leave a bunch of extraneous code/outputs in your solutions).

```
val <- c(255, 190, 55, 137, 138, 174)

mu <- 220
xbar <- 158.167
s <- 4427.767
n <- 6

t <- (xbar - mu) / sqrt(s/n)
sprintf("t = %f", t)
```

```
## [1] "t = -2.276164"
```

```
pvalue <- 2 * (1 - pt(abs(t), n - 1))
sprintf("pvalue = %f", pvalue)
```

```
## [1] "pvalue = 0.071882"
```

c. (3 points) Compute, by hand, and interpret a 95% confidence interval for the population mean. Write your confidence interval in "interval form", i.e. *(lower bound, upperbound)*, and interpret the interval in the context of this particular question. Show your work in the space indicated below.

- 95% confidence interval: (88.33601, 227.998)

CI upper_limit <- x +- (error * sqrt(s/n))

error = 1 - (0.05/2), 6 - 1 = 0.975,5 qt(0.975,5) = 2.570582

CI lower bound = 158.166667 - (2.571 * sqrt(4427.767/6)) = 88.33601 CI upper bound = 158.166667 + (2.571 * sqrt(4427.767/6)) = 227.998

- Interpretation:

We are 95% confident that the true population mean ascorbic acid content (measured in mg. per cc) of dairy cows raised in Benton County is between 88.33601 and 227.998

Show your work for this question below. If you write R code, please make sure the computed values are shown as outputs (And please don't leave a bunch of extraneous code/outputs in your solutions).

```
cv <- 1 - (0.05/2)
error <- qt(cv,5)

lowerbound <- xbar - (error * sqrt(s/n))
sprintf("lowerbound = %f", lowerbound)
```

```
## [1] "lowerbound = 88.336012"
```

```
upperbound <- xbar + (error * sqrt(s/n))
sprintf("upperbound = %f", upperbound)
```

```
## [1] "upperbound = 227.997988"
```

d. (2 points) In discussing the outcome of your hypothesis test in part b. with the researchers, they inform you that they believe that the population distribution is likely to be heavily (and I mean *heavily*) right-skewed. Why does this new knowledge about the population distribution make you doubt the validity of the test you conducted in part b.? Be specific and be sure to explain the underlying reasoning in your explanation.

If we have an outlier, it can make the distribution rightly skewed, then the mean value will be changed based on those outliers and we will not be able to get a good estimation of our distribution from the mean the value. T-test depends on mean value and when the mean value is not good enough, it looses its validity. This is the reason that makes me doubt the validity of the test you conducted in part b.

e. (2 points) Is this study an example of a randomized experiment or an observational study? How can you tell?

It is an observational study because the researcher is not influencing the ascorbic acid content.

Question 3 (12 points)

A European professional basketball league is trying to decide whether a new workout supplement should be banned for also being a "performance enhancing" drug. To test the workout supplement, they measure how many free-throws a person successfully makes while shooting a basketball through a standard, regulation height hoop in 4 minutes. For each of the 11 randomly selected individuals, the league measures the number of successful free-throws the person makes while *unmedicated* (before taking the new supplement) and again while *medicated* (after taking the drug). If more free throws are made while *medicated* than *unmedicated* then the league will ban the supplement. A variety of summaries are given below (Note: "(Medicated - Unmedicated)" denotes the pairwise differences for each observation):

	Unmedicated	Medicated	(Medicated - Unmedicated)
n	11	11	11
$\bar{X}$	20.73	23.55	2.82
s^2	21.42	29.27	10.56

a. (2 points) Using correct statistical notation, write out the null and alternative hypotheses needed to conduct a t-test which will help the league decide whether or not they should ban the supplement.

$H_0 : \mu_M - \mu_U = 0$ $H_A : \mu_M - \mu_U > 0$

The R output below shows some of the output from two different t-tests, one of which is the equal variance (independent) two-sample t-test and the other is for a paired t-test.

```
    Two Sample t-test

data:  Medicated and Unmedicated
t = 1.3128, df = 20, p-value = 0.1021
sample estimates:
mean of Medicated mean of Unmedicated
         23.54545            20.72727


    Paired t-test

data:  Medicated and Unmedicated
t = 2.8758, df = 10, p-value = 0.008252
sample estimates:
mean of the differences
               2.818182
```

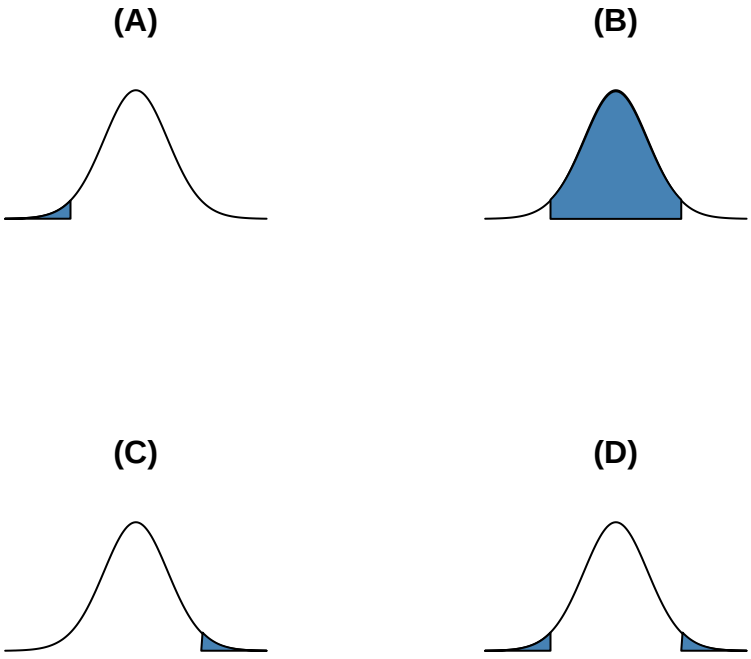
b. (2 points) Which one of the above tests is the most appropriate to use to answer the question of interest? Explain why?

Paired t-test. This is because the European professional basketball league is created 2 pairs: 1. Before testing 2. After testing

c. (2 points) Using the test output you chose in part b. and the hypotheses you stated in part a., write a conclusion for the test. Let $\alpha = 0.05$. Be sure to (1) write a sentence in regards to the outcome of your test, (2) a brief explanation as to why you chose that particular outcome, and (3) what the outcome of your test means in the context of the problem.

We reject the null hypothesis at 0.05 significance level because our p-value is less than alpha. We have evidence that more free throws were made while medicated than unmedicated, therefore the league will ban the supplement.

d. (2 points) Which of the following plots best represents the p-value associated with this particular hypothesis test? Here, assume the shaded region represents the p-value of the test.



The p-value is best represented by plot: (C)

e. (2 points) Based on the test you chose in part b., compute, by hand, a 95% confidence interval. Write your confidence interval in "interval form", i.e. *(lower bound, upperbound)* and show your work in the space indicated below. Note: You do not need to interpret this confidence interval.

- 95% confidence interval: (1.042340, 4.594024)

Show your work for this question below. If you write R code, please make sure the computed values are shown as outputs (And please don't leave a bunch of extraneous code/outputs in your solutions).

```
cv <- 1 - (0.05)
error <- qt(cv,10)
s <- 10.56
xbar <- 2.818182
n <- 11
```

```r
lowerbound <- xbar - (error * sqrt(s/n))
sprintf("lowerbound = %f", lowerbound)
```

```
## [1] "lowerbound = 1.042340"
```

```r
upperbound <- xbar + (error * sqrt(s/n))
sprintf("upperbound = %f", upperbound)
```

```
## [1] "upperbound = 4.594024"
```

f. (2 points) Explain why we "fail to reject" rather than "accept" a null hypothesis when our test statistic is less extreme than the critical value.

Accepting the null hypothesis would indicate that we have proven an effect doesn't exist. Failing to reject the null indicates that our sample did not provide sufficient evidence to conclude that the effect exists.