ST 411/511 - Methods of Data Analysis I

Final Exam

Chimdi Chikezie

Due: Thursday, March 17th 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)

The table below lists several of the statistical hypothesis tests that were covered during the course. For each of the scientific questions in parts (a) through (e), write the letters for **all** of the the tests (from the table) which could be used to address the stated scientific question of interest. Further, discuss how each test that you list differs, if at all, from the others in terms of the parameters they test in the hypotheses.

Letter of the test	Name of the Test
A	One-sample t-test
B	Two-sample t-test
C	Paired t-test
D	Wilcoxon rank-sum test
E	Sign test
F	Signed-rank test
G	Levene's test
H	ANOVA
I	Kruskal-Wallis

a. (2 points) Does the ascorbic acid content of dairy cows raised in Benton county tend to vary more than dairy cows raised in Jefferson county?

- Tests you could use to address this question: G
- Describe how each test that you listed differs, if at all, from the others in terms of the parameters they test in the hypotheses. If there's only one test you would use to address the question of interest, state the population parameter(s) used for that test's hypotheses.

Here in Levene's test, the parameter is the population variances/spread. $H_0 : \sigma_x^2 = \sigma_y^2$ $H_A : \sigma_x^2 \neq \sigma_y^2$

b. (2 points) Does the amount of ascorbic acid content for dairy cows raised in Benton county differ from the value 220 mg/cc?

- Tests you could use to address this question: A, E, F
- Describe how each test that you listed differs, if at all, from the others in terms of the parameters they test in the hypotheses. If there's only one test you would use to address the question of interest, state the population parameter(s) used for that test's hypotheses.

Here in the One-sample t-test, the parameter is the mean. We are testing if the mean of the population is equal, greater or less than a certain value.

while

Sign Test and Signed-rank test are used when the single or paired samples have outliers. Sign test's parameter is the median. Signed-rank test's parameter is the population pseudomedian.

c. (2 points) Do dairy cows raised in Benton county and Jefferson county have different amounts of ascorbic acid content?

- Tests you could use to address this question: B, D

- Describe how each test that you listed differs, if at all, from the others in terms of the parameters they test in the hypotheses. If there's only one test you would use to address the question of interest, state the population parameter(s) used for that test's hypotheses.

Here in the Two-sample t-test, the parameter of interest is the mean. We are testing if the mean of two independent populations are equal, greater or less than each other.

while

Wilcoxon rank-sum test, the parameter of interest is the an additive effect. It is mainly used when the Two-sample t-test contains extreme outliers.

d. (2 points) Do dairy cows raised in the 36 different counties in Oregon have different amounts of ascorbic acid content?

- Tests you could use to address this question: H, I
- Describe how each test that you listed differs, if at all, from the others in terms of the parameters they test in the hypotheses. If there's only one test you would use to address the question of interest, state the population parameter(s) used for that test's hypotheses.

In ANOVA, the parameter of interest is the mean of the populations. We want to know if the means are equal.

while in

Kruskal-Wallis test, our parameter is the ranks of the observations. We use this when we have not met the assumptions of Anova.

e. (2 points) How does the ascorbic acid content of dairy cows raised in Benton county change when they start with a grass-based diet and are then swapped to a grain-based diet?

- Tests you could use to address this question: C, E, F
- Describe how each test that you listed differs, if at all, from the others in terms of the parameters they test in the hypotheses. If there's only one test you would use to address the question of interest, state the population parameter(s) used for that test's hypotheses.

For the Paired T-test, our parameter is the mean. We are testing to know if the mean difference between pairs of measurements is zero or not.

while

Sign Test and Signed-rank test are used when the single or paired samples have outliers. Sign test's parameter is the median difference. Signed-rank test's parameter is the population pseudomedian.

Question 2 (10 points)

A study is conducted to examine the relationship between coffee consumption and the amount of exercise a person engages in each week. For the study, the researchers placed people into the following groups (based on their average coffee consumption), ranked from lowest coffee consumption to most coffee consumption per week: *One or fewer cups per week*, *two to six cups per week*, *one cup per day*, *two to three cups per day*, *four or more cups per day*. Here, “cups” means 8oz servings of coffee. Note: Some categories are “per week” while others are “per day”.

Scientific question of interest: *Do people who consume different amounts of coffee engage in different amounts of exercise?*

a. (2 points) Write out the null and alternative hypotheses the researchers should use if they intend to use an ANOVA testing procedure.

- Null hypothesis: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- Alternative hypothesis: $H_A : At - least - two - population - means - are - different.$

b. (3 points) Complete the ANOVA table below by filling in the “.” values with the correct answers. That is, within the table itself, delete the “.” values and replace them with the computed values. Show your work in the space indicated.

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F	p-value
Between	.	.	.	3.9058	.
Within	.	1593.06	.		
Total	135	.			

Source	Degrees of Freedom	Sum of Squares	Mean Squares	F	p-value
Between	6	289.4028	48.2338	3.9058	0.00128
Within	129	1593.06	12.3493		
Total	135	1882.4628			

Show your work below:

The total degrees of freedom is $n - 1 = 135 - 1 = 134$, so $n = 134$.

The within group degrees of freedom is $n - I$, and there are 5 different groups, so $134 - 5 = 129$

Mean of squares within group is $SSW/dfW = 1593.06/1593.06 = 12.349$, so $MSB = 12.349$

F is MSB/MSW, so $MSB = F * MSW = 3.9058 * 12.3493 = 48.2338$, then $MSB = 48.2338$

The between group degrees of freedom is $Tdf - Wdf$, so $136 - 131 = 4$

Degrees of Freedom between groups is $Tdf - dfW = 135 - 129 = 6$

Mean of squares between group is SSB/dfB, so $SSB = MSB * dfB = 48.2338 * 6 = 289.4028$, so $SSB = 289.4028$

```
F <- 3.9058
dfB <- 6
dfW <- 129
1 - pf(F, df1=dfB, df2=dfW)
```

```
## [1] 0.001283254
```

c. (2 points) Using a significance level of $\alpha = 0.05$, based on the hypotheses you wrote in part a. and the ANOVA table you filled out in part b., interpret the results of the ANOVA test and answer the Scientific question of interest.

We reject the null hypothesis for the alternative at 5% significance level. We do not have enough evidence to conclude that all the population means are equal.

This means that at least two groups of people on average do not consume the same amount of coffee.

d. (3 points) Based on the R output below, denoting the 95% family-wise confidence intervals using the Tukey-Kramer post-hoc procedure, list the comparisons that are statistically different than zero at the 5% significance level. Then, for each interval in your list (i.e. just the comparisons that are statistically different than zero), write an interpretation regarding how much more or less exercise one group does than the other, on average, using the `Estimate` of the difference.

```
Multiple Comparisons of Means: Tukey Contrasts

95% family-wise confidence level

Linear Hypotheses:
                                  Estimate lwr      upr
2-6 cups/week - <1 cup/week == 0  -1.06930 -4.07340  1.93480
1 cup/day - <1 cup/week == 0      -0.35640 -2.92502  2.21223
2-3 cups/day - <1 cup/week == 0    1.85223 -0.85066  4.55513
>4 cups/day - <1 cup/week == 0     1.92294 -0.66174  4.50762
1 cup/day - 2-6 cups/week == 0     0.71290 -2.09549  3.52130
2-3 cups/day - 2-6 cups/week == 0  2.92153 -0.01017  5.85323
>4 cups/day - 2-6 cups/week == 0   2.99224  0.16916  5.81532
2-3 cups/day - 1 cup/day == 0      2.20863 -0.27494  4.69219
>4 cups/day - 1 cup/day == 0       2.27933 -0.07503  4.63369
>4 cups/day - 2-3 cups/day == 0    0.07071 -2.42946  2.57087
```

- **Comparison(s) that are statistically different than zero:** $> 4cups/day - 2 - 6cups/week$ group is statistically different than zero.
- **Interpretation(s)**

With 95% confidence, the group of people that consume >4 cups/day consume about (0.16916 to 5.81532)oz servings of coffee on average more than the group that consume 2-6 cups/week.

Question 3 (10 points)

Below, you'll find numerical summaries from the study conducted in Question 2. To recap, a study is conducted to examine the relationship between coffee consumption and the amount of exercise a person engages in each week (in minutes). For the study, the researchers measured a random sample of people's average coffee consumption and categorized them as follows: *One or fewer cups per week*, *two to six cups per week*, *one cup per day*, *two to three cups per day*, *four or more cups per day.* Here again, "cups" means 8oz servings of coffee. Further, the researchers computed the sample mean and variance of the number of minutes spent exercising each week for each group.

	<1 cup/week	2-6 cups/week	1 cup/day	2-3 cups/day	>4 cups/day
Sample size	24.0	18.0	34.0	27.0	33.0
Sample mean	22.3	21.3	22.0	24.2	24.3
Sample variance	15.2	7.1	20.2	4.6	10.5

The researchers wish to test if the average of the means for people who drink **1 or more cups of coffee per day** tended to *exercise more*, on average, than the average of the means for people who drank **strictly less than 1 cup of coffee per day.**_

Scientific question of interest: *Do people who drink one or more cups of coffee* ***per day*** *exercise more on average than those who drink less?*

a. (3 points) In the equation below, replace the values of C1 through C5 with an appropriate set of coefficients the researchers could use to answer their question of interest using a linear combinations of the means procedure. Let μ_1 be the population mean for the `<1 cup/week` group, μ_2 be the population mean of the `2-6 cups/week` group, and so on. Further, write an appropriate *alternative* hypothesis you would use, in terms of γ, to address the scientific question of interest. Note: Please write the alternative hypothesis in terms of γ and not the population means, $\mu_i, i = 1, 2, 3, 4, 5$.

$$\gamma = -1/2\mu_1 + (-1/2)\mu_2 + 1/3\mu_3 + 1/3\mu_4 + 1/3\mu_5$$

$$H_A : \gamma = 1/3(\mu_3 + \mu_4 + \mu_5) - 1/2(\mu_1 + \mu_2) > 0$$

b. (4 points) Compute the values of (1) the test statistic, (2) the pooled sample variance, (3) the standard error of the test statistic, (4) the degrees of freedom, and (5) the p-value of the test for the alternative hypothesis you specified in part a. above. Show your work in the space indicated.

- Test statistic value: 2.6041
- Pooled sample variance: 12.15649
- Standard error: 0.6528054
- Degrees of freedom: 131
- p-value: 0.010274

Show your work below:

```
n1 <- 24
mu_1 <- 22.3
s1 <- 15.2

n2 <- 18
mu_2 <- 21.3
s2 <- 7.1

n3 <- 34
mu_3 <- 22.0
s3 <- 20.2

n4 <- 27
mu_4 <- 24.2
s4 <- 4.6

n5 <- 33
mu_5 <- 24.3
s5 <- 10.5

sp2 <- ((n1-1)*s1 + (n2-1)*s2 + (n3-1)*s3 + (n4-1)*s4 + (n5-1)*s5)/((n1-1) + (n2-1) + (n3-1) + (n4-1) +
sprintf("Pooled variance: %f", sp2)
```

```
## [1] "Pooled variance: 12.156489"
```

```
g <- ((mu_3 + mu_4 + mu_5)/3) - ((mu_1 + mu_2)/2)

df <- (n1-1) + (n2-1) + (n3-1) + (n4-1) + (n5-1)
sprintf("Degrees of freedom: %f", df)
```

```
## [1] "Degrees of freedom: 131.000000"
```

```
coe <- ((-1/2)^2)/24 + ((-1/2)^2)/18 + ((1/3)^2)/34 + ((1/3)^2)/27 + ((1/3)^2)/33
SE <- sqrt(sp2 * coe)
sprintf("Standard Error: %f", SE)
```

```
## [1] "Standard Error: 0.652805"
```

```
t <- (g-0)/SE
sprintf("Test Statistic: %f", t)
```

```
## [1] "Test Statistic: 2.604145"
```

```
pvalue <- 2 * (1 - pt(t, df))
sprintf("P-value: %f", pvalue)
```

```
## [1] "P-value: 0.010274"
```

c. (3 points) Using a significance level of $\alpha = 0.05$, interpret the results of the test you conducted in part b. to answer the Scientific question of interest

I reject the null hypothesis for the alternative at 5% significance level because the $P - value$ is less than α. We have evidence that mean time difference between the two groups is greater than zero. This implies that people who drink one or more cups of coffee **per day** exercise more on average than those who drink less.

Question 4 (15 Points)

A regression model is used to fit the amount of `gift_aid` (Y), in thousands of dollars, received by a random sample of $n = 60$ students at a prestigious liberal arts college using the reported `income` (X) of the students' parents, also in thousands of dollars, as the explanatory variable. Data summaries and limited output from the model summary in R are provided below.

Assume that the relationship between the amount of `gift_aid` and `income` of the students' parents is negative. That is, students who have parents with higher `income`s tend to receive *less* `gift_aid`. Further, recall that the values displayed below are in *thousands of US Dollars*.

Variable	Sample Mean	Sample Variance
income	44.29	741.6857
gift_aid	13.56	11.7380

```
Call:
lm(formula = gift_aid ~ income)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3275 -1.1525 -0.0211  1.0794  3.7987 

Residual standard error: 1.809 on 58 degrees of freedom
Multiple R-squared:  0.7259,    Adjusted R-squared:  0.7212 
F-statistic: 153.6 on 1 and 58 DF,  p-value: < 2.2e-16
```

(a) (4 points) Compute the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ and then write down the simple linear regression equation using correct statistical notation. Also, provide an interpretation as to what the slope means *in the context of the problem.* Remember, assume the relationship between the response and explanatory variable is negative. Show your work in the space indicated.

- $\hat{\beta}_0 = 14.2609$
- $\hat{\beta}_1 = -0.01583$
- Equation of the regression line:$\hat{\mu}(Y|X) = 14.2609 - 0.0158X$
- Interpretation of $\hat{\beta}_1$: A one unit increase in income of the students' parents is associated with a decrease of $\hat{\beta}_1 = -0.01583$ in mean amount of gift_aid.

Show your work below:

```r
Sy <- 11.7380
Xy <-  13.56
Sx <-  741.6857
Xx <- 44.29

beta1hat <- (Sy/Sx)*(-1)
beta1hat
```

```
## [1] -0.01582611
```

```
beta0hat <- Xy - beta1hat*Xx
beta0hat
```

```
## [1] 14.26094
```

(b) (4 points) Perform a level $\alpha = 0.05$ two-sided hypothesis test of the null hypothesis $H_0 : \beta_1 = 0$. State your conclusion of the test in the *context of the problem.* Show your work in the space indicated.

- Test statistic value: -1.830086
- p-value: 0.072377
- Conclusion: We fail to reject the null hypothesis at 5% significant level because the P-value is greater than α. There is no evidence to show that the there is a linear trend in the mean of amount of gift_aid as a function of income of the students' parents.

Show your work below:

```
n <- 60
resid.se <- 1.809
SEbeta1hat <- resid.se * sqrt(1 / ((n - 1) * Sx))
sprintf("SE(beta1hat): %f", SEbeta1hat)
```

```
## [1] "SE(beta1hat): 0.008648"
```

```
t <- (beta1hat - 0)/SEbeta1hat
sprintf("test statistic: %f", t)
```

```
## [1] "test statistic: -1.830086"
```

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

```
## [1] "P-value: 0.072377"
```

(c) (3 points) Compute the 95% confidence interval for the mean of `gift_aid` when `income = 75`. Interpret this interval *in the context of the problem.* Show your work in the space indicated.

- 95% confidence interval: (12.365738,13.781562)
- Interpretation: With 95% confidence, we expect to get a mean `gift_aid` of between (12.365738 to 13.781562) thousands of dollars when the `income` of the students' parents is 75 thousand of dollars.

Show your work below:

```
X <- 75
se.m <- resid.se * sqrt(1/n+(X-Xx)^2/((n-1)*Sx))

meany <- 14.2609 - 0.01583*X

a <- 0.05
te <- qt(1-(a/2),n-2)
CIU <- meany + te*se.m
CIL <- meany - te*se.m
sprintf("Confidence interval lower bound: %f", CIL)
```

```
## [1] "Confidence interval lower bound: 12.365738"
```

```
sprintf("Confidence interval upper bound: %f", CIU)
```

```
## [1] "Confidence interval upper bound: 13.781562"
```

(d) (2 points) In a sentence, provide an interpretation for the values of multiple R-squared and the correlation coefficient (r_{XY}) in the context of this particular question.

- Interpretation of multiple R-squared: This tells us that about 72% of the total variance in the response variable (gift_aid of students) is explained by the model.

- Interpretation of the correlation coefficient: This is measuring the linear association between two variables, the amount of `gift_aid` of students and income‘ of the students’ parents in a sample of pairs (X_i, Y_i).

(e) (2 points) Based on the linear model’s output from parts (a) through (d), what are some general statements you would make regarding the relationship between `income` and `gift_aid` and the use of linear regression models to describe that relationship? In other words, what do we learn about the relationship between our variables of interest and why is a linear regression model useful for characterizing this relationship?

Based on the linear model’s outputs and the computed values, I would say that there is a negative linear association between gift_aid and income. This is because I got a negative slope parameter value. Also, while playing with the values of X in the confidence interval, I discovered that the smaller the income value, the greater the gift_aid value and vice versa.

Linear regression model useful for characterizing this relationship because: 1. We had a large sample size that is normally distributed. 2. The variance of gift_aid were the same. 3. Gift_aid was shown to be a function of income because a 1-unit increase in income is associated with a decrease of $\hat{\beta}_1$ in the mean of gift_aid.