STAT 441/541: Final Exam
Name:

Please turn the exam in on D2L and include the R Markdown code *and* a PDF or Word file with output. Verify that all of the code has compiled. While the exam is open book and you can use any resources from class or freely available on the internet, this is strictly an individual endeavor and **you should not discuss the problems with anyone outside the course instructor including class members.** All resources, including websites, should be acknowledged.

Part A. Experimental Design Plan (24 points)

Suppose that you utilize your experimental design skills and start a business growing and selling sunflowers. Given that you live in Bozeman, you have limited space for a garden, but you do have a small area beside your house to construct a raised bed garden. Your goal is to select the combination of sunflower `species` (evening sun, ring of fire, velvet queen) and `watering_regime` (daily with watering bucket or every-other day with a hose) that optimizes the number of sunflowers you can sell. When building a raised garden bed, you realized that roughly 1/3 of your garden will be shaded by your neighbor's Sprinter Van.

Use this information to construct an experimental design plan. You only need to complete components that have points associated with them, but you'll need to use information provided for the other questions.

1. Define the objective of your study (2 point)

2. Define meaningful and measurable response (2 point)

3. Diagram treatment application process for a single run (2 points)

4. Identify experimental units (and if appropriate observational units) (2 points)

5. List sources of variation (2 points)

6. Perform pilot runs *Given how long it takes for sunflowers to bloom (around 75 days), you won't have time to perform a pilot study before planting your summer crop. However, assume that your research has suggested that a single seed can have about fifteen flowers on average and that a standard deviation for the number of flowers is around 7.*

7. Choose experimental design (i.e. randomization) (4 points) *Describe the randomization procedure and also include code that assigns treatments to your EUs*

8. Determine number of replicates required (4 points) *Regardless of your randomization approach, you can assume a completely randomized design with equal numbers of replicates for each treatment combination when determining the number of replicates you will need.*

9. Describe method(s) for data analysis (4 points) *Write out the statistical model and specify the R code that could analyze the data. (You won't able to run the code, but include it in an R chunk with {r, eval = F} in the header for the chunk.)*

10. Timetable and budget for resources to complete experiment (2 points) *You can assume a pack of sunflower seeds cost $4 and has 100 seeds. Watering your garden costs $1 per week for either the watering bucket or hose approach.*

Part B. Sample Size considerations (10 points)

1. (4 points) *Compare and contrast using power versus a desired precision for the standard error as tools to inform sample size consideration.*

2. (6 points) *Consider a model with one treatment (two levels). Suppose that without blocking σ_{CRD} is 10, but blocking controls the variability such that σ_{RCBD} is 5. For both scenarios, compute the total number of samples required such that the expected standard error of the contrast between the two groups is less than 1. (Hint: if $x \sim N(\mu_x, \sigma_x^2)$ and $y \sim N(\mu_y, \sigma_y^2)$, then $x - y \sim N(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$)*

Part C. Randomization (6 points)

While looking for your gardening gloves you find an old packet with 12 sunflower seeds. You also have 12 seeds left over from your experiment in part A. Using a 24 slot starter you plan to explore whether giving seeds La Croix vs tap water has any difference in the total number of flowers the plants produce. This question will explore the difference between a CRD and blocking (on old seeds vs. new seeds) with a GRCBD.

EUs	seed_type
1	old
2	old
3	old
4	old
5	old
6	old
7	old
8	old
9	old
10	old
11	old
12	old
13	new
14	new
15	new
16	new
17	new
18	new
19	new
20	new
21	new
22	new
23	new
24	new

I have written two functions `return_contrast_CRD()` and `return_contrast_GRCBD()` that will return the contrast between the La Croix and tap water. The functions require a vector of length two where each element is either `LaCroix` or `Tap` corresponding to that EU.

1. (6 points) For each approach (`CRD` and `GRCBD`) generate 1000 realizations and use a histogram to plot the distribution for the contrast from each scenario. Comment on the differences in the figures and the implications for the two sample regimes.

```
# Enter a vector of length 24 that includes a treatment for each
# each EU. The first entry will be for EU 1, and so on.
# I have set up the first CRD design for you.
# You'll need to do this 1000 times.
return_contrast_CRD(sample(rep(c('LaCroix', 'Tap'), each =12)))
```

```
## [1] 2.833333
```

```
# You'll also need to do 1000 replicates for the GRCBD.
# I've done one for you.
# If you don't have a balanced design with 6 of each treatment
# in each block, the function will return an error.

return_contrast_GRCBD(c(sample(rep(c('LaCroix', 'Tap'), each =6)),
                        sample(rep(c('LaCroix', 'Tap'), each =6))))
```

```
## [1] 3.333333
```

Part D. Analysis 1 (6 points)

treat1	treat2	y
A	A	8.353660
A	B	6.809591
A	A	11.198326
A	B	6.290848
A	A	9.958775
A	B	9.003898
A	A	9.694087
A	B	10.227664
A	A	6.858151
A	B	6.133519
B	A	15.814688
B	B	9.772760
B	A	11.421689
B	B	9.219301
B	A	14.312925
B	B	10.544373
B	A	12.139364
B	B	8.067579
B	A	14.209331
B	B	9.943417

```
lm(y ~ treat1 + treat2, data = D1) %>% summary()
```

```
##
## Call:
## lm(formula = y ~ treat1 + treat2, data = D1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9921 -0.9234 -0.2248  1.2876  3.1722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.8503     0.6550  15.039 2.97e-11 ***
## treat1B       3.0917     0.7563   4.088 0.000766 ***
## treat2B      -2.7948     0.7563  -3.695 0.001796 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.691 on 17 degrees of freedom
## Multiple R-squared:  0.6411, Adjusted R-squared:  0.5989
## F-statistic: 15.18 on 2 and 17 DF,  p-value: 0.0001649
```

1. (2 points) Interpret the output above using p-values and associated statistical significance language.

2. (2 points) Interpret the output above without using p-values and associated statistical significance language.

3. (2 points) Reflecting on the previous two questions, which approach do you envision yourself taking in the future. Why? *(Note: All thoughtful answers for this question will receive full credit)*

Part E. Analysis 2 (10 points)

For this question, we use a portion of a dataset from an article titled "Unified Meta-analysis: Using a Single Model to Estimate Treatment Effects of Multiple Interventions" In case you question the relevance of experimental design for modern statistics or data science, consider the following statement from the article

> The eminent computer scientist, Professor Michael Jordan at Berkeley, had this to say about what big data is: "…You might be excused of thinking [big data is] just a problem of building bigger computer systems and faster systems of treating data … it's a database problem. … That is not my perspective. … I think it's an integrated inference and computer science problem and in fact it's a deeper integration than anything we've seen before. It's the first time the fields are really being forced together at their foundations". Jordan goes on to argue that the "problem of our age" is to learn how to use these large data sets to infer causality. It's very easy to find spurious correlations in huge data sets (like EdX data sets). So how do we find authentic causality? The best way to do that is to run controlled experiments. We argue that what is important about big data is the ability to run big experiments. Randomized controlled experiments are the best way we have to determine if an independent variable has a causal relationship with a dependent variable. That is why all the big web-facing companies (Amazon, Bing, Google, Facebook, etc.) use online web-based experiments to guide product development.

The article presents a dataset containing an experiment performed in the mathematics software ASSISTments. This particular experiment focuses on exercises to find the expected value. The treatment was a "Skill

Builders assignment” designed to teach important concepts related to the task. The dataset contains four columns:

- Treatment: E for Experiment and C for Control
- PriorPercentCorrect: the percent of problems the student has previously completed correctly
- ProblemCount: Number of problems the student correctly answers
- complete: binary variable for whether the student sucessfully completed the assessment

1. (4 points) Create a figure to display how `ProblemCount` differs across the treatment and control groups. Make sure the figure has complete labels, titles and an informative caption.

2. (4 points) Analyze how `ProblemCount` is impacted by the treatment. Make sure to explore whether `PriorPercentCorrect` can explain additional variability in the response. Write a few paragraphs to summarize your results.

(*Note: the researchers are more interested in the complete variable, but this analysis would require knowledge about GLMs and logistic regression. In fact, you may notice that `ProblemCount` is an integer and a count-regression model such as Poisson regression might be more appropriate; however, just use a linear model framework for this question.*)

4. (2 points) Throughout class, we have talked about a scientifically meaningful difference. Is this idea captured in typical Null Hypothesis Significance Testing and p-values? Why or why not?

Part F. Experimental Design Plan, Part 2 for 541 only (10 points)

After telling your neighbor about your business plan for sunflowers, they offer to hire you to help with their tomatoes. It turns out that they have collected some data over the last few years. Their goal is to determine which tomato plants give the biggest tomatoes. They have four large planters two of which have been planted with Cherokee Purple and two which have been planted with Rosso Sicilian.

You asked about EUs and OUs and they gave you a puzzled look at then said, “I don’t know what that means, but I have a dataset you can analyze.” Below you can see a snapshot of the first 10 rows of the dataset.

```
## # A tibble: 10 x 6
##     year planter plant_number tomato_number tomato_weight species
##    <dbl>   <dbl>        <dbl>         <int>         <dbl> <chr>
##  1  2020       1            1             1          6.69 Cherokee Purple
##  2  2020       1            1             2          3.23 Cherokee Purple
##  3  2020       1            1             3          3.50 Cherokee Purple
##  4  2020       1            1             4          4.07 Cherokee Purple
##  5  2020       1            1             5          2.96 Cherokee Purple
##  6  2020       1            2             1          4.04 Cherokee Purple
##  7  2020       1            2             2          3.48 Cherokee Purple
##  8  2020       1            2             3          5.38 Cherokee Purple
##  9  2020       1            3             1          3.14 Cherokee Purple
## 10  2020       1            3             2          4.74 Cherokee Purple
```

1. (4 points) Based on what you can decipher from the dataset, what are the EUs in this study?

2. (2 points) Are there any OUs in the experiment? If so, what are they?

3. (4 points) In a couple of paragraphs, write a few recommendations for an improved experimental design.