STAT 506: Final Exam
Name:

Please turn in the exam to GitHub and include the R Markdown code and a PDF or Word file with output. Verify that all of the code has compiled and the graphics look like you think they should on your files, you are welcome to upload image files directly if they look distorted in the output file.

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 group mates or class members.** All resources, including websites, should be acknowledged.

For full credit, include your code and graphics for each question and create neat output by using options like `kable()` for tables and writing results in line with R commands.

Question 1. ANOVA (14 points)

A. (4 points) Write code to simulate data from a linear model with one categorical variable that has two levels. Assume this is a balanced design, with the same number of samples from each group. Use the reference case specification. Finally, fit a model to this dataset and verify you can recover the model parameters.

B. (4 points) Suppose $\sigma = .5$ and your goal is to estimate the contrast (β_1 in the reference case specification) between group 1 and group 2. How many samples are required to estimate the contrast between the group means such that there is a 90% probability the estimated standard error (of the contrast) is less than .1. You can assume a balanced design with equal samples sizes for each group.

You are welcome to solve this analytically or use simulation, but explain and defend your work.

c. (6 points) Consider the simplification of a cupcake dataset (collected by STAT 441 students) that contains measurements of cupcake heights from a designed experiment. For this question

```
cupcakes_q1 <- read_csv('CupcakeHeights.csv') %>%
  filter(Temp.F %in% c(325, 375)) %>%
  dplyr::select(Temp.F, Height.cm)
```

- **Temp.F:** temperature of oven
- **Height.cm** measured height of cupcake

Using the model framework from parts a and b, fit a linear model with two categorical variables for temperature to explore whether and how cupcake height depends on oven temperature. Present and defend your results and then write a short summary of your finding.

Question 2. Hierarchical Regression (18 points)

We have seen how hierarchical regression can improve group-level estimates through shrinkage. Hierarchical regression can also be used to control for observations that are not independent. (Remember the assumption

that $\epsilon \sim^{iid} N(0, \sigma^2)$) It turns out that the covariance (or correlation) between observations in the same group is not 0, but is related to σ and the σ terms on the random effects (more later).

This complete cupcake dataset includes information about the batch number of cupcakes that corresponds to pan the cupcakes were baked in. Cupcakes from the same batch are cooked in the oven together and would likely be correlated. Failing to account for this results in a phenomenon called pseudo-replication.

```
cupcakes_q2 <- read_csv('CupcakeHeights.csv') %>%
  filter(Temp.F %in% c(325, 375)) %>%
  mutate(Batch = case_when(Batch.Number == 1 & Temp.F == 325 ~ 1,
                           Batch.Number == 2 & Temp.F == 325 ~ 2,
                           Batch.Number == 3 & Temp.F == 325 ~ 3,
                           Batch.Number == 1 & Temp.F == 375 ~ 4,
                           Batch.Number == 2 & Temp.F == 375 ~ 5,
                           Batch.Number == 3 & Temp.F == 375 ~ 6)) %>%
  dplyr::select(Temp.F, Batch, Height.cm)
```

- **Temp.F:** temperature of oven
- **Batch** identifier for batch of cupcakes (cooked in same pan)
- **Height.cm** measured height of cupcake

A. (4 points) Create a graphic (or series of graphics) to illustrate the relationship between cupcake height, batch, and temperature. The graphic should include appropriate labels, titles, and an informative caption.

B. (2 points) Based on your figure, do cupcake heights in the same batch seem more similar than cupcake heights between batches (with the same temperature)? Why or why not?

C. (4 points) Let's account for the fact that cupcake heights within a batch are potentially correlated. Write out the notation for a hierarchical model that explores differences in cupcake heights across the batches of cupcakes. Define all of your notation. Note: in our housing dataset lingo this is akin to looking at differences in prices by zipcode.

D. (4 points) Fit the model, specified in part C and discuss your results. Specifically, comment on how the standard error associated with the difference in cupcake height between 325 degrees and 375 degrees changes (from Q1C) *and* how that influences your inferences.

E. (4 points) Another approach that is commonly employed to deal with pseudoreplication is to take the mean of all of the subsamples (in this case, cupcakes within each batch). Comment on the code below and discuss how it differs from the two analyses from Q2D and Q1C.

```
cupcakes_q2 %>% group_by(Batch, Temp.F) %>%
  summarize(Height.cm = mean(Height.cm), .groups = 'drop') %>%
   lm(Height.cm ~  factor(Temp.F), data = .) %>% summary()
```

```
##
## Call:
## lm(formula = Height.cm ~ factor(Temp.F), data = .)
##
## Residuals:
##        1        2        3        4        5        6
##  0.16667 -0.08333 -0.08333 -0.01667  0.05000 -0.03333
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.33333    0.06161  70.330 2.45e-07 ***
## factor(Temp.F)375  0.18333    0.08714   2.104    0.103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1067 on 4 degrees of freedom
## Multiple R-squared:  0.5253, Adjusted R-squared:  0.4067
## F-statistic: 4.427 on 1 and 4 DF,  p-value: 0.1032
```

Question 3. Correlation in Hierarchical Models (12 points)

As mentioned Q2, hierarchical models can account for correlation in observations. Hierarchical models are often rewritten in matrix notation such that

$$\begin{aligned}\underline{y} =& X\underline{\beta} + Z\underline{\theta} + \underline{\epsilon} \\ \underline{\epsilon} \sim& N(\underline{0}, \sigma^2 I) \\ \underline{\theta} \sim& N(\underline{0}, \Sigma_\theta^2)\end{aligned}$$

where $X\underline{\beta}$ are the fixed effects and $Z\underline{\theta}$ are random effects and $\underline{\epsilon}$ is uncorrelated error.

A. (2 points) Consider simulated data from the following hierarchical model

$$\begin{aligned}y_i \sim& N(\alpha_{j[i]}, \sigma^2) \\ \alpha_j \sim& N(\mu_\alpha, \sigma_\alpha^2)\end{aligned}$$

which can equivalently be written as

$$\begin{aligned}\underline{y} =& \underline{1}\mu_\alpha + Z\underline{\alpha} + \underline{\epsilon} \\ \underline{\epsilon} \sim& N(\underline{0}, \sigma^2 I) \\ \underline{\theta} \sim& N(\underline{0}, \sigma_\alpha^2) \\ \underline{\alpha} =& [\alpha_1, ..., \alpha_J]\end{aligned}$$

Z is an indicator matrix for groups.

```r
set.seed(04232022)
n <- 500
num_groups <- 10
mu_alpha <- 5
sigma_alpha <- 5
sigma <- 10
X <- rep(1, n)
z <- factor(sample(1:10, replace = T, n))
Z_mat <- model.matrix(~z - 1)
alpha <- rnorm(num_groups, sd = sigma_alpha)

y <- rnorm(rep(1, n) * mu_alpha + Z_mat %*% alpha, sigma )
```

Are the error terms ($\epsilon_i's$) in this model independent? How about the responses ($y_i's$)?

B. (4 points) The correlation in observations can be solved analytically or computationally. The code below gives a Monte Carlo estimate of correlation for observations within the same group and between groups. Use this information to derive how correlation between and within groups depends on σ and σ_α. (You can, but don't have to use analytical information to support your argument.)

```
sigma_alpha <- 1
sigma <- 1

num_replicates <- 100000

within_group <- cor(matrix(1, nrow = num_replicates, ncol = 2) +
  matrix(rnorm(num_replicates * 2, mean = 0, sd  = sigma_alpha),
         nrow = num_replicates, ncol = 2) +
  matrix(rnorm(num_replicates * 2, mean = 0, sd  = sigma),
         nrow = num_replicates, ncol = 2))
within_group
```

```
##            [,1]       [,2]
## [1,] 1.00000000 0.00195242
## [2,] 0.00195242 1.00000000
```

```
between_group <-cor(matrix(1, nrow = num_replicates, ncol = 2) +
  matrix(rnorm(num_replicates , mean = 0, sd  = sigma_alpha),
         nrow = num_replicates, ncol = 2) +
  matrix(rnorm(num_replicates * 2, mean = 0, sd  = sigma),
         nrow = num_replicates, ncol = 2))
between_group
```

```
##           [,1]      [,2]
## [1,] 1.0000000 0.5022108
## [2,] 0.5022108 1.0000000
```

C. (6 points) Now we will re-examine the residuals from the models fit in parts Q2d (hierarchical model for Batch) and Q1c (linear model, but you may need to use `cupcakes_q2` to extract `Batch`). Specifically create 2 figures that display residuals while highlighting each batch. Comment on your results and the implications related to assumptions that the residuals are independent.

Question 4. Predicting Binary Outcomes (16 points)

This question will explore classification trees using `rpart()` and logistic regression.

A. (4 points) Simulate 100 binary data points on the unit square ($x_1 \in [0, 1]$ and $x_2 \in [0, 1]$). Use `ggplot2` to plot these locations.

B. (4 points) With the data from part A, simulate a response using a logistic regression model where $\underline{\beta}$ = [-5, 5, 5].

Update the figure from part A to replace the points with estimated probabilities (to one decimal place). Also add a boundary line that corresponds to p = .5, in other words, and separates points most likely to be zeros or ones.

C. (4 points) Use the data from part B to fit three models: logistic regression, decision tree (`rpart`), and a random forest (`randomforest`). Briefly summarize the results from the logistic regression and the decision tree models.

D. (4 points) Generate another 1000 data points from this model. Use the models from part C to predict the binary outcome for these 1000 data points. Using classification error (the proportion of points incorrectly classified), comment on the predictive ability of the three models. Do the results match your expectation, why or why not?