# STAT 506: Midterm Exam
# Name:

Please turn in the exam to GitHub and include the R Markdown code and a PDF or Word file with output. Please 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.

## Short Answer Questions

For questions in this section, keep your answers concise. You are welcome to use a combination of prose, math, and pseudocode, but your responses should be well thought out and defended.

### 1. (4 points)

Describe statistical significance, in your own words.

### 2. (4 points)

How can the standard deviation of the data be used to characterize the uncertainty in a point estimate?

### 3. (4 points)

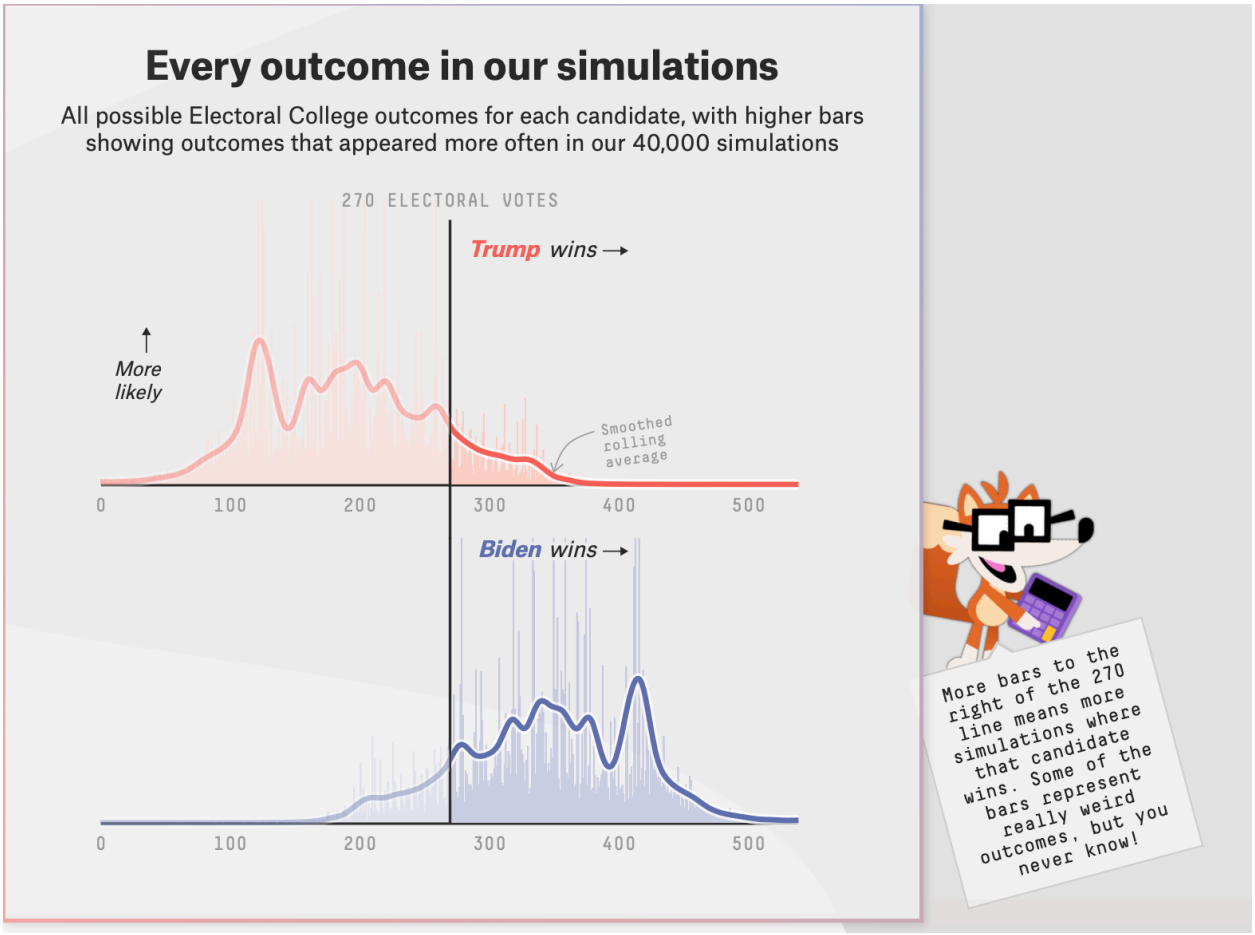
Convince a collaborator, say a scientist modeling butterfly movement, that using Bayesian analysis is a defensible approach.

### 4. (4 points)

Convince a collaborator, say a scientist modeling butterfly movement, that a classical analysis is a defensible approach.

**5. (4 points)**

Consider the following image, obtained from fivethirtyeight.com. Describe what the bars and curves on the figure represent *and* how the figure can be interpreted. You can assume your audience would be a STAT 216 student with an introductory knowledge of statistics.




.

Additional details about the figure and methods behind it are available at: https://projects.fivethirtyeight.com/2020-election-forecast/.

## Simulation Question

### 5. (10 points)

This question is focused on understanding an interaction with categorical variables. When answering this question, please include all code in the text.

a. Write out the mathematical model for a two-way Anova model with an interaction. (There should be at least 3 levels for one of the categorical variables, but the other can have two levels).

b. Simulate data from this model.

c. Create a data visualization that clearly depicts the interaction. Make sure to include appropriate titles, labels, and captions. Annotation may also be useful with this figure.

d. Fit the model and interpret the model coefficients.

## Reading and Critique

### 6. (10 points)

Read the article titled, "Survey data on students' online shopping behaviour: A focus on selected university students in Indonesia" (link: https://www.sciencedirect.com/science/article/pii/S2352340919314295)

a. Comment on the experimental design and how the results from the study can be generalized.

b. To the best of your ability, write out the mathematical notation that corresponds to the regression model displayed in Tables 7, 8, and 9.

c. Critique the following section.

> The hypothesis to be tested is as follow: Ho: There are no variables influencing online shopping behaviour.
>
> We see that the ANOVA produces P-value of the regression = 0.000, which is less than 0.05 significant level. This leads to the rejection of the null hypothesis, meaning that at least one of the predictors significantly influences the purchasing behaviour. The R-square is 47.26%, meaning that the predictors have an effect of 47.26% on onine shopping behaviour.
>
> The coefficients in Table 8 show the individual effect of each variable. We see that the P-values of POR, EJY, SIF and OAD are less than 0.05 significant level. This means that the purchasing behaviour is significantly infuenced by the perception of risk (POR), enjoyment (EJY), social influence (SIF) and online advertisment (OAD). Meamwhile, two other variables, i.e. trust and security (TAS) and quality of website (QOW), did not significantly influence the online shopping behaviour (see Table 9).

d. Based on the analysis (and assuming all assumptions are satisfied). Write a summary statement of the results in Table 9.

## Data Analysis

### 7. (10 points)

Using a candy dataset (https://math.montana.edu/ahoegh/teaching/stat446/candy-data.csv), define and fit a regression model to understand the relationship between `winpercent` and `pricepercent, chocolate,` and `caramel`. More insight into the data is available at https://fivethirtyeight.com/videos/the-ultimate-halloween-candy-power-ranking/.

a. Write out the model and define all of the coefficients.

b. Fit the model with software of your choice and print the results.

c. Construct a contrast to compare the expected win percentage for a candy that has chocolate, caramel and the average price percent, with a candy that just has chocolate, no caramel, and the average price percent.

d. Try to create the candy with the highest win percentage, then specify the levels of the predictors and create a predictive distribution for an individual type of candy with those features.