STA 279
Spring 2022

Exam 2

Dr. Evans
Exam 2

Last Name: _____ **First Name:** _____

I hereby state that I have not communicated with or gained information in any way from other students or any outside resource during this exam. I agree to abide by the rules stated below, and to abide by the Wake Forest Honor Code. All work is my own. I understand that any violation of this agreement will be reported to the Honor Council and will result, at minimum, in a 0 on this exam.

Signature : _____

All work on this exam must be your own.

1. You have 50 minutes to complete the exam.
2. Show all your work on the open ended questions in order to get partial credit. No credit will be given for open ended questions where no work is shown, even if the answer is correct.
3. You are allowed a calculator, however you may not share a calculator with another student during the exam. The calculator must be only a calculator, and may not be connected to the internet.
4. You are allowed to ask clarification questions to me, but you may not ask anyone else.
5. You are **not** allowed a cell phone, even if you intend to use it as a calculator or for checking the time. You are **not** allowed a music device or headphones, notes, books, or other resources.
6. You may **not** communicate with anyone other than myself during the exam.
7. Write clearly and be clear. Make it easy to find your answers.

Good luck!

Part 1 We have a client who is interested in modeling Y_i = the type of major a student i chooses to pursue at a large university. The university in the study has four types of majors: STEM, Business, Education, and Liberal Arts. We have a random sample of $n = 2500$ students from the university, and the table below shows the number of students from each type of major in the sample:

STEM	Business	Education	Liberal Arts
703	576	364	857

The client is interested to see if the choice of major is related to (1) whether or not a student took any AP classes in high school (Yes/ No) and (2) the student's High School GPA.

---

1. Based on the information provided, write down a population model for the relationship between choice of major (response) and AP classes and high school GPA (predictors). You may ignore potential interactions. Use appropriate notation.

The client fits the following model.

Model 1

```
Coefficients:
           (Intercept)  APYes       GPA
Business     -3.77       0.74       0.95
Education    -4.36       1.08       0.74
STEM         -5.48       1.79       1.03

Residual Deviance: 23.86116
```

---

2. Based on the output above, write down the fitted model for business vs. liberal arts.

3. For a student whose GPA is 3.5 and who has taken an AP class, how many times higher / lower is their predicted probability of choosing an education major versus a liberal arts major? Show your work and write your final answer in a sentence.

Model 1

```
Coefficients:
           (Intercept)  APYes       GPA
Business     -3.77       0.74       0.95
Education    -4.36       1.08       0.74
STEM         -5.48       1.79       1.03

Residual Deviance: 23.86116
```

---

4. What is the predicted probability of choosing a liberal arts major, for a student with a GPA of 3.5 who has taken an AP class? Show your work.

5. For a person who has taken an AP class, what is the minimum high school GPA they need in order for the model to predict that they are more likely to choose a business major over a liberal arts major? Show your work.

		Actual			
		Liberal arts	Business	Education	STEM
Predicted	Liberal arts	502	256	297	158
	Business	355	320	67	545
	Education	0	0	0	0
	STEM	0	0	0	0

To assess performance of the fitted model, you create this confusion matrix.

---

6. Is the model doing a good job predicting student majors? Your answer should include at least one summary measure of the confusion matrix, and should compare against the two types of random guessing discussed in class:

   - Randomly assign each student to one of the four majors, with a 1/4 probability for each major
   - Assign all students to the most common major

Part 2: We have data on the amount of time (in days, with decimals denoting part of a day) that it takes a cat to be adopted from the Texas animal shelter system. We have a random sample of 500 cats from 10 shelters, composed of 40 - 70 cats per shelter. Staff at the shelter system are interested in how (1) the number of pictures posted about a cat and (2) whether the cat is given a name impacts the amount of time it takes a chat to be adopted. Staff suspect that there is variation in adoption times from shelter to shelter, but the effects of posting pictures or naming cats are the same for each shelter.

The data has 500 rows and the following columns:

- `Shelters`: an ID for the shelter the cat was adopted from
- `NumPictures`: the number of pictures posted of the cat
- `Name`: whether the cat was given a name (1 = yes, 0 = no)
- `Days`: time it took the cat to be adopted

---

7. Based on the information we have so far, (1) write down a population model that is appropriate to try and (2) briefly explain why this model is a reasonable choice. Use appropriate notation, and explain what your subscripts (e.g., i and/or j) represent. You may assume all necessary conditions are met.

The staff fit the model you suggest and end up with the following output.

Model 1

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Shelters  (Intercept) 53.9     7.34
 Residual              109.2    10.45
Number of obs: 500, groups:  Shelters, 10

Fixed effects:
             Estimate   Std. Error
(Intercept)    30.50     0.67
NumPictures   -1.89      0.38
Name          -7.23      1.56
```

---

8. Does the fitted model suggest that named cats get adopted more quickly? Explain your reasoning.

9. Does the fitted model suggest there is systematic variation in adoption times between shelters, after accounting for the effects of pictures and naming? Calculate a statistic to support your conclusion.

The staff want to test whether there is a difference in adoption times for cats with names vs. cats without names.

---

10. Write down null and alternative hypotheses, in terms of one or more model parameters, which allow you to investigate whether there is a difference in adoption times for cats with vs. without names.

To test your hypotheses from Question 10, the staff fit a second model, which they will compare to the first model with a nested F test:

Model 2

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Shelters  (Intercept) 60.3     7.77
 Residual              115.7    10.76
Number of obs: 500, groups:  Shelters, 10

Fixed effects:
             Estimate   Std. Error
(Intercept)    35.80     0.55
NumPictures   -4.30      0.87
```

11. Give the numerator degrees of freedom, and upper and lower bounds on the denominator degrees of freedom, for the nested F test.

Instead of using the F distribution to calculate a p-value, you decide to use a parametric bootstrap.

---

12. Describe the additional steps needed to calculate a p-value for the hypotheses in Question 10, using a parametric bootstrap. Provide as much detail as you can, so that someone could turn your description into R code if they wanted to (you do not need to write code, though you may choose to if it helps you explain your procedure). Your description should include details like values for the parameters of the model you will simulate from, how many simulations you will use, how you will calculate a test statistic for each simulation, and how you will calculate a p-value from your bootstrap results at the end.

You are done!!! Whooo!!!!