**Psychology 610**

**HW 3**

Due Wednesday, September 27, 2023, 1:30pm on Canvas

* Select the Assignments menu and find HW Week 3
* Submit both an HTML file and a .Rmd file
* Make sure you can access to this folder ahead of the homework due date

**Part A. Reading Questions**

Answer in your R Markdown file.

1. Define Type I and Type II error. For each type, give an example of how a researcher could reduce the risk of committing the error.
2. In practical terms, what does the numerator of the *F* statistic equation represent? The denominator?
3. Statistical terms: What special name is given to the mean-squared-error from the mean-only model?

**Part B. Reliability and Validity Exercise**

Take the personality test found at <https://www.matthewbarr.co.uk/personality/>. This test identifies which character from the original *Star Wars* movies you are most like.

Answer in your R Markdown file.

1. What does it mean for a test to be reliable? Take the test a second time. Does it seem to be reliable?
2. What does it mean for a test to be valid? Does this test seem valid? Explain your answer referring to one of the four types of validity.

**Part C. Practice with ggplot2**

In ChatGPT, paste the following prompt: Fifty students answer the questions: (1) How many Badger shirts do you own? and (2) How many Badger football games did you attend? The variables are named badger\_shirts and games\_attended, respectively. Badger\_shirts has a minimum of 0 and a maximum of 20 and games\_attended has a minimum of 0 and a maximum of 10. Generate a fake dataset for me in R so that as badger\_shirts increases games\_attended increases. Also, include 6 outliers.

1. Copy the code that ChatGPT provides into your RMarkdown and run it. You don’t necessarily need to know what this code is doing. Note: If you want to save a dataset in R, you can use write.csv() to save it in your working directory. However, you won’t need to save this dataset.
2. Now, use the ggplot2 cheat sheet to create a quick and dirty scatterplot with the two variables. The plot should have a white background, meaningful x- and y-axis labels, and a title. Don’t worry about adding a best-fit line or error bands.

**Part D. Data analysis**

For this homework, we will use the “stray\_animals.csv” file. This dataset contains (fictitious) statistics about the number of stray animals (in thousands) in each of the 50 US states. The dataset contains information about the number of stray dogs, cats, and other animals measured in one year. The dataset also specifies whether the state participated in a nation-wide TNR (trap-neuter-return) program that was launched in an attempt to reduce the number of stray cats. In the data file, states that adopted the TNR program have a “1” in the tnr variable. States that did not adopt the program have a “0” in the tnr variable. Finally, the dataset also includes information about the number of animal shelters per 10,000 people in each state.

1. Read in the data and inspect them (summaries, structure, descriptive stats).
2. Create a quick and dirty plot that shows whether the number of stray dogs is related to the number of animal shelters in the state. This time, use the plot() function.
   1. Create the plot.
   2. Add a best fitting line.
   3. Does the relationship appear positive or negative? Strong, moderate, or weak?
3. Go through the following steps to create a composite score that reflects the overall numbers of stray animals, referred to as the “stray index” in these questions:
   1. Standardize the measures for stray dogs, stray cats, and other stray animals, individually.
   2. Check to make sure you standardized these variables correctly. Show the code.
   3. Create a new variable (strays\_z\_m) representing the average of these standardized scores.
   4. Comment on the researchers’ decision to standardize then average these scores. Does their decision make sense or not? How does this measure differ (in practical terms) from one in which the researchers average the scores without standardizing them? Type your answer to these questions in your RMarkdown-script.
4. Create two histograms of the number of stray cats. One for the states that took part in the program, and the other for the states that did not. Label your histograms properly.
5. What is the formula for each of the two models you should compare if you want to know whether the number of stray cats is different from zero? Type your answers to this question in your RMarkdown script (as plain text).
6. Compute the SSE for each of the two models in the previous question. Do so by using a step-by-step brute force method, like we did in lab. Remember to label clearly (e.g., sse\_1, sse\_2).
7. Compare the two models by calculating the appropriate *F* statistic and *p* value.
   1. Is estimating the mean worth the increased complexity of the model? Why? Interpret the result in plain English.
   2. Why is this model comparison not super informative (i.e., why are you not surprised 0 is a worse predictor for the number of stray cats per year? Why is the null model “stupid”?).
8. Sometimes learning stats or programming is difficult and you just need a little positive feedback. Figure out which package contains the praise() function, and complete all the steps you need in order to write a line of code using praise() to tell you something nice.

**Part C. Tell us how long it took you to complete this assignment.**