Problem set 4

due Monday, September 29, 2025 at 11:59am (noon!)

Instructions Upload your `.ipynb` notebook to gradescope by 11:59am on the due date. Please include your name, Problem set number, and any collaborators you worked with in a text cell at the top of your notebook. Please also number your problems in some way and include comments in your code to indicate what part of a problem you are working on.

Get help! If you need support working on your pset, see our week at a glance schedule for office hours and pset support times!

Dataset

Suppose you are studying brain responses to words. You measure EEG activity in response to word stimuli for both children and adults. You collect your data and store it in this CSV file: word_eeg_amplitudes.csv. For each participant, you collect the following variables:

- Subject: identifier for each participant
- Age_Group: whether the participant is a child or an adult
- Electrode: Fz, Cz, or Pz (common scalp recording sites)
- N400_Amplitude (μV): mean amplitude of the N400 ERP response to words
- Reading_Score: standardized reading comprehension score (extra variable not needed for this question)

The N400 is an event-related potential (ERP) component that appears as a negative-going deflection in the EEG signal around 400 ms after a stimulus. It is reliably elicited by meaningful words and is often larger (more negative) when words are unexpected or harder to process. Because it reflects semantic processing, it is widely used in language research.

Problem 1

Your first question is about developmental changes in language-related brain activity: **Is there a difference in median N400 amplitude between children and adults?** Start by exploring these variables with a `ggplot` boxplot. Then, use `infer` to construct the null

distribution and to compute the observed difference in `medians`. Using the `infer` way, visualize the null distribution and shade the p-value, including the observed difference in medians. Return the p-value. Should you reject the null hypothesis? Why or why not?

Problem 2

Your second question asks whether participants with stronger neural responses also show better language skills: **Is there a correlation between N400 amplitude and reading score?** Important note: the N400 is a negative-going ERP component, so more negative amplitudes indicate stronger responses. Explore this relationship with a `ggplot` scatter plot. Use `infer` to construct the null distribution and to compute the observed correlation. Using the `infer` way, visualize the null distribution and shade the p-value, including the observed correlation. Return the p-value. Should you reject the null hypothesis? Why or why not?

> ⚠ Short pset this week!
>
> To give you extra time to study for the exam, there are only 2 questions this week.