Problem set 3

due Monday, October 14, 2024 at noon

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

Problem 1

Suppose you are studying songbird brains. You measure the density of neurons in brain regious involved in song learning for both juvenile and adult songbirds. You collect your data and store it in this CSV file: songbird_neurons.csv. For each bird, you collect the following variables:

- **Subject**: identifier for each bird
- **Age_Group**: whether the bird is a juvenile or adult
- **Brain_Region**: HVC, RA, or Area X, three regions known to be involved in song learning.
- **Neuron_Density (neurons/mm^3)**: the number of neurons per cubic millimeter, which measures how densley packed neurons are in the specified brain region.
- **Song_Complexity (syllables)**: the total number of distinct syllables produced in the bird's song, which serves as a measure of how complex the bird's song is.

Your first question is about developmental changes in the songbird brain: **Is there a difference in median neuron density between juvenile and adult birds?** Start by exploring these variables with a botplot. Then, use `infer` to construct the null distribution and to compute the observed difference in medians. 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 birds with denser neural circuits produce more complex songs: **Is there a correlation between neuron density and song complexity?** Explore this relationship with a scatter plot. Use `infer` to construct the null distribution and to compute the observed correlation. 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?

Problem 3

Explore the songbird data by plotting each of the explanatory variables in the dataset against neuron density. Create a separate plot for each variable use appropriate geoms for the type of variable (e.g., scatter plots, boxplots, or — a new geom — violin plots). Based on your exploration, do you think that brain region explains a meaningful amount of variation in neuron density? What about song complexity? Age group? If not, consider simplifying your model by leaving one or more of these out. Justify your decision in terms of parsimony and model interpretability.

Problem 4

Based on your decisions in problem 3, specify the linear model that expresses neuron density as a weighted sum of the input variables you selected. Specify the model first as a formula, using Google Colab's LaTeX ($$) formatting. Then, specify and fit the model with ~~`infer`~~ `lm()`. Use the `predict()` function to calculate the predicted neuron densities (y) based on your model. Finally, plot your data and model together.

> Hint
>
> Google Colab can render LaTex math expression if you put them in between two dollar signs `$y = w_1x_1 + w_2x_2$` becomes $y = w_1 x_1 + w_2 x_2$. Here is an equation builder to help you generate what you need: https://editor.codecogs.com/

Problem 5

Suppose your roommate is keeping a bunch of plants in your apartment. You notice that the plants exposed to more light seem to be taller, and — as an emerging data scientist — you record these data in a csv file: polynomial_plants.csv. Explore the relationship between light_exposure and plant_height across different plant species by plotting the data using an appropriate geom. Then, specify, fit and compare polynomial models of increasing degrees (linear, quadratic, and cubic) to the data. Start by specifying and fitting a simple linear model. Next, specify and fit second- and third-degree polynomial models, and visualize each using geom_smooth(). Which best captures the relationship between light_exposure and plant_height? For each model, make sure you specify as a mathematical expression first in LaTex, then use `infer` to specify and fit the model.

> Hint
>
> We can express polynomials in R in a few ways. But importantly we are **expanding the input space** of the model by adding inputs. Consider a cubic polynomial, which we can express like this:
>
> - `y ~ 1 + x + I(x^2) + I(x^3)`
>
> or with this shorthand, which expresses the same model:
>
> - `y ~ 1 + poly(x, 3)`
>
> Both of the model specifications above include:
>
> - an intercept: 1
> - a linear term: `x`
> - a quadratic term: `I(x^2)`
> - a cublic term: `I(x^3)`.
>
> The `poly(x, 3)` way is just shorthand for `y ~ x + I(x^2) + I(x^3)`. But importantly, neither are equivalent to:
>
> - `y ~ 1 + x + I(x^3)`
>
> Which is missing the quadratic term `I(x^2)`. When we use the `I()` way, we have to explicitly add every term to the model.

Problem 6

Using the animal_brain_body_size.csv data we were working with in class, start by exploring the relationship between these two variables with `ggplot`. Then, specify and fit a linear model with `infer`, choosing to apply a log transformation either to the variables before fitting the model or directly in the model specification. After specifying the model, create a plot that includes the original data points and the fitted model line, ensuring that your plot reflects the transformed scales if you applied a log transformation.

Challenge (optional)

How might you use what you already know about the `infer` package to estimate the uncertainty around the free parameters you estimated for the model in Problem 4? Can you figure out how to use the `infer` pipeline to construct the sampling distribution of the parameter estimates, construct a confidence interval, and visualize? What about a hypothesis test? Can you construct a null distribution and shade the p value for model parameters as well? Hint: here are some examples to get you started.