Problem set 2

due Monday, September 23, 2024 at noon

> ⚠ Under construction
>
> Pset is still under constructrion; this is a sneak peak

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

Using the provided dataset of 1,000 babies, import the CSV file using the `readr` package from the `tidyverse` family. Handle any common issues (like missing values or incorrect data types) if they arise. Then, use `mutate()` to add a new column called `Age_First_Word`, which samples from a guassian distribution with a mean of 13 months and a standard deviation of 1.5 months. Finally, use `arrange()` and `head()` to show the 10 babies who spoke their first word the earliest. Then do the same to show the 10 babies who spoken their first word the latest.

- simulated-first-words.csv

Problem 2

Using the `Age_First_Word` column you created, plot a histogram to visualize the distribution of ages at which babies spoke their first word. Choose an appropriate bin width to best represent the data. Make sure to adjust the plot's readability using a built-in theme of your choice and include a suitable `base_size` font. Then, use `group_by()` and `summarize()` to calculate descriptive parametric statistics (mean, median, and standard deviation) for `Age_First_Word`, grouped by gender. Include `n()` in your call to summarize to count the number of babies per group.

Problem 3

Using the `infer` package, construct a bootstrap sampling distribution for the `Age_First_Word` (or `First_Word_Age` if renamed) to estimate the typical age babies say their first word. Use at least 1,000 resamples to build the distribution. Next, visualize the distribution with a histogram and shade the 95% confidence interval on the plot. Finally, calculate and report the standard error of this bootstraped distribution.

Problem 4

Suppose we are only interested in studying the "late talkers," defined as babies who spoke their first word after 15 months. Using the `dplyr` package (also part of `tidyverse`), first use `select()` to keep only the columns `Baby_ID`, `Gender`, and `Age_First_Word`. Then, use `rename()` to change `Age_First_Word` to `First_Word_Age`. Finally, use `filter()` to show only the babies who spoke their first word after 15 months, focusing your analysis on late talkers.

Problem 5

Determine whether there is a difference in mean age of first word production in late talking babies by sex. Use `infer` to construct the null distribution and to compute the observed difference in medians. Visualize the null distribution and shade the p-value on the plot, including a line that identifies the observed difference in medians. Print the p-value your analysis obtained. Should you reject the null hypothesis? Why or why not?

Problem 6

Using `mutate()`, create a new variable called `Gestational_Age` that represents the gestational age of each baby in weeks (for all babies, not just the late talkers). Assume gestational ages follow a normal distribution with a mean of 40 weeks and a standard deviation of 2 weeks. Add this variable to your dataset by sampling from this distribution. After adding the variable, visualize the distribution of `Gestational_Age` using a histogram, and use `group_by` and `summarise` to calculate both paramteric and non-parametric descriptive statistics for this variable.

Problem 7

Suppose you hypothesize that the age at which babies produce their first word is correlated with their gestational age. Explore this relationship with a scatter plot. Use `infer` to construct the null distribution and to compute the observed correlation. Visaulize the null distribution and shade the p-value, including the observed correlation. Print the p-value obtained by your

analyis. Why is the null hypothesis? Should you reject the null hypothesis? Why or why not?