Problem set 2

due Monday, September 23, 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

Using the provided dataset of 1,000 babies (simulated-first-words.csv), import the CSV file using the `readr` package from the `tidyverse` family. Use the `na` argument in your `read_*()` function to handle the missing values in the `First Word` column. Use a `dplyr` verb to remove the spaces in the column names.

Problem 2

Use `mutate()` to add a new column to the data you imported called `Age_First_Word`, which samples from a guassian distribution with a mean of 13 months and a standard deviation of 1.5 months. Visualize the distribution with a density plot. Use one of R's built-in functions for working with probability distributions to determine by what age will 80% of babies in the dataset have spoken their first word (probability of 0.8)? 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.

Problem 3

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 parametric descriptive statistics (central tendency and variability) for `Age_First_Word`, grouped by gender. Include `n()` in your call to summarize to count the number of babies per group.

Problem 4

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. Quantify the spread of the distribution with standard error. Next, visualize the distribution with a histogram and shade the standard error on the plot.

Problem 5

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 select only the columns `ID`, `Gender`, and `Age_First_Word`. Then, rename the `Age_First_Word` column to `First_Word_Age`. Finally, filter the data to include only the babies who spoke their first word after 15 months, focusing your analysis on late talkers.

Problem 6

Using the `infer` package, construct a bootstrap sampling distribution to estimate the *median* age your "late talkers" say their first word. Use at least 1,000 resamples to build the distribution. Quantify the spread of the distribution with a confidence interval. Next, visualize the distribution with a histogram and shade the ci on the plot.