


# Problem set 3

due Monday, September 22, 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!

 Warning: Avoid redundant loading

You will need the `tidyverse` library. Recall that Colab comes with this library already installed, and `tidyverse` includes `tibble`, `readr`, and `ggplot`. Avoid redundant loading.

## Problem 1

Using the provided dataset of 1,000 babies (`simulated-first-words.csv`), import the CSV file with the `readr` package and make sure missing values in the First Word column are treated as NA. Use a `dplyr` verb to remove spaces in the column names, then use `mutate()` to add a new column called `Age_First_Word` by sampling from a Gaussian distribution with a mean of 15 months and a standard deviation of 1 months. Use one of R's built-in probability distribution functions to determine by what age 5% of babies will have spoken their first word. Finally, use `filter()` to display all of the babies who spoke their first word by that point.

## Problem 2

Using the `Age_First_Word` column you created, plot a histogram with an overlaid density curve to visualize the distribution of ages at which babies spoke their first word. Adjust the plot's readability using a built-in theme of your choice. Adjust the `base_size` of the font, the histogram `bin_width`, `color`, and `fill`. Then, use `group_by()` and `summarize()` to calculate nonparametric 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 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 3,000 resamples to build the distribution. Quantify the spread of the distribution with a confidence interval. Next, use the `infer` way to visualize the distribution with a histogram and shade the confidence interval on the plot.

### Problem 4

Suppose we are only interested in studying the “early talkers,” defined as babies who spoke their first word before 11 months. Using `dplyr`, select only the columns `ID`, `Gender`, and `Age_First_Word`. Then, filter the data to include only the babies who spoke their first word before 11 months. Generate a plot of your choice to visualize the data.

### Problem 5

Using the `infer` package, construct a bootstrap sampling distribution to estimate the *median* age your “early talkers” say their first word. Use at least 1,000 resamples to build the distribution. Quantify the spread of the distribution with standard error. Next, use the `infer` way to visualize the distribution with a histogram and shade the `se` on the plot.