

# Problem set 2

due Tuesday, October 03, 2023 at 11:59pm



## Estimated time: 6 hours

Allocate about **1 hour per problem**, though some will take longer than others. You may need more time if programming is completely new to you, or less if you have some experience already.

**Instructions** Upload your .ipynb notebook to gradescope by 11:59pm on the due date.

- Note that each problem will be graded according to this [rubric](#). Solutions that include packages or functions not covered in this course will receive a score no higher than 2.
- You may collaborate with any of your classmates, but you must write your own code/solutions, understand all parts of the problem, and name your collaborators.
- You should also cite any outside sources you consulted, like Stack Overflow or ChatGPT, with a comment near the relevant lines of code (see example below). Recycled code that has not been cited will be considered plagiarism and receive a zero.

```
# code here was inspired by user2554330 on stack overflow:  
# https://stackoverflow.com/questions/69091812/is-everything-a-vector-in-r
```

## Problem 0

not graded

Create a new colab R notebook. Please include the title “Problem set 2”, your name, the date, and any collaborators somewhere at the top.

## Problem 1

Suppose you want to simulate data for the age at which typically developing babies produce their first word. Create a tibble with a single column for age that includes 500 samples generated from the Gaussian probability distribution with a mean age of 12 months and a standard deviation of 1 month. Plot your data with a density plot. Given this Gaussian distribution, what is the probability that a baby says their first word by the time they are 10 months? Return this value with one of R's functions for probability distributions.

## Problem 2

The dataset below includes data simulated from [Ritchie et al 2018](#), including the total brain volume and sex of 5216 subjects. Suppose you are interested in just one sex. Filter the data to include only male or only female participants (data scientists choice!). Explore these data with a histogram and use `summarise()` to report both parametric and nonparametric summary statistics.

Note that your research assistant has accidentally coded sex female as NA. You'll need to solve this issue with your dplyr skills!

```
"http://kathrynschuler.com/datasets/brain_volume.csv"
```

## Problem 3

Continuing with your filtered brain volume data, use `infer` to construct the bootstrap sampling distribution of the mean brain volume and quantify the spread of the distribution with standard error. Create a visualization that includes a histogram of the distribution and a shaded confidence interval with the standard error you computed.

## Problem 4

Suppose you now want to know whether there is a difference in *median* brain volume between the sexes. 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 5

Suppose you asked each participant to think of a number (integer) between 1 and 100. Assume each participant is equally likely to select a given number. Add a column to the brain volume data that samples from this distribution for all 5216 participants. Coerce this column to an integer if necessary. Explore this new variable with a histogram and summary statistics appropriate for this distribution.

## Problem 6

Suppose you now want to know whether participant brain volume is correlated with the random number they selected. 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?