CSDE 502 2021 Winter
Assignment 4
Integrating R Markdown
Instructor: Phil Hurvitz
phurvitz@uw.edu

Due Date: 2021-02-04 09:00 AM

Instructions:

1. Create an R Markdown Rmd file and render it to a self-contained HTML file. Use the output format **bookdown::html_document2** in the YAML header so you can use various methods of cross referencing. Feel free to use any of the code examples provided or any other resources at your disposal.
2. Upload both the Rmd and HTML files to the course Canvas site (https://canvas.uw.edu/courses/1434040). Upload two separate files rather than creating a zip file.

Guidelines:

Make sure your document includes at least the following:

- Your name and contact information
- Date of creation
- A table of contents
- Sequentially numbered section headers for each question
- Any additional R Markdown elements to make your document easier to navigate, read, and understand. All tables and figures should be captioned and cross-referenced.
- Source code for the Rmd at the end of the document
- For any data driven values reported in the text (e.g., mean of a vector), use the construction
  `` `r somecode` ``
- Pay attention to the number of decimal places you report

Explanation:

This assignment will build your skills in writing R functions. Functions are particularly useful if you will be running the same set of operations multiple times on different data sets that have the same structure. For example if you wanted to make identical bar plots from the Add Health race and ethnicity variables, it would be more efficient to write a function rather than copying-and-pasting a block of code and making a series of small edits.

This exercise uses an example of bootstrapping: sampling *with replacement* from a sample of size *n*, many samples of size *n*.

Figure 1 shows simulated data for a hypothetical study asking graduate students to rate graduate school on a pain scale from zero to 10, where zero is no pain and 10 is the worst pain imaginable. You are going to resample from this data and look at the distribution of the mean.

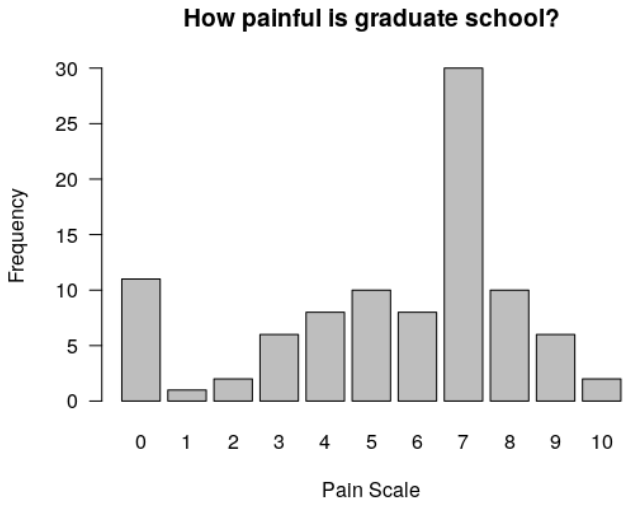

Figure 1: Survey results on pain of graduate school

Create the data using the following R statement:

```
gradpain <- c(rep(0,11), rep(1,1), rep(2,2), rep(3,6), rep(4,8), rep(5,10),
              rep(6,8), rep(7,30), rep(8,10), rep(9,6), rep(10,2))
```

FYI: The plot was created with the following R statement:

```
barplot(table(gradpain), las=1, ylab="Frequency", xlab="Pain Scale",
        main="How painful is graduate school?")
```

Answer the following questions:

1. How many graduate students are in the sample? Use R code to determine this.

2. What is the sample mean?

Box 1
Create a function, with these arguments:
1. the vector of data: "d.vec"
2. the size of the sample: "n"

The function will sample ***with replacement*** a sample of size "n" from the vector "d.vec". The function will return a *list* that contains
1. the size of the sample
2. the mean of the sample

Box 2
Use `set.seed(7)` then run your function passing in the "gradpain" vector calculated above and a sample size of `length(gradpain)`. Use a loop to do this 100 times and store all 100 returned means.

3. What is the mean of these 100 means?

4. What is the standard deviation of these 100 means?

Box 3
Write another function that performs the steps listed in Box 2. That should be a function with these arguments:

1. the vector of data: "d.vec"
2. the size of the sample: "n"
3. the number of samples: "num.samples"

The function should sample ***with replacement*** a sample of size "n" from the vector "d.vec" and does this "num.samples" times.

The function should return a *list* that contains

1. the size of each sample
2. the number of samples
3. a vector of length num.samples with the mean of each sample
4. the mean of the means
5. the standard deviation of the means
6. the 95% confidence interval around the mean

Run your function with the three arguments

```
d.vec = gradpain, n = length(gradpain), num.samples = 100
```

5. What does your function return for the mean of means?

6. What does your function return for the standard deviation of means?

7. What does your function return for the 95% confidence interval around the mean?