COMPSCIX 415.2 Homework 9/Final

Robert Clements

In this assignment/final...

We will review these topics:

- Sampling distributions and standard error
- Functions and loops
- AB testing
- Linear regression

Remember to work on this independently. If you have questions you can reach me on Canvas or through email at robert_clements@berkeley.edu.

Remember to make your document look good, which means you may need to change some settings on the figure sizes and locations, use markdown syntax to create headings or to format your text (use the cheatsheet), and you may want to play with the different themes.

Use complete sentences, and **divide your work in a logical way**. Remember that the whole point of doing reproducible analysis in R Markdown is so that a complete stranger can take your results, *understand them*, and reproduce them.

Remember to save and knit often. Commit when you've completed a big chunk of work or when you are done for the day and will be resuming later.

What to Turn In

For this assignment you have two choices:

You can upload a pdf document (you will have to install latex);
You can upload a standalone html file.

To complete this final you will need...

Access to the internet.
The `tidyverse` and `broom` packages installed.
RStudio and git/Github.

To start your assignment

1. Go to File -> Recent Projects

2. Click on the Project that you created during Homework 1. This project should be the one that is already under git version control.

3. RStudio will switch to that project and the Git pane should appear.

4. Go to File -> New File and choose **R Markdown**.

5. Change the title ("COMPSCIX 415.2 Homework 9/Final") and add your name and the date to the YAML header.

6. Save the file in the same folder (or create a subfolder) with your other HW assignments and give it the name **firstname_lastname_final.Rmd**.

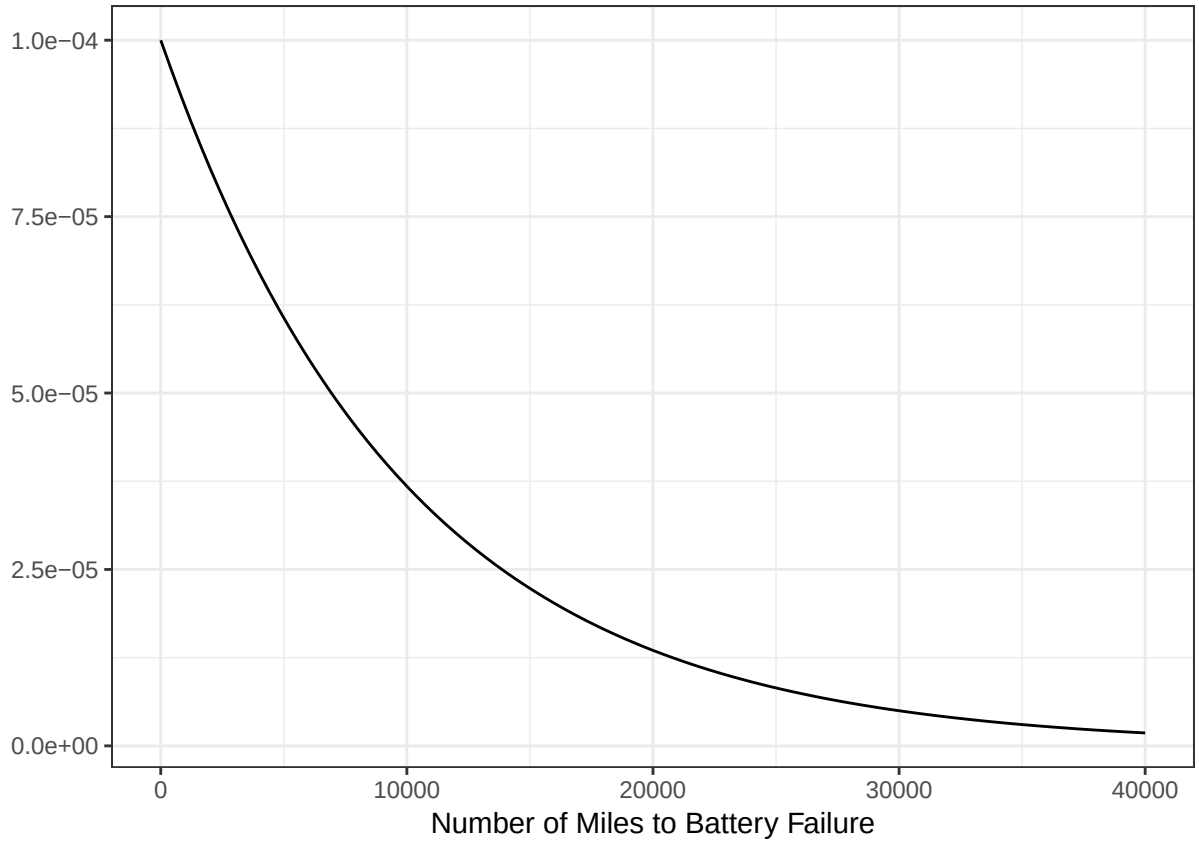
7. Knit your document into an html or pdf document.
8. Go to the Git pane and commit both your Rmd and html (or pdf) files by clicking on the checkboxes next to the file names and hitting the Commit button. Write a useful message, and hit the commit button.

Exercises (Total Points - 20)

Exercise 1 - Sampling Distributions, Functions and For Loops (10 points)

Recall that the distribution of the sample mean is approximately a Normal distribution, and that the standard error is $\frac{\sigma}{\sqrt{n}}$. This holds true regardless of the distribution of our population.

For this problem, assume that the number of miles that a particular car can run before its battery wears out is exponentially distributed with an average of 10,000 miles. The exponential distribution looks like this:



The exponential distribution has a rate parameter that controls how quickly the distribution decays and defines what the mean and standard deviation will be. In our case the `rate = 1/10000`, the `mean = 10000`

and the `standard deviation = 10000`. You can sample from this exponential distribution in R using this code:

```
# sample size
samp_size <- 100
# set the rate parameter
samp_rate <- 1/10000

# take sample
rexp(n = samp_size, rate = samp_rate)
```

STEP 1

Write an R function that does the following:

- Takes a sample of size `samp_size` from this exponential distribution (`samp_size` is an input parameter for the function)
- Calculates the mean of that sample
- Calculates the standard deviation of that sample
- Returns the calculated mean and standard deviation as a list

Helper code

```
samp_fun <- function(samp_size, samp_rate) {

  ...your code here...

  stats <- list(samp_avg = samp_avg, samp_std_dev = samp_std_dev)
  return(stats)
}
```

STEP 2

Then write a loop that does this:

- Runs the above function 1000 times, with `samp_size = 50` and `samp_rate = 1/10000`
- Saves all of the sample means in a vector called `sample_means`, and all of the sample standard deviations in a vector called `sample_sds`

STEP 3

Then

- plot your sample means as a histogram
- output the standard deviation of your sample means
- calculate the theoretical standard error ($\sigma = 10000$, n = sample size)
- calculate the mean of the sample standard deviations and use this to calculate the empirical standard error

STEP 4

Repeat **STEP 2** and **STEP 3** using a sample size of `5000`.

Exercise 2 - Linear Regression (5 points)

For this exercise we will return to the House Prices prediction dataset that we used for HW 7. You should have already downloaded the `train.csv` dataset before, but if you didn't you can download it from Canvas in this week's module.

Load the `train.csv` dataset into R and fit a regression model with:

- `y = SalePrice`
- Features: `LotArea`, `OverallQual`, and `ExterQual`

Answer these questions:

- Use the broom package to output the coefficients and the R-squared
- Interpret the coefficient on `LotArea`
- Interpret the coefficient on `ExterQualGd`
- Compare this model to the model we fit in HW 7 with `GrLivArea`, `OverallQual`, `Neighborhood`. Which is the better fitting model?

Exercise 3 - AB Testing (5 points)

Download the `ab_test_data.csv` file from Canvas. This file contains two columns: `version` and `conversion`. Each row is a visitor to a webpage. The `version` column tells us which version of the webpage the visitor saw, and the `conversion` column is a binary value and equals `1` if the visitor converted (0 otherwise).

We want to perform an AB test on this data to see if the conversion rates are different for the two versions of the webpage.

Answer these questions:

a. What proportion of visitors converted for each version of the webpage?
b. Perform the AB test in R. What is the p-value for the AB test (hypothesis test of proportions)?

Turn in your completed final

This week you should turn in your final by uploading it to Canvas by Friday Aug 10 at Noon.