

# Lab 1: Data Manipulation, Random Number Generation

July 7, 2020

Today's agenda: Manipulating data objects; using the built-in functions, doing numerical calculations, and basic plots; reinforcing core probabilistic ideas.

0. Open a new R Markdown file; set the output to HTML mode and “Knit”. This should produce a web page with the knitting procedure executing your code blocks. You can edit this new file to produce your homework submission.

## Background

The exponential distribution is defined by its cumulative distribution function

$$F(x) = 1 - e^{-\lambda x}$$

The R function `rexp` generates random variables with an exponential distribution.

```
rexp(n=10, rate=5)
```

```
## [1] 0.23363183 0.13336094 0.01150605 0.20757636 0.89988099 0.21226629  
## [7] 0.21183161 0.11614725 0.90649537 0.06865349
```

produces 10 exponentially-distributed numbers with rate ( $\lambda$ ) of 5. If the second argument is omitted, the default rate is 1; this is the **standard exponential distribution**.

## Part I

1. Generate 200 random values from the standard exponential distribution and store them in a vector `exp.draws.1`. Find the mean and standard deviation of `exp.draws.1`.
2. Repeat, but change the rate to 0.1, 0.5, 5 and 10, storing the results in vectors called `exp.draws.0.1`, `exp.draws.0.5`, `exp.draws.5` and `exp.draws.10`.
3. The function `plot()` is the generic function in R for the visual display of data. `hist()` is a function that takes in and bins data as a side effect. To use this function, we must first specify what we'd like to plot.
  - a. Use the `hist()` function to produce a histogram of your standard exponential distribution.
  - b. Use `plot()` with this vector to display the random values from your standard distribution in order.
  - c. Now, use `plot()` with two arguments – any two of your other stored random value vectors – to create a scatterplot of the two vectors against each other.

4. We'd now like to compare the properties of each of our vectors. Begin by creating a vector of the means of each of our five distributions in the order we created them and saving this to a variable name of your choice. Using this and other similar vectors, create the following scatterplots:
  - a. The five means versus the five rates used to generate the distribution.
  - b. The standard deviations versus the rates.
  - c. The means versus the standard deviations.

For each plot, explain in words what's going on.

## Part II

5. R's capacity for data and computation is large to what was available 10 years ago.
  - a. To show this, generate 1.1 million numbers from the standard exponential distribution and store them in a vector called `big.exp.draws.1`. Calculate the mean and standard deviation.
  - b. Plot a histogram of `big.exp.draws.1`. Does it match the function  $1 - e^{-x}$ ? Should it?
  - c. Find the mean of all of the entries in `big.exp.draws.1` which are strictly greater than 1. You may need to first create a new vector to identify which elements satisfy this.
  - d. Create a matrix, `big.exp.draws.1.mat`, containing the the values in `big.exp.draws.1`, with 1100 rows and 1000 columns. Use this matrix as the input to the `hist()` function and save the result to a variable of your choice. What happens to your data?
  - e. Calculate the mean of the 371st column of `big.exp.draws.1.mat`.
  - f. Now, find the means of all 1000 columns of `big.exp.draws.1.mat` simultaneously. Plot the histogram of column means. Explain why its shape does not match the histogram in problem 5b).
  - g. Take the square of each number in `big.exp.draws.1`, and find the mean of this new vector. Explain this in terms of the mean and standard deviation of `big.exp.draws.1`. *Hint:* think carefully about the formula R uses to calculate the standard deviation.