Homework 1: Let's Learn Exponentially

CS 201-A01, Spring 2021

25 January 2021

This homework was adapted from Lab 1 of Carnegie Mellon University's 36-350 course[1].

General instructions for homeworks: Upload both (i) the R Markdown file and (ii) the `nb.html` file to eCampus. You should give the commands to answer each question in its own code chunk. Each answer must be supported by written statements as well as any code used. Include your name in the `author` field of your header.

Note: Your responses must be supported by both textual explanations and code you use to produce your result. Just examining your various objects in the "Environment" section of RStudio is insufficient – you must use scripted commands. If you are unsure what I mean by this, **ask me**.

R Notebook Test

0. Open a new R Notebook file. Delete everything below the second set of `---`. Click "Preview". This should generate a `.nb.html` file, which is the R Notebook produced from your `Rmd` file. You can edit the `Rmd` file to produce the files for your homework submission.

Part I: Background and "Let Me Google That For You"

The exponential distribution is used to model phenomena that can only take positive values with comparatively "light" right tails. The exponential distribution can be defined by its cumulative distribution function

$$F(x) \triangleq P(X \leq x) = \begin{cases} 1 - e^{-\lambda x} & : x \geq 0 \\ 0 & : \text{otherwise} \end{cases},$$

where λ is the **rate parameter** of the exponential distribution.

1. Wikipedia has excellent entries on most of the parametric[2] distributions from statistics. Search for "Exponential distribution" using Google, and click on the search result corresponding to Wikipedia's entry on the exponential distribution. Find the table to right, giving a summary of many of the key properties of the exponential distribution.
    a. What is the mean of the exponential distribution?
    b. What is the variance of the exponential distribution? Its standard deviation?

Part II: Small Data

The R function `rexp` generates random variates from an exponential distribution. For example,

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

produces 10 exponentially-distributed numbers with rate (λ) of 5. If the second argument is omitted, the default rate is 1; this is the "standard exponential distribution''.

[1] Shalizi, C. R. and Thomas, A. C. (2014), "Statistical Computing 36-350: Beginning to Advanced Techniques in R", http://www.stat.cmu.edu/ cshalizi/statcomp/14

[2] A parametric distribution is a probability distribution determined by a *finite* number of constants called **parameters**.

2. 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` using `mean()` and `sd()`.

3. Repeat the previous problem, but change the rates 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`. Find the mean and standard deviation of each new vector.

4. 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.

5. 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 III: Big Data

6. R can generate a **very large** vector of random variates and perform computations on that vector in the blink of an eye.
   a. Generate 1.1 million numbers from the standard exponential distribution (`rate=1`) and store them in a vector called `big.exp.draws.1`. Calculate the mean and standard deviation of `big.exp.draws.1`.
   b. 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.
   c. Use `?matrix` to bring up the documentation for the `matrix()` function. Read about the first four arguments of `matrix()`. What are the first four arguments and what do they do?
   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. Use `?apply` to bring up the documentation for the `apply` function. Read the **Description**, **Usage**, and **Arguments** sections of the documentation for `apply`.
   f. Read the first example with header `## Compute row and column sums for a matrix:` in the **Examples** section of the documentation for `apply`.
   g. Find the means of all 1000 columns of `big.exp.draws.1.mat` simultaneously using `apply` and store this is a vector named `sample.means`. **Hint:** Make sure that `sample.means` is a vector of length 1000.