# STAT 135 Lab 1

January 26, 2015

## Introduction

To complete this lab, you will need to have access to R and RStudio. If you have not already done so, you can download R from `http://cran.cnr.berkeley.edu/`, and RStudio from `http://www.rstudio.com/`. You will have the 2 hours in this lab to begin the lab questions, although you are not required to turn in your lab as they will not be graded. There will be 3-4 assignments throughout the semester (worth a total of 20%) which require knowledge of R, hence if you do not finish the lab during the section, you are encouraged to finish the lab in your own time.

## Question 1

The content of this question refers to the file "Question1.R" located in `https://github.com/rlbarter/STAT135`. This R script contains an example of some code which generates a Bernoulli(0.5) random variable by manipulating a Uniform(0,1) random variable. This question involves editing the code in the "Question1.R" file in order to generate a Bernoulli(0.3, 7) random variable.

1. Edit the code provided to generate a Binomial(0.3, 7) random variable.
   Hint: there are several ways to do this, however, for a particularly simple way, you only need to:
   - change one number in each line of code, and
   - add a final line of code which involves the `sum()` function

   (Recall that a Binomial($p$,$n$) random variable can be considered as the sum of $n$ Bernoulli($p$) random variables).

2. Using your code from the previous part, use a `for` loop to generate 10,000 Binomial$(0.3, 7)$ random variables. Plot a histogram, using `hist()`, of the generated numbers (use this as a check to make sure that your code is working).

3. Use the inbuilt `R` function, `rbinom()`, to generate 10,000 Binomial$(0.3, 7)$ random variables, and plot a histogram of the generated numbers. Compare with the histogram generated in the previous question.

## Question 2

Suppose that there are $n < 365$ people in a room. Ignoring leap years (i.e. ignoring February 29), we are interested in finding the probability that no two people have the same birthday. Assume that all birthdays are

equally probable so that the probability of a given birthday for a person chosen from the entire population at random is 1/365.

1. Show that the probability that no two people have the same birthday is given by

$$P(\text{no two people have the same birthday}) = \frac{365!}{(365-n)!365^n}$$

2. Hence, calculate the probability that two or more people out of a group of $n$ *do have the same birthday.* Hint: identify first the complementary event.

3. Suppose that $n = 50$. Verify this formula in `R` by taking $n$ samples from a Uniform(1,365) distribution (using the `runif()` function, be aware that this function generates a *continuous* uniform RV, so can you think of a way to convert this to a discrete version?) and identifying whether or not two or more people have the same birthday. Repeat this $N = 1000$ times using a `for` loop to estimate the probability that two or more people have the same birthday.

## Question 3

Suppose we have a population $x_1, x_2, ...., x_N$, with population mean $\mu$ and population variance $\sigma^2$. Denote the distinct values assumed by the population members by $a_1, a_2, ..., a_m$, and denote the number of population members that have the value $a_k$ by $n_k$, $k = 1, ..., m$. Then given a sample $X_1, ..., X_n$ drawn from this population, we have that each $X_i$ is a discrete random variable with probability mass function

$$P(X_i = a_k) = \frac{n_k}{N}$$

1. Show that the expected value of $X_i$ is equal to the population mean. That is, show that $\mathrm{E}X_i = \mu$.

2. Show that the variance of $X_i$ is equal to the population variance. That is, show that $\mathrm{Var}(X_i) = \sigma^2$.

To complete the remainder of this exercise, you will need to download the "jester" dataset which can be found at `https://github.com/rlbarter/STAT135`. This dataset contains data from 14,116 users (rows) who have rated each of 30 jokes (columns). The ratings fall between -10 (not at all funny), and 10 (extremely funny). The text for the jokes are provided in the 'jokes' folder.

3. Using the `read.csv()` command, define an object called `jester` containing a data frame corresponding to the data found in the jester.csv file (ensure that your working directory contains the jester.cv file).

4. Read through some of the jokes, and plot histograms of a joke that you found funny and a joke that you did not (note that you can plot histograms using the `hist()` function). Comment on the distribution of ratings for each joke.

5. Calculate and print the mean and standard deviation (in `R`) of the ratings for each joke (hint: the `apply()` function will be very useful here). What can you conclude about the overall response to jokes 43 and 91?

6. Using `R`, calculate the correlation and covariance between the ratings for jokes 43 and 91. Comment on the results.

7. Prove that

$$-1 \leq \rho \leq 1$$

where $\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$ is the correlation between $X$ and $Y$. Hint: to show the lower bound, consider

$$0 \leq \text{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right)$$

and a similar inequality for the upper bound.

8. Using the results of part 6, verify that $\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$, where $X$ corresponds to the ratings for question 43, and $Y$ corresponds to the ratings for question 91.

## Some R tips

For a (free) comprehensive introduction to R, see `http://cran.r-project.org/doc/manuals/r-release/R-intro.pdf`. However, if you're not sure how to use a particular function, you can use the `?` command. For example, to see how to use the `runif()` command, you can type `?runif`.