# Pol Sci 630: Problem Set 3 - Comparisons and Inference

Prepared by: Jan Vogler (jan.vogler@duke.edu)

Due Date: Tuesday, September 15th, 2015, 10 AM (Beginning of Class)

**Note 1: It is absolutely essential that you show all your work, including intermediary steps, and comment on your R code to earn full credit. Showing all steps and commenting on code them will also be required in future problem sets.**

**Note 2: Please use a *single* PDF file created through knitr to submit your answers. knitr allows you to combine R code and LaTeX code in one document, meaning that you can include both the answers to R programming and math problems. Also submit the source code that generates the PDF file (i.e. either .Rnw or .Rmd files)**

**Note 3: Make sure that the PDF files you submit do not include any references to your identity. The grading will happen anonymously. You can submit your answer at the following website:** `http://ps630-f15.herokuapp.com/`

## R Programming

### Problem 1

**Do the following in R:**

**a)** Create a variable $x$ that is a sequence from 1 to 1000 (intervals of 1). Then, create a variable $y$ that is linearly dependent on $x$. Now evaluate their covariance and correlation. Interpret the results. Then, create a second variable $y2$ that is linearly dependent on $x$ but additionally has some normally distributed error. Now evaluate their covariance and correlation. Interpret the results. Finally, interpret the difference between the covariances and correlations.

Hint: In order to create the randomly distributed error, you have to first create a vector of length 1000 with random draws from a normal distribution, then add this vector to $y2$.

**b)** Write a function that returns the correlation between two vectors in R. For this purpose, do not use the covariance (*cov*) or the correlation (*cor*) functions that are built into R. Instead, please refer to the lecture and text book for the mathematical definition of covariance and correlation and emulate those calculations in your function. Demonstrate that your function works by comparing it to the built-in *cor* function in R.

**c)** Copy the function you created for problem B. Now integrate two error messages. The first message should appear if you plug in two vectors of different lengths. The second message should appear if you plug in a non-numeric vector. Demonstrate that your function works by, first, plugging in two vectors of different lengths and, second, plugging in a vector consisting of characters.

## Problem 2

**Do the following in R:**

**a)** Create two vectors of length 50 with random draws from a Poisson distribution. The mean of the first vector should be 10, the mean of the second vector should be 12. Conduct a two-sample t-test for those two vectors in R.

Hint: If you don't know how to draw from a Poisson distribution, try to find out through a search engine.

**b)** Write a function that allows you to perform a two-tail, two-sample t-test, using two vectors as input. In your calculation, assume that the two groups have different variances. The function is meant to return a t-value. Test the function by plugging in the two vectors you created for part a of this problem and see whether you get the same result as with the built-in *t.test* function.

Note: To solve this problem, you're allowed to use built-in functions inside your function, except for the t.test function itself.

**c)** Include an error message if any vector plugged in is not numeric.

## Problem 3

**Do the following in R:**

**a)** Load the *swiss* dataset in R via the command *data(swiss)*. According to the documentation this is data on "Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888".

**b)** We are interested in how the level of "Education" is correlated with other measures. Use the lm command to run a regression of education on all other variables in the dataset. Note that R is case-sensitive.

**c)** Interpret the results for every variable in the linear model. Be specific about the magnitude and direction of the linear relationships and the meaning of p-values. Say at which levels of significance each relationship is statistically significant.

Bonus question: What can you say about causality with respect to the statistical relationships identified? Explain your answer.

# Probability Theory: Covariance and Correlation

## Problem 4

**Do the following problems. Show every step. For all of the following problems, be aware of the following mathematical definitions:**

**1.** $Cov(X,Y) = \mathbb{E}(X*Y) - \mathbb{E}(X)*\mathbb{E}(Y)$
**2.** $\mathbb{E}(X) = \sum_{i=1}^{n} x * Pr(X = x)$
**3.** $\mathbb{E}(X*Y) = \sum_{i=1}^{n} x * y * Pr(X = x, Y = y)$

**a)** This problem is taken from Pitman (1993) Probability: Let $(X,Y)$ have uniform distribution on the four points $(-1,0),(0,1),(0,-1),(1,0)$. Show that X and Y are uncorrelated. Then prove that they are not independent.

**b)** This problem is taken from Pitman (1993) Probability: Let X have uniform distribution on -1,0,1 and let $Y = X^2$. Are X and Y uncorrelated? Are X and Y independent? Explain carefully.

**c)** This problem is taken from Pitman (1993) Probability: Let $X_1$ and $X_2$ be the numbers on two independent fair six-sided die rolls, $X = X_1 - X_2$ and $Y = X_1 + X_2$. Show that X and Y are uncorrelated, but not independent.

Hint: For this problem, it might be most convenient to use a for loop in R to calculate the covariance. The solution will be based on an R code, too. It is also possible to calculate this manually though.