MSAN 504 — Probability/Statistics — Summer 2015
Homework Three

1. Download the sets of data `AAPLSP50020022006.csv` and `AAPLSP50020072011.csv`. Convert the price data in these CSV files to return data, i.e., compute the returns on both Apple stock and the Standard and Poor's Index according to the logarithmic return formula $\log(p_t/p_{t-1})$, where by log we mean the natural logarithm. Then test $H_0 : \rho_{\texttt{20022006}} = \rho_{\texttt{20072011}}$ against the two-sided alternative at the usual level of significance $\alpha = 0.05$.

2. Create R code that implements the acceptance-rejection method. Use this method to simulate the variates from a gamma distribution with parameters $\alpha = 2$ and $\beta = 1.5$. Choose an exponential distribution as the prospective, or auxiliary, random variable, and make sure that the means of the two variables match one another. What is the optimal constant c used to govern the acceptance or rejection of prospective realizations from the exponential random variable? To generate 10000 realizations from the gamma distribution, how many realizations must you generate from the auxiliary exponential random variable? Upload your R code to Canvas. In your write-up, include a histogram of the realizations you generated for the target gamma random variable.

3. Modify your code for #4 to create realizations from a standard normal random variable. Use a Cauchy random variable, i.e., a random variable with density function $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$ for $-\infty < x < \infty$, as the prospective random variable. What is the optimal constant c used to govern the acceptance or rejection of prospective realizations from the Cauchy random variable? To generate 10000 realizations from the standard normal distribution, how many realizations did you have to generate from the auxiliary Cauchy random variable? In your write-up, include a histogram of the realizations you generate for the target standard normal random variable.

4. Download the data set `EffectOfGenderBodyTemperaturesAndRestingHeartRate.csv`. Use the sign test to determine whether or not the true mean body temperature (over all genders) is equal to 98.6 degrees Fahrenheit. Formulate statements of null and alternative hypothesis carefully, being sure to state a conclusion in the context of the research question.

5. Reconsider the data set `DartsVersusExperts.csv`. Follow the same data preparation procedures you used in the last homework assignment. However, now use the signed rank test to test the null hypothesis that there is no outperformance by investment professionals of stock selections chosen at random by journalists at the *Wall Street Journal*.

6. We found that males and females have statistically distinguishable body temperatures when we used a two-sample student t test. Explore this issue again, being sure that you carefully formulate statements of null and alternative hypothesis, using the Mann-Whitney U test, i.e., the rank-sum test.

7. Use R's native functionality for simulating realizations from a chi-squared distribution with 1 degree of freedom. Create 5000 realizations of three independently-generated chi-squared random variables with one degree of freedom (i.e., a total of 15,000 realizations). Add together the three independently-generated realizations, to obtain a total of 5000 realizations. Plot the result in a histogram, and overlay the density function of a chi-squared random variable with three degrees of freedom. Explain what you're examining.

8. **A Randomization Test for a Correlation Coefficient.** In the following exercise, we will learn how to execute a **randomization test** of the null hypothesis $H_0 : \rho = 0$.

8(a). In research by I.T. Elo, G. Rodriguez, and H. Lee, published in 2001 in the *Proceedings of the Annual Meeting of the Population Association of America*, a random sample of all live births occurring in Philadelphia, Pennsylvania in 1990 was obtained. The researchers studied the interactions between variables like the mother's race, smoking habits, educational level, as well as the "gestate age" (estimated number of weeks after conception before the baby was born) and the baby's birth weight (measured in grams). We could explore the usefulness of the mother's educational level (measured in years) as a linear predictor of the baby's birth weight. Do better-educated mothers tend to have heavier (and therefore possibly healthier) babies? Download the data set `MotherEducationBirthWeight.csv` from Dropbox and import it into R. Then, compute the sample correlation between the two sets of data. Superficially, does it seem like a mother's education is related to the birth weight of her baby?

Like every other statistic that is subject to sampling error, your estimate of the true correlation between the number of years of formal education possessed by the mother and the birth weight of the baby is subject to sampling error. We could run a hypothesis test using the material from Chapter 7 in Hogg and Tanis, i.e., use the so-called "normality-based" approach to hypothesis testing. Another response to this situation would be to directly simulate the behavior of $\widehat{\rho}$ when the null hypothesis is true, and then to compare the original $\widehat{\rho}$ to the simulated results. The next few steps will walk you through this process.

8(b). Create a function that preserves the data set but smashes the relationship between the independent data (mother's educational level) and dependent data (baby's birth weight). To do so, fix the column of birth weights but randomly permute the mother's educational level. Once you scramble the column with educational level, "scotch tape" the two columns back together and compute the statistic $\widehat{\rho}$ for the "scrambled" or "permuted" data set. Notice now that the independent variable assignments are effectively randomized against the birth weights, i.e., the null hypothesis $H_0 : \rho = 0$ has been forced to be true.

8(c). Execute the function you created in part (b) a total of 25,000 times and make a histogram of the resulting $\widehat{\rho}$. Comment on the shape of the distribution. We call this distribution the **randomization distribution** under $H_0 : \rho = 0$.

8(d). Plot the value of $\widehat{\rho}$ for the original, non-permuted data set as a vertical line and overlay it on the histogram you created for part (c). Compute the relative proportion of the $\widehat{\rho}$ generated by the permutations that are as extreme, or more extreme than, the value of $\widehat{\rho}$ from the original, un-permuted set of data. Call this quantity the **empirical p value**. If the alternate hypothesis is one-tailed, i.e., if $H_1 : \rho > 0$ or $H_1 : \rho < 0$, then compare this empirical p value to the significance threshold α and reject H_0 if it less than α. If the alternate hypothesis is two-tailed, i.e., if $H_1 : \rho \neq 0$, then compare **twice** this empirical p value to the significance threshold α and reject H_0 if the doubled empirical p value is less than α. Write a conclusion on the context of the implied research question. Make sure to state why you chose a one-tailed or two-tailed alternate hypothesis.

9. A market researcher investigates the effects of a lavender odor on customer behavior in a small restaurant. Lavender is a relaxing odor. The researchers also looked at the effects of lemon, a stimulating odor. In the file `OdorTimesInRestaurant.csv`, you will find the time (in minutes) that customers spend in the restaurants when there is no odor (the first column) and when lemon odor was present (the second column). Perform an appropriately-chosen nonparametric test to determine whether or not customers tend to linger for less time when they are exposed to the lemon odor (i.e., the alternate hypothesis is one-sided).

10. To investigate water quality, on August 8, 2010, the *Columbus Dispatch* took water samples at 20 Ohio state Park swimming areas. Those samples were taken to laboratories and tested for fecal coliform, which are bacteria found in human and animal feces. An unsafe level of fecal coluform means there's a higher chance that disease-causing bacteria are present and more risk that a swimmer will become ill. Ohio considers it unsafe if a 100-milliliter sample (about 3.3 ounces) of water contains more than 400 coliform bacteria. The fecal coliform levels in these swimming areas are in the file `FecalColiformLevels.csv`. Select and use an appropriate nonparametric test to determine if $H_0 : \mu = 400$ should be rejected in favor of an appropriately-chosen one-sided alternative.