MSAN 504 — Probability/Statistics — Summer 2016 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 $H_0 : \rho_{\mathsf{20022006}} = \rho_{\mathsf{20072011}}$ 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 related to male and female resting heart rates and body temperatures. Test the null hypothesis that men and women have the same average body temperature at the usual level of significance. Select an appropriate alternative hypothesis. Show all calculations. Be sure to make an appropriate argument as to whether or not the variances of the body temperatures belonging to men and the body temperatures belonging to women can be regarded as equal. **Data Set:** `EffectOfGenderBodyTemperaturesAndRestingHeartRate.csv`

5. From the data set in problem #4, compute a 90% confidence interval for the true mean difference in male and female resting heart rates. Notice that you must also make an argument concerning whether or not we are in a "variance equal" or a "variances unequal" situation here.

6. People gain weight when they take in more energy from food than they expend. James Levine and his collaborators at the Mayo Clinic investigated the link between obesity and energy spent on daily activity. Twenty health volunteers were chosen who don't exercise. Ten were chosen that were lean and ten others were chosen who were mildly obese but otherwise healthy. Sensors were attached to the subjects that more or less monitored all of their movements for a 10-day period. The subjects' times spent standing/walking, sitting, and lying down were measured. The following sample statistics related to the number of minutes spent standing/walking were computed:

Group	n	$\overline{x}$	s
Group 1 (lean)	10	525.751	107.121
Group 2 (obese)	10	373.269	67.498

Is it appropriate to assume the variances of these two groups are equal? Why or why not? Test the null hypothesis that lean and obese folks spend the same number of minutes standing or walking at the $\alpha = 0.01$ level.

7. Compute a 95% confidence interval for the true difference in the number of minutes spent standing/walking by lean and obese people, using the information from problem #7.

8. A data set is available that has the daily returns of IBM, Microsoft, Home Depot, Exxon Mobil, and Apple. At the 10% significance level, test whether or not Exxon Mobil and Home Depot have statistically indistinguishable variances on their daily stock returns. **Data Set:** `DailyReturnsForFiveStocks.csv`

9. Recall how to compute a confidence interval for the ratio of two variances. Then take the above data and compute a 95% confidence interval for the true ratio of the variance of Exxon Mobil daily stock returns to the variance of Home Depot stock returns.

10. Download the data set related to contagious yawns. Is there evidence, at the $\alpha = 0.02$ significance level, in favor of the null hypothesis that there is no difference in yawning rates between those who are exposed to a "yawn seed" and those receive no such stimuli? **Data Set:** `EffectOfSeeingAYawnOnYawning.csv`

11. For the same experiment which, hilariously, was performed not on humans but on tortoises, compute an 80% confidence interval for the true difference in yawning rates if subjects (tortoises, in this case) do and do not receive the "yawning seed." **Data Set:** `EffectOfSeeingAYawnOnYawningOnTortoises.csv`

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

12(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.

12(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.

12(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 $H_0 : \rho = 0$.

12(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.