Homework 1

Due Monday July 9, 2012 *(before class starts!)*

The assignment looks long, but it shouldn't take you more than 6 hours to complete. The purpose of homework 1 is to get you introduced to R and to review concepts we learned in lecture 1 (population/sample), 2 (summarizing data), and 3 (properties of sample statistics). Bonus questions are optional and partial credit will be given to those who attempt them *honestly*.

1. Your instructor wants to know how his age is perceived amongst the students, staff, and faculty at UPenn. To conduct this study, he asks Stat 431 students how old he is and records the response. Question 5 will take you roughly 30 minutes. Question 6 will take you about an hour to complete.

   (a) What is the population for this study? Briefly justify your response.

   (b) What comprises the sample for this study? Is the sample representative of the population? Why or why not?

   (c) What parameters is your instructor interested in? Name at least two parameters of interest and explain why they are meaningful.

   (d) What statistics should your instructor use to estimate the parameters of interest?

2. Please download the Facebook data set from the course website (fb.txt). This data set is a small subset of your instructor's friends on Facebook. This question will guide you through using R.

   (a) Open the data set in R! This may be the most frustrating question or the easiest question in this entire homework. How many individuals are in the data? How many measurements are taken per individual?

   (b) Summarize the number of friends your instructor's friends have on Facebook with summary statistics (e.g. mean, standard deviation) and other relevant visuals.

   (c) Is the number of friends normally distributed? Justify your answer with relevant plot(s).

   (d) How many non-US students are your instructors friends with? How many females :) are your instructor friends with, in comparison to males? What type of data are you summarizing in this question?

   (e) Is there any relation between the number of friends and the year in which your instructor's friends graduated from high school? Explain with relevant plots and numerical summaries.

   (f) *(Bonus question)* Come up with an R script or any script to pull this kind of data from Facebook. Present this data to your instructor for awesome brownie points!

3. *This is a follow-up to question 2.* Your instructor wants to reformat how the years are written in the Facebook data. Specifically, he wants to write them as "10" instead of "2010". However, he is concerned whether reformatting the data this way will be a problem when he's computing summary statistics. Show *mathematically* that reformatting them will not change the following summary statistics: sample variance and sample correlation (like question 2e). Test your mathematical reasoning by computing the sample variance/sample correlation for the original and formatted years

   Hint: Let's consider sample variance. Let X_i denote the unformatted scale and X_i' denote the formatted scale. You must show that the sample variance from $(X_1, ..., X_n)$ are identical to that from $(X_1', ..., X_n')$

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i' - \bar{X}')^2$$

   Work out the relation between X_i and X_i' and show that $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ leads to $\frac{1}{n-1}\sum_{i=1}^{n}(X_i' - \bar{X}')^2$.

4. *(Bonus Question)* Show that any linear transformation of your data from $(X_1, ..., X_n), X_i = \begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} \in \mathbb{R}^2$ to $(Y_1, ..., Y_n), Y_i = \begin{pmatrix} Y_{i,1} \\ Y_{i,2} \end{pmatrix} \in \mathbb{R}^2$ where $Y_{i,1} = a_1 + b_1 X_{i,1}$ and $Y_{i,2} = a_2 + b_2 X_{i,2}$ for constants a_1, a_2, b_1 and b_2 gives the same sample correlation. This property is known as *linear invariance*

   Fun fact: This mathematical fact has huge consequences when you are cleaning up your data or converting your measurements from one scale to another. You may have come across this scenario when you were doing data analysis for your research projects.

5. In this question, you'll explore how skewness and tail-behavior presents itself in a normal QQ plot and get a taste of simulating data. When simulating data, make sure you generate at least 100 data points.

   (a) Simulate a uniform distribution between 0 and 1 in R and make a normal QQ plot.

   (b) Simulate a right-skewed and left-skewed data and make a normal QQ plot. Describe any patterns you see.

   (c) Simulate a fat-tailed and skinny-tailed data which are both symmetric. Make a normal QQ plot and describe any patterns you may see.

6. *This question hopes to inspire you how awesome mathematical theory can be in statistics. In particular, your instructor wants to demonstrate that a health combination of numerical analysis (i.e. computing sample means) along with mathematicla justification can go a long way in solving real-life problems.*

   Quinnipiac Univeristy Polling Institute conducts polls on a wide range of issues. During every Presidential election season, Quinnipiac polls frequently in "swing states" to gauge the likelihood of a candidate getting elected. For example, the poll asks registered voters the following question:

   If the election for President were being held today, and the candidates were Barack Obama

the Democrat and Mitt Romney the Republican, for whom would you vote?
a. Obama
b. Romney
c. Someone else
d. Wouldn't vote
e. Don't know/NA

In this question, we're going to predict the margin between Obama and Romney on the next polling result (due to be out around July 9th)

(a) What is the populuation under study? What is the parameter of interest? What should our statistic aim to estimate?

(b) Suppose we assume the poll is conducted correctly and the sample is representative (i.e. we collect responses from an i.i.d sampled individual and record his/her preference as X_i where X_i can take on values "a","b","c","d", and "e". If you were the statistician working at Quinnipiac, how would you summarize this data?

(c) To make it easier to conduct the mathematical analysis, we'll assume that there are only two choices, Obama and Romney, and X_i would take on values 1 for Obama and 0 for Romney from now on. Show that, with large sample size, the sample proportion of those who would vote for Obama would converge to the true proportion of individuals who would vote for Obama in Pennsylvania. Show that this holds even if the population is not normally distributed (see lecture 3 slides for details) .

(d) What is the limiting distribution of the sample proportion? Prove your answer with mathematical theory. *Hint: use the Central Limit Theorem!*

(e) Come up with a statistic that estimates the margin between Obama and Romney. Prove that your statistic, with large enough sample size, is close to the true margin between Obama and Romney for *any* distribution of the population. Also, prove that your statistic is unbiased for the margin.

(f) What is the limiting distribution of your sample statistic in (e)? Make sure this distribution does not depend on the distribution of the population!

(g) Quinnipiac had the following margins for the past three pollings (see Table 1).

Polling date	Obama	Romney	Margin	Sample size (for Obama and Romney)
Quinnipiac (6/25/2012)	45%	39%	6	$n = 1,052$
Quinnipiac (6/10/2012)	46 %	40 %	6	$n = 857$
Quinnipiac (5/1/2012)	47 %	39%	7	$n = 1,016$

Table 1: Table of previous polls done by Quinnipiac

Quinnipiac used the difference in proportions as its estimate for the margin. Assuming that the true proportion of Obama and Romney supporters are p_O and p_R, respectively, for all three surveys, which estimate of the margin has the lowest variance? Why?

Hint: Note that $p_O + p_R$ does not equal to 1 and that p_O and p_R are TRUE proportions, not SAMPLE proportions! To tackle this problem, first, we can treat them as $p_O + p_R + p_{NA} = 1$ where p_{NA} represents individuals who did not choose Obama or Romney. From here, go back to the example about balls and urns from Stat 430 to derive the necessary estimates for variance

(h) (Bonus question) Come up with a statistic that combines different estimates of the true margin bewteen Obama and Romney (like in (g)). Prove that your statistic is unbiased. Based on your statistic, can you suggest the optimal sample size for the next Quinnipiac poll?