Homework #1

1. A questionnaire was sent to 34,000 randomly chosen individuals living in USA asking about their current and past smoking habits. For those who answered the questionnaire, they are tracked in the next 10 year for the death rate. Their data was pulled from the health record database so that if they died, the death is recorded with the diagnosis of the cause of the death. The death rates due to heart disease among the smokers and nonsmokers are then compared to see if smoking causes heart disease.

What is the targeted population? What is the sampling population? What is the sample? What are the parameters we are interested in? What are the statistics? Is this an observational study or controlled experiment? Is it possible to draw conclusion from this study that smoking causes heart disease? Why and why not?

2. Each month, University of Michigan conducts phone interviews (starting 2014, the interviews are for cell phones only) for at least 500 households in continental United States (i.e., excluding Alaska and Hawaii). Based on the answers to a 50 questions questionnaire, a consumer sentiment score is calculated for each household. The mean score is published in the index of consumer sentiment by University of Michigan.

What is the targeted population? What is the sampling population? What is the sample? What is parameter we are interested in? What is the statistic?

3. Mr. Ironheart came in to rescue the Troubleland Company. He immediately announced a massive layoff the day he took over as the Chairman. He said that the layoff is necessary because the payroll of this company has run out of control. In the last two years, he claimed, *the average payroll had increased by 17%*. The employee union's President denounced the layoff decision, and countered that *the average employee's income increased merely 1% over the last two years*, far below the inflation. "The best way to cut the payroll," the President said, "is to fire (Mr. Ironheart) himself, who is paid over 80 million dollars a year to take over the CEO position." Which of these two is using the mean and which is using the median when discussing the payroll increase? And who do you think is right?

4. A researcher believes that a college education is a waste of time. To prove this, she collected data on the cumulated wealth of every member of the Harvard Class of 1977 (and those dropouts supposed to graduate in 1977) in 25 years after graduation (wealth gained in the period of year 1977-2002). The average cumulated wealth of the dropouts is much higher than that of the graduates. She concludes from this data that finishing the college education is not as important as being admitted to the college. What do you think? (Bonus point: can you guess why the average wealth of the dropouts is higher than the graduates?)

====== Following are problems to do in R ==========

5. Collect your response time data for the mini-project. (See lab1 instruction). Then produce the following:

a) A scatterplot of your response times versus the observation order;
b) A stratified scatterplot of your response times in three phrases each of length 10;
c) A printout of the program you used in producing the above two plots.
d) From your plots, do the response times seem to come from a stationary process?

6. Analyze the salary data set, contained in the 'salary.txt' file which give the weekly salaries and gender for production workers in Anaheim, California. (This is not the 'fsalary.txt' file used in the example.)

(a) Produce a histogram of the salaries using R default setting.
(b) Produce a histogram use your own break points (at least 15 intervals).
(c) Produce a boxplot of the salaries.
(d) Produce side by side boxplots of the salaries in two gender groups.
(e) Produce summary statistics (as those in 'psych' package) of the salaries as one group, and also summary statistics within each gender.
(f) Submit the two histograms, two boxplots, and the summary statistics, as answer for each part (a-e). From the above plots, which histogram (a) or (b) shows the feature in the data set better? Do the salaries for men and women seem to be the same?
(g) Can you use one number to summarize the center of the distribution of the salaries? If yes, what is the number? If not, why?
(h) Are there any outliers in the data? If there are, can you identify them?

7. Answer the following questions for the percentages of low birth weight infants contained in file unicef.txt under the variable name lowbwt.
(Hint: you may run into trouble importing the data. The first row in unicef.txt contains an extra '*' in it. You may edit the unicef.txt file to delete the '*'. Also, the file contains missing data, denoted by '.'. R code the missing values as 'NA'. Set na.strings = "." inside read.table() to deal with these missing values.)

(a) Produce a histogram and a boxplot.

(b) Does the data appear to be skewed? If so, skewed to right or to the left?

(c) Are there any outliers?