USP 634 Data Analysis I Spring, 2017
Assignment 1: Descriptive Statistics; Probability and Distribution
Due Tuesday, 4/25. Submit your homework to course D2L dropbox

The purpose of this assignment is for you to:

- Identify variables as numerical and categorical and apply appropriate descriptive statistics to each type of variables;
- Explore and describe the relationship between a pair of variables;
- Critique research design and sampling methods;
- Apply basic knowledge of probability and distribution;
- Develop familiarity with statistics software of your choice and your own dataset.

You are encouraged to explore your dataset and the software beyond just the requirements of this assignment.

1. (70 points) Descriptive analysis for your dataset

1) Determine the type for all variables (if total number of variables in your dataset is less than 6; otherwise select 6 variables including both continuous and categorical variables) in your dataset. Select 2 continuous and 2 categorical variables for the following exercises.
2) For each of the 2 continuous variables, calculate these summary statistics:
    a. mean;
    b. mode;
    c. median;
    d. range;
    e. interquartile range;
    f. variance;
    g. standard deviation.

    And then show with appropriate graphs and describe the distribution of each variable. Which of these two variables resembles the normal distribution more closely?
3) For the variable more normally distributed in 2), calculate the Z-scores for all the observations in your dataset. Next, choose two observations and identify the Z-scores for them. Assuming that this variable is normally distributed (even if it isn't), what proportion of observations would be predicted to lie between these two Z-score values? How does this prediction vary from the actual number of observations, and why?
4) Show with appropriate graphs and describe the relationship between the two continuous variables. Are they dependent? If so, positively or negatively?
5) For each of the 2 categorical variables, show with appropriate graphs and tables and describe its distribution.
6) Show with appropriate graphs and describe the relationship between the 2 categorical variables. Are they dependent?
7) Select one continuous variable and one categorical variable, show with an appropriate graph and describe the relationship between the two variables.

2. (10 points) What is the probability of rolling two 5s with two fair dice? What is the probability of rolling snake eyes (two ones) twice in a row, followed by a four and a six, followed by a score adding to 10?

3. (5 pts) In 2005, the average annual ozone levels in Smogsville were normally distributed with a daily mean of 100 ppb (parts per billion) and a standard deviation of 25 ppb. How many days in 2005 were smog levels either above 75 ppb (their air quality standard) or below 50 ppb?

4. (15pts) Inferring the direction and existence of causal relationships from observational data is plagued by selection bias, reverse causality, and confounding variables (a third variable or a number of other variables, influence both explanatory and response variables). The following empirical patterns have been cited in press reports as potential evidence of causal relationships.

- Oakland is considering a Fresh Food Financing program that incentivizes grocery stores to locate in East Oakland. This program is based on studies showing that residents of neighborhoods without stores selling fresh foods have an unhealthy diet.
- Two percent of residents in Fresno, CA bike to work while eight percent bike in Berkeley. Berkeley has 50 more miles of bike lanes on their roads than Fresno. Therefore, if Fresno were to add more bike lanes its bike ridership would increase.
- A recent study in Minneapolis found that people who live in neighborhoods where the majority of houses have porches are more likely to talk to their neighbors at least once a week in comparison with people who live in neighborhoods where there are few porches. To encourage social cohesion in neighborhoods, Minneapolis is therefore considering a new grant program to help people add porches to their houses.

All three empirical patterns are seen in observational (non-experimental) data. Can you apply any of the criticisms of non-experimental empirical results to these three examples? If these criticisms were true, how do they alter interpretation of these patterns?