USP 634 Data Analysis I Spring, 2018
Assignment 1: Descriptive Statistics; Probability and Distribution
Due Tuesday, 4/24. Submit your homework to D2L/Assignments for USP 634

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 of your dataset

First determine the type for all variables in your dataset (if total number of variables in your dataset is less than 6; otherwise select 6 variables including both continuous and categorical variables). Select 2 continuous (you may use ordinal categorical variables if you don't have any any numeric variables) and 2 categorical variables for the following exercises.

1) 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.

2) Show with appropriate graphs and describe the distribution of each continuous variable. Which of these two variables resembles the normal distribution more closely? How can you tell?
3) Show with appropriate graphs and describe the relationship between these two continuous variables. Are they dependent? If so, positively or negatively?
4) For each of the 2 categorical variables, show with appropriate graphs and tables, and describe its distribution.
5) Show with appropriate graphs and describe the relationship between the 2 categorical variables. Are they dependent?
6) Select one continuous variable and one categorical variable, show with an appropriate graph and describe the relationship between the two variables.
7) Among the exercises above, what do you think is the most interesting relationship you find between a pair of variables in your data? Why?

2. (10 points) Data collected at elementary schools in DeKalb County, GA suggest that each year roughly 25% of students miss exactly one day of school, 15% miss 2 days, and 28% miss 3 or more days due to sickness.

a) What is the probability that a student chosen at random doesn't miss any days of school due to sickness this year?
b) What is the probability that a student chosen at random misses no more than one day?
c) What is the probability that a student chosen at random misses at least one day?
d) If a parent has two kids at a DeKalb County elementary school, what is the probability that neither kid will miss any school? Note any assumption you must make to answer this question.
e) If a parent has two kids at a DeKalb County elementary school, what is the probability that both kids will miss some school, i.e. at least one day? Note any assumption you make.
f) If you made an assumption in part (d) or (e), do you think it was reasonable? If you didn't make any assumptions, double check your earlier answers.

3. (5 pts) The average daily temperature in Orangetown is approximately normally distributed with a mean of 75F and a standard deviation of 15F. What is the probability for the average daily temperature to be below 55F or above 90F, which may damage orange crop? How many days in a year do the average daily temperature stay in the range of 55-90F?

4. (15pts) Inferring the direction and existence of causal relationships from observational data can be 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?