Problem Set 1: Review basic Concepts

This problem set is meant to refresh some basic ideas and concepts you ought to have learned in your previous statistics classes. For every question, I have indicated the relevant section in the textbook for your reference. Some of these questions are very open-ended and you may want to be creative. There may be no one right/wrong answer.

Problem 1

A study was conducted about the smoking habits of some UK residents. Below is a data frame displaying a portion of the data collected in this study. A blank cell indicates that data for that variable was not available for a given respondent.

						amount	
	sex	age	marital_status	gross_income	smoke	weekend	weekday
1	Female	61	Married	2,600 to 5,200	No		
2	Female	61	Divorced	10,400 to 15,600	Yes	5	4
3	Female	69	Widowed	5,200 to 10,400	No		
4	Female	50	Married	5,200 to 10,400	No		
5	Male	31	Single	10,400 to 15,600	Yes	10	20
...	...	...	...	...	...		
1691	Male	49	Divorced	Above 36,400	Yes	15	10

a. How many variables (list them) and how many observations/cases are in the data frame? (see section 1.2)

b. For each variable identified in (a) above, state whether it is numerical or categorical. If numerical, state whether it is continuous or discrete. If categorical, state whether it is nominal or ordinal. (see section 1.2.2)

c. Assuming that someone's **age** affects the **amount** of cigarettes they smoke per day, which variable is the explanatory variable and which is the response variable? (see section 1.2.3)

d. What visualizations would you use to display the relationship between **age** and **amount** of cigarettes smoked per day? Which variable would you place on the y-axis and which one on the x-axis? Does it matter what axis you place the variables? (see section 1.3.1)

Problem 2

The first step in conducting research is to identify topics or questions that are to be investigated. A clearly laid out research question is helpful in identifying what subjects or cases should be studied and what variables are important. It is also important to consider how data are collected so that the data are reliable and help achieve the research goals. Suppose you are conducting a study trying to answer the following research question: ***Does a new drug (call it drug X) reduce the number of deaths in patients with severe heart disease?***

a. What type of study (experimental or observational) would you conduct? Explain why? (see section 1.4)

b. Briefly describe how you would set up your study and the data you would collect to answer the above question. In your description describe the variables and their types (i.e., categorical or numerical). (see sections 1.2.1 - 1.2.2 and section 2.2)

c. Suppose you wanted to conduct your study in Tompkins County. Explain what your population of interest? What would be your sample and how would you obtain it? (see section 2.1 and 2.2)

d. What are some potential ethical concerns you would need to consider before conducting your study?