MIDS 203: Exploring and Analyzing Data
Summer 2015
Lab 1

Please construct a report that addresses all parts of Lab #1. This report should be a pdf file. At the bottom of this report, include an appendix with your R-script.
NOTE: This report should be one single document.

You will receive a web-based link from bCourses. This link will include all questions from Part 1 as well as questions (where applicable) from Parts 2ab.

As a suggestion, complete your report first, and include your answers for all parts in the report. For multiple choice questions, only the letter is required and nothing more than that is going to be read.

Then access the web-based link from bCourses and answer all of the questions found there. Finally, be sure to upload your report to bCourses (you will be able to do so at the end of the bCourses quiz/assignment).

Note that your report must be generated, or look as if it is generated, by R Markdown.

Part 1. Multiple Choice (5 points each)

1. In a survey, a question asks how many pets you have, with four possible responses: 0, 1 to 2, 3 to 5, and 6 or more. What type of variable does this question produce?
   a) Interval
   b) Dichotomous
   c) Normal
   d) Ratio
   e) Ordinal

2. Many scientists view the concept of gender as a wide spectrum based on biological and social factors. In light of this, the choice to measure the concept of gender using a strict male/female dichotomy is an example of:
   a) measuring a potentially interval or ordinal variable as a binary variable
   b) measuring a nominal variable as an interval variable
   c) giving priority to psychological over social variables
   d) the ecological fallacy
   e) giving priority to conceptual over operational variables

3. Which of the following is a benefit of using standard deviation as a measure of dispersion?
   a) Standard deviations are unaffected by outlying data points.

b) The chance that a single draw from a population falls within one standard deviation of the mean is always the same for any population.
c) Standard deviations are unaffected by multiplying a variable by a constant.
d) Standard deviations can be directly compared to the individual deviation of one data point away from the mean.

4. Your friend thinks that gender and place of residence influence perceptions about health care legislation. She obtains a list of voters in the state of North Carolina and divides the list into subpopulations of men who live in small towns, men who live in large towns, women who live in small towns, and women who live in large towns. She then randomly selects 500 people from each subpopulation. What kind of sampling procedure is your friend using?

a) Non-probability quota sampling
b) stratified random sampling
c) social network sampling
d) cluster sampling
e) nonrandom sampling
f) systematic random sampling

5. Suppose that weekly beer consumption among the San Francisco population is normally distributed, with a mean of 50oz. Which of the following is more likely to occur:

a) Choosing one San Franciscan at random and finding that they drink over 70oz of beer a week.
b) Choosing 100 San Franciscans at random and finding that they drink an average of over 52oz of beer a week.
c) a and b are equally likely.
d) It depends on the standard deviation of the population.

6. Why is the Central Limit Theorem Important to data scientists?

a) For large samples, it guarantees that a sample mean approaches the true population mean.
b) For large samples, it suggests that the normal distribution is a good model for the distribution of the mean and other statistics.
c) For large samples, it suggests that the sample distribution of a variable approaches the population distribution of that variable.
d) When a population distribution is normal, it tells us that the sampling distribution of the mean will also be normal.

7. When measuring a single metric variable in a population, increasing your sample size tends to lead to which of the following outcomes:

a) A larger variance of the sample distribution of the variable.
b) No change in the standard error of the mean.

c) A smaller variance of the sampling distribution of the mean.
d) A smaller standard deviation of the population distribution.
e) A less-normal sampling distribution of the mean.
f) None of the above.

8. Say you collected data from 15 fellow MIDS classmates. One of your measures included age. Surprisingly, all 15 of your fellow classmates are of the same age: 30 years old. Given this information, which of the following statements below is false.

a) The mean, median, and mode for your age variable are equal.
b) The distribution of your age variable is unimodal.
c) The variance and standard deviation of your age variable are both exactly zero.
d) The distribution of your age variable is platykurtic.
e) The sum of squared errors for your age variable is zero.

9. Out of 100 coins, 99 are normal and 1 has two heads. One coin is pulled out at random. Let H be the event that it is the trick coin. The prior belief is $P(H) = 0.01$. The random coin is flipped and comes up heads. Call this event A_1. Compute the posteriori belief that the coin is the trick coin, $P(H|A_1)$. Which of the following is true?

a) $0 \leq P(H|A_1) < .01$
b) $.01 \leq P(H|A_1) < .02$
c) $.02 \leq P(H|A_1) < .05$
d) $.05 \leq P(H|A_1) < .50$
e) $.50 \leq P(H|A_1) \leq 1.0$

Part 2a. Variable Manipulations (10 points each)

10. Load the data found in the file, GDP_World_Bank.csv into a new dataframe. Notice that this dataset includes GDP statistics for the years 2010, 2011, and 2012. Create a new variable, gdp_growth, that equals the increase in GDP from 2011 to 2012 (the nominal increase, not a fractional one). What is the mean of your new variable?

11. Create a histogram of your new variable from question 10. Is it normally distributed? Describe its shape in terms of the distribution properties from the class.

12. Create a new Boolean variable, high_growth, that equals TRUE if a country's gdp growth is higher than the mean. How many countries have above average growth, and how many have below average growth? Explain this result in terms of the shape of the gdp_growth distribution.

Part 2b. Data Import (25 points)

13. Find one new metric country-level variable from some public source, and merge it into your dataset. Possible sources include the World Bank (http://

data.worldbank.org/indicator), the United Nations (http://data.un.org), and the World Health Organization (http://apps.who.int/gho/data/view.main). We would be especially happy, however, if you find a new data source we don't know about. Please identify your source clearly in comments. Most likely, you will need to import your data as a csv or tab-delimited text file. You will then need to view your new dataframe to diagnose any errors that happened during the import. Most common problems can be corrected by opening the file in a spreadsheet program and editing some entries.

Once you have successfully merged your dataframes, create one more graph that you think is interesting that involves your new variable. Write a few sentences about what the graph shows.