Summer 2017
DA 460
Data Analysis with Software and Programming
Winnie Li

Take Home Midterm Exam
July 2017

Name (Print):_____

I acknowledge and accept the Honor Code

Signature:_____

Score: _____/150

Instructions:

- Please complete this exam in a Word document, and save it as ***DA460_Midterm_XXXXX***, where XXXXX is the first five letters of your last name. Make sure you put down the problem # clearly.
- This exam is Open Book, Open Notes. ***You are required to use R and SAS to solve this questions. Make sure you include the code/command, as well as the relevant output.***
- You have 48 hours to complete this exam. Exam must be submitted through Canvas Assignment Tool by **11:59pm of Sunday, 7/16/17 (Pacific Time).**
- Round to the THIRD decimal place, unless otherwise noted in the instruction.
- **PLEASE SHOW ALL YOUR WORK COMPLETELY AND CLEARLY!!!**

☺ Good Luck ☺

Part 1:

Apply R to answer the following questions. Make sure you include clear headings (e.g., Midterm Exam Part 1 -- R, or Midterm Exam Part 2 -- SAS). For each part of the question, make sure you include the command line/code, then paste relevant output/results, and also comment on the output/results as needed (to answer the questions). Note: You need Handout 4 Materials too!

Problem 1:

Researchers did investigation on the situation of smoking in Great Britain and got the sample data set **smoking.csv (or smoking.txt)**, read the data set and answer the following questions.

1. Download ***smoking.csv (or smoking.txt)*** and read corresponding data into R. Example command in R: MyData <- read.csv(file="path/***TheDataIWantToReadIn***.csv", header=TRUE, sep=",") Note: use forward slash "/" instead of backward slash"\" in the path. Make sure to include the code/command.
2. How many observations are there in this data set? How many variables, and what are they? What is the 300th observation of nationality? Include both the code/command and the output/graph.
3. Create a numerical summary for age and compute the interquartile range. Compute the relative frequency distribution for gender. How many males are in the sample? Include both the code/command and the output/graph.
4. Using numerical summaries and a side-by-side box plot, determine if male smokers are as old as female smokers. Include both the code/command and the output/graph.
5. Create a bar chart or frequency table for maritalStatus, what is the proportion for Divorced, Single, Married, and Widowed, respectively? What can you interpret from these numbers? Include both the code/command and the output/graph.

Problem 2:

Apply R simulation to answer the following questions:

1. Suppose we're flipping an unfair coin that we know only lands heads 30% of the time. Please simulate this flip 10 times, what is the proportion of heads? If you simulate this flip 100 times, what is the proportion of heads now? Include both the code/command and the output/graph.
2. Suppose we're flipping an unfair dice and the corresponding probability of landing 1, 2, 3, 4, 5, and 6 is 0.05, 0.1, 0.15, 0.2, 0.3, and 0.2, respectively. If you simulate this flip 10 times, what is the proportion of land on side 5? Simulate this flip 100 times, what is the proportion of side 5 now? Include both the code/command and the output/graph.
3. Compare the proportions in each questions above, what conclusion can you draw? Does the number of simulations affect the proportions? If so, how? Please explain in details.

Problem 3:

Data set **countyComplete.csv (or countyComplete.txt)** shows the population information from all counties in US, apply this data set to answer the following questions:

1. Download ***countyComplete.csv (or countyComplete.txt)*** and read corresponding data into R. Example command in R: MyData <-read.csv(file="path/***TheDataIWantToReadIn***.csv", header=TRUE, sep=",") Note: use forward slash "/" instead of backward slash"\" in the path. Make sure to include the code/command
2. Make a histogram of pop2010, how can you descript its distribution, bell-shaped or normal? Is it right skewed or left sewed? Include both the code/command and the output/graph.
3. Create a new subset named **Washington** which contains only the observations of Washington, and then make a histogram of pop2010, how can you describe its distribution, bell-shaped or normal? Is it right skewed or left sewed? Compare this with question 1. Include both the code/command and the output/graph.
4. Based on the subset **Washington**, make a normal probability plot of pop2010. Do all of the points fall on the line? How does this plot compare to the probability plot of the original data? Include both the code/command and the output/graph.
5. Suppose the variable pop2010 has a normal distribution, what is the probability that pop2010 is greater than 102,410? What is the probability that pop2010 is between 190,000 and 1,000,000? Include both the code/command and the output/graph.

Part 2:

Apply SAS to answer **ALL part 1 problems** (use the same source data). Example command in SAS to read corresponding dataset (.csv file). *Note: replace "path/datasest.csv" with the actual file path and file name.*

```
proc import datafile="path/dataset.csv"
     out=mydata
     dbms=csv replace;
     getnames=yes;
run;
```

Make sure you include clear headings (e.g., Handout 2 R or Handout 2 SAS). For each part of the question, make sure you include the command line/code, then paste relevant output/results, and also comment on the output/results as needed (to answer the questions)

Part 3:

Save your file as ***DA460_MidtermExam_XXXXX.docx (or .pdf)*** where ***XXXXX*** is the first five letters of your last name, and submit it through the Assignment Tool.