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

Take Home Final Exam
August 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_Final_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.***
- Exam must be submitted through Assignment Tool by **11:59pm of Sunday, 8/6/17 (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., Final Exam Part 1 -- R, or Final 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 8 Materials!

Problem 1:

Read the data set **run10.csv (or run10.txt)**, and answer the following questions.

1. Download ***run10.csv (or run10.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. Calculate the population mean and standard deviation for `divTot`. Apply `rep` function and `for` loop to collect 50 simple random samples of size 100 from `divTot`, and then use these stored statistics to calculate 50 confidence intervals of 95% confidence level for population mean. Include both the code/command and the output/graph.
3. Apply function `plot_ci` to display all the 50 confidence intervals in question (2). What proportion of your confidence intervals include the true population mean? Is this proportion consistent with the confidence level? Why or why not? Include both the code/command and the output/graph.
4. Make a side-by-side boxplot of `gender` and `divTot`. What does the plot highlight about the relationship between these two variables? Include both the code/command and the output/graph.
5. Calculate and a 95% confidence intervals for the difference between the mean of male `divTot` and the mean of female `divTot`, and interpret. Include both the code/command and the output/graph.
6. Conduct a hypothesis test at 95% significant level evaluating whether the mean of male `divTot` is different from the mean of female `divTot`? Make sure you indicate the hypotheses, the test rest, and the conclusion(s) clearly. Include both the code/command and the output/graph.

Problem 2:

Read the data set **smoking.csv (or smoking.txt)**, and answer the following questions. Make sure to include both the code/command and the output/graph:

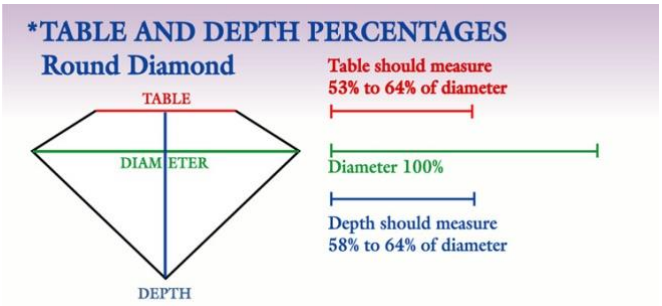
1. Download ***smoking.csv (or smoking.txt)*** and read corresponding data into R. Example command in R: MyData<- read.csv(file="path/***TheDataIWantToReadIn***.csv",

header=TRUE, sep=","). Note: use forward slash "/" instead of backward slash"\" in the path. Make sure to include the code/command.

2. Create a new data frame that contains only the rows in ***smoking.csv*** associated with **male smokers** (Male for `gender` and Yes for `smoke`). Then, calculate the proportion of Divorced in `maritalStatus`. Include both the code/command and the output/graph.
3. Check if the conditions for inference are reasonable. If so, apply `inference` function to calculate the standard error and construct a 95% confidence interval for the proportion of Divorced in `maritalStatus`. Include both the code/command and the output/graph.
4. Based on the R output in question (3), what is the standard error and the margin of error for the estimate of the proportion of Divorced in `maritalStatus`?
5. Use simulation to show how the proportion affect the margin of error. Describe the relationship between the proportion and the margin of error. Include both the code/command and the output/graph.
6. Apply `for` loop to simulate the process of drawing 300 samples of size 1000 from a population with a true Divorced proportion of 0.3. For each of the 3000 samples, compute $\hat{p}$ and then plot a histogram to visualize the distribution. Describe the sampling distribution of the sample proportions. Be sure to note the center, spread, and shape clearly. Include both the code/command and the output/graph.

Problem 3:

Data set **diamonds.csv (or diamonds.txt)** shows the price information of diamonds, apply this data set to answer the following questions. Make sure to include both the code/command and the output/graph.



1. Download ***diamonds.csv (or diamonds.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. What type of plot is the most appropriate one to display the relationship between `price` and `carat`? Plot this relationship using the variable `carat` as the predictor. Does the relationship look linear? Include both the code/command and the output/graph.
3. If the relationship looks linear in question (2), what is the correlation coefficient? Interpret this result. Include both the code/command and the output/graph
4. Fit a simple linear model (use `carat` to predict `price`). Using the estimates from the R output, write down the regression equation. What are the y-intercept and the slope? Interpret the regression line. Is variable `carat` significant? Why or why not? What is the coefficient of determination? Interpret the results. Include both the code/command and the output/graph.

5. Fit a multiple linear model with `carat`, `depth`, and `table` as independent variables, and variable `price` as dependent variable. Using the estimates from the R output, write down the regression equation. What are the y-intercept and the slopes? Interpret the regression line. Are all the independent variables in the model significant? Why or why not? What is the coefficient of determination? Interpret the result, and compare it with question (4). Include both the code/command and the output/graph.
6. Using backward-selection and p-value as the selection criterion, determine the best model. You do not need to show all steps in your answer, just the output for the final model. Also, write down the final regression equation and interpret it. Include both the code/command and the output/graph
7. What are the conditions to validate the model? Apply appropriate plots to check these conditions. Are all conditions passed? Is the model valid? Why or why not? Include both the code/command and the output/graph.
8. Based on your final model, what are the most important factors that influence the selling price of a diamond? Why?

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;
```

Note: You may also access the data set via
Launch SAS → Highlight sasuser.v94 folder in left column (Server Files and Folders) → Click Upload icon to upload your data from your hard drive → use `proc import` to access the data.

Make sure you include clear headings (e.g., Problem 1 - 1 R or Problem 3 - 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_FinalExam_XXXXX.docx (or .pdf)*** where ***XXXXX*** is the first five letters of your last name, and submit it through the Assignment Tool.