STAT 341: Assignment 2 - Fall 2020

Ryan Browne

46 Marks, Due: Friday, October 9 at 10:00am

NOTES

Your assignment must be submitted by the due date listed at the top of this document, and it must be submitted electronically in .pdf format via Crowdmark/LEARN. This means that your responses for different questions should be in separate .pdf files. Your .pdf solution files must have been generated by R Markdown. Additionally:

- For mathematical questions: your solutions must be produced by LaTeX (from within R Markdown). Handwritten and scanned/photographed solutions will not be accepted and you will receive zero points.
- For computational questions: R code should always be included in your solution (via code chunks in R Markdown). If code is required and you provide none, you will received zero points.
- For interpretation question: plain text (within R Markdown) is fine.

Organization and comprehensibility is part of a full solution. Consequently, points will be deducted for solutions that are not organized and incomprehensible.

- You will submit your solutions in the form of one pdf file per question through LEARN For example, for Q1 you should submit one pdf file containing the solution to the first question only. Failing to follow the formatting instructions may result in your whole paper or individual questions receiving a grade of 0%.

Question One - 26 Marks - MNIST Database and Graphical Attributes

Here we will use a subset of the Modified National Institute of Standards and Technology database (MNIST) as an introduction to R and data exploration. A description of the data can be found at http://yann.lecun.com/exdb/mnist/

Again, we look at a subset of the Modified National Institute of Standards and Technology database (MNIST) as an introduction to R and data exploration.

- Load the digits from files `one500.csv` and `two500.csv` .

a) One feature to distinguish the one's from the two's might be the sum of the brightness values. This similar to the amount of ink used if you were to write an one or two. Use both populations for this part.
   i) **[3 Marks]** Construct a boxplots and histograms of this feature for each digit. For the histograms pick a resonable number of bins. Compare and constrast the histogram and boxplots.

ii) **[1 Mark]** Is a boxplot a good represntation of the populations? Why?

iii) **[3 Marks]** Construct two quantile plots for this feature. What characteristic do these quantile plots exhibit?

b) Another feature could be the brightness of a particular pixel. From the two csv files extract the brightness from pixel or column 403. You should have 500 brightness values from the two populations.

   i) **[3 Marks]** Construct a boxplots and histograms of this feature for each digit. For the histograms pick a resonable number of bins. Compare and constrast the histogram and boxplots.

   ii) **[1 Mark]** Is a boxplot a good represntation of the populations? Why?

   iii) **[3 Marks]** Construct two quantile plots for this feature. What characteristic do these quantile plots exhibit?

c) **[1 Marks]** What is the implicit assumption for a boxplot to be a "good" representation of a population?

d) Comparing the Freedman–Diaconis and Scott's rule for the number of bins. Use only the digit 1 population and pixel or column 406 for this part.

   i) **[4 Marks]** Calculate the number of bins required for Freedman–Diaconis and Scott's for the number of bins. Then compare the number of bins generated by the function 'hist' when using the argument `breaks="FD"` and `breaks="Scott"`. Display the results in a table.

   ii) **[3 Marks]** Construct 4 histograms using the number of bins from part i) in a 2×2 grid.

   iii) **[3 Marks]** Construct a histogram which is a "good" represenation of this population along with a quantile plot. What characteristic does the quantile plot exhibit.

e) **[1 Mark]** Using R, find the top 5 pixels (denoted by column) with the largest difference in averages between the group of one's and two's.

Question Two - 10 Marks - A Bump Rule Multiple-Choice Question

Construct two multiple-choice questions for the two bump rules, one for 1-dimensional rule and another for 2-dimensional rule. For an motivating see the question in the file "Power_Transformation_Sample_Question.pdf". Some requirements of your solution are given below:

- For each part, give the correct answer and explain why it is correct and why the other choices are incorrect;
- Show the before-and-after comparison for each option;
- You are limited to three pages in total;
- The format does not need to be the same as shown in "Power_Transformation_Sample_Question.pdf". For example, your are not required to have two columns.

Rubric

Criteria	Descriptor	Marks
Format	Organization & LaTex	/3
Question	Difficulty, Clarity and Creativity	/3
Solution	Justification and Clarity	/4

Question Three - 10 Marks - Influence

In your own words summarize the concept of influence based on subsection 2.2.3 - Influence & Sensitivity.

- You are limited to 2 pages.
- Your summary should include a combination of formulas, full sentences and an example.
- From your summary, the reader should be able to know the definition and understand influence conceptually. In addition, know how to apply it to an arbitrary attribute and learn how to interpret its values.

Rubric

Criteria	Descriptor	Marks
Format	Organization & LaTex	/3
Writing	Clarity & Grammar	/2
Content	Coverage, Depth, Relevant Terminology used and Example	/5