STAT 341: Assignment 1 - Spring 2020

Name

108 Marks, Due: Friday, June 5 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.

Question 1: Evaluating the population range - 25 points

Consider the population $\mathcal{P} = \{y_1, \ldots, y_N\}$. The population range is

$$a(\mathcal{P}) = a(y_1, \ldots, y_N) = y_{(N)} - y_{(1)}$$

and hence a measure of spread In this question you will investigate several of its properties.

- Note: **5 Marks** are allocated for formating and organization.

(a) **[3 points]** Determine whether the range is location invariant, location equivariant, or neither.

(b) **[3 points]** Determine whether the range is scale invariant, scale equivariant, or neither.

(c) **[3 points]** Determine whether the range is location-scale invariant, location-scale equivariant, or neither.

(d) **[3 points]** Determine whether the range is replication invariant, replication equivariant, or neither.

(e) **[3 points]** Derive the sensitivity curve for the range, given a population $\{y_1, y_2, \ldots, y_{N-1}\}$.

(f) **[3 points]** For the population below, plot the sensitivity curve from part (e) for $y \in [-7, 7]$. You may find the `sc()` function from class useful.

```
set.seed(341)
pop <- rnorm(10000)
```

(g) **[2 points]** Given all that you have learned in parts (a) - (f), state one thing that is *good* about thie range attribute and one thing that is *bad* about the range attribute.

Question 2: Write a plot-making function [5 points]

Write a function called `matrix.plot()` that takes in a single input (called `df`), that is an $N \times m$ data frame containing *numeric* data. This function should produce as its output an $m \times m$ matrix of plots where:

- the diagonal plots contain histograms of the columns of `df`
- the upper triangle of plots are scatter plots between all pairs of columns of `df`
- the lower triangle of plots report the correlation coefficients between the pairs of columns of `df`
- all plots should be labelled with the headings provided in `df`

Question 3: Spotify Top 30 Analysis - 25 points

Spotify, the popular music streaming service, organizes and classifies songs based on a wide range of properties (variates):

Variate	Description
`genre`	the genre of the track
`year`	the release year of the recording (note that due to vagaries of releases, re-releases, re-issues and general madness, sometimes the release years are not what you'd expect)
`bpm`	beats per minute - the tempo of the song
`energy`	the higher the value the more energetic the song
`danceability`	the higher the value the easier it is to dance to the song
`loudness`	the higher the value the louder the song
`liveness`	the higher the value the more likely the song is a live recording
`valence`	the higher the value the more positive the mood of the song
`duration`	the duration of the song (in seconds)
`acousticness`	the higher the value the more acoustic the song is
`speechiness`	the higher the value the more spoken words the song contains
`popularity`	the higher the value the more popular the song is

Available for us to study is the population of $N = 300$ Billboard Top 30 songs from 2010 - 2019 (inclusive). In addition to the song's title and artist, measurements on each of the 12 variates listed in the table above have also been recorded for each of these songs. This data is available in the `spotify.csv` file.

- Note: **4 Marks** are allocated for formating and organization.

(a) **[2 points]** Using the `matrix.plot()` function you developed in Question 2, produce the summary graphic for `duration`, `acousticness`, and `speechiness`.

(b) **[3 points]** Considering all variates (except `genre` and `year`), which three are most strongly correlated with valence? For each variate, explain the nature of its linear relationship with valence.

(c) **[1 point]** Using R, determine which are the Top 5 songs with the highest valence.

(d) **[2 points]** Using R, determine which song is the highest and lowest danceability.

(e) **[4 points]** Let y_1 denote the how acoustic (`acousticness`) a song sounds, and let $a(\mathcal{P}) = \bar{y}$ be the

attribute of interest. Define the influence of song u on $a(\mathcal{P})$ to be:

$$\Delta(a, u) = |a(y_1, \ldots, y_{u-1}, y_u, y_{u+1}, \ldots, y_N) - a(y_1, \ldots, y_{u-1}, y_{u+1}, \ldots, y_N)|$$

Construct an influence plot of Δ vs. observation number and identify the song with the largest influence on the average `acousticness` attribute.

(f) **[3 points]** Repeat part e) using `speechiness`. What is the song with the most influence?

g) **[3 points]** Add the two measures of influence from part e) and f) into a single measure. What is the song with the most influence?

(h) **[3 points]** Using R, determine which artists have appeared in the Billboard Top 30 four times. For each of these artists state the number of times they have appeared and calculate the average valence score of their songs.

Question 4 - 10 Marks

By searching the web, find a public dataset that constitutes a sample. For this data, provide the following:

- A description of the data (define what is a unit and two variate(s) that have been recorded)
- A justification for why the dataset is indeed a population (as opposed to a sample)
- A URL to access the data

Some places you might consider looking:

- Kaggle
- UCI Machine Learning Repository
- r/datasets
- data.gov
- KDnuggets

Rubric

Criteria	Descriptor	Marks
Data	Creativity was an interesting or unique dataset chosen	/1
Description	Clarity & Correct Justification	/6
Justification	Correctness	/2
URL	Provided?	/1

Question 5 - 7 Marks

From Section 2.2 Explicit Attributes and including all the subsections construct a true/false question and explain why it is true or false.

Rubric

Criteria	Descriptor	Marks
Question	Concept, Difficulty, Clarity and Creativity	/4
Explanation	Correct Justification and Clarity	/3

Question 6 - 20 Marks

In your own words summarize the concept of influence or sensitivity based on subsection 2.2.3 "Influence_Sensitivity".

- You are limited to 1 to 2 pages.
- You are recommended to use a combination of formulas, full sentences and an example.
- This question can be Rmarkdown.

Rubric

Criteria	Descriptor	Marks
Format	Organization	/5
Writing	Clarity & Grammar	/5
Content	Coverage, Depth, Relevant Terminology used and Example	/10

Question 7 - 16 Marks

Demonstrate the two **bump** rules for using power transformations on simulated data.

- For each rule use at least two plots to demonstrate a before and after transformation.
- Clearly explain why that direction was chosen.

- One page per **bump**

Rubric for each **bump** rule

Criteria	Descriptor	Marks
Data	Description & Reproducibility	/2
Format	Grammar, Punctuation and Organization	/2
Explanation	Correct Justification, Clarity, Relevant Terminology used	/4