STAT 341: Assignment 4 - Spring 2020

Name

70 Marks, Due: Friday, July 24 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 unless otherwise specified. 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 receive zero points.
  - **Exception** any functions used in the notes or function glossary can loaded using `echo=FALSE` but any other code chunks should have `echo=TRUE`. e.g. the code chuck loading `gradientDescent` can use `echo=FALSE` but chunks that call `gradientDescent` should have `echo=TRUE`.
- 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 1 - 36 Marks

In this question you will be analysing data for major league baseball players who were notably big "Hitters". The data are from the ISLR package; see help("Hitters") for information on the variates.

- First we need to get the data without any missing values.

```
library(ISLR)
# We want the baseball salary data from this package
data("Hitters")
# IMPORTANT: We will work only with players having complete records:
C_Hitters <- na.omit(Hitters)
# head(C_Hitters)
```

- And you will need Rcode for Pearson's second skewness coefficient (median skewness) given by

$$3 \times [\overline{y} - \text{median}_{\mathcal{P}}(y)] / SD_{\mathcal{P}}(y)$$

```
sdn <- function(z) {
    N = length(z)
    sd(z) * sqrt((N - 1)/N)
}

skew <- function(z) {
    3 * (mean(z) - median(z))/sdn(z)
}
```

- A commonly used transformation when $y > 0$ is the family of power transformations which is indexed by a power α. Define this transformed variable to be

$$T_\alpha(y) = \begin{cases} y^\alpha & \alpha > 0 \\ \log(y) & \alpha = 0 \\ -(y^\alpha) & \alpha < 0 \end{cases}$$

```
powerfun <- function(x, alpha) {
    if (sum(x <= 0) > 1)
        stop("x must be positive")
    if (alpha == 0)
        log(x) else if (alpha > 0) {
        x^alpha
    } else -x^alpha
}
```

- We define the attribute α implicitly such that

$$3 \times \left[\overline{t}_\alpha - \text{median}_{\mathcal{P}}(t_\alpha)\right] / SD_{\mathcal{P}}(t_\alpha) = 0$$

  i.e. the value of the power transformation such that the transformed variable has zero sknewness.

- **Note**
  - This questions is related to sample exercises question 1.12 An Implicitly defined Skewness Attribute, it might be helpful to review that question.
  - Here we will focus on the the number of career runs denoted as `CRuns`,

a) The population - Using the number of career runs denoted as `CRuns`,
   i) **[2 Marks]** Construct a histogram.
   ii) **[1 Mark]** Calculate mean and Pearson's second skewness coefficient.
   iii) **[2 Marks]** If we apply the power transformation using α as the power we can change the skewness. Using the `uniroot` function find the value of α which makes the skewness of the power-transformed variable equal to zero.

   iv) **[3 Marks]** Using the value of α from part (iii), calculate the skewness on the power-transformed variable and construct a histogram of the power-transformed variable.

v) **[2 Marks]** Write a function named `attr3` that takes in a population or sample of variates and outputs the mean, skewness and the value of α which makes the skewness of the power-transformed variable equal to zero. Apply `CRuns` to this function.

b) **[5 Marks]** Sampling Distribution of the Attributes - Select $M = 1000$ samples of size $n = 50$ without replacement. i.e. construct $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{1000}$.

- For each sample apply the `attr3` function. Then construct three histograms (in a single row) of the sample error for each attribute.

c) A Sample and the Bootstrap - Using the the variable `CRuns` and following sample

```
sam = c(220, 97, 256, 241, 137, 83, 140, 186, 34, 135, 50, 191, 213, 216,
    58, 91, 244, 263, 240, 51, 258, 254, 224, 89, 62, 86, 247, 5, 166,
    81, 61, 136, 217, 157, 207, 47, 124, 25, 260, 32, 160, 114, 246, 143,
    57, 261, 70, 2, 110, 181)
```

i) **[1 Mark]** Based on the sample, calculate the three attributes of interest.

ii) **[4 Marks]** Construct two histograms; one of the raw values and another of the power-transformed variable `CRuns` using the value of α from part c i).

iii) **[5 Marks]** Bootstrap - By resampling $\mathcal{S}$ with replacement, construct $B = 1000$ bootstrap samples $\mathcal{S}_1^\star, \mathcal{S}_2^\star, \ldots, \mathcal{S}_{1000}^\star$ and calculate the three attributes of interest on each bootstrap sample. Then construct three histograms (in a single row) of the bootstrap sample error for each attribute.

iv) **[3 Marks]** Calculate standard errors for each sample estimate and then construct a 95% confidence for the population attributes using the percentile method.

d) **[8 marks]** Sampling Properties of the Bootstrap - For each of three attributes of interest estimate the coverage probability when using the percentile method. Give a standard error for estimate the coverage probability and a conclusion.

Question 2 - 16 Marks

Compare two sub-populations. Your comparison should include:

- a description of the context and the two sub-populations,
- compare the sub-populations using at least two attributes (but you are required to consider multiple testing),
- numerical and graphical summarizes,
- a conclusion.
- You comparison should be limited to 1 to 2 pages.

Your solution should be **in your own words**, but as motivating examples, see from the Inference exercises:

- 1.4 Comparing Sub-populations in Fire Emblem Heroes
- 1.8 Comparing male and female final grades
- 1.9 Comparing Midterm to final grades
- 1.7 City of Baltimore, Crime & Safety Rates for (2010-2014)

Rubric

Criteria	Descriptor	Marks
Population/Attributes	Description and Difficulty	/4
Format	Clarity, Organization and LaTeX	/4
Comparision	Description, Results and Graphic	/4
Discussion/Summary	Justification and Relevant Terminology used	/4

Question 3 - 10 Marks

In your own words summarize the subsection **4.4.2b-Bootstrap_t_Confidence_Interval**

- You are recommended to use a combination of formulas, full sentences an example.
- You may incorporate subsection 4.4.2c-The_Double_Bootstrap but is not required.
- You are limited to 1 to 2 pages.

Rubric

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

Question 4 - 8 Marks

In the context of regression describe how to perform the parametric bootstrap.

- There is a 1 page limit.

Rubric

Criteria	Descriptor	Marks
Format	Clarity, Organization	/2
Writing	Grammar & Punctuation, Clarity	/2
Content	Correctness & Relevant Terminology Used	/4