STAT 341: Assignment 2 - Fall 2020

Ryan Browne

XX Marks, Due: Friday, October 2 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.

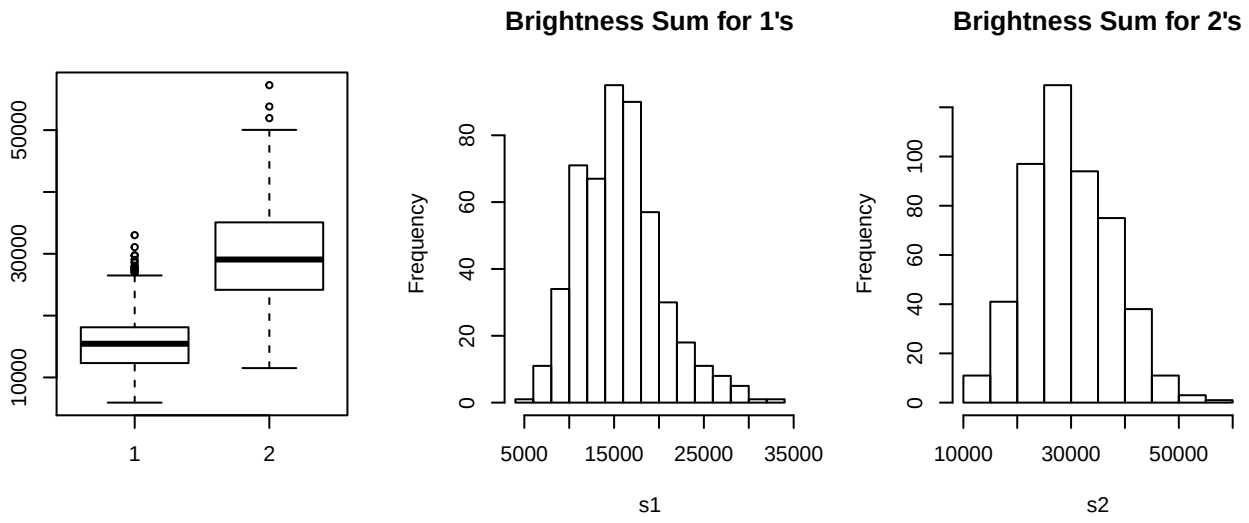
- Load the digits from files `one500.csv` and `two500.csv` .

```
x1 = as.matrix(read.csv("one500.csv", header = TRUE))
x2 = as.matrix(read.csv("two500.csv", header = TRUE))
```

a) One feature to distinguish the one's from the two's might be the all the brightness values. This similar to the amount of ink used if you were to write an one or two.

b) **[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.

```
s1 = apply(x1, 1, sum)
s2 = apply(x2, 1, sum)

par(mfrow = c(1, 3))
boxplot(s1, s2, names = c(1, 2), main = "")
hist(s1, main = "Brightness Sum for 1's")
hist(s2, main = "Brightness Sum for 2's")
```



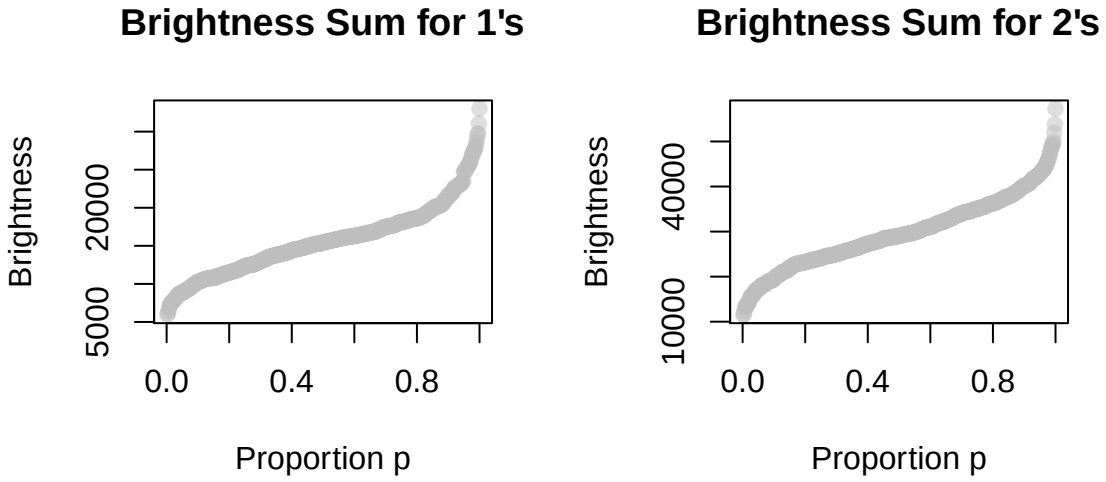
- The histogram for the digit 1 is a bit skewed.
- The histogram for the digit 2 is a fairly symmetric.
- Both boxplot and histogram convey the same information.

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

Yes, they are unimodal.

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

```
N <- length(s1)
par(mfrow = c(1, 2))
plot((1:N)/N, sort(s1), pch = 19, col = adjustcolor("grey", alpha = 0.5),
    xlim = c(0, 1), xlab = "Proportion p", ylab = "Brightness", main = "Brightness Sum for 1's")

plot((1:N)/N, sort(s2), pch = 19, col = adjustcolor("grey", alpha = 0.5),
    xlim = c(0, 1), xlab = "Proportion p", ylab = "Brightness", main = "Brightness Sum for 2's")
```

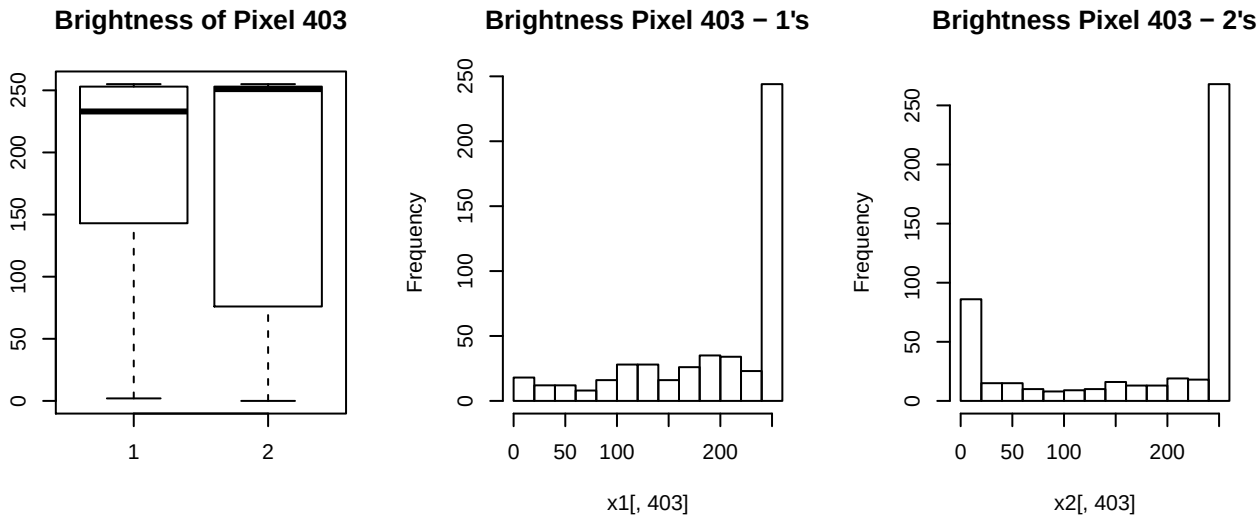


- Both have a "S" curved and exhibit some skewness.

b) Another feature could be the brightness of a particular digit. Use column 403 for this part.

c) **[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.

```
par(mfrow = c(1, 3))
boxplot(x1[, 403], x2[, 403], names = c(1, 2), main = "Brightness of Pixel 403")
hist(x1[, 403], main = "Brightness Pixel 403 - 1's")
hist(x2[, 403], main = "Brightness Pixel 403 - 2's")
```



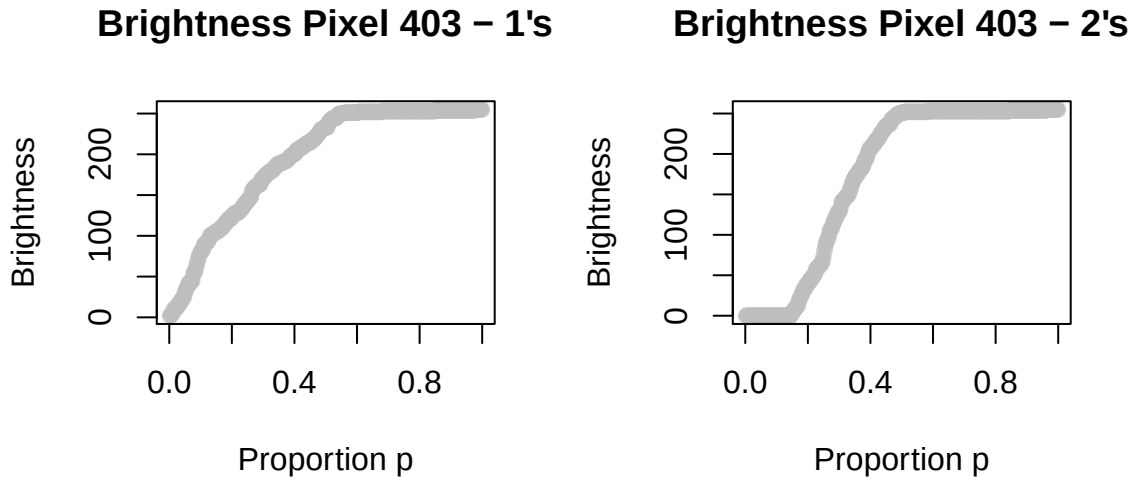
- The histogram for the digit 1 looks uniform expect for a spike at the high brightness.
- The histogram for the digit 2 is bimodal.
- Both boxplots and histograms do not convey the same information.

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

No, beause one is bimodal and the other has a spike at the high brightness.

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

```r
N <- length(s1)
par(mfrow = c(1, 2))
plot((1:N)/N, sort(x1[, 403]), pch = 19, col = adjustcolor("grey", alpha = 0.5),
    xlim = c(0, 1), xlab = "Proportion p", ylab = "Brightness", main = "Brightness Pixel 403 - 1's")

plot((1:N)/N, sort(x2[, 403]), pch = 19, col = adjustcolor("grey", alpha = 0.5),
    xlim = c(0, 1), xlab = "Proportion p", ylab = "Brightness", main = "Brightness Pixel 403 - 2's")
```



- The 1's have an approximately linear increase (indicating uniform density) and then a spike at the high brightness.
- The 2's have a spike at the high and low brightness. It is linear in the middle indicating uniform density.

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

- Unimodality

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

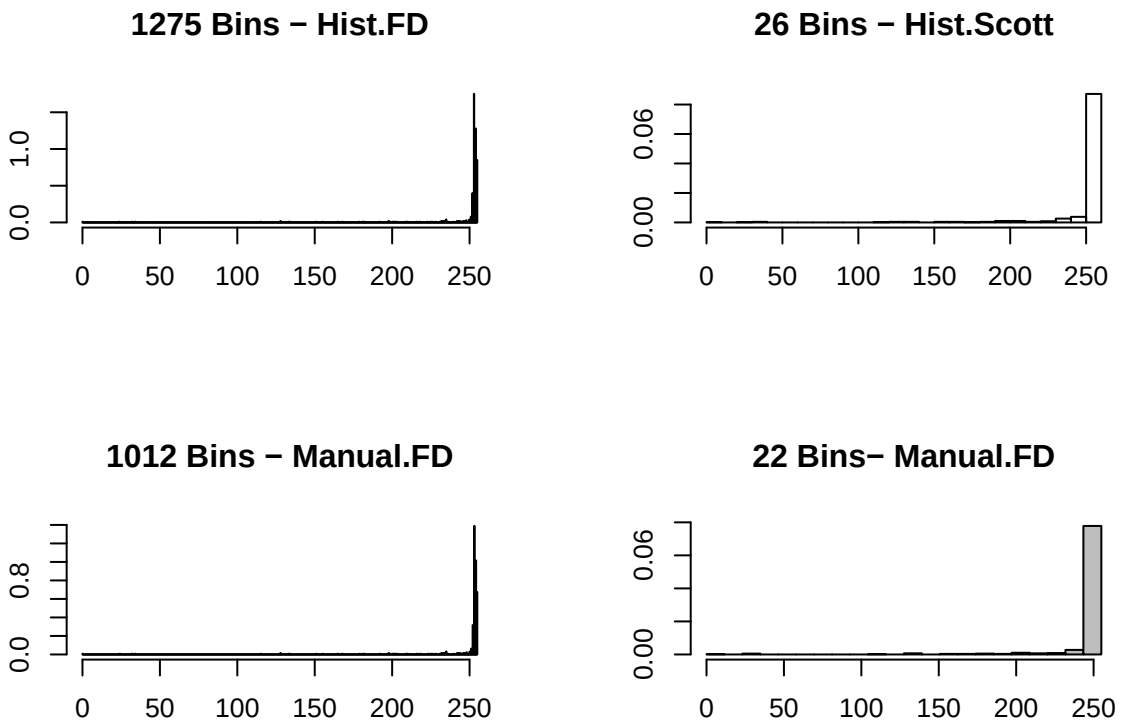
e) **[4 Marks]** Calculate the number of bins required for Freedman–Diaconis and Scott's for the number of bins. Then in addtion, compare the number of bins generated by the function 'hist' using the agruement `breaks="FD"` and `breaks="Scott"`. Display the results in a table.

```r
temp.FD = hist(x1[, 406], breaks = "FD", plot = FALSE)
hist.FD = length(temp.FD$breaks) - 1

binSize.FD = 2 * IQR(x1[, 406])/(500)^(1/3)
binNum.FD = diff(range(x1[, 406]))/binSize.FD

temp.Scott = hist(x1[, 406], breaks = "Scott", plot = FALSE)
hist.Scott = length(temp.Scott$breaks) - 1

binSize.Scott = 3.5 * sd(x1[, 406])/(500)^(1/3)
binNum.Scott = diff(range(x1[, 406]))/binSize.Scott

temp = matrix(0, 2, 2)
dimnames(temp) = list(c("Hist", "Manual"), c("FD", "Scott"))
temp[, 1] = c(hist.FD, binNum.FD)
temp[, 2] = c(hist.Scott, binNum.Scott)
ceiling(temp)
```

```
##          FD Scott
## Hist   1275    26
## Manual 1012    22
```

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

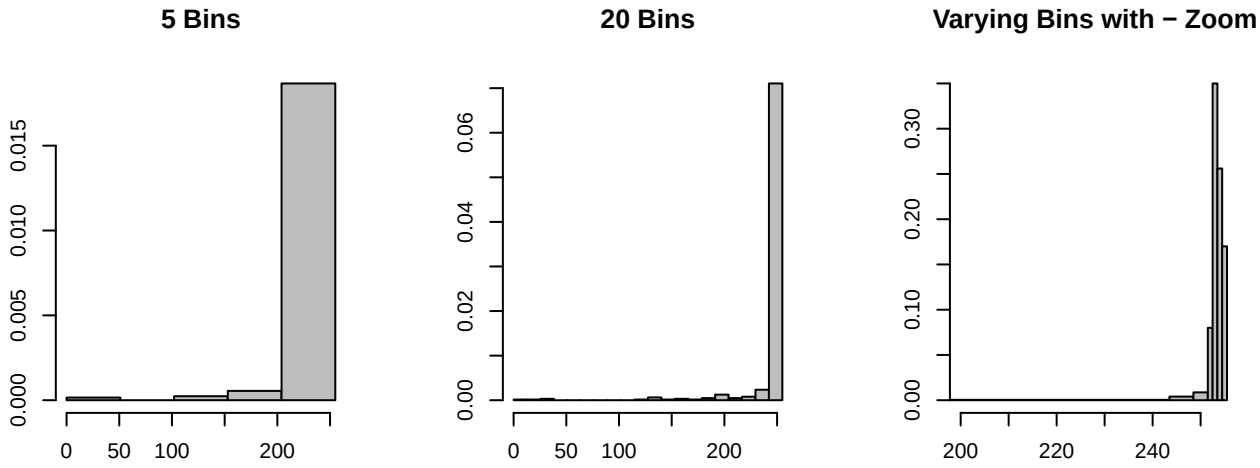
```
par(mfrow = c(2, 2))

hist(x1[, 406], breaks = "FD", prob = TRUE, main = "1275 Bins - Hist.FD",
    xlab = "", ylab = "")
hist(x1[, 406], breaks = "Scott", prob = TRUE, main = "26 Bins - Hist.Scott",
    xlab = "", ylab = "")

hist(x1[, 406], breaks = seq(0, 255, length.out = 1012 + 1), prob = TRUE,
    main = "1012 Bins - Manual.FD", col = "grey", xlab = "", ylab = "")
hist(x1[, 406], breaks = seq(0, 255, length.out = 22 + 1), prob = TRUE,
    main = "22 Bins- Manual.FD", col = "grey", xlab = "", ylab = "")
```



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.

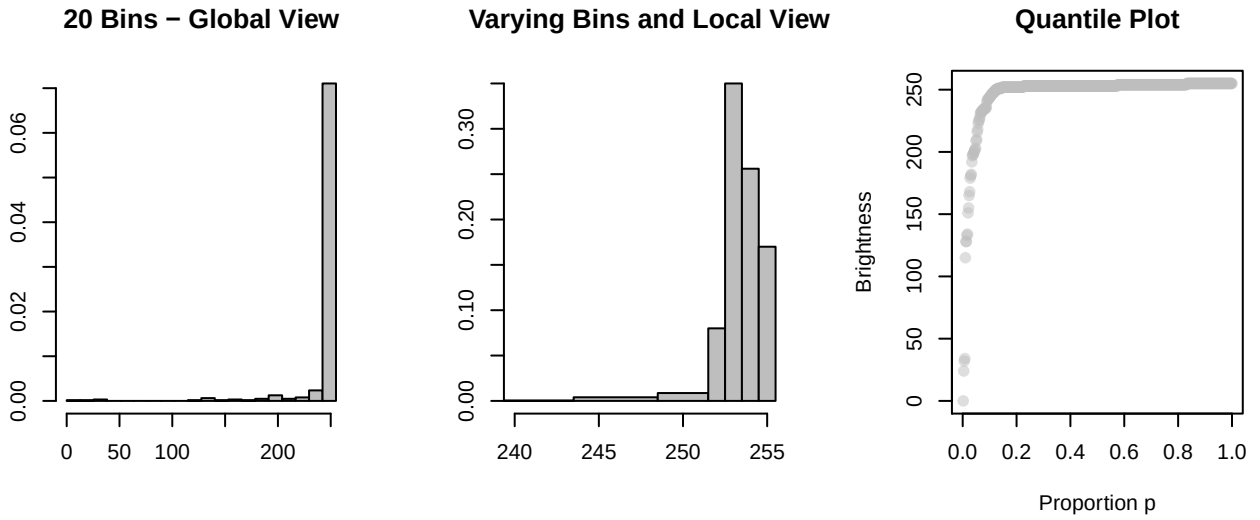
- Lot of histograms will work.

```
par(mfrow = c(1, 3))
hist(x1[, 406], breaks = seq(0, 255, length.out = 5 + 1), prob = TRUE,
    main = "5 Bins", col = "grey", xlab = "", ylab = "")
hist(x1[, 406], breaks = seq(0, 255, length.out = 20 + 1), prob = TRUE,
    main = "20 Bins", col = "grey", xlab = "", ylab = "")
hist(x1[, 406], breaks = c(0, 244, 249, 252:256) - 1/2, prob = TRUE, main = "Varying Bins with - Zoom",
    col = "grey", xlab = "", ylab = "", xlim = c(200, 256))
```



- One where I zoom in on the area of high density.
- Approximately 10% of the brightness values are less or equal to 244.

```
par(mfrow = c(1, 3))
hist(x1[, 406], breaks = seq(0, 255, length.out = 20 + 1), prob = TRUE,
    main = "20 Bins - Global View", col = "grey", xlab = "", ylab = "")
hist(x1[, 406], breaks = c(0, 244, 249, 252:256) - 1/2, prob = TRUE, main = "Varying Bins and Local Vie
    col = "grey", xlab = "", ylab = "", xlim = c(240, 256))
plot((1:500)/500, sort(x1[, 406]), pch = 19, col = adjustcolor("grey",
    alpha = 0.5), xlim = c(0, 1), xlab = "Proportion p", ylab = "Brightness",
    main = "Quantile Plot")
```



- The quantile has a large section of increase and a large section of flatness.
- It has a rotated L shape.

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.

```
order(abs(apply(x1, 2, mean) - apply(x2, 2, mean)), decreasing = TRUE)[1:5]
```

```
## [1] 408 407 409 437 436
```