STAT 341: Assignment 1 - Fall 2020

Name

54 Marks, Due: Friday, September 25 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%.

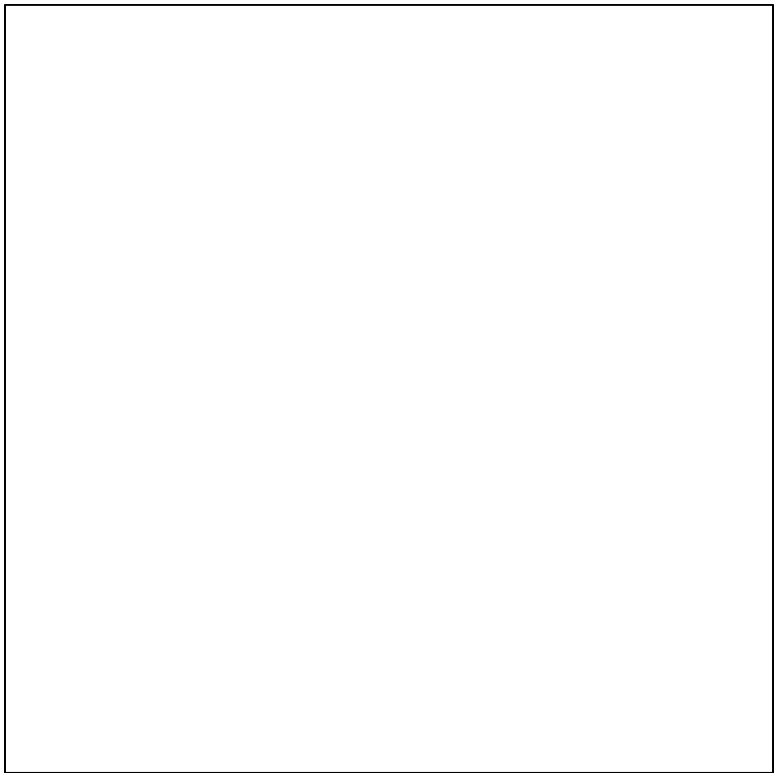
Question One - 17 Marks - MNIST Database

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/

> The MNIST database of handwritten digits, available from this page, has a training set of 60,000 examples, and a test set of 10,000 examples. It is a subset of a larger set available from NIST. The digits have been size-normalized and centered in a fixed-size image.
>
> It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.

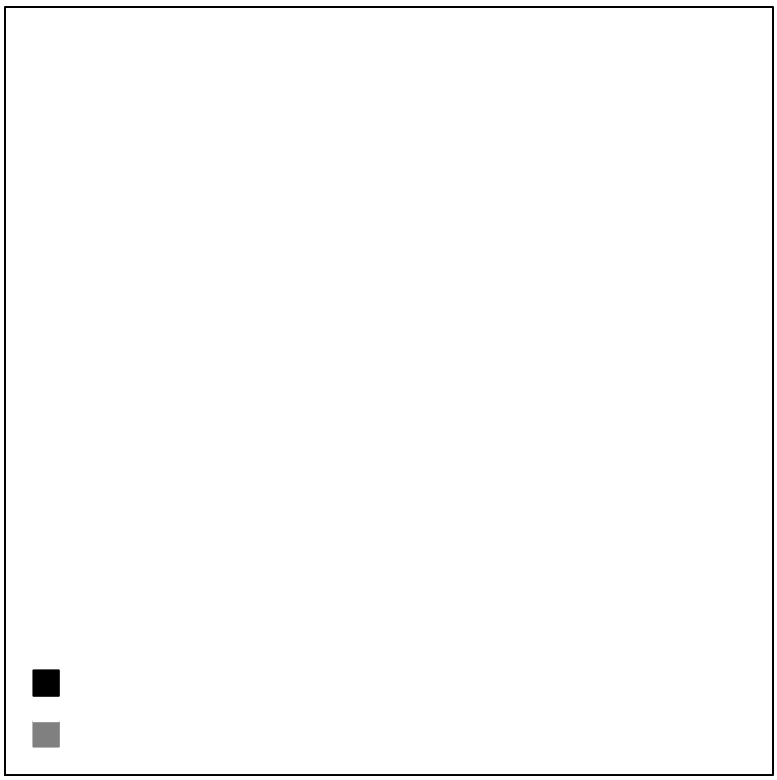
a) **[2 Marks]** Write a function initializeDigitPlot that accepts no parameters and creates a blank plot with x and y coordinate ranges of 0.5 to 28.5, no borders, and no axes. Test this function.

```r
initializeDigitPlot <- function() {
  par(mar=c(0,0,0,0) )
  plot(NA, xlim=c(0.5,28.5), ylim=c(0.5,28.5) , xaxt="n", yaxt="n")
}
initializeDigitPlot()
```



b) **[2 Marks]** Write a function drawBox that accepts three parameters (x, y, brightness) that draws a solid-filled 1x1 box centred at (x, y) onto a preexisting plot with colour equal to the greyscale value. The grayscale value is between zero and 1, where 0 is white, 1 is black and the values inbetween are differents levels of grey. Test this function using the part a) and the parameters `(1,1,0.5)`, `(1,2,0)` and `(1,3,1)`. Use the commands `polygon` and `gray`.

```r
drawBox <- function(posg=NULL) {
  xpos = c(-0.5, 0.5, 0.5, -0.5)
  ypos = c(-0.5, -0.5, 0.5, 0.5)

  polygon( posg[1] +xpos , posg[2] + ypos, col=gray(1-posg[3]), border = gray(1-posg[3]))
  return(invisible())
}

initializeDigitPlot()
drawBox(posg=c(1,1,0.5))
drawBox(posg=c(1,2,0))
drawBox(posg=c(1,3,1))
```



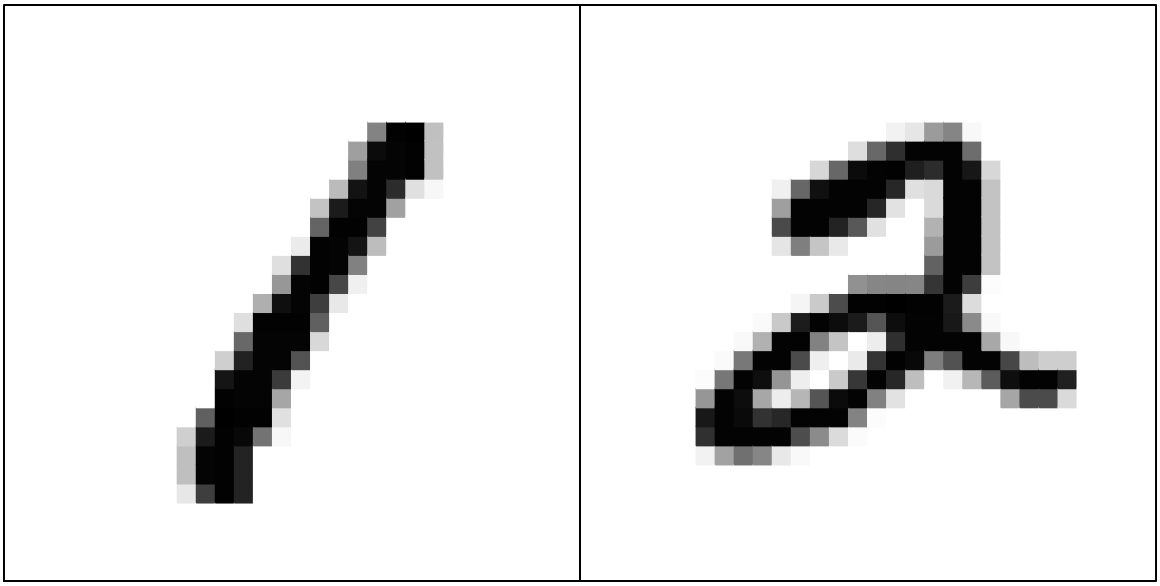
c) **[2 Marks]** Write a function `drawDigit` that accepts a numeric matrix of dimension 28×28 that calls initializeDigitPlot and drawBox to draw the matrix. The values of the matrix are brightness values.

```
drawDigit <- function(colx=NULL) {
  initializeDigitPlot()
  setx = rep(1:28, each=28)
  sety = rep(1:28, 28)
  setxyt = cbind(setx, sety, as.numeric(colx)/255)

  #for (i in 1:length(setx)) drawBox(setxyt[i,])
  temp = apply(setxyt, 1, drawBox)
  return(invisible())
}
```

d) **[2 Marks]** Import the files `one100.csv` and `two100.csv` digits file. These files contain one hundred 1's and one hundred 2's with $28 \times 28 = 784$ columans. Each row has brightness values of a digit made from the consecutive columns of a matrix. Use the `drawDigit` function to draw the first two digits from the files `one100.csv` and `two100.csv` side by side in a 1×2 grid.

```
d1 = as.matrix(read.csv("one100.csv", header=TRUE))
d2 = as.matrix(read.csv("two100.csv", header=TRUE))
```

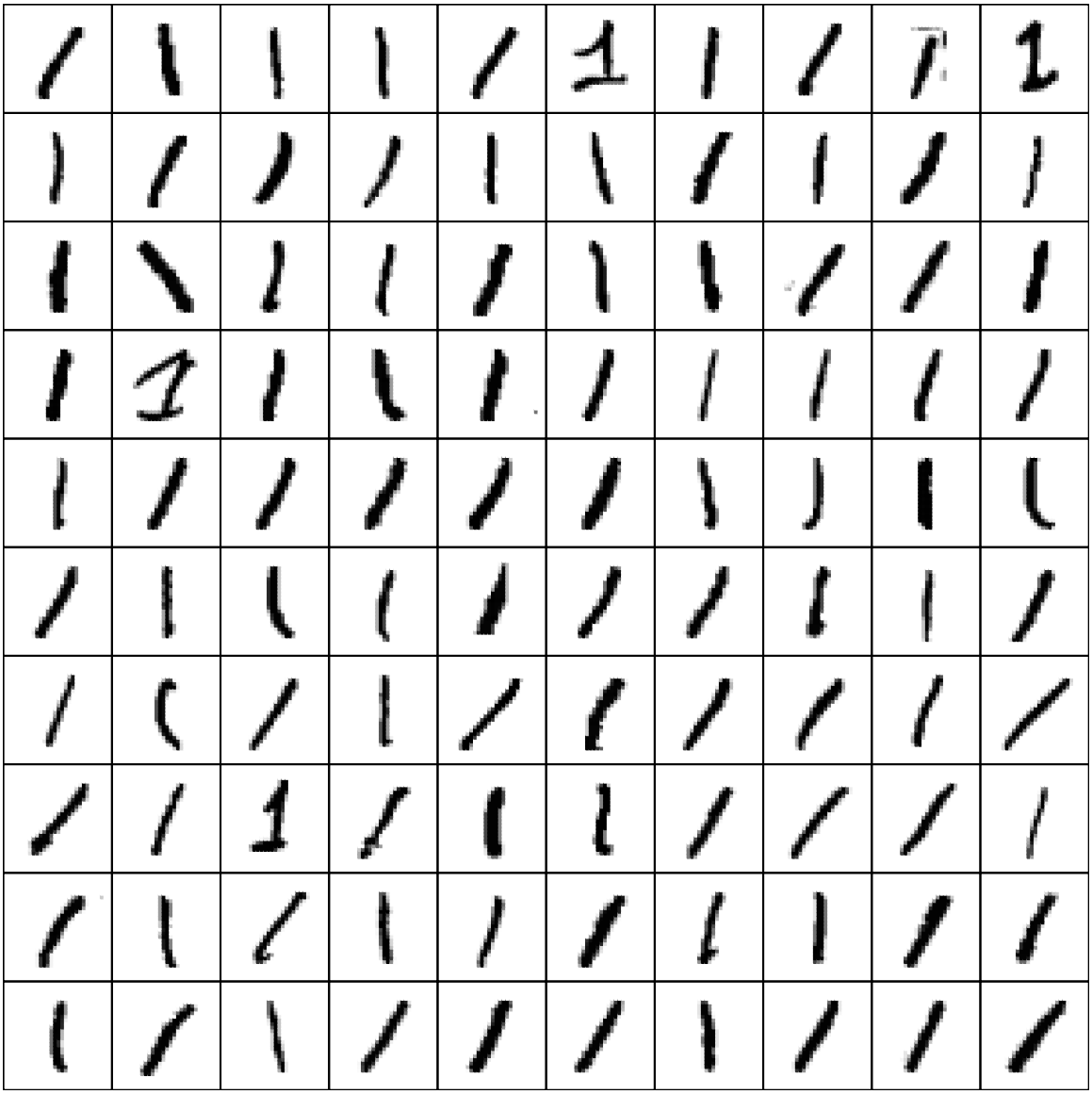
```
par(mfrow=c(1,2), mar=c(0,0,0,0))
drawDigit( matrix(d1[1,], 28, 28) )
drawDigit( matrix(d2[1,], 28, 28))
```



- It's okay if the digit is transposed or flipped. But the colour shouldn't be inverted.

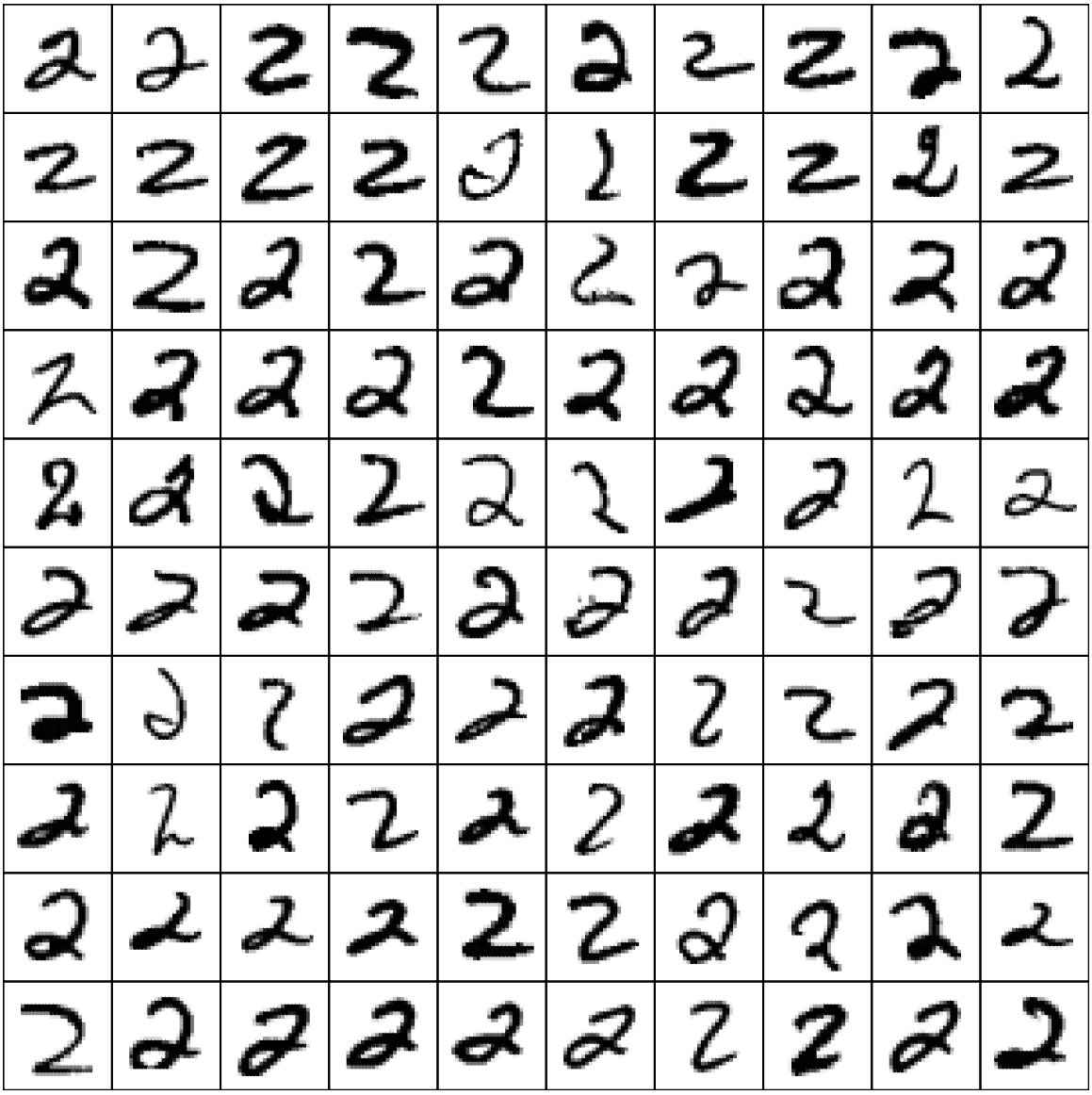
e) [**1 Mark**] Plot all the digits from `one100.csv` in a 10×10 grid.

```
par(mfrow=c(10,10), mar=c(0,0,0,0))
for (j in 1:100) drawDigit( matrix(d1[j,], 28, 28) )
```



f) [**1 Mark**] Plot all the digits from `two100.csv` in a 10×10 grid.

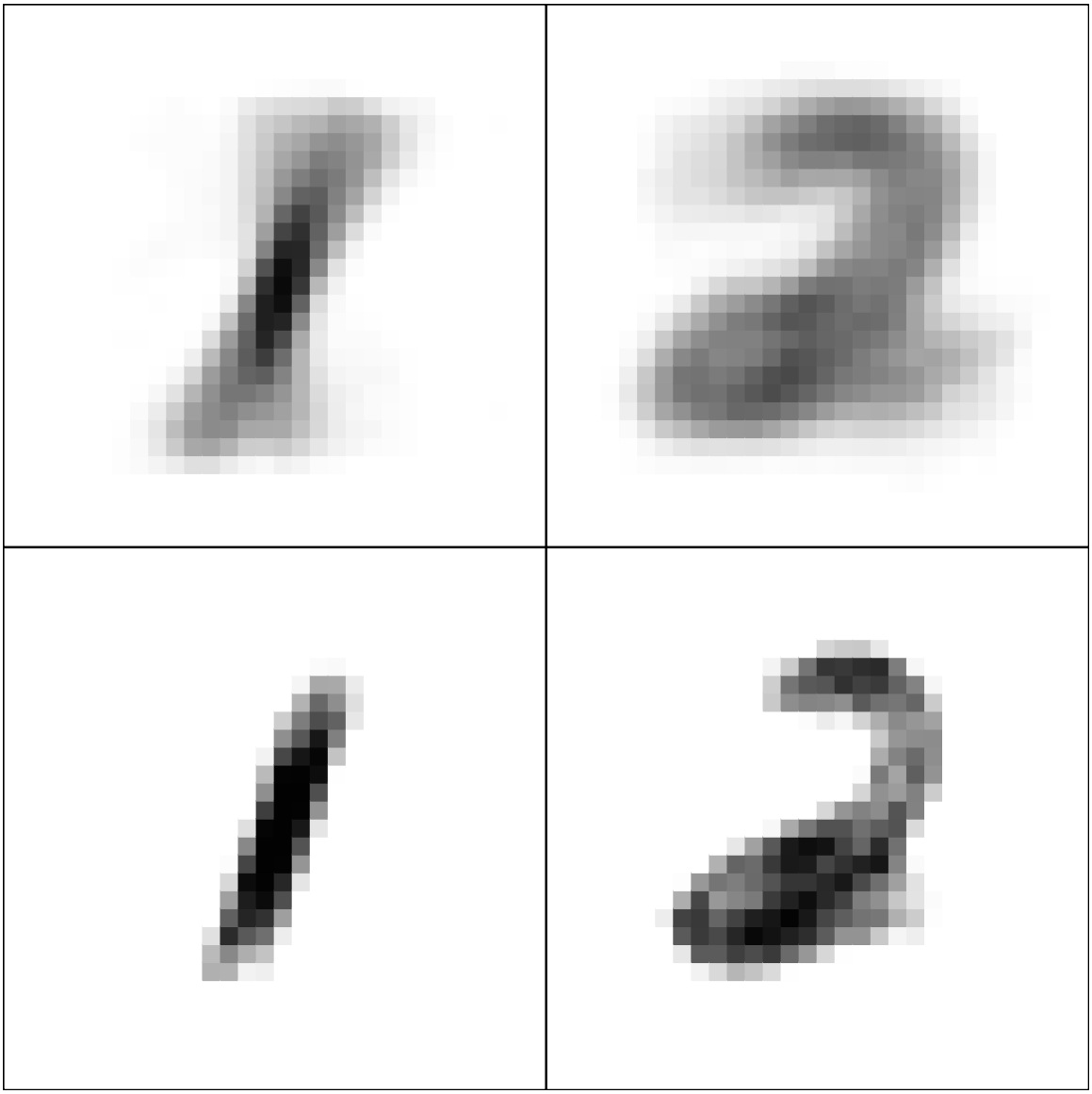
```
par(mfrow=c(10,10), mar=c(0,0,0,0))
for (j in 1:100) drawDigit( matrix(d2[j,], 28, 28)  )
```



g) **[4 Marks]** Summarize each of these populations by taking the average and median of the brightness within each pixel. Then plot the four resulting 28×28 matrices in a 2×2 grid using your `drawDigit` function. Compare and constrast using the average and median to summarize the images.

```
m1  = apply(d1, 2, mean)
m2  = apply(d2, 2, mean)

median1  = apply(d1, 2, median)
median2  = apply(d2, 2, median)
```

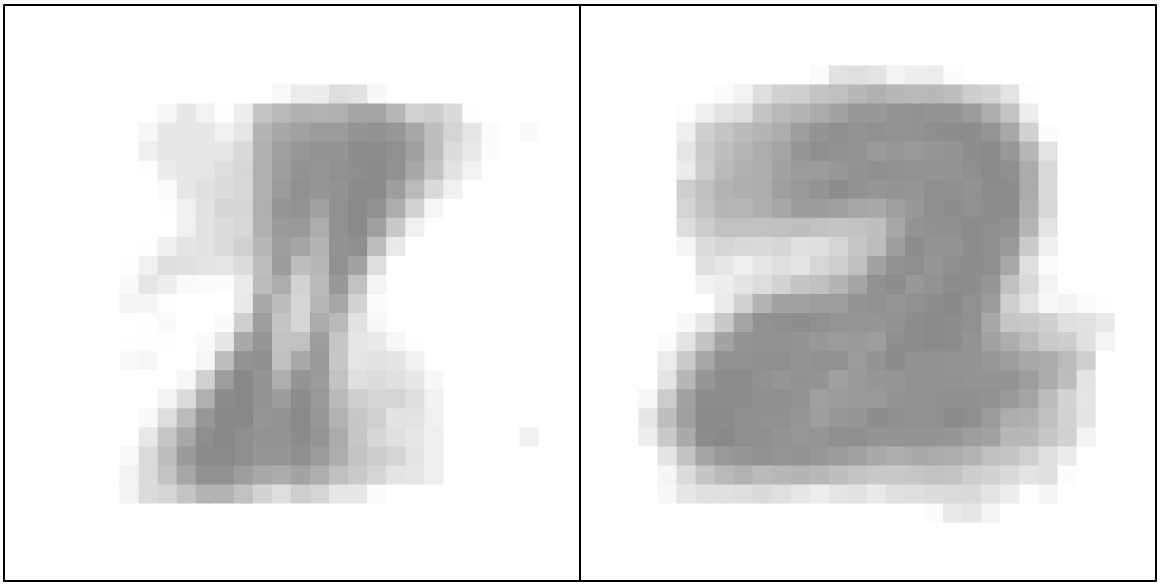
```
par(mfrow=c(2,2), mar=c(0,0,0,0))
drawDigit(m1)
drawDigit(m2)

drawDigit(median1)
drawDigit(median2)
```



- In all the plot we can sort see the digits.
- The median shows the digit 1 slightly better than the average.
- The average shows the digit 2 slightly better than the median

h) **[3 Marks]** Summarize the variability of these populations by taking the standard deviation of the brightness within each pixel. Then plot the two resulting 28×28 matrices in a 1×2 grid using your `drawDigit` function. Comment on the variability.

```
sd1  = apply(d1, 2, sd)
sd2  = apply(d2, 2, sd)

par(mfrow=c(1,2), mar=c(0,0,0,0))
drawDigit(sd1)
drawDigit(sd2)
```



- They digit both have pixels of zero variabiltiy surrounding the images.
- The two seems to have uniform variability around the average image.

- The one seems to almost uniform variability around the average image except for the middle pixels which have low variability and high averages.

Question Two - 17 Marks - Properties of Pearson's moment coefficient of skewness

Consider the population $\mathcal{P} = \{y_1, \ldots, y_N\}$. Pearson's moment coefficient of skewness is

$$a(\mathcal{P}) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^3}{[SD_{\mathcal{P}}(y)]^3}$$

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

- Note: Some marks in each part allocated to formating and organization.

(a) **[3 points]** Determine whether this skewness coefficient is location invariant, location equivariant, or neither.

From $\mathcal{P}^\star = \{y_1 + b, \ldots, y_N + b\}$, the average is location equivariant and the standard deviation is location invariant, we have

$$a(\mathcal{P}^\star) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u + b - \overline{y} - b)^3}{[SD_{\mathcal{P}^\star}(y)]^3} = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^3}{[SD_{\mathcal{P}}(y)]^3} = a(\mathcal{P})$$

The attribute is location invariant.

(b) **[3 points]** Determine whether this skewness coefficient is scale invariant, scale equivariant, or neither.

From $\mathcal{P}^\star = \{m \times y_1, \ldots, m \times y_N\}$, the average is scale equivariant and the standard deviation is scale equivariant, we have

$$a(\mathcal{P}^\star) = \frac{\frac{1}{N}\sum_{u\in\mathcal{P}}(m \times y_u - m \times \overline{y})^3}{[SD_{\mathcal{P}^\star}(y)]^3} = \frac{\frac{m^3}{N}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^3}{[m^3 SD_{\mathcal{P}}(y)]^3} = a(\mathcal{P})$$

The attribute is scale invariant.

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

We have that the attribute location invariant and scale invariant, so we need to show it also location-scale invariant. Let $\mathcal{P}^{\star\star} = \{m \times y_1 + b, \ldots, m \times y_N + b\}$ and $\mathcal{P}^\star = \{m \times y_1, \ldots, m \times y_N\}$ then

$$\begin{aligned} a(\mathcal{P}^{\star\star}) &= a(\mathcal{P}^\star) \quad \text{(The attribute is location invariant)} \\ &= a(\mathcal{P}) \quad \text{(The attribute is scale invariant)} \end{aligned}$$

The attribute is location-scale invariant.

(d) **[3 points]** Determine whether this skewness coefficient is replication invariant, replication equivariant, or neither.

Let $\mathcal{P}^k$ represent the population $\mathcal{P}$ replicated k times and using the noting that the average is replication invariant and the standard deviation with

$$N$$

is replication invariant we have

$$a(\mathcal{P}^k) = \frac{\frac{1}{Nk}\sum_{u\in\mathcal{P}^k}(\ y_u - \overline{y})^3}{[SD_{\mathcal{P}^k}(y)]^3} = \frac{\frac{k}{Nk}\sum_{u\in\mathcal{P}}(y_u - \overline{y})^3}{[SD_{\mathcal{P}}(y)]^3} = a(\mathcal{P})$$

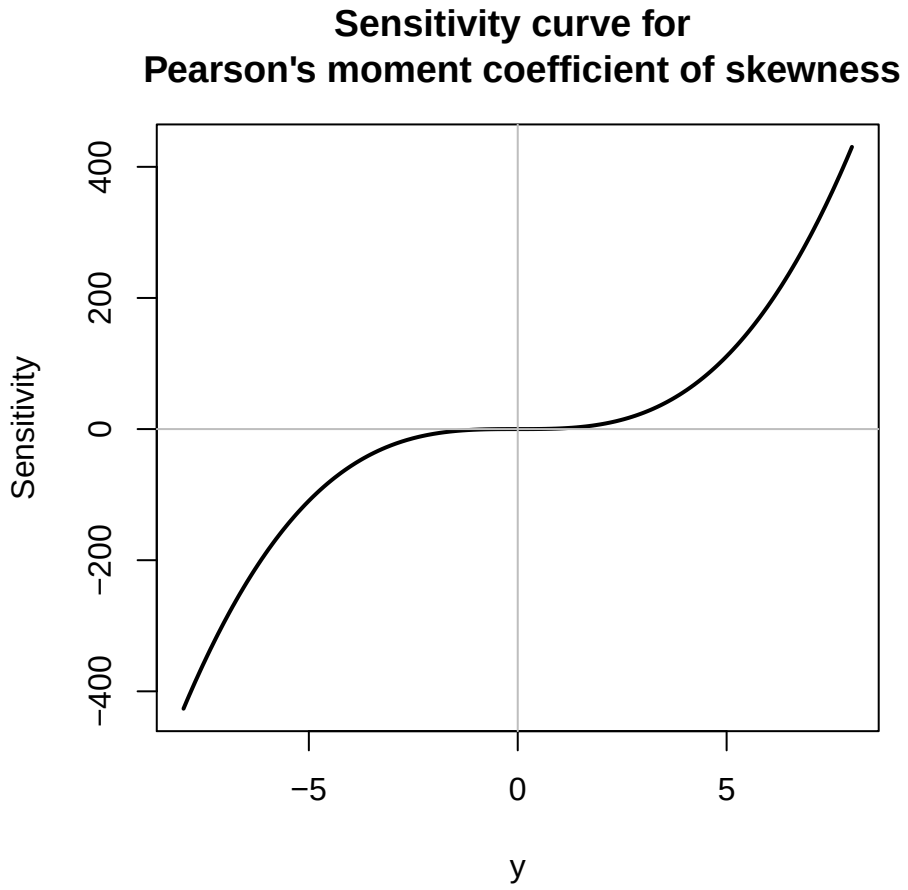
When using

$$N$$

for the standard deviation, the attribute is replication invariant.

(e) **[3 points]** For the population below, plot the sensitivity curve of $a(\mathcal{P})$ for $y \in [-8, 8]$. You may find the `sc()` function from lecture useful.

```r
set.seed(341)
pop <- rnorm(1001)

sdn <- function( y.pop ) {
  N = length(y.pop)
  sd(y.pop)*sqrt( (N-1)/(N) )
}

skewness <- function(pop=NULL) {
  ybar = mean(y)
  mean( (pop - ybar )^3 )/sdn(pop)^3
}

sc = function(y.pop, y, attr, ...) {
  N <- length(y.pop) +1
  sapply( y, function(y.new) {  N*(attr(c(y.new, y.pop),...) - attr(y.pop,...))  } )
}


y <- seq(-8,8, length.out=1000)

plot(y, sc(pop, y, skewness), type="l", lwd = 2,
     main="Sensitivity curve for \n Pearson's moment coefficient of skewness",
     ylab="Sensitivity")
abline(h=0, v=0, col="grey")
```



(f) **[2 points]** Given all that you have learned in parts (a) - (f), state one thing that is *good* about the pearson's moment coefficient of skewness attribute and one thing that is *bad* about the range attribute.

Good: It is location-scale invariant which preferred for measures of asymmetry. **Bad:** It seems to sensitve to large outliers because its sensitivity curve is unbounded.

Question Three - 10 Marks - Finding and describing a population.

By searching the web, find a public dataset that constitutes a population. Provide the following:

- Give a description of the data **in your own words**. Then justify why the dataset is indeed a population as opposed to a sample.
- A URL to access the data.
- Define what is an unit and describe two variate(s) that have been recorded.
- Give a single graphical display of the populaton.
- Describe some interesting attribute or feature of this population.

Some places you might consider looking:

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

Rubric

Criteria	Descriptor	Marks
Data/URL	Creativity: Was an interesting or unique dataset chosen, provided?	/2
Description	Clarity & Justification	/2
Unit/Variate	Description, Correctness and Graphic	/4
Feature	Description and Justification	/2

Question Four - 10 Marks - Review

A student (Ryan Browne) was given the following question.

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

- You are limited to 1 to 2 pages.
- Your solution should use a combination of formulas, full sentences and an example.

Rubric

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

- The student's solution is provided on LEARN in the file "Sensitivity_Question.pdf". Line numbers have been added to each page to help refer to certain parts of the report.

a) **[3 Marks]** Give 3 examples of improper formatting, LaTex or grammar

Here are some examples but there might others not on this list.

- Page 1, line 13, the population size "N-1" should be typeset as LaTeX as $N-1$.
- Page 1, line 15, the sequence of typset equations should better presented. Specifically, there is required punctuation before $\mathcal{P}^*$. It should also be aligned, with $\mathcal{P}$ and $\mathcal{P}^*$ introduced properly.
- Page 1, "variate of with the value" is grammatically incorrect and "sensitivity cruve" is misspelled.
- Page 2, line 9, the order statistic $y_{(1)}$ is missing the bracket around the index.
- Page 2, line 17, the comparison operator ">=" should be $\geq$.
- Page 2, lines 20-22, the align environment does not have alignment points correctly delimited. (More simply, the equal signs do not line up.)
- Page 2, in the code chunk, the code flows off the page and is unreadable.
- Page 3, the figure does not have an informative y-axis label.
- Page 3, line 19, the grammar for the sentence is incorrect. ("is unbounded" should be "being unbounded".)
- The horizontal line at y = 0 could be made more visible.
- They student went over the page limit.

b) **[2 Marks]** Give 2 examples of errors in content or gaps in knowledge.

Here are some examples but there might others not on this list.

- Page 1, the definition of $\mathcal{P}$ is incorrect.
- Page 1, line 15, the current fraction is a bit squished. The equation could be improved by using the LaTex command dfrac instead of frac.
- Page 1, line 16, $a(P)$ should be mentioned before defining SC.
- Page 2, lines 10-13, the solution is incorrect; $y < y_{(N-1)}$ does not guarantee $y_{(1)} = y$.

- Page 2, lines 14-16, the solution is unfounded; $a\left(y_1, \ldots, y_{N-1}\right)$ is not shown to be y (nor would it be correct).
- Page 2, line 8, the definitions of $\mathcal{P}$ and $\mathcal{P}^*$ are inconsistent with page 1. The second equality in both definitions is non-sensical, and the lack of y in $\mathcal{P}^*$ is incorrect.
- Page 2, lines 18-19, the solution is incorrect; y_{N-1} is not guaranteed to be $y_{(1)}$.
- Page 2-3, the example data is inadequately described.

c) **[2 Marks]** Did the student summarize the concept of sensitivity in their own words? Briefy explain how and provide some evidence or examples.

The style and format resembles the examples given the lectures. Some notable things are

- Page 1, lines 20-23, this is copied verbatim from the notes. In particular, the attribute has been assumed to be the average.
- The student seems be for e.g. Page 2, line 6, the section is "Minimum", but the definition is for a maximum.

d) **[1 Mark]** Give one suggestion on how to improve the report.

Fixing any of the errors above would improve the report. Some high level suggestions are

- They student could condense the summarize to reach the 2 page limit.
- The description of the data example could be expanded.

When to writing a summarize in your words try the following

- Review the lecture and take your own notes,
- Write a draft summary or reprot after waiting an a hour or maybe even the next day.
- While writing your a draft summary do not look at the lecture notes but maybe review your own written notes.
- After finishing your draft wait another an a hour or maybe even the next day before editing.
- While editing it open the notes and fix any possible error or omissions.

e) **[2 Marks]** Give this report a grade and briefy describe why this grade is appropriate.

A low grade is appropriate because the report lacks originality and does not seems to be in the student's own words. In additon, there are numerous formatting, LaTex or gramm errors.