MS 204 Exam I

Your Name Here

General instructions for Midterms:

- Create a new Markdown file
- Change the heading to include your author name
- Save the R Markdown file (named as: [MikeID]-[MidtermI].Rmd – e.g. "mlopez-MidtermI") to somewhere where you'll be able to access it later (zip drive, My Documents, Dropbox, etc)
- Your file should contain the code/commands to answer each question in its own code block, which will also produce plots that will be automatically embedded in the output file. **Plots and answers must show up as code output to receive full credit for any relevant questions**
- Each answer must be supported by written statements (unless otherwise specified) as well as any code used
- A printed HTML or PDF copy of this midterm is due in class on Tuesday at 5:30 PM. I will not answer any questions on this exam after 12:00 noon on Tuesday I am avaialable over email through that point, and also in my office prior.
- Each student must abide by Skidmore's honor code. You may not use the internet to solicitating answers or communicate with other students, or to check for any help on any question. Your book, class notes, activities, and labs are sufficient for finishing this exam.

Part I: short answer (4 points)

For each of the following examples, explain in no more than 2 sentences. If you respond in 3 or more sentences, I will only grade the first two sentences.

1. Describe the difference between standard deviation and standard error.
2. Identify why randomized designs make it easier to infer cause and effect than observational studies.
3. Describe the difference between ordinal categorical and nominal categorical variables, and provide one reason why it could be important to distinguish between these two types.
4. Why does `pnorm(0) = 0.5`?
5. Identify one aspect of a distribution that shows up in a boxplot but not in a histogram, and identify what shows up in a histogram but not in a boxplot.
6. SAT math scores are normally distributed with mean 500 and standard deviation 100. Identify the score corresponding to the top 35 percent of SAT math scores.

For questions (7) - (10), answer using the following example.

A line of best fit between student test score given hours studied is as follows:

$\hat{testscore} = 75 + 0.5 * hours$

7. Identify and interpret the slope estimate.
8. Shelby studied 3.3 hours and scored a 85. Estimate Shelby's residual.
9. A statistician identifies that it is unlikely the observed link between hours studied and test performance is due to chance (e.g, it is statistically significant). Do you think confounding is possible? If so, name a possible confounding variable.
10. Is reverse causation possible? Briefly justify your answer.

Part I: long answer (5 points each)

For this question, we are going to use the airlines data set.

```
library(oilabs)
library(mosaic)
library(tidyverse)
data("nycflights")
```

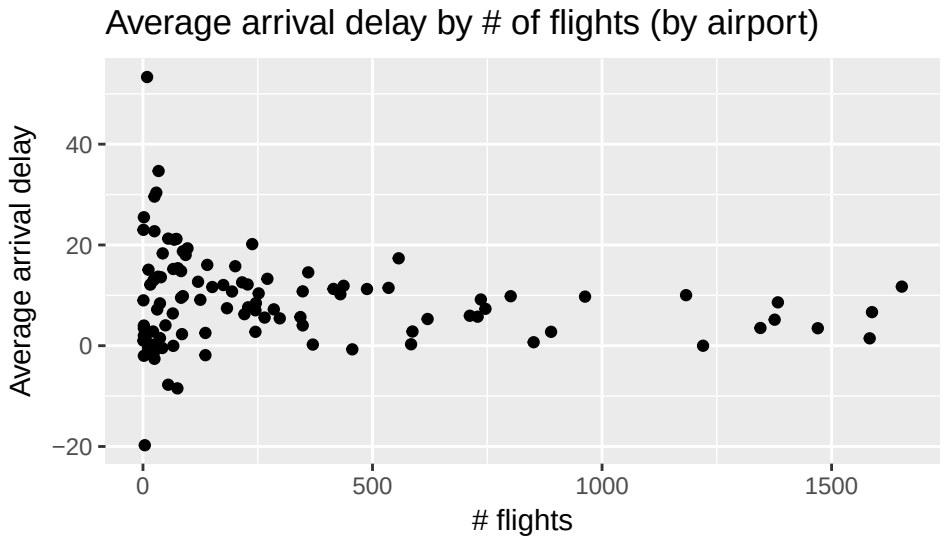
1. Identify the information provided (the output) when running the following code.

```
flight.sum <- nycflights %>%
  group_by(dest) %>%
  summarise(y.var = mean(arr_delay, na.rm = TRUE), n = n()) %>%
  arrange(-y.var) %>%
  filter(n > 200) %>%
  head(3)

flight.sum
```

```
## # A tibble: 3 x 3
##    dest    y.var     n
##   <chr>    <dbl> <int>
## 1   RIC 20.16387   238
## 2   BNA 17.35189   557
## 3   MCI 15.77612   201
```

2. Produce the following graph (*hint*: the code above will help you get started). Note that each point reflects a destination airport.



3. Why does the plot (in Question 2) look like a funnel shape? In other words, why is there more noise earlier in the data but not later? Should we not be surprised?

4. One airport that does not boast a particularly impressive performance is Richmond (`RIC`). Look within each month to determine if average delay times at Richmond seem to be tied to the month in which the flight left.

5. The president of the Richmond Airport Association hires you to look through the data to see if there are any flaws in the calculations above. Briefly (and without using additional code), explain why averages may be inappropriate for measuring delays, and propose a better metric.

Part II: long answer (4 points each)

Note: Q9 worth 3 points each

We'll stick with the airlines data set.

Given flight conditions and after accounting for a few other confounding variables, the Federal Aviation Association makes note that one would have expected departure delays on exactly 29 percent of flights between Boston and the NYC airports.

A deputy in Boston investigates the data.

```
nycflights %>%
  filter(dest == "BOS") %>%
  mutate(is.delay = dep_delay > 0) %>%
  summarise(n.delays = sum(is.delay, na.rm = TRUE),
            n.flights = n(),
            delay.rate = n.delays/n.flights)
```

```
## # A tibble: 1 x 3
##   n.delays n.flights delay.rate
##      <int>     <int>      <dbl>
## 1      475      1470  0.3231293
```

1. Write the null and alternative hypotheses in words or symbols.
2. Should we consider a one or two proportion test?
3. Is a one sided or two sided test most appropropriate for this example?
4. Develop a chance model that represents 1000 samples of delay outcomes from 1470 flights that we could expect under the null distribution.
5. Graph the chance model, and visually identify if the observed delay probabilities are surprising or not that surprising.
6. Estimate a p-value for this data.
7. Interpret your $p - value$.
8. Summarize your findings using non-technical terms (no more than 2 sentences, accessible for someone outside of statistics)
9. Why might a representative from Laguardia airport (origin `LGA`) complain about your findings in the parts above?

Part IV

Write the Skidmore honor code (handwritten): While taking this examination, I have not witnessed any wrongdoing, nor have I personally violated any conditions of the Skidmore College Honor Code.