# IMT 573: Problem Set 5 - Statistics

Stephen V. Tucker

Due: Friday, November 11, 2022

**Collaborators:**

**Instructions:** Before beginning this assignment, please ensure you have access to R and RStudio.

1. Download the `problemset5.Rmd` file from Canvas. Open `problemset5.Rmd` in RStudio and supply your solutions to the assignment by editing `problemset5.Rmd`.
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment. In particular, note that Stack Overflow is licenses as Creative Commons (CC-BY-SA). This means you have to attribute any code you refer from SO.
4. Partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. But please **DO NOT** submit pages and pages of hard-to-read code and attempts that is impossible to grade. That is, avoid redundancy. Remember that one of the key goals of a data scientist is to produce coherent reports that others can easily follow. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

6. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit PDF`, rename the knitted PDF file to `ps5_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.
7. Collaboration is often fun and useful, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.

**Setup:** In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
```

### Problem 1: Overbooking Flights

Airlines frequently overbook passengers on flights. You are hired by *Air Nowhere* to recommend the optimal overbooking rate for their flights. Air Nowhere is a small airline that uses a 100-seat plane to carry you from Seattle to, well, nowhere. The tickets cost $100 each, so a fully booked plane generates $10,000 in revenue. The sales team has found that the probability of passengers who have paid their fare actually showing up is

98% and that showing up for each passenger can be considered independent. The additional costs associated with finding alternative solutions for passengers who are refused boarding are $500 per person.

1. Which distribution would you use to describe the actual number of passengers who show up for the flight? *Hint: read the Open Intro Statistics (OIS) chapter on distributions.*

   **Answer: Binomial as we can measure how by number of successes.**

2. Assume the airline never overbooks (i.e. it only sells 100 tickets per flight). What is the expected profit? Expected profit means expected income or revenue from ticket sales minus the expected costs related to alternative solutions.

   **Answer: 1 x 100 x 10 - 500 per ind = Expected profit**

3. Now assume the airline sells 101 tickets for the 100 seats on each aircraft. What is the probability that all 101 passengers will show up? *Hint: note that passengers showing up or not is a binary outcome. What probability distribution would you go for under this scenario?*

   - P = binomial probabilitiy
   - x or k = number of times for a specific outcome
   - p = probability of success on a single trial
   - q = probability of failure on a single trial
   - n = number of trials

   **problem set**

   - K = 101
   - probability of success (p) = 98%
   - n = 1

```
probability_1 <- dbinom(101,1, .98) # There is a 0 probability that exist that all will show up to

print(probability_1)
```

```
## [1] 0
```

4. Now assume the airline sells 102 tickets for the 100 seats. What is the probability that all 102 passengers show up?

   **solution same application as 3.**

   **Answer**

```
probability_2 <- dbinom(102,1, .98) # 0
print(probability_2)
```

```
## [1] 0
```

5. What is the probability that 101 passengers – still one too many – will show up when 102 seats have been sold?

   - 99% success rate

**Answer**

```
probability_3 <- dbinom(101,1, .98)
```

6. What does it mean that the probability of passengers showing up is independent? Why is it important in this case? Is this realistic - why or why not?
    - In this case it means that we are running the probability of each individual person/ticket will show up. This is important because treating it as such that all show up as a group will then it will increase the odds thus raising the probability.

Note: some of the expressions may be hard to write analytically. Feel free to use R for calculations but be sure to show the code and explain what you are doing.

### Problem 2: The Normal Distribution

In this problem, we will explore data and ask whether it is approximately normal. We will consider two different datasets, one on height and one on research paper citations.

#### (a) Let's start with the human height data.

1. How should human height be measured (e.g. What type of variable should you use? Should it be continuous or discrete? Positive or negative?...)?
    - In measuring human height we should use numeric data specifically interval continuous. the standard is inches, we should not be able to measure negative inches as well as it should be continuous.
2. Read the `fatherson.csv` dataset into R. It contains two columns - one for a father's height and one for their son's height (in cm). Let's focus on the father's height for a moment (variable `fheight`). Provide basic descriptive and summary stats of this variable (e.g. What do descriptive stats look like? How many observations do we have? Do we have any missing data?...)

```
fatherson <- read.delim("~/Downloads/fatherson.csv")
```

3. Compute the mean, median, mode, standard deviation and range of the fathers' heights. Discuss the relationships between these numbers. Is mean larger than median? What does this imply? Is mean larger than mode? By how much (in relative terms)? What does this suggest? How does standard deviation compare to mean?

**Father mean, median & mode**

- Mean: 171.93
- Median: 172.1
- mode: 175 in count of 13
- standard deviation 6.972346
- range: 149.9-191.6

mean and median are relatively close, but mean is slightly higher by 17cm, this implies that the middle value is distribution is representative of 0 and that The relationship that exist between these numbers are

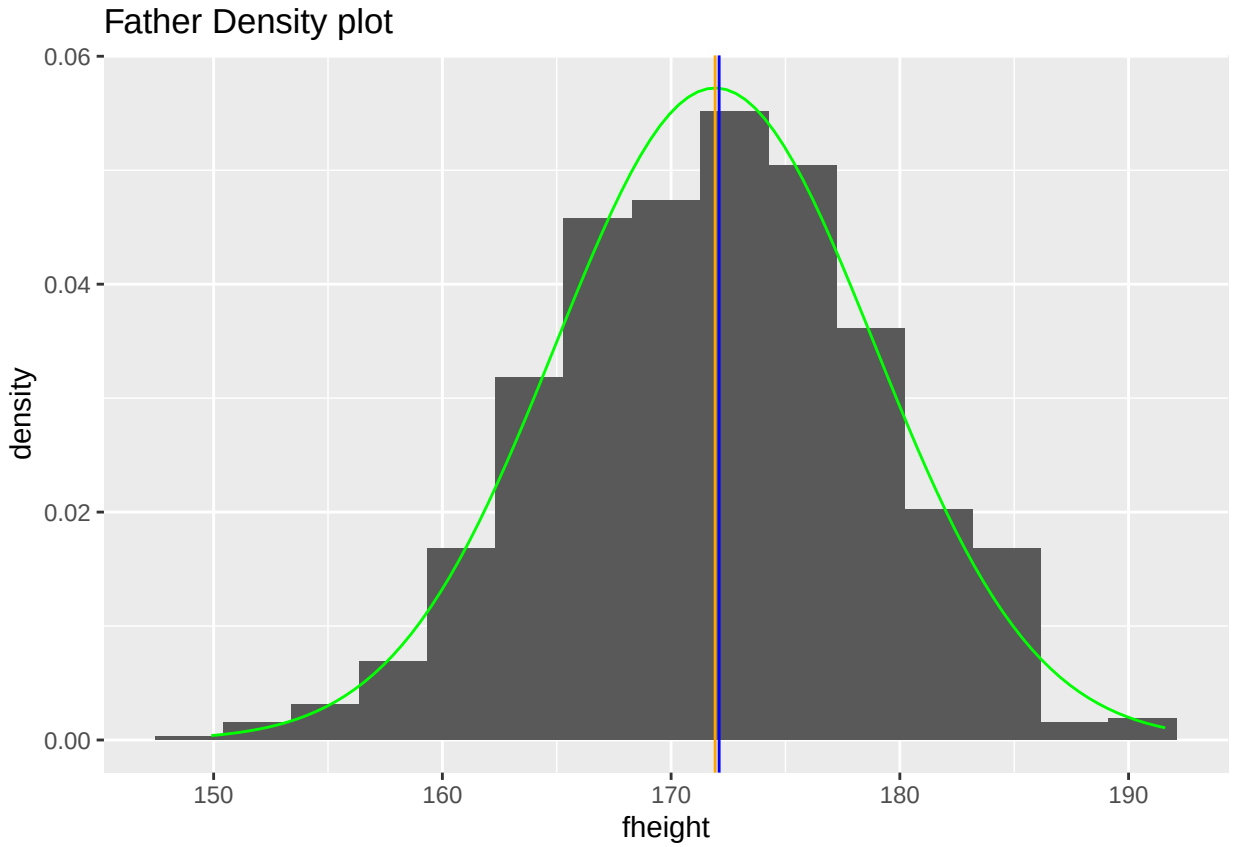
```
#Father: Mean, Median Mode
mean_father <- mean(fatherson$fheight) # mean - 171.93

median_father <- median(fatherson$fheight) # median- 172.1

## Father mode
```

```r
mode_father <- group_by(fatherson,fheight)
#
mode_father <- summarise(mode_father, count = n()) # 175 count of 13
#
mode_father <- arrange(mode_father, desc(count))

## standard deviation

std <- sd(fatherson$fheight) # 6.972346

# range

range(fatherson$fheight) # 149.9 191.6
```

```
## [1] 149.9 191.6
```

4. Plot a histogram of the data. On the same plot, overlay a plot of the normal distribution with the same mean and standard deviation as the father's height data (use transparency, add an outline, or both). Additionally, indicate the mean and median of the data using vertical lines of different colors and indicate these on the legend. What do you find? Are the histogram and the plot of the normal distribution similar?

```r
# Density plot graph
ggplot(data=fatherson, aes(x=fheight)) +
  geom_histogram((aes(y=..density..)),bins=15) +
  stat_function(
    fun = dnorm,
    args = list(mean = mean(fatherson$fheight), sd = sd(fatherson$fheight)),
    col = 'green'
 ) +
  geom_vline(xintercept = mean_father,col= 'orange') +
  geom_vline(xintercept = median_father,col= 'blue') +
  ggtitle("Father Density plot")
```




**(b) Next, let's take a look at the number of citations of research papers.**

1. How should citation counts (i.e. the number of times that a paper is referenced by other papers) be measured (e.g. What type of variable should you use? Should it be continuous or discrete? Positive or negative?...)?
    - We should measure citation counts as a continuous and ratio.
2. Read the `mag-in-citations.csv` data. This is data from the Microsoft Academic Graph for citations of research papers and it contains two columns: paper id and number of citations. We only care about citations here. Provide basic descriptive and summary stats of this variable as you did with the height data.
    - The variables that are present are paper ID, and a citations. one identifies a document the other counts total number of times its been cited.
    - basic summary statistics of this data set shows the that the citations data frame has extremes being that the min is 0 1st quartile is 1.00 and the third is 12.00 and the max is 18682.00. This is indicative that the data may have outlier's.

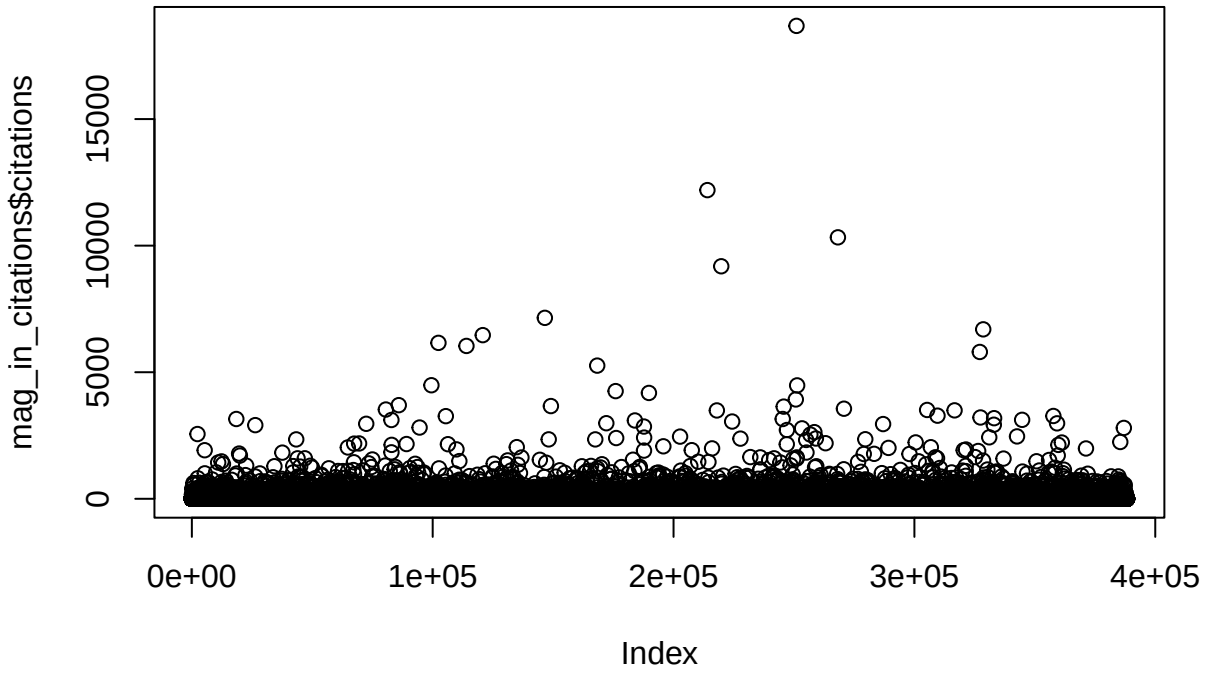
```
mag_in_citations <- read_csv("mag-in-citations.csv")
```

```
## Rows: 388258 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): paperId, citations
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
summary(mag_in_citations)
```

```
##     paperId            citations
##  Min.   :1.304e+04   Min.   :    0.00
##  1st Qu.:1.981e+09   1st Qu.:    1.00
##  Median :2.074e+09   Median :    3.00
##  Mean   :1.955e+09   Mean   :   15.61
##  3rd Qu.:2.278e+09   3rd Qu.:   12.00
##  Max.   :2.794e+09   Max.   :18682.00
```

```
# plot to see the data points
plot(mag_in_citations$citations)
```




3. mag_in_citations_mode <- arrange(mag_in_citations_mode, desc(citations))
4. Compute mean, median, mode, standard deviation and range of the citations.Discuss the relationships between these numbers. Is mean larger than median? What does this imply? Is mean larger than mode? By how much (in relative terms)? What does this suggest? How does standard deviation compare to mean?
    - The mean is 15.61223 meaning the average number
    - The median is 3 meaning middle, and is useful in giving information on the distribution since the mean and median are so wide apart this signals that the distribution is not even.
    - Mode is 18682: the citation number of this mode is 2034269086
    - range 0 18682

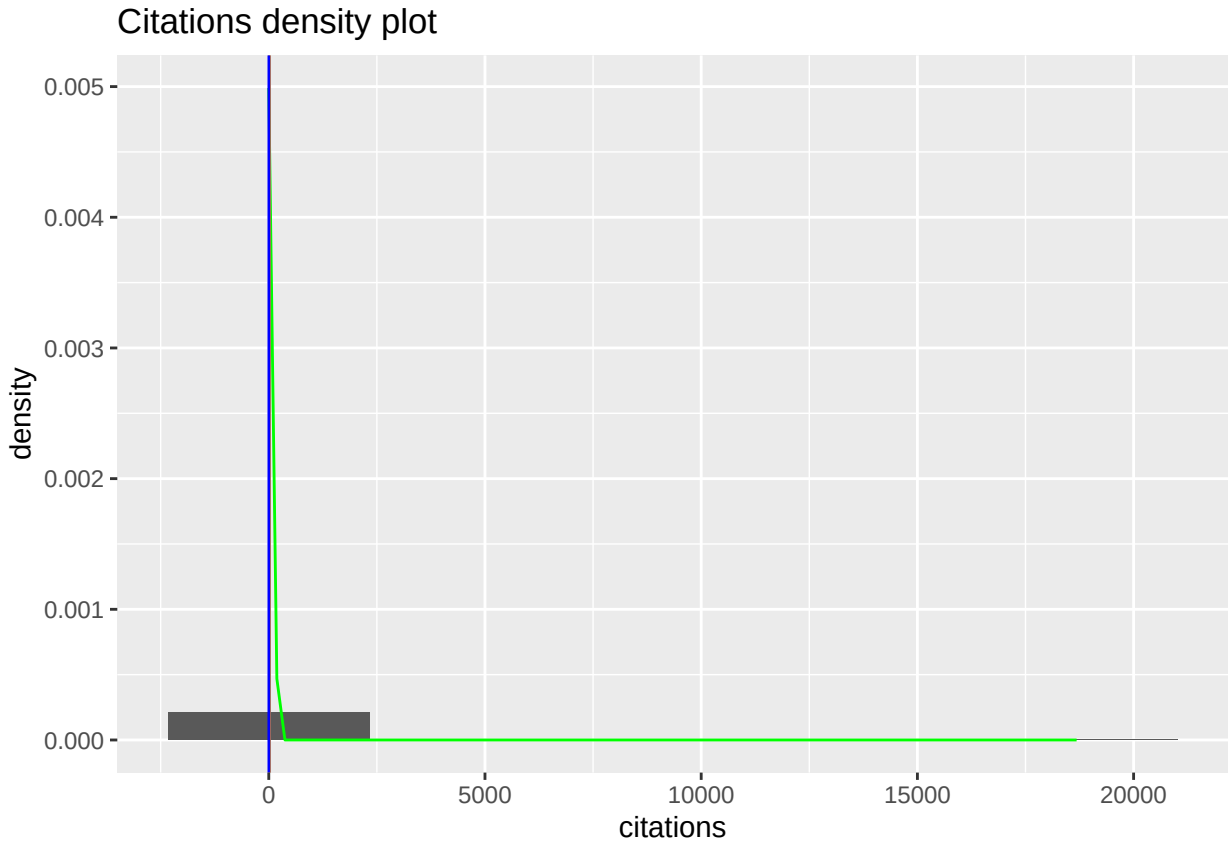
```
## mean
mag_in_citations_mean <- mean(mag_in_citations$citations) # 15.61223
#
mag_in_citations_median <- median(mag_in_citations$citations) # 3

#### What do i need to do to get the keep paper identifier the same.
```

```
mag_in_citations_mode <- group_by(mag_in_citations,citations)
#
mag_in_citations_mode <- arrange(mag_in_citations_mode, desc(citations))
# i.d = 2034269086:  count - 18682

## Standard deviation

mag_sd <- sd(mag_in_citations$citations) # 78.39079

# Range

range(mag_in_citations$citations) #  0 18682
```

```
## [1]     0 18682
```

5. Plot a histogram of the data. On the same plot, overlay a plot of the normal distribution with the same mean and standard deviation as the father's height data (use transparency, add an outline, or both). Additionally, indicate the mean and median of the data using vertical lines of different colors and indicate these on the legend. What do you find? Are the histogram and the plot of the normal distribution similar?

```
ggplot(data=mag_in_citations, aes(x=citations)) +
       geom_histogram(aes(y = ..density..),bins=5) +
      stat_function(
        fun = dnorm,
        args = list(mean = mean(mag_in_citations_mean), sd = mag_sd),
        col = 'green'
     ) +
      geom_vline(xintercept = mag_in_citations_mean,col= 'orange') +
      geom_vline(xintercept = mag_in_citations_median,col= 'blue')+
   ggtitle("Citations density plot ")
```

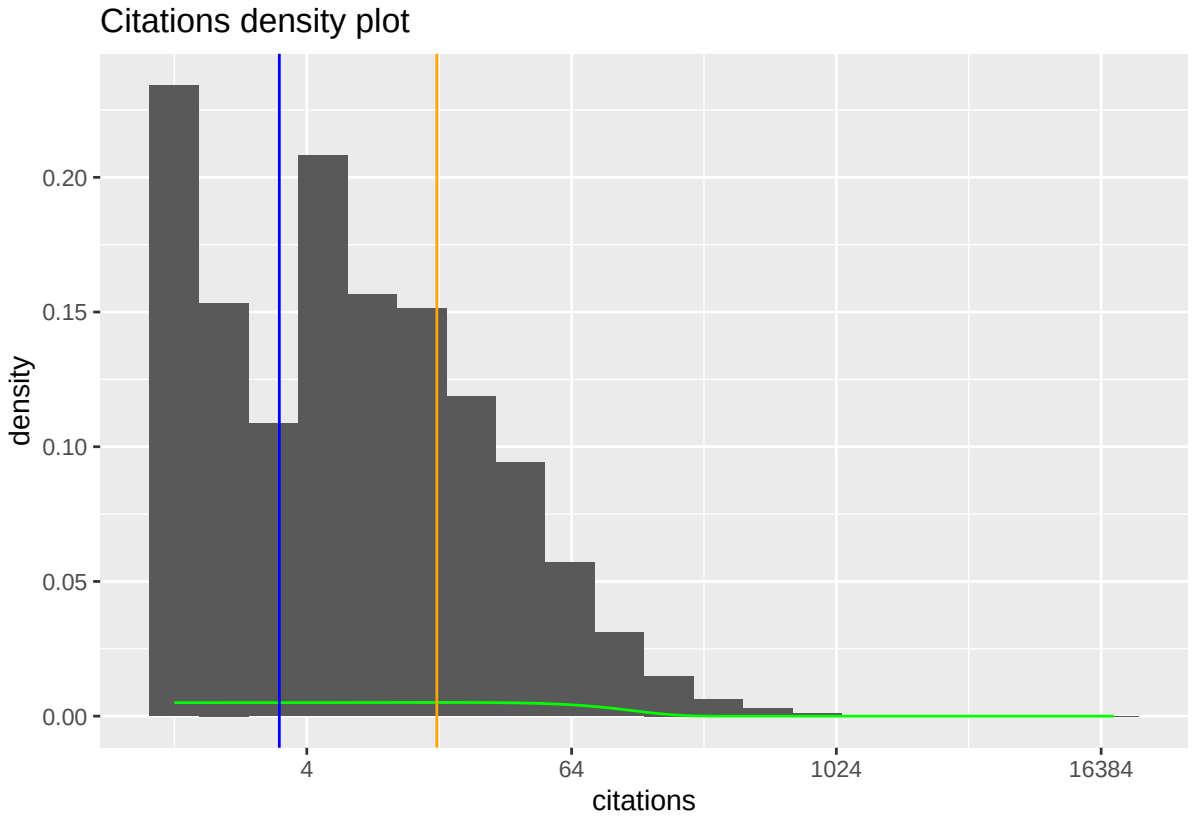



6. Now, repeat part 4 but using what is called a "log-log" transformation - plotting the x and y axes for the citation data on a logarithmic scale. As before, overlay a plot of the normal distribution by scaling the mean and standard deviation of the normal distribution to be the log of the mean and log of the standard deviation of the citation data. What do you see?

```
ggplot(mag_in_citations, aes(x=citations)) +
  geom_histogram(aes(y = ..density..),bins=20) +
  stat_function(
    fun = dnorm,
    args = list(mean = mag_in_citations_mean, sd = mag_sd),
    col = "green"
 ) +
  scale_x_continuous(trans = 'log2')+
  geom_vline(xintercept =mag_in_citations_mean,color= 'orange') +
  geom_vline(xintercept =mag_in_citations_median,color= 'blue')+
   ggtitle("Citations density plot ")
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
## Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 84550 rows containing non-finite values (stat_bin).
```




**(c) Comment on your findings from part (a) and part (b). Be sure to compare the two cases. That is, seeing how well (or not) the heights and the citations data align with the normal distribution. What are your thoughts on these datasets and do the findings make sense with respect to what we'd expect to see concerning heights and influence (as measured by citations) in the real world?**

- In my findings the heights fit the model better according to this data set the. The citations data have a right tail distribution as well. In regards of the heights and the counts I am not sure whats going with the citation distribution a