# IMT 573: Problem Set 5 - Bayes Theorem & Distributions

Vighnesh Misal

Due: Tuesday, November 5, 2019

*Collaborators: Ashish Anand*

*Instructions:*

Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Server.

1. Download the `problemset4.rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problemset4.rmd` in RStudio and supply your solutions to the assignment by editing `problemset4.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. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.

4. Collaboration on problem sets is fun and useful, and we encourage it, 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.

5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.

6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are 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 withouth errors you can do so with the `eval=FALSE` option.

7. 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 `ps4_ourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

*Setup*

Load any R packages of interest here.

```
library('dplyr')
library('tidyverse')
library('ggplot2')
```

**NOTE: You do not need to perform all calculations in R. Writing them in LaTeX and/or plain text is completely fine. However, be sure your work is readable and understandable. If you do solve problems programmatically, clearly describe your approach and what you are doing.**

### Problem 1: Overbooking Flights

You are hired by Air Nowhere to recommend the optimal overbooking rate. It is a small airline that uses a 100-seat plane to carry you from Seattle to, well, nowhere. The tickets cost $100 each. The sales team has found that the probability, that the passengers who have paid their fare actually show up is 98%, and individuals showing up can be considered independent. The additional costs, associated with finding an alternative solutions for passengers who are refused boarding are $500 per person.

*(a) Which distribution would you use to characterize the actual number of people who show up for flights?*

## Binomial Distribution with N = 100, p = 0.98

*(b) Assume the airline never overbooks. What is the expected revenue from a full flight in this scenario? Expected revenue is the expected income from ticket sales minus expected costs related to alternative solutions.*

```
100*100
```

```
## [1] 10000
```

*(c) Now assume the airline sells 101 tickets for 100 seats on a given flight. What is the probability that all 101 passengers will show up?*

```
dbinom(101,101,0.98)
```

```
## [1] 0.1299672
```

*(d) What are the expected profits (where profits are revenue - expected additional costs) when the airlines sells 101 tickets for 100 seats? Would you recommend overbooking or selling just the right number of tickets per flight?*

```
101*100 - 500*pbinom(100,101,0.98,lower.tail = FALSE)
```

```
## [1] 10035.02
```

*(e) Now assume the airline sells 102 tickets for 100 seats on a given flight. What is the probability that all 102 passengers show up?*

```
dbinom(102,102,0.98)
```

```
## [1] 0.1273678
```

*(f) What is the probability that 101 passengers - still one too many - will show up when 102 tickets are sold for a given flight?*

```
dbinom(101,102,0.98)
```

```
## [1] 0.265133
```

*(g) Would it be advisable to sell 102 tickets, 101 tickets, or 100 tickets for a given flight if the airline wanted to maximize revenue? (i.e. which has the highest expected revenue: selling 100, 101, or 102 tickets?*

```
100*100
```

```
## [1] 10000
```

```
101*100 - 500*pbinom(100,101,0.98,lower.tail = FALSE)
```

```
## [1] 10035.02
```

```
102*100 - 500*pbinom(100,102,0.98,lower.tail = FALSE)
```

```
## [1] 10003.75
```

*(h) What is the optimal number of seats to sell for the airline? How much are expected profits the expected profits in this case?*

```
101*100 - 500*pbinom(100,101,0.98,lower.tail = FALSE)
```

```
## [1] 10035.02
```

**The airline should sell 101 tickets in order to maximize their revenue.**

*(g) What does it mean to state that individuals showing up for a flight are independent? Why is this important in this case?*

**Binomial distribution can only be used for events that have only two possible outcomes with a fixed probability for each outcome i.e. they should be mutually exclusive and independent. So we can use a binomial distribution to model the behavior of passengers if their arrival for taking the flight is independent.**

### Problem 2: Asking Data Science Questions: Crime and Educational Attainment

For a given exam, there is a multiple-choice question with four (mutually exclusive) options. On average, 80% of the students know the answer. Among those who know the answer, 10% answer incorrectly due to exam stress.

*(a) If a student gets the answer correct, what is the probability that they actually know the material?*

**P(correct_ans) = 0.25**

**P(incorrect_ans) = 0.75**

**P(studenk_knows) = 0.8**

**P(student_doesnt_know) = 0.2**

**P(incorrect_ans | student_knows) = 0.1**

**P(correct_ans | student_knows) = 0.9**

**We have to calculate P(student_knows | correct_ans)**

**P(student_knows | correct_ans) = P(student_knows AND correct_ans)/P(correct_ans)**

**P(student_knows AND correct_ans) = P(student_knows)*P(correct_ans | student_knows) = 0.8 x 0.9 = 0.72 # P(correct_ans) = P(student_knows)*P(correct_ans | student_knows) + P(student_doesnt_know)*P(correct_ans | student_doesnt_know) = (0.8 x 0.9) + (0.2 x 0.25) = 0.77 # P(student_knows | correct_ans) = 0.72 / 0.77 = 72/77 = 0.93506493506**

Be sure to describe and outline each step in your calculations.

### Problem 3: Histograms and distributions

In this problem, you will be examining human height and citation counts for research papers (separately).

*(a) What kind of measure is human height (nominal, ordinal, interval, ratio)? How should it it be measured (continuous, discrete; positive, negative, either)?*

## Height is ratio scale measure

## Height is continuous like most numerical data that we measure

*(b) Read in the "fatherson.csv" data. The data consists of two columns: father's height and son's height (both in cm). Let's focus on fathers' heights (). How many observations are there? Are there any missing values?*

```
dataset <- read.csv('C:/Users/Administrator.UWIT-
J5ODA8OL5A/Downloads/fatherson.csv', sep = "", header = TRUE)

unique(is.na(dataset$fheight))

## [1] FALSE

length(dataset$fheight)

## [1] 1078

unique(is.na(dataset$sheight))

## [1] FALSE

length(dataset$sheight)

## [1] 1078
```

## There are 1078 observations

## There are no missing values in either of the columns

*(c) Compute the mean, median, standard deviation, and range of the heights. Discuss the relationship between these numbers. Is the mean larger than the median? What does this suggest? Would calculating the mode give a useful descriptive statistic? Why or why not? How does standard deviation compare to mean?*

```
mean(dataset$fheight)

## [1] 171.9252

mean(dataset$sheight)

## [1] 174.4572

median(dataset$fheight)
```

```
## [1] 172.1
```

```
median(dataset$sheight)
```

```
## [1] 174.3
```

```
sd(dataset$fheight)
```

```
## [1] 6.972346
```

```
sd(dataset$sheight)
```

```
## [1] 7.150713
```

```
range(dataset$fheight)
```

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

```
range(dataset$sheight)
```

```
## [1] 148.6 199.0
```

```
Mode = function(x){
    ta = table(x)
    tam = max(ta)
    if (all(ta == tam))
         mod = NA
    else
         if(is.numeric(x))
    mod = as.numeric(names(ta)[ta == tam])
    else
         mod = names(ta)[ta == tam]
    return(mod)
}
```

```
Mode(dataset$fheight)
```

```
## [1] 175.4
```

```
Mode(dataset$sheight)
```

```
## [1] 170.0 174.2
```

## The mean(171.9252) for fheight is less than the median (172.1) but not by much and this generally means that data is negatively skewed.

#The mean (174.4572) for sheight is slightly greater than the median (174.3). A mean greater than the median generally means data is postively skewed.

#Calculating the mode would give us a list of values that appear with the highest frequency in the dataset. Although since the dataset consists of numerical values that are continuous,

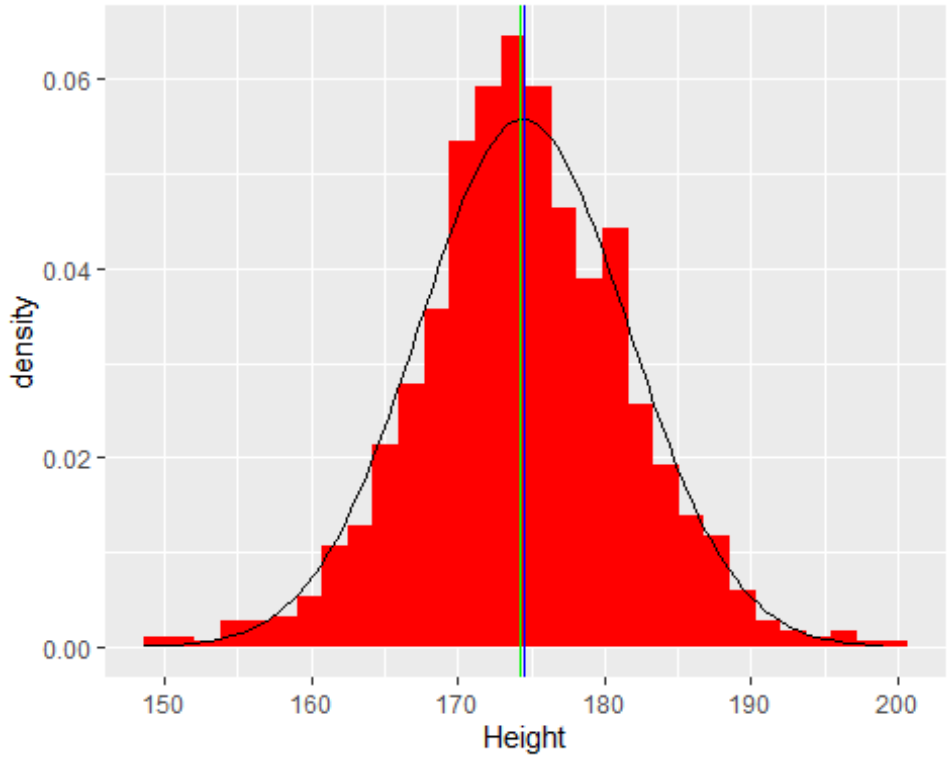
mode won't be of much use. Instead we should use modal interval to find the interval in which most of the values lie.

#Generally a data that fits a normal curve will have 68% density within one standard deviation of the mean. Since the above data fits that model, that is the relation between mean and standard deviation.

*(d) 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 data. Additionally, indicate the mean and median of the data using vertical lines of different colors. What do you find? Are the histogram and the density plot similar?*

```
ggplot(dataset, aes(x = sheight)) +
    geom_histogram( aes(y = ..density..),fill = "red") +
    stat_function(
        fun = dnorm,
        args = with(dataset, c(mean = mean(dataset$sheight), sd =
sd(dataset$sheight)))
    ) +
    scale_x_continuous("Height") + geom_vline(xintercept =
mean(dataset$sheight), color = 'blue') + geom_vline(xintercept =
median(dataset$sheight), color = 'green')
```

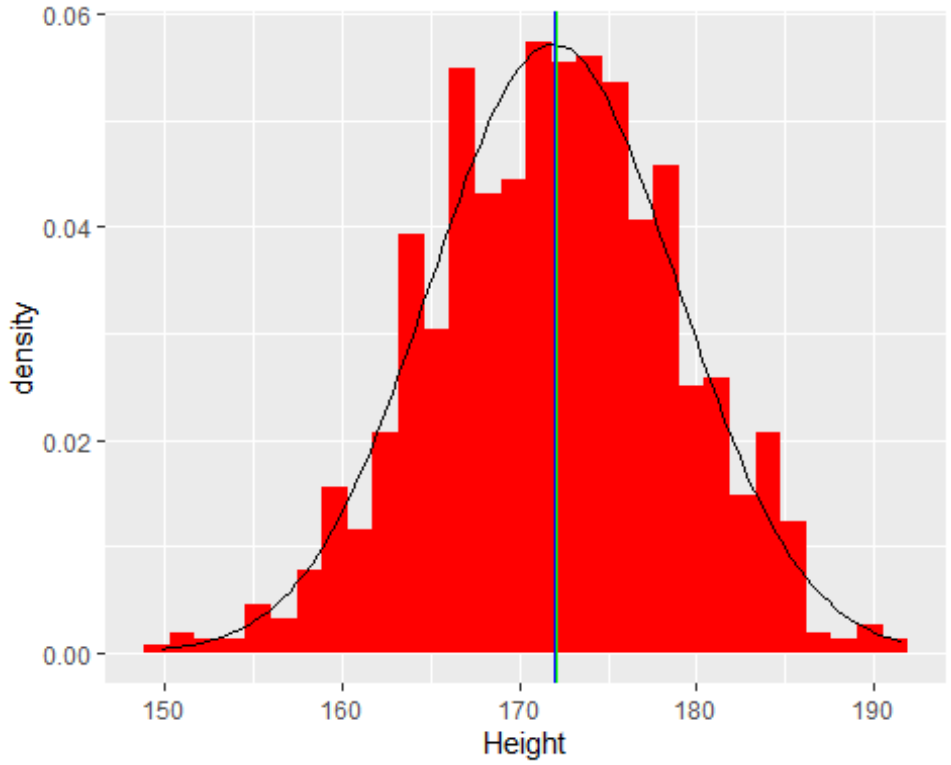
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(dataset, aes(x = fheight)) +
    geom_histogram( aes(y = ..density..),fill = "red") +
```

```
    stat_function(
        fun = dnorm,
        args = with(dataset, c(mean = mean(dataset$fheight), sd =
sd(dataset$fheight)))
    ) +
    scale_x_continuous("Height") + geom_vline(xintercept =
mean(dataset$fheight), color = 'blue') + geom_vline(xintercept =
median(dataset$fheight), color = 'green')

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




# Yes the histogram does fit the density plot rather snugly for the son's as well as the father's height.

*(e) What kind of measure is the citation counts for research papers (i.e. the number of times that a paper is referenced by other papers)? How should it be measured?*

## It is a ratio scale measure that can be measured as a continuous variable.

*(f) Read in the "mag-in-citations.csv" data. This is Microsoft Academic's Graph for citations of research papers and it contains two columns: paper id and the number of citations. We will only look at the number of citations. How many observations are there? Are there any missing values?*

```
dataset2 <- read.csv("mag-in-citations.csv",sep=",", header = TRUE)
unique(is.na(dataset2$citations))
```

```
## [1] FALSE
```

```
length(dataset2$citations)
```

```
## [1] 388258
```

## There are no missing values

## These are 388258 observation

*(g) Compute the mean, median, standard deviation, and range of the citations. Discuss the relationship between these numbers. Is the mean larger than the median? What does this suggest? Would calculating the mode give a useful descriptive statistic? Why or why not? How does standard deviation compare to mean?*

```
mean(dataset2$citations)
```

```
## [1] 15.61223
```

```
median(dataset2$citations)
```

```
## [1] 3
```

```
sd(dataset2$citations)
```

```
## [1] 78.39079
```

```
range(dataset2$citations)
```

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

```
Mode(dataset2$citations)
```

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

## The mean(15.61223) for citation is significantly greater than the median (3) and this generally means that data is positively skewed

#Calculating the mode would give us a list of values that appear with the highest frequency in the dataset. Although since the dataset consists of numerical values that are continuous, mode won't be of much use. Instead we should use modal interval to find the interval in which most of the values lie.

#Generally a data that fits a normal curve will have 68% density within one standard deviation of the mean. Since the above data fits that model, that is the relation between mean and standard deviation.

*(h) Calculate the 90th percentile for the citation data. How does this compare to the maximum value of the citation data? Calcualte the 10th percentile for the citation data. How does this compare to the minimum value of the citation data? What does this all suggest with respect to the shape of the distribution of citation counts?*

```
quantile(dataset2$citations, c(  .10,  .90))
```

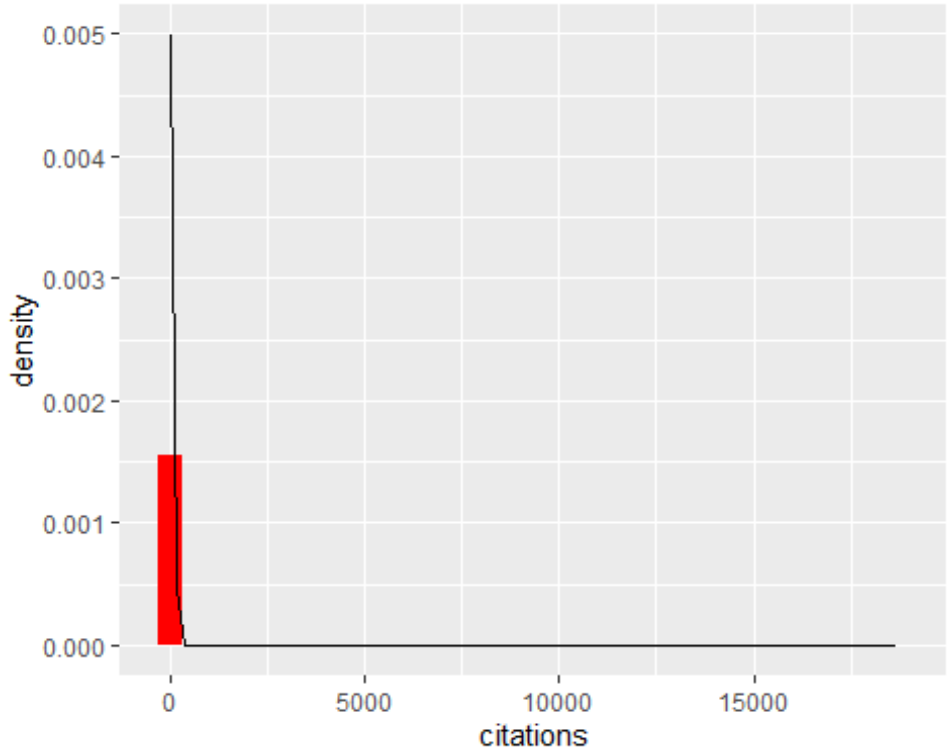
```
## 10% 90%
##   0  34
```

## The minimum value of the data(0) corresponds with the 10 percentile value (0) while the 90 percentile value (34) is quite low compared to the highest value (18682) within the dataset. This tells us that the dataset is heavily skewed towards the left.

*(i) 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 data. Additionally, indicate the mean and median of the data using vertical lines of different colors. What do you find? Are the histogram and the density plot similar? Now try this with what is called a "log-log" transformation (i.e. plotting the x and y axes on a logarithmic scale)*

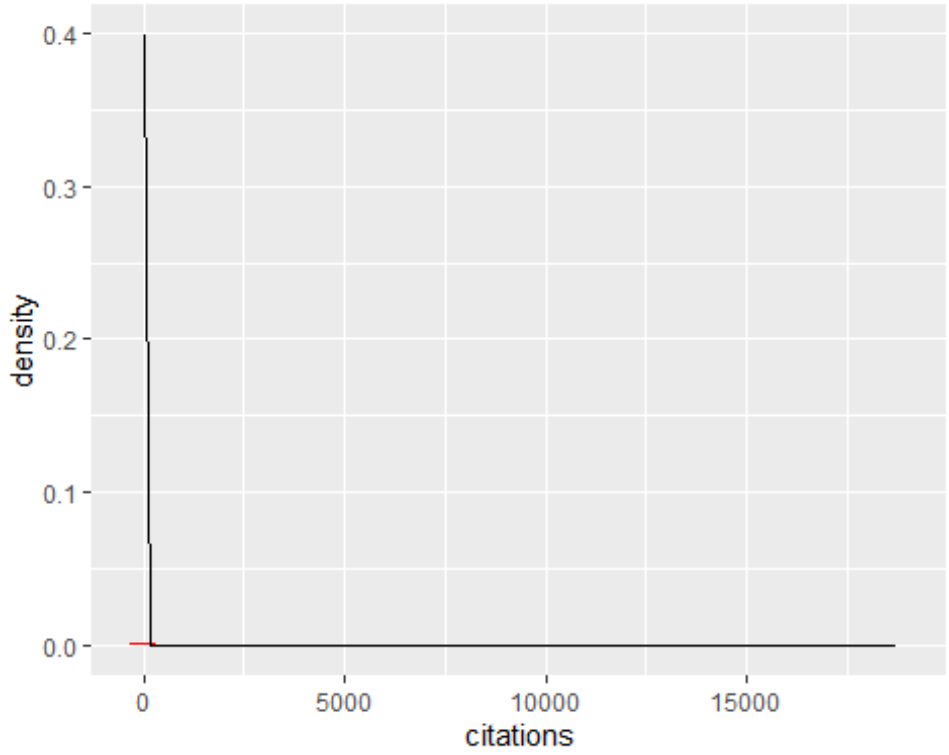
```
ggplot(dataset2, aes(x = citations)) +
    geom_histogram( aes(y = ..density..),fill = "red") +
    stat_function(
        fun = dnorm,
        args = with(dataset2, c(mean = mean(dataset2$citations), sd =
sd(dataset2$citations)))
    )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(dataset2, aes(x = citations)) +
    geom_histogram( aes(y = ..density..),fill = "red") +
    stat_function(
        fun = dnorm,
        args = with(dataset2, c(mean = mean(dataset2$citations), sd =
sd(dataset2$citations))) + scale_x_continuous(trans = "log2")
    )

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




*(j) Seeing how well (or not well) that the heights and the citations datasets 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)?*

**There are a lot of published papers with 0 citations and this is reflected in the normal plot of the data. The data is highly skewed to the left which matches the normal distribution for the histogram plot.**