IMT 573: Problem Set 5 - Statistical Theory

Divya Gaurav Tripathi

Due: Wednesday, November 06, 2019

Collaborators:

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 `problemset5.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. 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. 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 not 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 *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. (Note: I am also using the `include=FALSE` option here to not include this code in the PDF, but you need to remove this or change it to `TRUE` if you want to include the code chunk.)
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 `ps5_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup:

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

```r
# Load standard libraries
library(tidyverse)
library(tidyr)
library(modeest)
library(ggplot2)
library('dplyr')
```

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, so a fully booked plane generates $10,000 revenue. The sales team has found that the probability, that the passengers who have paid their fare actually show up is 98%, and individual show-ups can be considered independent. The additional costs, associated with finding an alternative solutions for passengers who are refused boarding are $500 per person.

Question: Which distribution would you use to describe the actual number of show-ups for the flight? Hint: read OIS ch 3 about distributions.

In my opinion, it follows Binomila distribution. A person showing up or not showing up is a Bernoulli random variable with exactly two possible outcomes- showing up(which we consider success, Psuccess = 0.98) and not showing up(which we consider failure, Pfailure = 0.02). If k people show up for n seats, it can be said as k successes in n trials. So it satisfies the condition to follow a Binomial distribution, as mentioned in the OIS book: (1) The trials are independent. (2) The number of trials, n, is fixed. (3) Each trial outcome can be classified as a success or failure. (4) The probability of a success, p, is the same for each trial.

Question: Assume the airline never overbooks. What is it's expected revenue? Expected revenue means expected income from the ticket sales, minus the expected costs related to alternative solutions.

There is no overbooking, so we would assume that 100 seats are sold, the revenue from 100 seats = 100 * $100 = $10000. The airpline has to take some loss only if more people show up than the number of seats, which is a case of overbooking, but we are not doing it.

So the net revenue = $10000

Question: Now assume the airline sells 101 tickets for 100 seats. What is the probability that all 101 passengers will show up?

We have 101 passengers, which means that we have 101 trials. Each trial can be a success(Psuccess = 0.98) or a failure(Pfailure = 0.02). The probability of all 101 passenger show up = Probability of 101 successes and 0 failure = (Psuccess)1̂01 = (0.98)1̂01 = 0.129

Question: What are the expected profits (= revenue − expected additional costs) in this case? Would you recommend overbooking over selling just the right number of tickets?

Let us talk about worst case. \textcolor{blue}{The total money airline receives in this case = 101 * $100 = 10100.} However it only has 100 seats, so it has to spend $500 for an extra passenger. So net profit = $10100-$500 = $9600. This is for worst case scenario which when all 101 passengers show up.

Let us find it for average case.The probability of all 101 passengers showing up is only 0.129. So the expected revenue should be 101*$100- 0.129*1*$500 = $10035.5

\textcolor{blue}{I would recommend overbooking because, the probability of 101 passengers showing up is only 12.9%.} \textcolor{blue}{So its only in only 12.9% instances the airline would have to spend $500 to make arrangements for extra one passenger.} For the other cases the airline would earn more revenure than what they would have earned if they had sold only 100 tickets.We had found the expected revenue = $10035.5

Question: Now assume the airline sells 102 tickets. What is the probability that all 102 passengers show up?

We have 102 passengers, which means that we have 102 trials. Each trial can be a success(Psuccess = 0.98) or a failure(Pfailure = 0.02). We have 102 trials(consisting of 102 success and 0 failure). The probability of all 102 passenger show up = Probability of 102 successes and 0 failure = (Psuccess)1̂02 = (0.98)1̂02 = 0.127

Question: What is the probability that 101 passengers – still one too many – will show up?

We have sold 102 tickets, and have to find the probabilty of 101 people showing up, which means we have 101 successes and 1 failure in 102 trials. So the probability of 101 people showing up =(102 Choose 101)*((Psuccess)1̂01)*(Pfailure1̂) = (102)*(0.98)1̂01 * 0.02 = 0.2651

Question: Would it be advisable to sell 102 tickets, i.e. is the expected revenue from selling 102 tickets larger than from selling 100 and 101 tickets?

Let us find the best case scenario where 102 tickets sold and 100 people show up. The total revenue of selling 102 tickets and only 100 people showing up = 102*$100= $10200.

Let us find the worst case scenario when 102 tickets sold and 102 people show up.The total revenue if a flight sells 102 tickets and 102 passengers show up =102*$100 - 2*$ 500 = $9200

Let us find the average case. The probability of 102 sold and all 102 passengers showing up = 0.127(as we had calculated above). The probablity of 102 tickets sold and 101 people show up is 0.2651 (as we had calculated above). Let us calculate the probability in average case when 102 tickets and considering average cases, which means considering probablity of 102 people show up or 101 people show up. That is the average case. So revenue in average case = 102*$100 - (0.127*2*$500 +0.2651*1*$500)= $9940.45

If they sell 102 tickets, and only 100 people show up, its the best case and revenue is $10200. If the sell 102 tickets and 102 people show up, its worst case and revenue is $9200. Lets talk about average case. If they sell 102 tickets, considering probabilities of 102 people showing up and 101 people showing up, in average case expected revenue = $9940.45. In average case the revenue is less when they sell 102 tickets compared to selling 100 or 101 tickets. So I would not recommend selling 102 tickets.

Question: What is the optimal number of seats to sell for the airline? How big are the expected profits?

The expected average revenue when they sell 100 tickets = $10000. The expected average revenue when they sell 101 tickets = $10035.5. The expected average revenue when they sell 102 tickets = $9940.45

So we see that the revenue starts decreasing if they sell more than 101 tickets. The optimum number of tickets to sell = 101.

Question: What does it mean that the show-ups are independent? Why is it important? Hint: read about independence in OIS 2.1.6 (2017 version).

Show-ups are independent. It means that one person showing up or not showing up for the flight has no impact on another person showing or not showing up for the flight. It is importnat because, then it becomes a Bernouli trial and we can apply Binomial distribution to estimate our revenue from selling tickets.

Note: some of the expressions may be hard to write analytically. Feel free to use computer for the calculations, just 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 of research paper citations.

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

```
heightsdataset <- read.csv("fatherson.csv",
                         stringsAsFactors=FALSE)

str(heightsdataset)
```

```
## 'data.frame':    1078 obs. of  1 variable:
##  $ fheight.sheight: chr  "165.2\t151.8" "160.7\t160.6" "165\t160.9" "167\t159.5" ...
```

There is only one column in heightsdataset, which has father's and son's height sepertaed by a white space. It has been read as a charecter variable. We would try to split it into two columns.

```
seperateheights <- separate(heightsdataset, fheight.sheight,
                c("fatherheight", "sonheight"), sep = "\\b\\s\\b",
                remove = TRUE)
#We seperated the height column in two columns on white space.

str(seperateheights)
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ fatherheight: chr  "165.2" "160.7" "165" "167" ...
##  $ sonheight   : chr  "151.8" "160.6" "160.9" "159.5" ...
```

We have seperated the origibal heights column into two sepertae columns for father’s heights and son’s heights. The height columns are char variable . we want to convert them to numeric.

```
heights_tidydata <- seperateheights %>% mutate_if(is.character, as.numeric)
str(heights_tidydata)
```

```
## 'data.frame':    1078 obs. of  2 variables:
##  $ fatherheight: num  165 161 165 167 155 ...
##  $ sonheight   : num  152 161 161 160 163 ...
```

Lets find the summary of heights data

```
summary(heights_tidydata)
```

```
##   fatherheight     sonheight
##  Min.   :149.9   Min.   :148.6
##  1st Qu.:167.1   1st Qu.:170.0
##  Median :172.1   Median :174.3
##  Mean   :171.9   Mean   :174.5
##  3rd Qu.:176.8   3rd Qu.:179.0
##  Max.   :191.6   Max.   :199.0
```

```
mode_fatherheights <- mlv(heights_tidydata$fatherheight, method = "mfv",na.rm = FALSE)
mode_fatherheights
```

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

```
#Its the mode of father's heights

mode_sonheights <- mlv(heights_tidydata$sonheight, method = "mfv",na.rm = FALSE)
mode_sonheights
```

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

```
#Its the mode of son's heights

fatherheights_sd = summarise(heights_tidydata, fatherheights_sd = sd(heights_tidydata$fatherheight))
#This give the standard deviation of father's heights
fatherheights_sd
```

```
##   fatherheights_sd
## 1         6.972346
```

```
fatherheights_variance = fatherheights_sd*fatherheights_sd
#This gives the variance of father's heights
fatherheights_variance
```

```
##   fatherheights_sd
## 1         48.61361
```

What level of measurement (nominal, ordered, difference, ratio) does this data on human height use? How should it be measured (e.g. continuous, discrete, positive. . .)?

The level of measurement is ratio because if its 0, it means absence of quantity being measured. Height should be a continuous variable because it can take continuous values of intigers and decimals.

Read the `fatherson.csv` dataset into R. It contains two columns, father's height and son's height, (in cm). Let's focus on father's height for a moment, (variable `fheight`). Provide a basic description of this variable, for example how many observations do we have? Do we have any missing data?

We have 1078 rows of a column which contains father's heights and son's heights. There is no missing data. The heights of father and son are seperated by a white space. we seperated the heights into two seperate columns. Human heights follow a continuous distribution, as they take continuous values.

Compute mean, median, mode, standard deviation and range of the heights. Discuss the relationship between these numbers. Is mean larger than median? Than mode? By how much (in relative terms)? How does standard deviation compare to mean?

For father's heights:mean = 171.9, median :172.1, min= 149.9, max = 191.6, mode = 175.4, standard deviation = 6.97,variance = 48.61. For son's heights: mean=174.5,median = 174.3,min = 148.6, max = 199.0,mode= 170.0 and 174.2

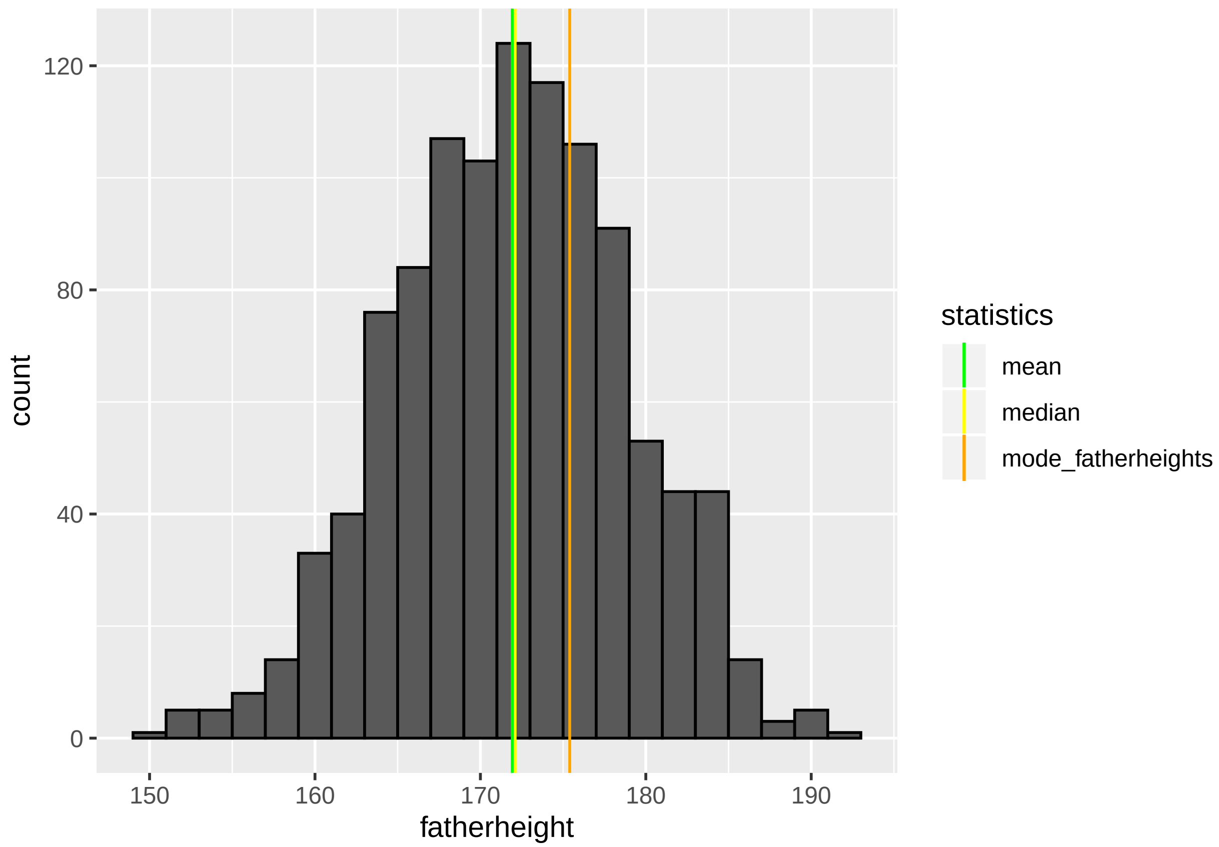
We are talking about father's heights. The mean(171.9) is slightly smaller than the median(172.1). It means the data is slightly left skewed, as most data points are towars the left of the mean. The standard deviation is quite small than the mean, which means the deviation of data is not very large.

Hint: there is no built-in method to computing sample modes in R. Several packages provide a way to do it, for example try `modeest::mlv` (installed on the server). However, as height is a continuous variable, there are many ways to compute it. Take a look at the corresponding documentation. You may experiment with a few options and pick one, for instance the *naive* method or write your own!

Plot a histogram of the data. Add to this histogram: (1) a plot of normal distribution with the same mean and standard deviation as the data, (2) the sample mean, median, and mode. You can use vertical lines of different colors to do this. What do you find? Are the histogram and the density plot similar?

Ler us try to plot the histogram.

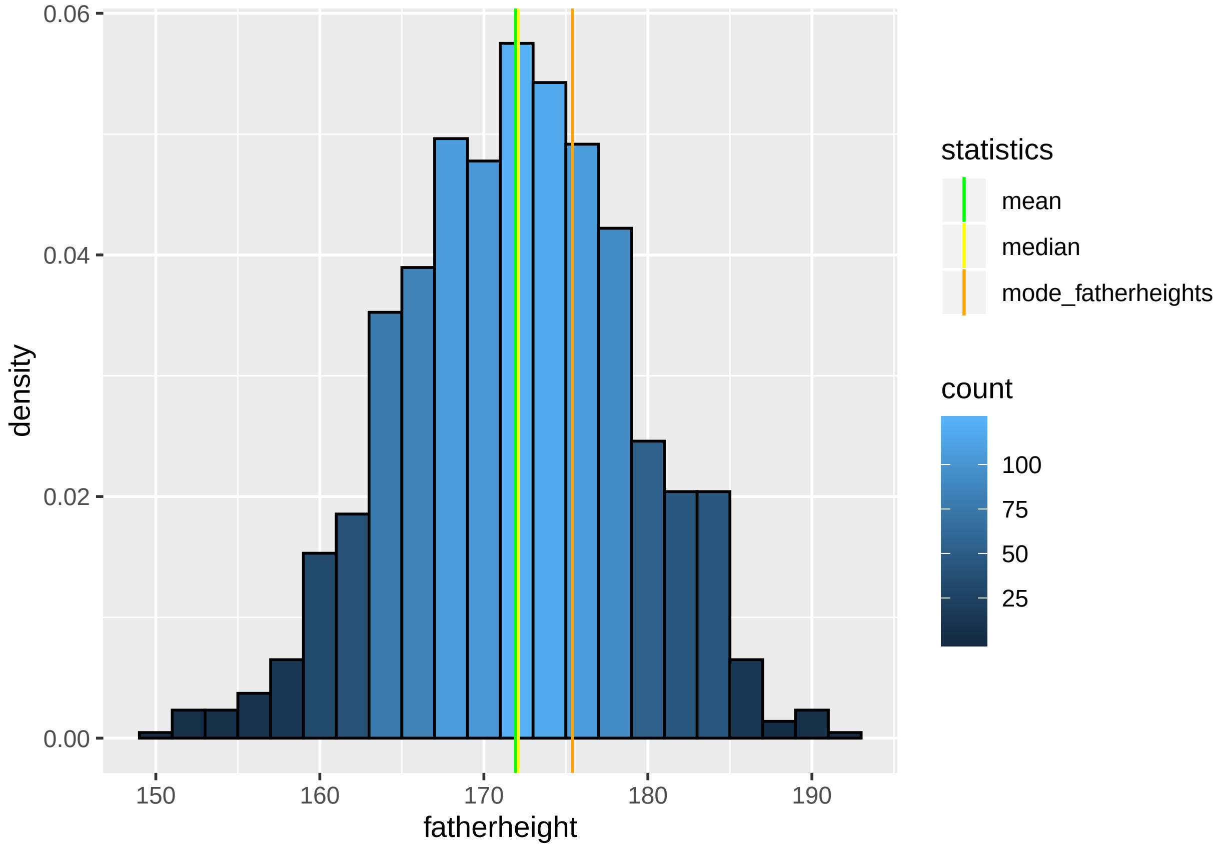
```r
ggplot(heights_tidydata, aes(x=fatherheight)) +
  geom_histogram(binwidth=2, colour="black") +

  geom_vline(data=heights_tidydata, aes(xintercept =
                mean(heights_tidydata$fatherheight),color = "mean")) +

geom_vline(data=heights_tidydata, aes(xintercept =
                      median(heights_tidydata$fatherheight), color ="median")) +

  geom_vline(data=heights_tidydata, aes(xintercept =
              mode_fatherheights,color = "mode_fatherheights")) +

  scale_color_manual(name = "statistics", values =
  c(median = "yellow", mean = "green", mode_fatherheights = "orange"))
```



```
#We plotted a histogram and drew lines for mean, median and mode.
```

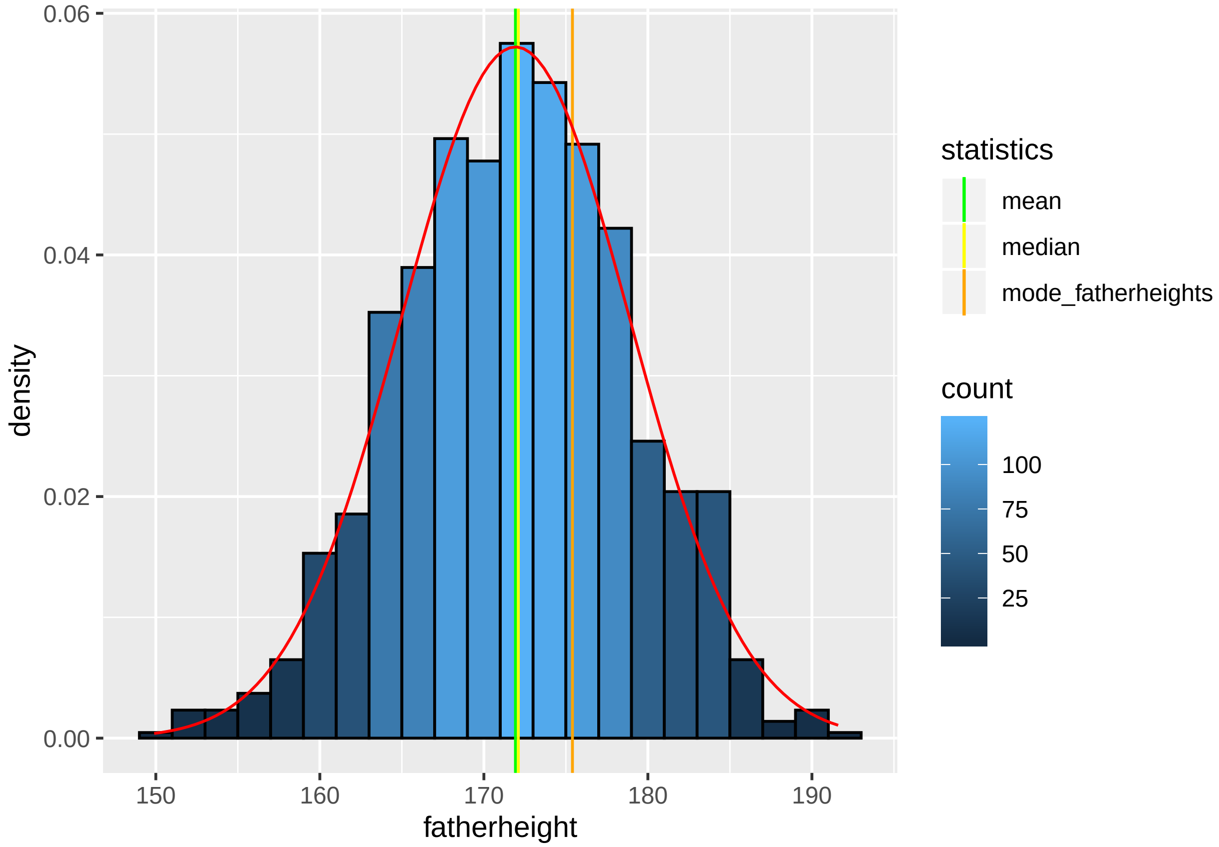
Let us try to plot the density plot of father's heights.

```
plot_fatherheights <-ggplot(heights_tidydata, aes(x=fatherheight)) + geom_histogram(binwidth=2,
colour="black",
aes(y=..density.., fill=..count..)) +

  geom_vline(data=heights_tidydata, aes(xintercept =
            mean(heights_tidydata$fatherheight),color = "mean")) +

geom_vline(data=heights_tidydata, aes(xintercept =
            median(heights_tidydata$fatherheight), color ="median")) +

  geom_vline(data=heights_tidydata, aes(xintercept =
                  mode_fatherheights,color = "mode_fatherheights")) +

  scale_color_manual(name = "statistics", values =
  c(median = "yellow", mean = "green", mode_fatherheights = "orange"))

#We made a density plot.
plot_fatherheights
```



Let us try to add a normal curve(on the density plot) of data having mean = 171.9, and standard deviation= 6.97.

```
normalplot <- plot_fatherheights+ stat_function(fun=dnorm,
                         color="red",
                         args=list(mean=mean(heights_tidydata$fatherheight),
                                  sd=sd(heights_tidydata$fatherheight)))

#We have added a normal plot with the mean and standard deviation
#as found from the father's heights column
normalplot
```



The histogram and density plot are similar in shape. The histogram represents count on Y axis, where as the density plot represents distribution of numeric variable on Y axis. We can put the normal curve(with required mean and standard deviation) on the density plot.

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

```
citations <- read.csv("mag-in-citations.csv")

citations_df <- as.data.frame(citations)

str(citations_df)
```

```
## 'data.frame':    388258 obs. of  2 variables:
##  $ paperId  : num  4090687 6537979 7484482 9444380 14056478 ...
##  $ citations: int  2 2 4 3 5 2 1 39 9 1 ...
```

```
#table(citations_df$citations)
#We commented it for readibility.

summary(citations_df)
```

```
##     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
```

```
citations_df1<- citations_df
#We saved the data citations_df into another data frame "citations_df1", just in case
#we might need a copy of it.

mode_citations_df <- mlv(citations_df1$citations, method = "mfv",na.rm = FALSE)
mode_citations_df
```

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

```
#This tells the mode of citations

citations_df_sd = summarise(citations_df1,citations_sd = sd(citations_df1$citations))
citations_df_sd
```

```
##   citations_sd
## 1     78.39079
```

```
#This tells the standard deviation of citations data.
```

The oroginal data has maximum count for 0 number of citations and it is also the mode.

What kind of measure is this? What kind of valid values would you expect to see (continuous, discrete, positive, . . .)

The data frame tells which research paper id(num type) has how many citations(int type). The level of measurement is ratio because if its 0, it means absence of quantity being measured. These can only take discrete positive intiger values. So they are discrete data type.

Read the `mag-in-citations.csv` data. This is 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 descriptives of this variable: how many observations do we have? Do we have any missing observations?

We have total 388,258 rows for different paper ids. We do not have any NA, white space or missing data.

Compute mean, median, mode, standard deviation and range of the heights. Discuss the relationship between these numbers. Is mean larger than median? Than mode? By how much (in relative terms)? How does standard deviation compare to mean?

Hint: here you do not want to use any smoothing as we are measuring discrete counts. Use the plain "most frequent value", `method="mfv"` if using the `modeest` package.

I am assuming the question in this part is not about height, but about citations. The following is the summary statistic for citations:Min.=0.00,1st Qu.= 1.00, Median=3.00, Mean=15.61,3rd Qu.=12.00,Max.=18682.00,standard deviation = 78.39,Mode = 0.

The mean is 5 times greater than the median, it means the data is strongly right skewed.The mean(15) quite larger than mode(0). The standard deviation is around 5 times that of the mean, which means the data has a lot of variation.

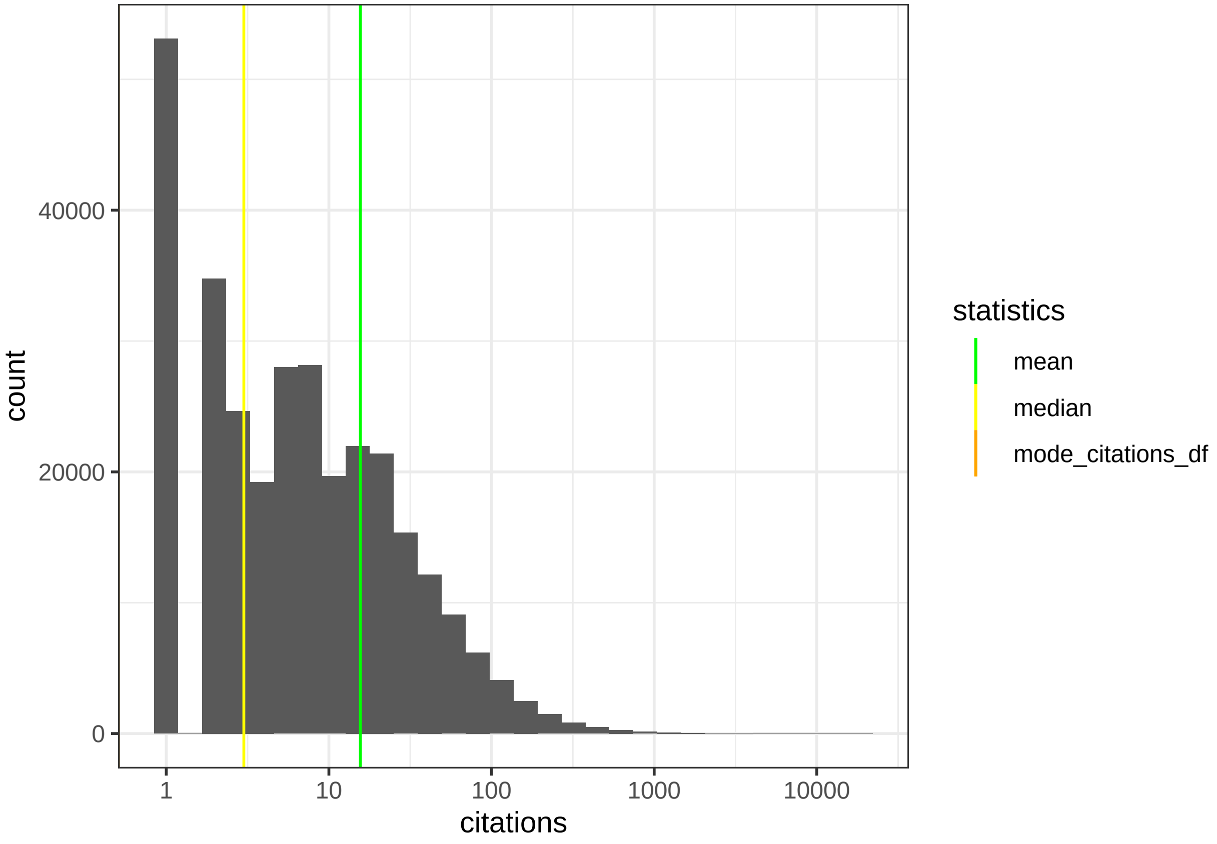
Plot a histogram of the data. Add to this histogram: (1) a plot of normal distribution with the same mean and standard deviation as the data, (2) the sample mean, median, and mode. You can use vertical lines of different colors to do this. What do you find? Are the histogram and the density plot similar?

Hint: You might try experimenting with log-log scale for the histogram. \end{enumerate}

Let us try to plot the histogram.

```
citationhistogram <- ggplot(citations_df1, aes(x = citations)) +
  geom_histogram() + scale_x_log10()  +
    theme_bw()+
  geom_vline(data=citations_df, aes(xintercept =
                      mean(citations_df1$citations),color = "mean")) +

geom_vline(data=citations_df, aes(xintercept =
          median(citations_df1$citations), color ="median")) +

  geom_vline(data=citations_df, aes(xintercept =
        mode_citations_df ,color = "mode_citations_df")) +

  scale_color_manual(name = "statistics", values =
      c(median = "yellow", mean = "green", mode_citations_df = "orange"))

#We plotted a histogram and drew lines for mean, median, mode.

citationhistogram
```

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

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

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

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

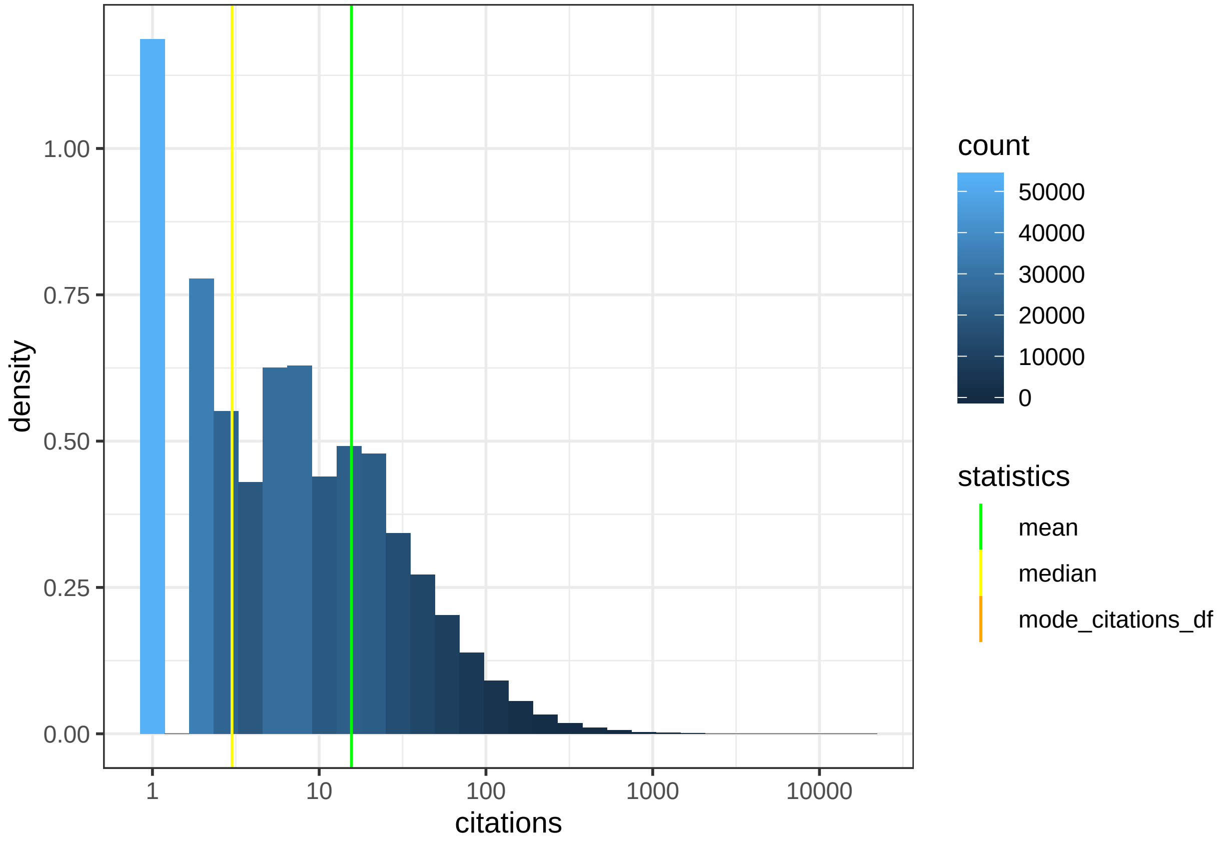


Please note that mode = 0, so the orange line depocting the mode coincides the y axis.

Let us try to plot the density plot.

```
citation_density <- ggplot(citations_df1, aes(x = citations)) + geom_histogram(aes(y=..density.., fill=
  scale_x_log10()  +
     theme_bw()+

  geom_vline(data=citations_df, aes(xintercept =
                      mean(citations_df1$citations),color = "mean")) +

geom_vline(data=citations_df, aes(xintercept =
                median(citations_df1$citations), color ="median")) +

  geom_vline(data=citations_df, aes(xintercept =mode_citations_df ,color = "mode_citations_df")) +

  scale_color_manual(name = "statistics",
  values = c(median = "yellow", mean = "green", mode_citations_df = "orange"))

#We plotted the density plot.

citation_density
```

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

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

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

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

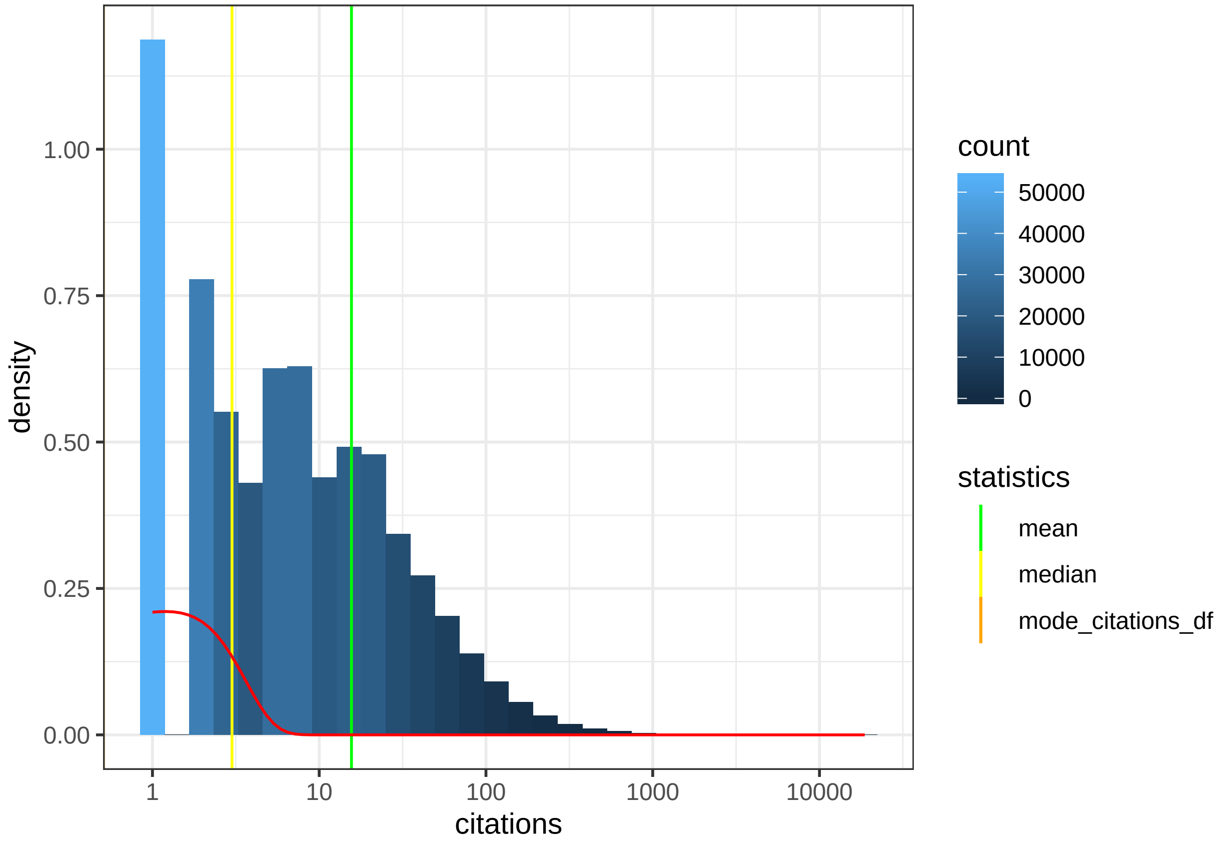


Please note that mode = 0, so the orange line depocting the mode coincides the y axis.

Lets try to add normal plot on our density plot.

```
normalplot_citation <- citation_density+ stat_function(fun=dnorm,
                         color="red",
            args=list(mean=log10(mean(citations_df1$citations)),
            sd=log10(sd(citations_df1$citations))) )

#We have added a normal plot with the log10 of mean and
#standard deviation as found from the citations column
normalplot_citation
```

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

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

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

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

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



Please note that mode = 0, so the orange line depocting the mode coincides the y axis.

The histogram and density plot are similar in shape. The histogram represents count on Y axis, where as the density plot represents distribution of numeric variable on Y axis. We can put the normal curve(with required mean and standard deviation) on the density plot.

(c) Comment on your finding from part (a) and part (b). Be sure to compare the two cases.

In part a, the data set had a slight left skew, and the variation in data was very less. Also the data was continuous, so we had no missing parts or gaps in histgram, and it was very smooth. In part b, the data set had a very strong right skew and the variation in data was extreme, with many outliers. So we had to plot the histogram on a logarathmic scale. The data was discrete, and not continuous, so there were breaks in histogram and it as not smooth. ### Extra Credit: Guessing on Multiple Choice Tests

In the exam, there is a multiple-choice question with four (mutually exclusive) options. In average, 80% of the students know the answer, but event those who know, still answer it wrong in 10% of time because of the exam stress.

If a student get's the answer right, what is the probability that she actually knows the material? Hint: read OIS 2.2, in particular 2.2.7 (2017 version).

We can solve it using the Baye's theorem.

Probability of knowing correct answer = P(knows) = 0.8 Probability of not knowing answer = P(notknow) = 0.2 Probability of marking incorrect when knowing answer = P(incorrect|know)= 0.1 Probability of marking correct when knowing answer = P(correct|know) = 0.9 Every question has 4 options. If a student does not know the answer, he would guess while marking it. So probability of marking correct when not knowing answer = P(correct|notknow) = 0.25 Probability of marking incorrect when not knowing answer = P(incorrect|notknow) = 0.75

The student has marked the correct answer. We have to find that given he has marked corect answer, what is the probability he knew the answer.

We have to find P(know|correct).

According to the Baye's theorem:

P(know|correct) = (P(correct|know)*P(knows)) / (P(know)*P(correct|know) + P(notknow)*P(correct|notknow))

= (0.9*0.8)/((0.8*0.9) + (0.2*0.25)) = 0.935 So the probability that the student knows the material given that he marked it correct = 0.935