IMT 573: Problem Set 5 - Statistics

Xinyi Esther Yang

Due: Tuesday, November 10, 2020

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)
library(statip)
library(stats)
library(installr)
```

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.

1. Which distribution would you use to describe the actual number of show-ups for the flight? Hint: read OpenIntro Statistic (OIS) ch on distributions.
2. Assume the airline never overbooks. What is it's expected profit? Expected profit means expected income/revenue from the ticket sales, minus the expected costs related to alternative solutions.
3. Now assume the airline sells 101 tickets for 100 seats. What is the probability that all 101 passengers will show up?
4. What are the expected profits (= revenue − expected additional costs) in this case? Would you recommend overbooking over selling just the right number of tickets?
5. Now assume the airline sells 102 tickets. What is the probability that all 102 passengers show up?
6. What is the probability that 101 passengers – still one too many – will show up?
7. 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?
8. What is the optimal number of seats to sell for the airline? How big are the expected profits?
9. What does it mean that the show-ups are independent? Why is it important in this case?

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.

Q1.RESPONSE:

I would use the binomial distribution to describe the actual number of show-ups for the flight.

```
q4 <- 100*101-500*1
cat(paste("Q2. When there’s no overbook, the expected profit is",100*100,"\n",
          "Q3. The probablity that all 101 passengers will show up is", 0.98^101,"\n",
          "Q4. The expected profits in this case(all 101 passengers will show up) is", q4,
          ". I would recommend overbooking because the probability that we need to pay for additional c
```

```
## Q2. When there’s no overbook, the expected profit is 10000
##  Q3. The probablity that all 101 passengers will show up is 0.129967164776858
##  Q4. The expected profits in this case(all 101 passengers will show up) is 9600 . I would recommend o
```

reference

```
paste("Q5. The probablity that all 102 passengers will show up is", 0.98^102)
```

```
## [1] "Q5. The probablity that all 102 passengers will show up is 0.127367821481321"
```

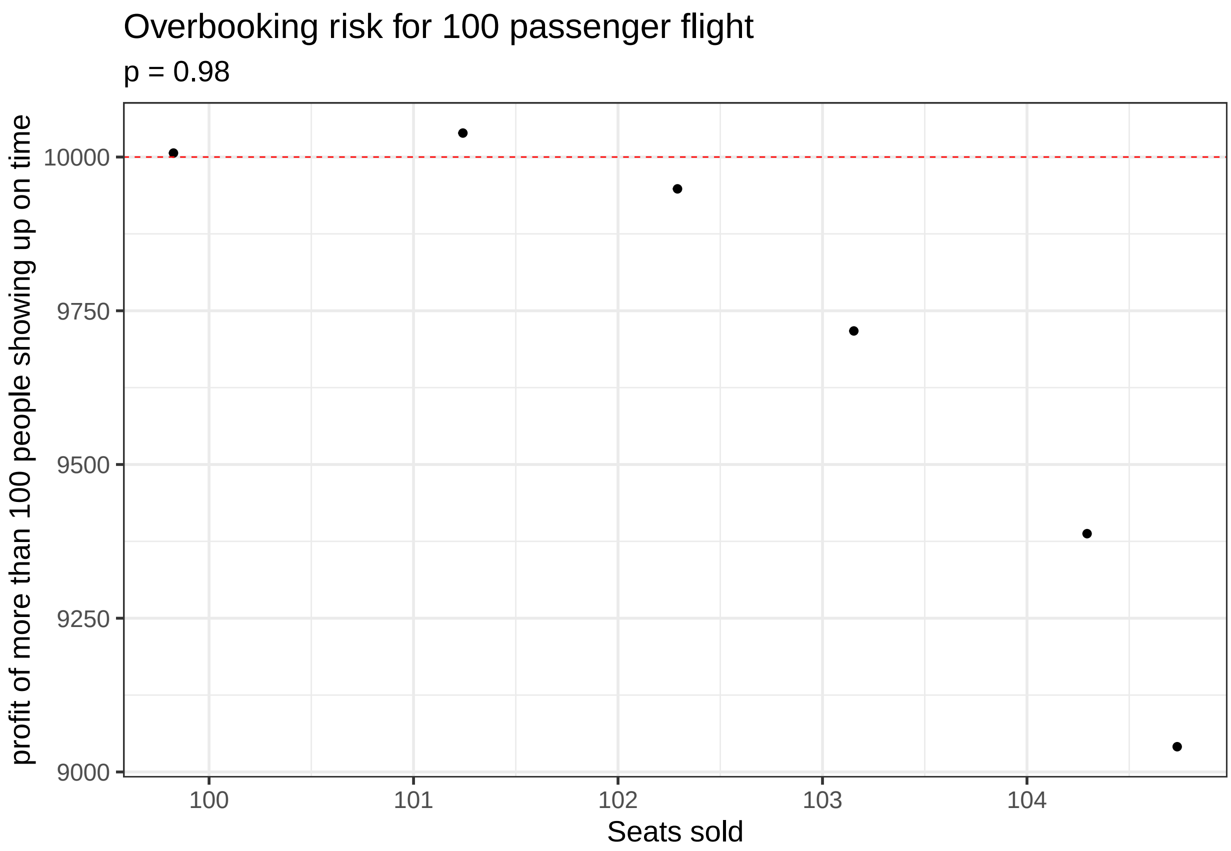
```
paste("Q6. The probablity that 101 out of 102 passengers will show up is", dbinom(101, 102, 0.98))
```

```
## [1] "Q6. The probablity that 101 out of 102 passengers will show up is 0.265133016144791"
```

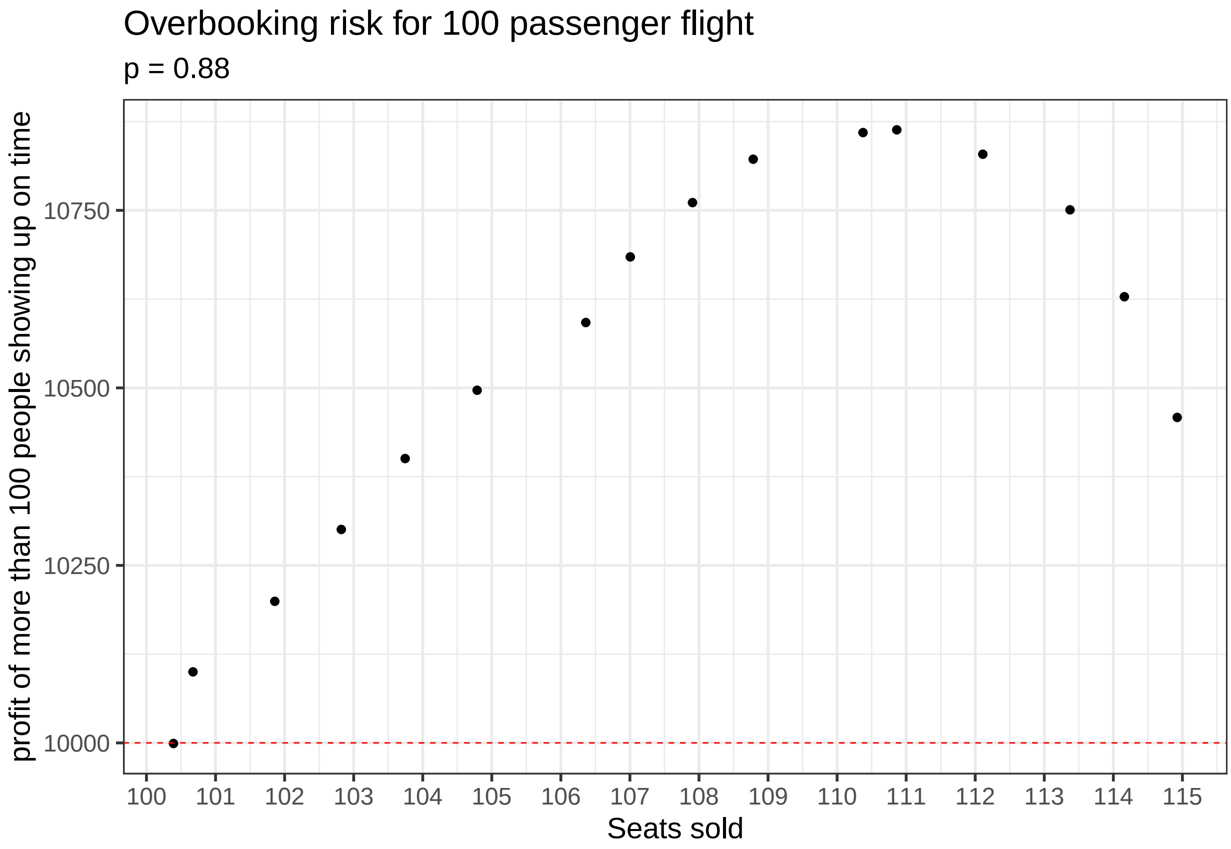
```
profits <- dbinom(101, 102, 0.98)*(102*100-500) +
        dbinom(102, 102, 0.98)*(102*100-500*2) +
        pbinom(100, 102, 0.98)*102*100
paste("Q7. The expected revenue to sell 102 tickets is", profits)
```

```
## [1] "Q7. The expected revenue to sell 102 tickets is 9940.06567044629"
```

```
#Q8 to find the optimal number
# create a dataframe
overbook_risks <- data.frame(’seats_sold’ = 101:115,
                            ’Profits’ = 1:15)

# calculate the expected profits for each sale plan
n = 1
p = 0.98
for (i in seq(101,115)){
        profits <- pbinom(100, i, p)*i*100
        for (j in seq(i-100)){
                r <- dbinom(100+j, i, p)*(i*100-500*j)
                profits <- profits + r
        }
        overbook_risks$Profits[n] <- profits
        n <- n+1
}
# add the baseline for comparison
overbook_risks <-overbook_risks %>% add_row(’seats_sold’=100, ’Profits’=10000, .before = 1)

# make the plot of the first 5 rows because the rest of the dataframe makes it difficult
# to see the nuance difference between the profits
# reference: above this chunk
head(overbook_risks) %>% ggplot() +
  geom_jitter(aes(seats_sold,
                Profits),
            size = 1) +
  geom_hline(yintercept = 10000,
             color = ’red’,
             linetype = ’dashed’,
             size = 0.25) +
  theme_bw() +
  labs(x = ’Seats sold’,
       y = ’profit of more than 100 people showing up on time’,
       title = ’Overbooking risk for 100 passenger flight’,
       subtitle = ’p = 0.98’) +
        scale_x_continuous(breaks = seq(99, 115))
```



```
# test another possibility to see the pattern
# and check if there would be an optimal number far away from the 100 baseline
overbook_risks2 <- data.frame('seats_sold' = 101:115,
                              'Profits' = 1:15)
n = 1
p = 0.88
for (i in seq(101,115)){

        profits <- pbinom(100, i, p)*i*100
        for (j in seq(i-100)){
                r <- dbinom(100+j, i, p)*(i*100-500*j)
                profits <- profits + r
        }
        overbook_risks2$Profits[n] <- profits
        n <- n+1
}
overbook_risks2 <- overbook_risks2 %>% add_row('seats_sold'=100, 'Profits'=10000, .before = 1)

overbook_risks2 %>% ggplot() +
  geom_jitter(aes(seats_sold,
              Profits),
            size = 1) +
  geom_hline(yintercept = 10000,
             color = 'red',
             linetype = 'dashed',
             size = 0.25) +
  theme_bw() +
  labs(x = 'Seats sold',
       y = 'profit of more than 100 people showing up on time',
       title = 'Overbooking risk for 100 passenger flight',
       subtitle = 'p = 0.88') +
        scale_x_continuous(breaks = seq(99, 115))
```



RESPONSE:
Q7. The expected revenue to sell 102 tickets is less than selling 101 tickets. So it is not advisable to sell 102 tickets.

Q8. The optimal number of seats is 101.

```
paste("Q8. The expected profits is", overbook_risks$Profits[2])
```

```
## [1] "Q8. The expected profits is 10035.0164176116"
```

Q9 RESPONSE: The independence of the show-ups means the occurrence of one show-up does not affect the probability of another show-up. Only in this case can we use analyze the data and implement the binomial distribution method, or there would be bias.

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.

1. What kind of measure is human height (nominal, ordinal, interval, ratio)? How should it be measured (e.g. continuous, discrete, positive, negative...)?
2. 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?
3. 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)? What does this suggest? How does standard deviation compare to mean?
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 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?

```r
fatherson <- read.csv('fatherson.csv.bz2', sep = "\t")
```

```r
# number of observations
length(fatherson$fheight)
```

```
## [1] 1078
```

```r
# check missing or empty data
sum(is.na(fatherson$fheight))
```

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

```r
sum(fatherson$fheight=="")
```

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

```r
# statistical parameters
summary(fatherson$fheight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   149.9   167.1   172.1   171.9   176.8   191.6
```

```r
# the difference of these values
median(fatherson$fheight)-mean(fatherson$fheight)
```

```
## [1] 0.1747681
```

```r
mfv(fatherson$fheight)-mean(fatherson$fheight)
```

```
## [1] 3.474768
```

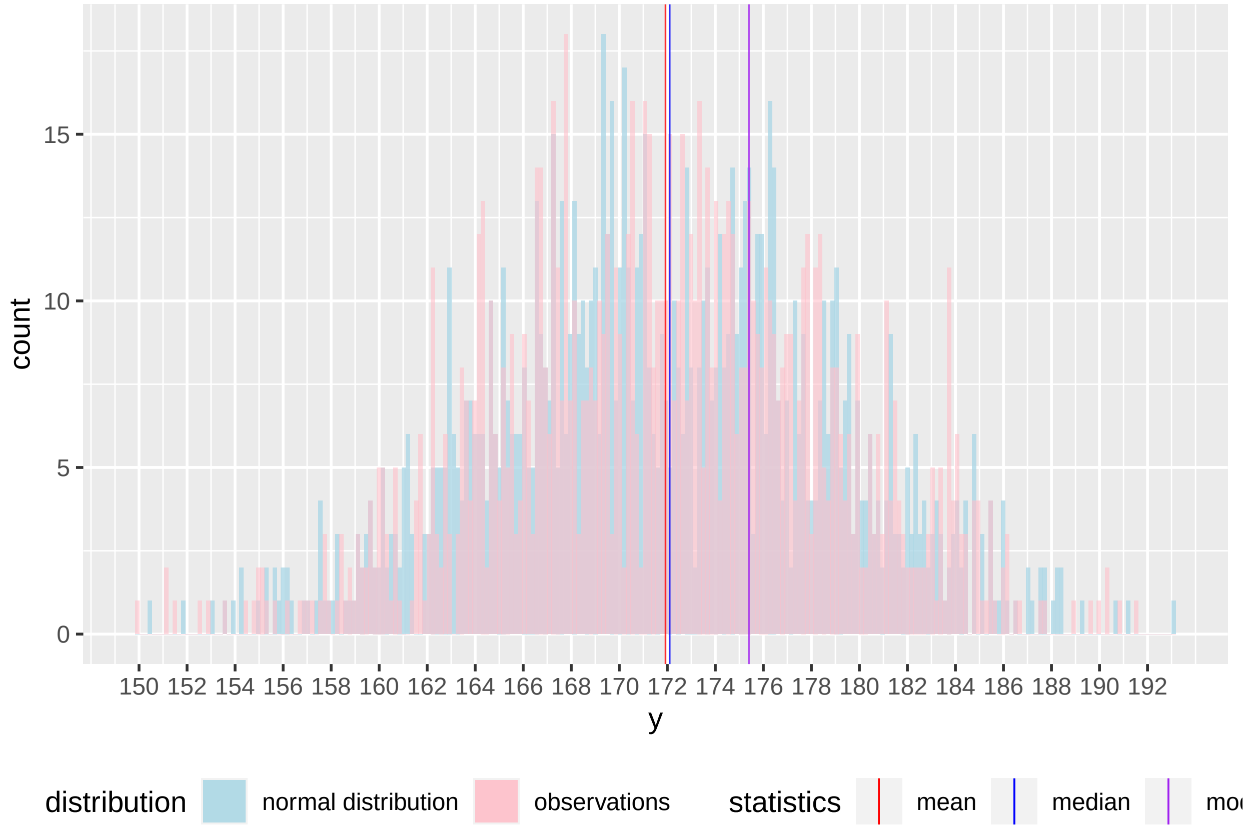
```r
# use cat function to print each result in a single line
cat(paste("There are", length(fatherson$fheight),"observations in the dataset with no missing or empty e
          "The mean value is", mean(fatherson$fheight),". \n",
          "The median value is", median(fatherson$fheight),".\n",
          "The mode value is", mfv(fatherson$fheight),". \n",
          "The standard deviation is", sd(fatherson$fheight),". ","\n",
          "The range of the heights is", range(fatherson$fheight)[1],
          "to", range(fatherson$fheight)[2],". ") )
```

```
## There are 1078 observations in the dataset with no missing or empty entry. 
##  The mean value is 171.925231910946 . 
##  The median value is 172.1 .
##  The mode value is 175.4 . 
##  The standard deviation is 6.97234580524201 .  
##  The range of the heights is 149.9 to 191.6 .
```

RESPONSE:

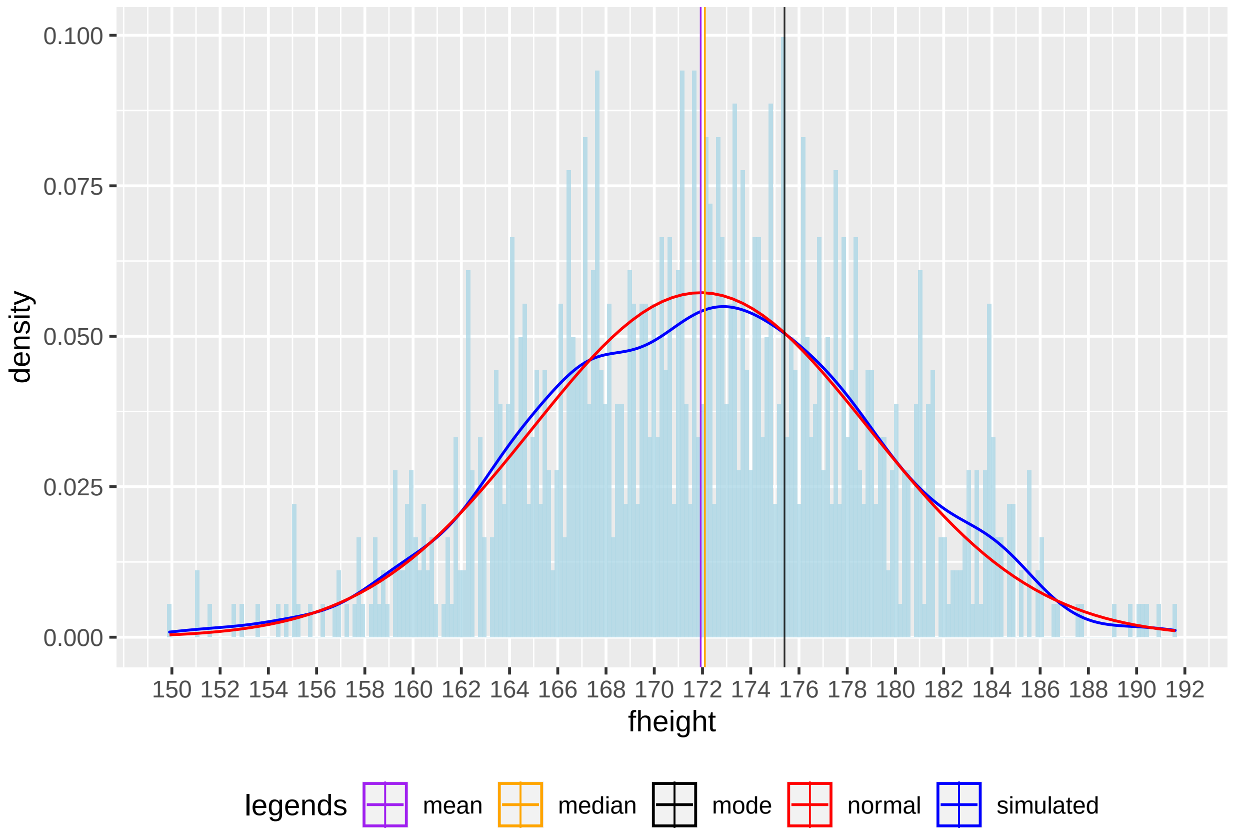
The mean value is smaller than median by approximately 0.2 and mode by approximately 3.5. The data is slightly left-skewed. The standard deviation represents how wide the spread of the data is and it describes the how far the data are away from the mean value. The smaller the standard deviation is, the more uniform the data are, the closer the data are to their mean value.

```
# create a normaldistribution with the same mean and sd
y <- rnorm(1078, mean = mean(fatherson$fheight), sd = sd(fatherson$fheight))


# plot them together
ggplot(fatherson) +
  geom_histogram(aes(y,fill="normal distribution"),bins=250, ,alpha=0.8)+
  geom_histogram(aes(fheight,fill="observations"),bins=250,,alpha=0.6)+
  scale_fill_manual(name="distribution",values=c("lightblue","pink"))+
  geom_vline(aes(xintercept = mean(fheight),color="mean"),size=0.3,alpha=0.8)+
  geom_vline(aes(xintercept = median(fheight),color="median"),size=0.3,alpha=0.8)+
  geom_vline(aes(xintercept = mfv(fheight),color="mode"),size=0.3,alpha=0.8)+
  scale_x_continuous(breaks=seq(as.integer(min(y)),as.integer(max(y)),by = 2))+
  scale_color_manual(name = "statistics", values = c(median = "blue", mean = "red",mode="purple"))+
  theme(legend.position="bottom")
```



reference

```
# not sure what kind of plot the question is looking for
# so I plot another density plot
ggplot(fatherson,aes(x=fheight)) +
  geom_histogram(aes(y=..density..),fill="lightblue",bins = 250,alpha=0.8)+
  geom_density(aes(color="simulated"))+
  stat_function(aes(color = "normal"), fun = dnorm,
                args = list(mean = mean(fatherson$fheight), sd = 6.97))+
  geom_vline(aes(xintercept = mean(fheight),color="mean"),size=0.3)+
  geom_vline(aes(xintercept = median(fheight),color="median"),size=0.3)+
  geom_vline(aes(xintercept = mfv(fheight),color="mode"),size=0.3,alpha=0.8)+
  scale_x_continuous(breaks=seq(as.integer(min(y)),as.integer(max(y)),by = 2))+
  scale_color_manual(name = "legends", values =
                       c(simulated="blue", normal="red",median = "orange", mean = "purple",mode="black"
  theme(legend.position="bottom")
```



RESPONSE:

The two plots are almost overlapped. They are similar, but we can see more pink on the left tail compared to the right tail of the plot. It means the distribution of the father height data is left-skewed.

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

1. 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 (e.g. continuous, discrete, positive, negative...)?
2. 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?
3. Compute mean, median, mode, standard deviation and range of the citations. Discuss the relationship

between these numbers. Is mean larger than median? Than mode? By how much (in relative terms)? What does this suggest? How does standard deviation compare to mean?

4. 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)

```r
magcite <- read.csv('mag-in-citations.csv.bz2')

# descriptive analysis of the data
# types of variables
str(magcite)
```

```
## '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 ...
```

```r
# missing value
sum(is.na(magcite$citations))
```

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

```r
# empty value
sum(str_length(magcite$citations)==0)
```

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

RESPONSE
There are 388258 observations of 2 variables and no missing or empty value.

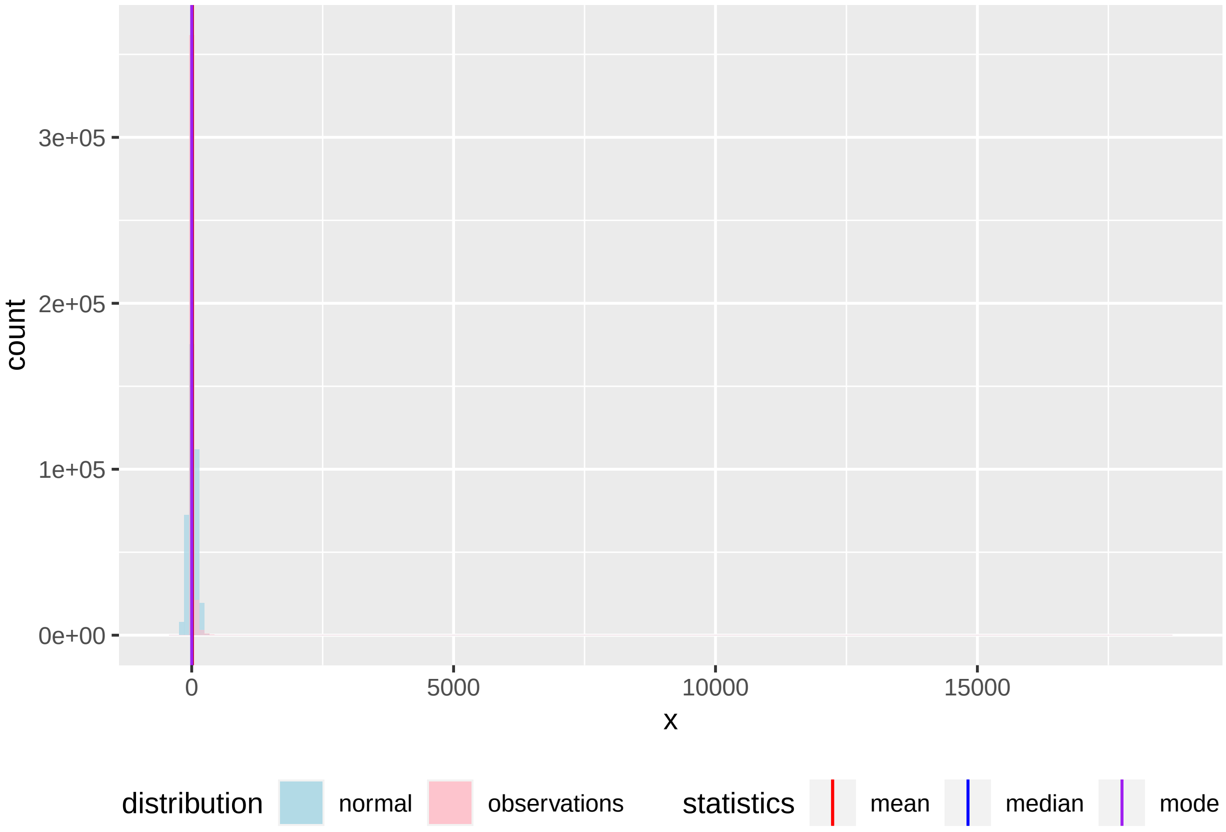
```r
mean <- mean(magcite$citations)
median <- median(magcite$citations)
mode <- mfv(magcite$citations)
sd <- sd(magcite$citations)
range <- range(magcite$citations)

cat(paste("mean is",mean,". median is",median,". mode is",mode,". \nstandard deviation is",sd,". \nThe
```

```
## mean is 15.612226921274 . median is 3 . mode is 0 .
## standard deviation is 78.3907906963468 .
## The range of values is from 0 to 18682 .
```

RESPONSE
The median is much smaller than the mean, by almost 13. It means the distribution is right-skewed. The mode is 0 which means papers that have not been cited are the most frequent observations in the dataset. The standard deviation is high which means the data spread widely. Most of the data is far away from the mean value. We can also see the range is wide.

reference

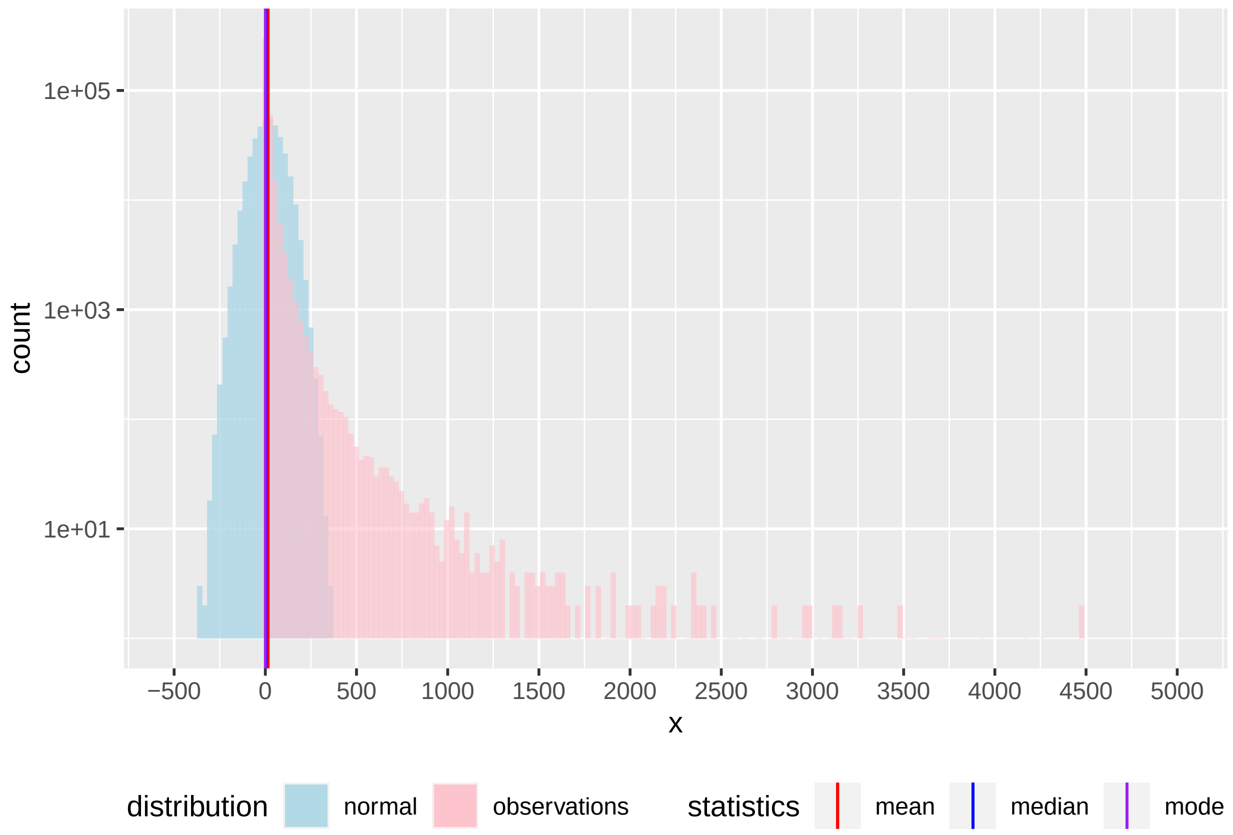
```r
x <- rnorm(nrow(magcite),mean = mean, sd = sd)


p <- ggplot()+
        geom_histogram(aes(x,fill="normal"),bins = 200, alpha=0.8)+
        geom_histogram(data=magcite,aes(citations,fill="observations"),bins = 200,alpha=0.6)+
        scale_fill_manual(name="distribution",values=c("lightblue","pink"))+
        geom_vline(aes(xintercept = mean,color="mean"),size=0.5)+
        geom_vline(aes(xintercept = median,color="median"),size=0.5)+
        geom_vline(aes(xintercept = mode,color="mode"),size=0.5)+
        scale_color_manual(name = "statistics", values = c(median = "blue", mean = "red",mode="purple"))
        theme(legend.position="bottom")

p
```



RESPONSE:
The plot could hardly be seen. The data are crowded in the left bottom corner of the plot.

```
# log the y axis
p + scale_y_log10()+
  scale_x_continuous(limits =c (-500,5000),breaks= seq(-5000,5000,by = 500))
```

```
## Warning: Removed 11 rows containing non-finite values (stat_bin).

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

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

## Warning: Removed 173 rows containing missing values (geom_bar).

## Warning: Removed 86 rows containing missing values (geom_bar).
```



RESPONSE:
With the y axis logged, the plot looks much more better than the previous one. The two plots are very different from each other. They are totally in different shapes. The citations plot is significantly right-skewed with a extremely long tail. I even cut off the tail beyond 5000 to focus on the left side of the plot . The most of the data are on the left side of the distribution. The plot shows the wide spread of the data.

(c) Comment on your finding from part (a) and part (b). Be sure to compare the two cases. That is, 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)? RESPONSE:
The distribution of the fathers' height data given in part (a) resembles the normal distribution but slightly left-skewed.
The distribution of the citations data given in part (b) hardly aligns with the normal distribution. This is maybe because a few the papers are basis of further research and development of science. Therefore, they cause more influence than others. Or maybe most of the papers only took a tiny step further in the area that might also have little influence. Here I wonder whether this observation is independent to each other as one citation of one paper must be cited by another paper. While the value human heights is random even for twins who share almost identical DNAs, they cannot have absolutely same heights. According to the central limits theorem, when we have enough data, we can see the distribution of human heights is nearly the normal distribution. I think the findings make sense with my expectations.