IMT 573: Problem Set 5 - Statistics

Jay Kuo

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:

```r
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.

```r
# 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 OpenIntro Statistics (OIS) chapter on distributions.*

Question

- Questions are answered correctly in their entirety

Answer

1 passenger:

$$X_i \sim B(1, 0.98).$$

100 passenger:

$$Y = \sum_{i=1}^{100} X_i \sim B(100, 0.98).$$

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.

Question

- The expected profit is:

Answer

$$100 \times E(Y) - 500 \times \ left[100 - E(Y) \ right] = 8800.$$

```
100*(100*0.98) - 500*(100 - 100*0.98)
```

```
## [1] 8800
```

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?*

Answer

$$Y = \sum_{i=1}^{101} X_i \sim B(101, 0.98)$$

The probability that all 101 passenges will show up is

$$P(Y = 101) = \binom{101}{101} (0.98)^{101} (1 - 0.98)^0 = 0.1299672$$

```
dbinom(101, size = 101, prob = 0.98)
```

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

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

Answer

$$Y = \sum_{i=1}^{102} X_i \sim B(102, 98)$$

The probability that all 102 passengers show up is

$$P(Y = 102) = \binom{102}{102} (0.98)^{102} (1 - 0.98)^0 = 0.1273678$$

```r
dbinom(102, size = 102, prob = 0.98)
```

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

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

Answer

102 passengers:

$$Y = \sum_{i=1}^{102} X_i \sim B(102, 0.98)$$

The probability that all 101 passengers show up is

$$P(Y = 101) = \binom{102}{101} (0.98)^{101} (1 - 0.98)^{102-101} = 0.265133$$

```r
dbinom(101, size = 102, prob = 0.98)
```

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

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?

Answer

- Independent: passenger i cannot impact on the attendance of passenger j.
- If independence does not exist, the covariance structure needs to be adjusted.
- Independence may not be realistic.
- Counterexample: A couple plans to go abroad, but they encounter an emergency event so they cannot show up.

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?…)?

Question

- Provides basic description / exploratory analysis of fheight variable.

Answer

human height: continuous; positive.

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?…)

Question

- Checks for missing data accross entire dataset.

Answer

```
# import the data
fatherson <- read.csv("/Users/Jay/Desktop/J/Working on/MSIM/IMT573/problem set/problemset5/fatherson.cs
                      header = FALSE, skip = 1, sep = "")

colnames(fatherson) <- c("fheight", "sheight")

# summary statistics
summary(fatherson)
```

```
##     fheight         sheight
##  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
```

```
# structure
str(fatherson)
```

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

```
# check missing values
is.na(fatherson) %>%
  apply(2, sum)
```

```
## fheight sheight
##       0       0
```

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?

Question

- Computes mean, median, mode, standard deviation, and range of the heights for both fathers and sons.

Answer

```
sum_stat <- function(x){
  res <- data.frame(
    Mean = mean(x),
    Median = median(x),
    Mode = names(sort(-table(x)))[1],
    SD = sd(x),
    Range = max(x) - min(x))
  return(res)

}

# apply(fatherson, 2, sum_stat)
fatherson$fheight %>%
  sum_stat()
```

```
##       Mean Median  Mode       SD Range
## 1 171.9252  172.1 175.4 6.972346  41.7
```

Question

- Discusses the relationship between these numbers and answers all questions.

Answer

fheight: mean < median; left-skewed; mean < mode; the dispersion above mean is smaller than below mean. mean, SD: (171.9 +- 6.9) contains 67% population if normal.

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 citation 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?

Question

- Plots histogram of the data with appropriate title and labeld axes.
- Overlays a plot of the normal distribution with the same mean and standard deviation as the data.
- Indicates mean and median of the data using vertical lines of different colors

Answer

```
ggplot(fatherson, aes(x = fheight)) +
  geom_histogram(aes(y = ..density..), color = 1, fill = 'white', bins = 20)+
  labs(title = "Histogram of fheight (fathers' heights)")+
  # normal distribution
  stat_function(fun = dnorm, args = list(mean = mean(fatherson$fheight), sd = sd(fatherson$height)))+
  # mean
  geom_vline(aes(xintercept = mean(fatherson$height), color = "mean")) +
  geom_vline(aes(xintercept = median(fatherson$fheight), color = "median"))+
  # legend
  scale_color_manual(name = "statistics", values = c(mean = "blue", median = "red"))
```

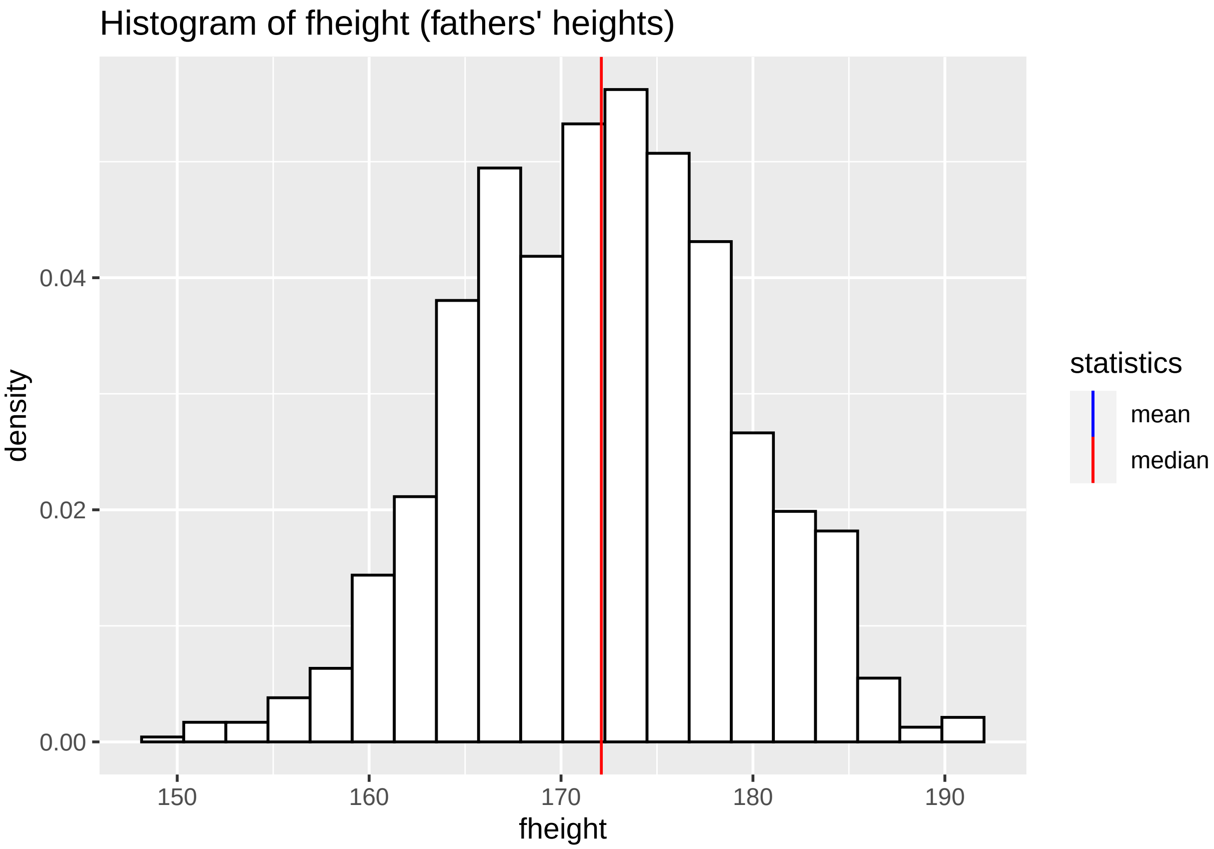
```
## Warning in mean.default(fatherson$height): argument is not numeric or logical:
## returning NA

## Warning: Use of `fatherson$height` is discouraged. Use `height` instead.

## Warning: Use of `fatherson$fheight` is discouraged. Use `fheight` instead.

## Warning: Removed 101 row(s) containing missing values (geom_path).

## Warning: Removed 1078 rows containing missing values (geom_vline).
```



Question

- Discusses findings including similarities or differences.

Answer

The distribution seems symmetric. The histogram and the plot of the normal distribution are similar.

(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?…)?

Question

- Correctly answers how citation counts are measured.

Answer

citation counts: discrete; positive

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.

Question

- Performs a basic exploratory analysis of data looking at data types, missing values, counts, ranges, or other important characteristics.

Answer

```
# import the data
mag_in_citations <- read.csv("/Users/Jay/Desktop/J/Working on/MSIM/IMT573/problem set/problemset5/mag-i

# summary statistics
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
```

```
# structure
str(mag_in_citations)
```

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

```
# check missing values
is.na(mag_in_citations) %>%
  apply(2, sum)
```

```
##   paperId citations 
##         0         0
```

```
# count range
range(mag_in_citations$citations)
```

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

3. 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?

Question

- Computes mean, median, mode, standard deviation and range of the citations.

Answer

Use the function I previously created

```
sum_stat(mag_in_citations$citations)
```

```
##       Mean Median Mode       SD Range
## 1 15.61223      3    0 78.39079 18682
```

Question

- Discusses the relationship betweeen the numbers, answering all questions.

Answer

mean > median: right skewed mean > mode: the dispersion above mean is larger than below mean. mean / SD: The result of mean / SD is close to 0.2; small

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?

Question

- Plots histogram of the data appropriate with title and labeled axes.
- Overlays a plot of the normal distribution with the same mean and standard diviation as the data.
- Indicates mean and median of the data using vertical lines of different colors

Answer

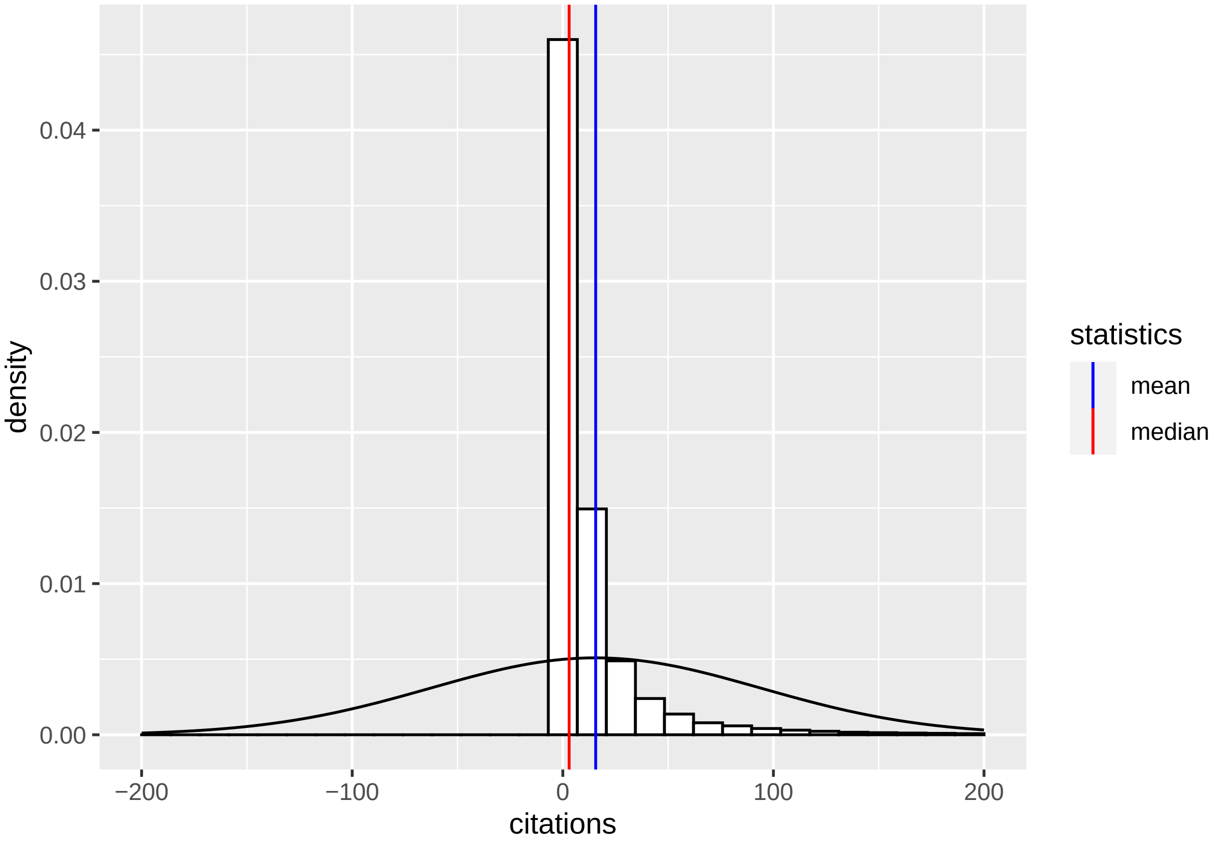
```r
ggplot(mag_in_citations, aes(x = citations)) +
  # histogram
  geom_histogram(aes(y= ..density..), color = 1, fill = 'white') +
  xlim(-200, 200) +
  # normal distribution
  stat_function(fun = dnorm, args = list(mean = mean(mag_in_citations$citations), sd = sd(mag_in_citati
  # mean
  geom_vline(aes(xintercept = mean(mag_in_citations$citations), color = "mean"))+
  geom_vline(aes(xintercept = median(mag_in_citations$citations), color = "median"))+
  # legend
  scale_color_manual(name = "statistics", values = c(mean = "blue", median = "red"))
```

```
## Warning: Use of `mag_in_citations$citations` is discouraged. Use `citations` instead.
## Use of `mag_in_citations$citations` is discouraged. Use `citations` instead.

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

## Warning: Removed 3148 rows containing non-finite values (stat_bin).

## Warning: Removed 1 rows containing missing values (geom_bar).
```



The histogram and the plot of the normal distribution are entirely different.

5. 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?

Question

- Discusses findings including similarities or differences.
- Plots data on log-log scale with appropriate title and labeld axes.
- Indicates mean and median of the data using vertical lines of different colors

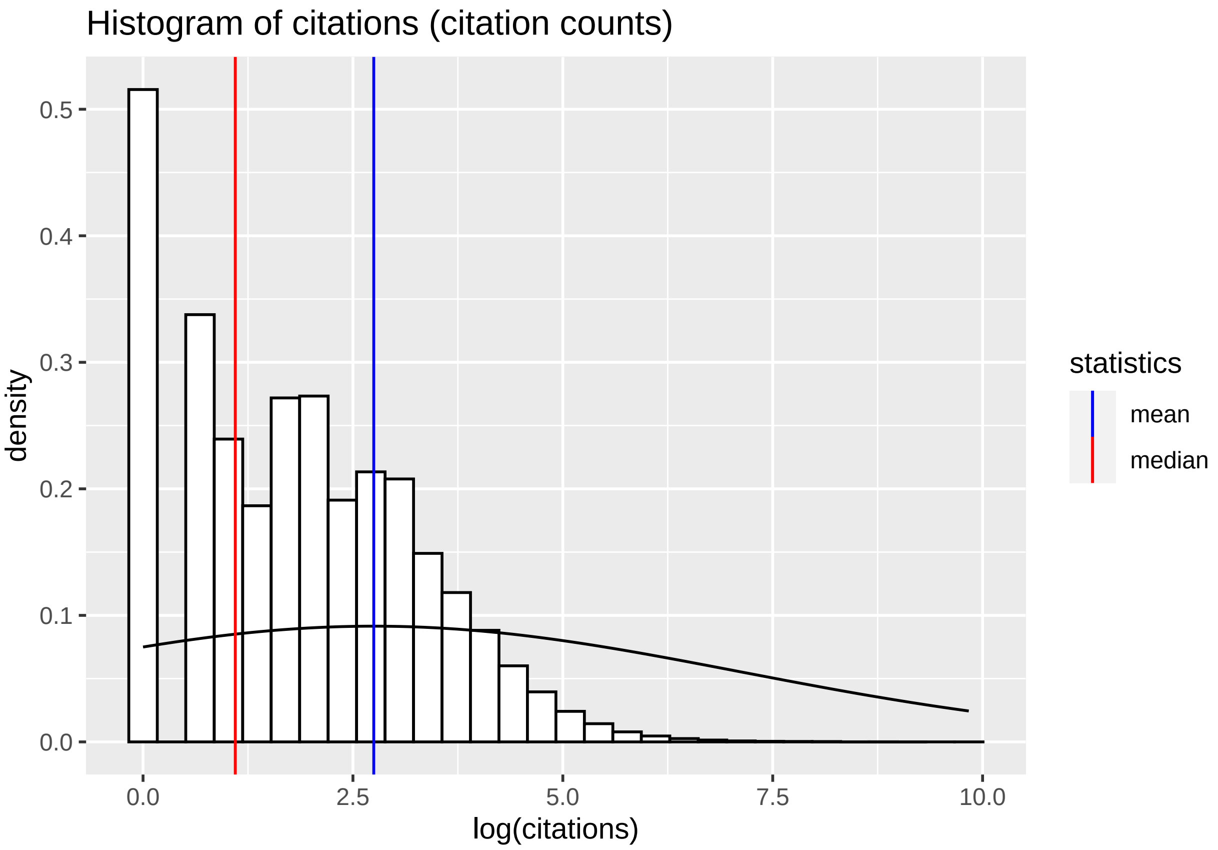
Answer

```
#mag_in_citations$log_citations <- ifelse(mag_in_citations$citations == 0, 0, log(mag_in_citations$cita

ggplot(mag_in_citations, aes(x= log(citations))) +
  # histogram
  geom_histogram(aes(y = ..density..), color = 1, fill = "white", bins = 30) +
  labs(title = "Histogram of citations (citation counts)")+
  # normal distribution
  stat_function(fun = dnorm, args = list(mean = log(mean(mag_in_citations$citations)), sd = log(sd(mag_i
  geom_vline(aes(xintercept = log(mean(mag_in_citations$citations)), color = "mean")) +
  geom_vline(aes(xintercept = log(median(mag_in_citations$citations)), color = "median")) +
  scale_color_manual(name = "statistics", values = c(mean = "blue", median = "red"))
```

```
## Warning: Use of `mag_in_citations$citations` is discouraged. Use `citations` instead.
## Use of `mag_in_citations$citations` is discouraged. Use `citations` instead.
```

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



Question

- Discusses findings including similarities or differences.

Answer

The histogram is more similar to normal distribution than before.

(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? Question

- Discusses relationship to normal distribution for parts a and b

Answer

Part A: normal distribution is valid. Part B: normal distribution is invalid

Question

- Discusses whether findings make sense in regard to expectations and provides reasoning as to why or why not.

Answer

Part A: It is natural phenomenon. Part B: Commonly, the citation count is zero

Question Answer