IMT 573: Problem Set 5 - Statistical Theory

Srushti Chaukhande

Due: Tuesday, November 5, 2024 by 10:00PM PT

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 `problem_set5.Rmd` file from Canvas or save a copy to your local directory on RStudio Server. Open `problem_set5.Rmd` in RStudio and supply your solutions to the assignment by editing `problem_set5.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)
```

Problem 1: Overbooking Flights (18pts)

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. (2pts) Which distribution would you use to describe the actual number of show-ups for the flight? Hint: read OIS ch 3 about distributions.
2. (2pts) 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. (2pts) Now assume the airline sells 101 tickets for 100 seats. What is the probability that all 101 passengers will show up?
4. (2pts) 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. (2pts) Now assume the airline sells 102 tickets. What is the probability that all 102 passengers show up?
6. (2pts) What is the probability that 101 passengers – still one too many – will show up?
7. (2pts) 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. (2pts) What is the optimal number of seats to sell for the airline? How big are the expected profits?
9. (2pts) 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).

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.

Solution 1 Insert Response

1. To describe the actual number of show-ups for the flight, we can use a binomial distribution. Here's why:

- Independent Events: Each passenger shows up independently of others.
- Fixed Number of Trials: The number of trials (passengers with tickets) is fixed at 100.
- Two Possible Outcomes: Each ticket holder either shows up or doesn't.
- Constant Probability of Success: The probability of any individual passenger showing up is constant at 98% (or 0.98). In this case, the binomial distribution B(n=100,p=0.98) would model the number of passengers who actually show up. This distribution helps estimate the likelihood of different show-up counts, allowing you to calculate the probability of various scenarios to determine the optimal overbooking rate.

```
# Setting parameters
n_seats <- 100         # Number of seats (trials)
prob_show_up <- 0.98  # Probability of a passenger showing up

# Generating a sample of number of show-ups for demonstration
# Using rbinom to simulate show-ups for a single flight with 100 seats
set.seed(42)  # Setting seed for reproducibility
show_ups <- rbinom(1, size = n_seats, prob = prob_show_up)

# Display the number of show-ups
show_ups
```

```
## [1] 96
```

2. To calculate the airline's expected profit without overbooking, we'll break down the components of expected revenue and expected costs.

- Step 1: Expected Revenue Ticket Price: Each ticket is sold for $100. Expected Show-Ups: With a probability of 98% for each passenger to show up on a 100-seat plane, the expected number of passengers who actually show up is: Expected Show-Ups = 100×0.98 = 98 Expected Revenue: Therefore, the expected revenue from the passengers who show up is: Expected Revenue=100×100=10,000 dollars
- Step 2: Expected Costs Since the airline is not overbooking, no passengers are denied boarding, meaning there are no costs associated with finding alternative solutions.
- Step 3: Expected Profit The expected profit can then be calculated as: Expected Profit=Expected Revenue−Expected Costs=10,000−0=10,000dollars So, the expected profit for the airline without overbooking is $10,000.

3. If the airline sells 101 tickets and each passenger has a 98% chance of showing up, we need to calculate the probability that all 101 ticketed passengers will show up.

Since the show-up events are independent, the probability that all 101 passengers show up can be calculated as: P(all 101 show up)=0.98^101 The probability that all 101 passengers will show up is approximately 0.13, or 13%.

```
# Parameters
prob_show_up <- 0.98  # Probability of each passenger showing up
n_tickets <- 101      # Total tickets sold

# Calculating the probability that all 101 passengers show up
prob_all_show_up <- prob_show_up^n_tickets

# Display the probability
prob_all_show_up
```

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

4. To determine whether overbooking by selling 101 tickets is more profitable, let's break down the expected revenue and the expected additional costs.

- Step 1: Expected Revenue Since each ticket costs $100: Expected Revenue=101×100=10,100dollars

- Step 2: Expected Additional Costs Probability of Needing Alternative Solutions: The airline incurs additional costs only if more than 100 passengers show up. This happens with a probability of: P(all 101 passengers show up)=0.98^101 0.13 Cost of Alternative Solutions: If 101 passengers show up but only 100 seats are available, 1 passenger will be denied boarding. The cost of rebooking or compensating that passenger is $500.

Expected Additional Costs: Thus, the expected additional costs are: Expected Additional Costs=P(all 101 show up)×500 Let's calculate the expected additional costs and the final expected profit.

The calculations give: Expected Revenue: $10,100 Expected Additional Costs: Approximately $65 Expected Profit: Approximately $10,035

Recommendation With overbooking, the expected profit is $10,035, which is slightly higher than the expected profit of $10,000 when selling only 100 tickets. Recommendation: Yes, overbooking by one ticket (selling 101 tickets) is likely more profitable, with a slight increase in expected profit. This strategy carries minimal risk of additional costs and can boost profitability without significantly impacting customer experience.

5. P(all 102 show up)=0.98^102 = 0.127 (or about 12.7%).

```
# Parameters
prob_show_up <- 0.98  # Probability of each passenger showing up
n_tickets <- 102      # Total tickets sold

# Calculating the probability that all 102 passengers show up
prob_all_show_up <- prob_show_up^n_tickets

# Display the probability
prob_all_show_up
```

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

6. Finding the probability that exactly 101 passengers show up when 102 tickets are sold :

```
# Given values
total_passengers <- 102     # Total tickets sold
k_passengers_showing <- 101  # Passengers that show up
p_show_up <- 0.98          # Probability of a passenger showing up

# Calculate the probability using dbinom
prob_101_show_up <- dbinom(k_passengers_showing, total_passengers, p_show_up)

# Print the result
prob_101_show_up
```

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

7. Selling 100 Tickets: Expected Revenue=100×100=10,000 Expected Additional Costs: $0 (no overbooking) Expected Profit: $10,000

Selling 101 Tickets: Expected Revenue=101×100=10,100 Expected Additional Costs: Probability of 101 show-ups and associated costs :

```
prob_101_show_up <- dbinom(101, size = 101, prob = 0.98)
expected_costs_101 <- prob_101_show_up * 1 * 500  # Compensation for 1 overbooked
#passenger
expected_profit_101 <- 10100 - expected_costs_101
print(expected_profit_101)
```

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

Selling 102 Tickets: Expected Revenue=102×100=10,200 Expected Additional Costs: Probability of 101 show-ups:

```
prob_101_show_up <- dbinom(101, size = 102, prob = 0.98)
expected_costs_102_101_show_up <- prob_101_show_up * 1 * 500  # Compensation for 1 overbooked
#passenger

prob_102_show_up <- dbinom(102, size = 102, prob = 0.98)
#Compensation for 2 overbooked passengers
expected_costs_102_102_show_up <- expected_costs_102_101_show_up + prob_102_show_up * 2 * 500

print(expected_costs_102_101_show_up)
```

```
## [1] 132.5665
```

```
print(expected_costs_102_102_show_up)
```

```
## [1] 259.9343
```

```
expected_profit_102_101_show_up <- 10200 - expected_costs_102_101_show_up
expected_profit_102_102_show_up <- 10200 - expected_costs_102_102_show_up

print(expected_profit_102_101_show_up)
```

```
## [1] 10067.43
```

```
print(expected_profit_102_102_show_up)
```

```
## [1] 9940.066
```

Revenue of selling 102 tickets is definitely higher than selling 100 or 101 tickets but the profit depends on additional costs incurred.

8.

```
# Given values
ticket_price <- 100        # Price per ticket
cost_per_denied <- 500      # Cost per denied boarding

# Function to calculate expected profits for different numbers of tickets sold
calculate_expected_profit <- function(total_passengers) {
  p_show_up <- 0.98
  # Expected Revenue
  expected_revenue <- total_passengers * ticket_price

  # Expected Additional Costs
  if (total_passengers == 100) {
    expected_additional_costs <- 0
  } else if (total_passengers == 101) {
    prob_all_show_up <- p_show_up^total_passengers
    expected_additional_costs <- prob_all_show_up * cost_per_denied
  } else { # total_passengers == 102
    prob_all_show_up <- p_show_up^total_passengers
    prob_101_show_up <- dbinom(101, total_passengers, p_show_up)
    expected_additional_costs <- prob_all_show_up *
      cost_per_denied + prob_101_show_up * cost_per_denied
  }

  # Expected Profit
  expected_profit <- expected_revenue - expected_additional_costs
  return(expected_profit)
}

# Calculate expected profits for 100, 101, and 102 tickets
profit_100 <- calculate_expected_profit(100)
profit_101 <- calculate_expected_profit(101)
profit_102 <- calculate_expected_profit(102)

# Print the expected profits
profit_100
```

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

```
profit_101
```

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

```
profit_102
```

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

```
# Find the optimal number of tickets to sell
optimal_profit <- max(profit_100, profit_101, profit_102)
optimal_tickets <- which.max(c(profit_100, profit_101, profit_102)) + 99  # Adjusting
#index

# Print optimal number of tickets and expected profit
optimal_tickets
```

```
## [1] 101
```

```
optimal_profit
```

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

Optimal tickets to be sold are 101 with optimal profit of 10035.02

9. When we say that the show-ups are independent, it means that the probability of one passenger showing up does not affect the probability of another passenger showing up. In statistical terms, for any two passengers A and B :

P(A shows up B shows up)=P(A shows up)×P(B shows up)

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. (28pts)

1. (3pts) What kind of measurement (nominal, ordered, difference, ratio) does human height use? How should it be measured (e.g. continuous, discrete, positive...)?
2. (4pts) 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. (8pts) 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: 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!
4. (13pts) 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?

Solution 2 (a)

1. **Solution:**

Insert Response

Human height is measured using a ratio scale and should be measured as a continuous variable. Let's break this down: Measurement Scale: Ratio Human height is measured on a ratio scale for the following reasons: - It has a true zero point: A height of 0 cm or 0 inches represents the absence of height, which is a meaningful concept. - It allows for meaningful ratios: We can say that someone who is 180 cm tall is twice as tall as someone who is 90 cm tall. - It has equal intervals: The difference between 150 cm and 160 cm is the same as the difference between 170 cm and 180 cm. - It includes all the properties of nominal, ordinal, and interval scales.

Measurement Type: Continuous Height should be measured as a continuous variable because: It can take any value within a range: Height can be measured with arbitrary precision (e.g., 175.6 cm, 175.63 cm, 175.632 cm, etc.). It is not limited to discrete steps: Unlike discrete variables that can only take specific values, height can be any real number above zero. It allows for infinite possible values: There are infinitely many possible height measurements between any two values.

2. **Solution:**

Insert Response

```r
# Read the CSV file
data <- read.delim("fatherson.csv")

# Basic description of the fheight variable
summary_fheight <- summary(data$fheight)
num_observations <- length(data$fheight)
missing_data <- sum(is.na(data$fheight))

# Print results
print(summary_fheight)
```

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

```r
cat("Number of observations:", num_observations, "\n")
```

```
## Number of observations: 1078
```

```r
cat("Number of missing values:", missing_data, "\n")
```

```
## Number of missing values: 0
```

We have 1078 observations and there is no missing data.

3. **Solution:**

Insert Response

```r
# Extract father's height
fheight <- data$fheight

# Compute statistics
mean_height <- mean(fheight)
median_height <- median(fheight)
sd_height <- sd(fheight)
range_height <- range(fheight)

# Compute mode using modeest package
```

```r
library(modeest)
mode_height <- mlv(fheight, method = "mfv")  # Most frequent value method

# Print results
cat("Mean:", mean_height, "\n")
```

```
## Mean: 171.9252
```

```r
cat("Median:", median_height, "\n")
```

```
## Median: 172.1
```

```r
cat("Mode:", mode_height, "\n")
```

```
## Mode: 175.4
```

```r
cat("Standard Deviation:", sd_height, "\n")
```

```
## Standard Deviation: 6.972346
```

```r
cat("Range:", range_height[1], "to", range_height[2], "\n")
```

```
## Range: 149.9 to 191.6
```

- Mean is slightly lower than median. Absolute difference: -0.17 cm
- Mode is the largest of 3. Absolute difference: mode - mean = -3.47 cm
- Approximately 68% of the fathers' heights are expected to fall within one standard deviation of the mean (between 164.96 cm and 178.90 cm), assuming a normal distribution.

4. **Solution:**

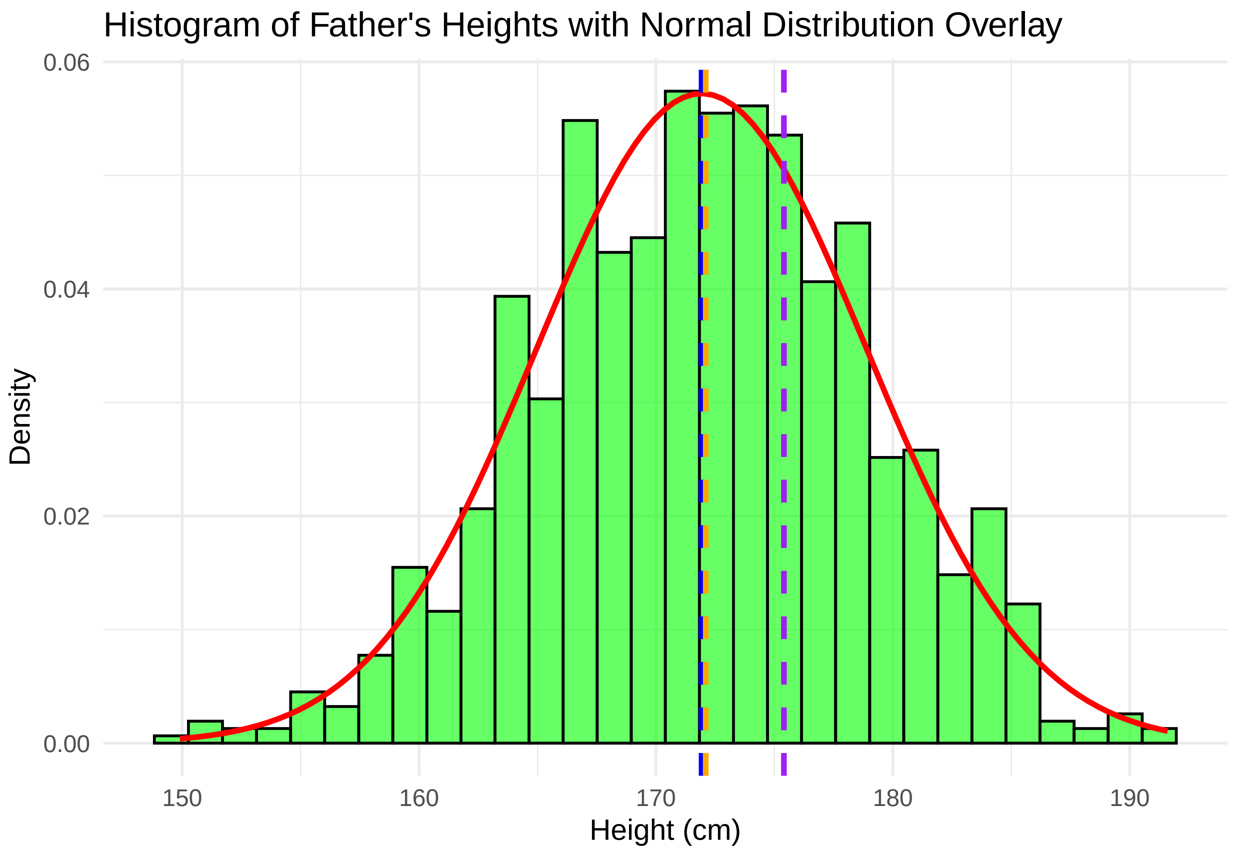
Insert Response

```r
# Load necessary libraries
library(ggplot2)
library(modeest)

# Extract father's height
fheight <- data$fheight

# Calculate statistics
mean_height <- mean(fheight)
median_height <- median(fheight)
mode_height <- as.numeric(names(which.max(table(fheight))))  # Most frequent value method
sd_height <- sd(fheight)

# Create the plot
ggplot(data, aes(x = fheight)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "green", alpha = 0.6,
                 color = "black") +
  stat_function(fun = dnorm, args = list(mean = mean_height, sd = sd_height),
                color = "red", size = 1) +
  geom_vline(xintercept = mean_height, color = "blue", linetype = "dashed", size = 1) +
  geom_vline(xintercept = median_height, color = "orange", linetype = "dashed", size = 1) +
  geom_vline(xintercept = mode_height, color = "purple", linetype = "dashed", size = 1) +
  labs(title = "Histogram of Father's Heights with Normal Distribution Overlay",
       x = "Height (cm)",
       y = "Density") +
  theme_minimal() +
  scale_color_manual(name = "Statistics",
                     values = c("blue", "orange", "purple", "red"),
                     labels = c("Mean", "Median", "Mode", "Normal Distribution")) +
  guides(color = guide_legend(override.aes = list(linetype = c("dashed",
                                                               "dashed", "dashed", "solid"))))
```



The histogram and density plot are quite similar. The close alignment between the actual data distribution and the theoretical normal distribution suggests that father's heights in this dataset can be well-approximated by a normal distribution.

(b) Next, let's take a look at the number of citations of research papers. (40pts)

1. (3pts) What kind of measure is this? What kind of valid values would you expect to see (continuous, discrete, positive, ...)

2. (4pts) 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. (8pts) 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.

4. (13pts) 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?

5. (12pts) Plot a histogram of the data on a log-log scale. Add to this histogram vertical lines representing the sample mean, median, and mode (use different colors for each). How does the distribution appear on a log-log scale? Are the mean, median, and mode close together or spread out? What does this tell you about the data's distribution?

Solution 2 (b)

1. **Solution:**

Insert Response

Type of Measure: This is a count measure, specifically the number of citations for each paper. Valid Values: Discrete: The number of citations is always a whole number (integer). We don't see fractional citations in this dataset. Non-negative: Citations cannot be negative. The minimum possible value is 0, which represents a paper that has not been cited. Unbounded on the upper end: There's no theoretical maximum number of citations a paper can receive.

2. **Solution:**

Insert Response

```
citations_data <- read.delim("mag-in-citations.csv.bz2", sep=',')
# Basic descriptives of the 'citations' variable
num_observations <- nrow(citations_data)
missing_values <- sum(is.na(citations_data$citations))

# Print results
cat("Number of observations:", num_observations, "\n")
```

```
## Number of observations: 388258
```

```
cat("Number of missing values:", missing_values, "\n")
```

```
## Number of missing values: 0
```

3. **Solution:**

Insert Response

```
# Compute statistics
mean_citations <- mean(citations_data$citations)
median_citations <- median(citations_data$citations)
mode_citations <- mlv(citations_data$citations, method = "mfv")
sd_citations <- sd(citations_data$citations)
range_citations <- range(citations_data$citations)

# Print results
cat("Mean:", mean_citations, "\n")
```

```
## Mean: 15.61223
```

```
cat("Median:", median_citations, "\n")
```

```
## Median: 3
```

```
cat("Mode:", mode_citations, "\n")
```

```
## Mode: 0
```

```
cat("Standard Deviation:", sd_citations, "\n")
```

```
## Standard Deviation: 78.39079
```

```
cat("Range:", range_citations[1], "to", range_citations[2], "\n")
```

```
## Range: 0 to 18682
```

```
# Compute relative differences
mean_median_diff <- (mean_citations - median_citations) / median_citations * 100
mean_mode_diff <- (mean_citations - mode_citations) / mode_citations * 100
sd_mean_ratio <- sd_citations / mean_citations * 100

mean_median_diff
```

```
## [1] 420.4076
```

```
mean_mode_diff
```

```
## [1] Inf
```

```
sd_mean_ratio
```

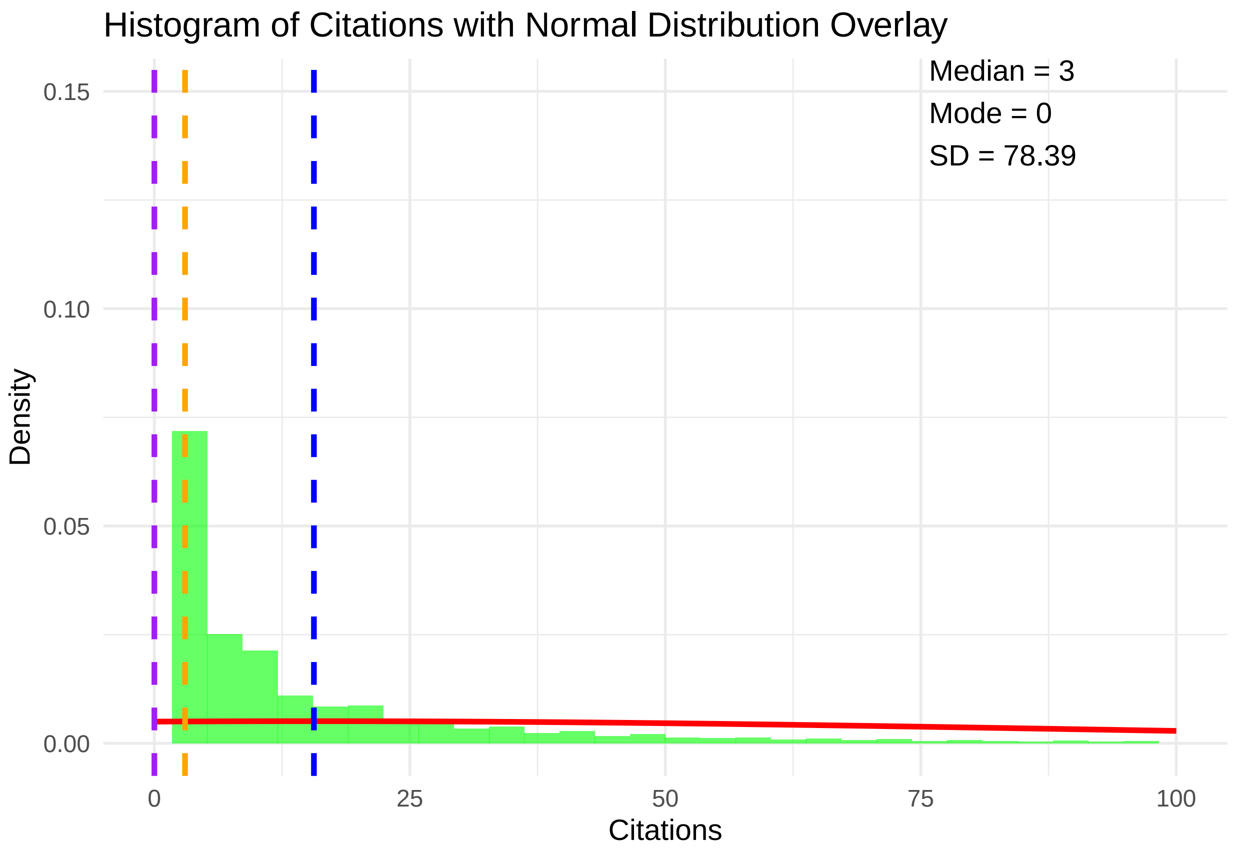
```
## [1] 502.1115
```

The mean (15.61) is larger than the median (3.0). The mean (15.61) is larger than the mode (0). The standard deviation is 502.11% of the mean.

4. **Solution:**

Insert Response

```r
# Create the plot
ggplot(citations_data, aes(x = citations)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = "green", alpha = 0.6) +
  stat_function(fun = dnorm, args = list(mean = mean_citations, sd = sd_citations),
                color = "red", size = 1) +
  geom_vline(xintercept = mean_citations, color = "blue", linetype = "dashed",
             size = 1) +
  geom_vline(xintercept = median_citations, color = "orange", linetype = "dashed",
             size = 1) +
  geom_vline(xintercept = mode_citations, color = "purple", linetype = "dashed",
             size = 1) +
  scale_x_continuous(limits = c(0, 100)) +  # Adjust x-axis limit for better visibility
  labs(title = "Histogram of Citations with Normal Distribution Overlay",
       x = "Citations",
       y = "Density") +
  theme_minimal() +
  annotate("text", x = 75, y = 0.15,
           label = paste("Mean =", round(mean_citations, 2), "\n",
                         "Median =", median_citations, "\n",
                         "Mode =", mode_citations, "\n",
                         "SD =", round(sd_citations, 2)),
           hjust = 0)
```



The histogram shows a highly right-skewed distribution of citations. The vast majority of papers have very few citations, with a long tail extending to the right for papers with many citations.

The red line representing the normal distribution with the same mean and standard deviation as the data does not fit the actual distribution well. This indicates that the citation data is not normally distributed.
5. **Solution:**

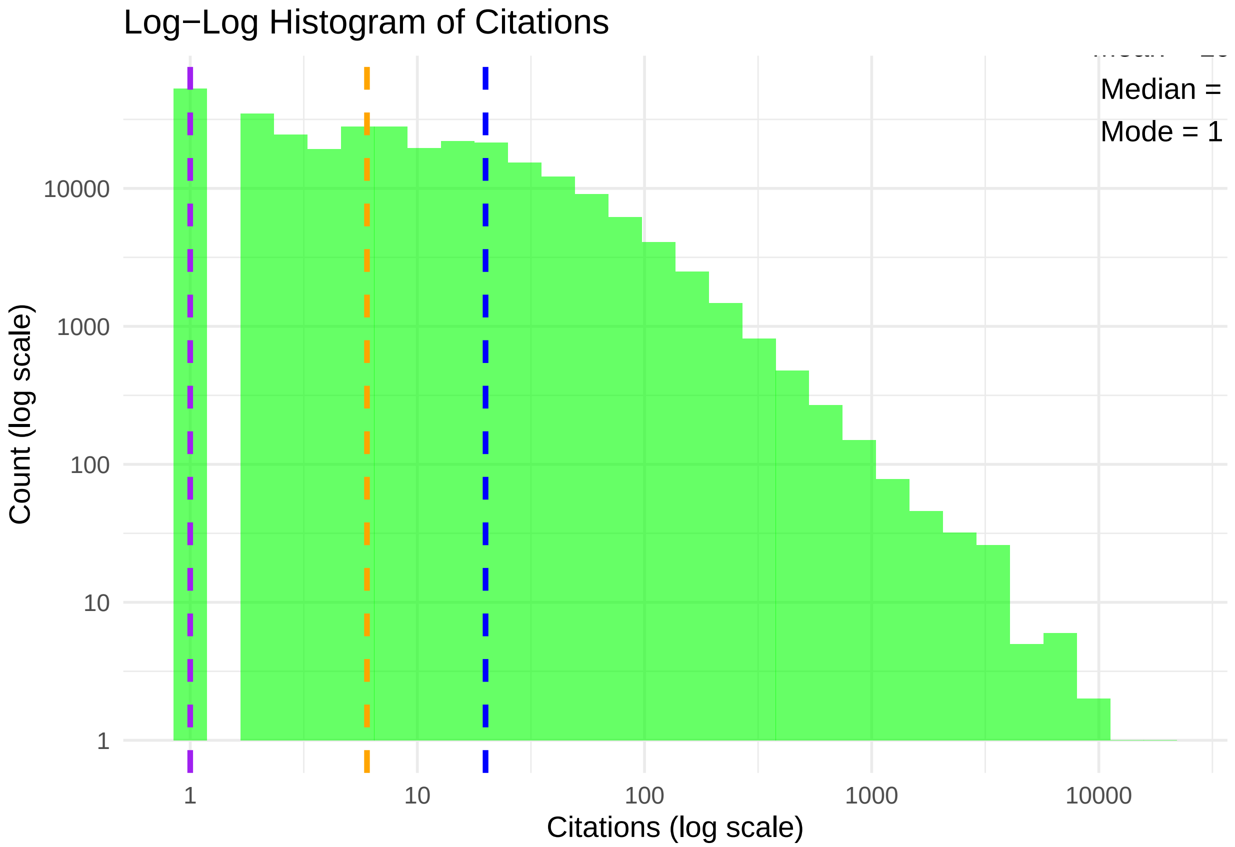
Insert Response

(c) (6 pts) Comment on your finding from part (a) and part (b). Be sure to compare the two cases. Solution:

Insert Response

```r
# Remove zero citations to avoid log(0)
citations_nonzero <- citations_data$citations[citations_data$citations > 0]

# Calculate statistics
mean_citations <- mean(citations_nonzero)
median_citations <- median(citations_nonzero)
mode_citations <- mlv(citations_nonzero, method = "mfv")
# Create the plot
ggplot(data.frame(citations = citations_nonzero), aes(x = citations)) +
  geom_histogram(bins = 30, fill = "green", alpha = 0.6) +
  scale_x_log10() +
  scale_y_log10() +
  geom_vline(xintercept = mean_citations, color = "blue", linetype =
               "dashed", size = 1) +
  geom_vline(xintercept = median_citations, color = "orange", linetype =
               "dashed", size = 1) +
  geom_vline(xintercept = mode_citations, color = "purple", linetype =
               "dashed", size = 1) +
  labs(title = "Log-Log Histogram of Citations",
       x = "Citations (log scale)",
       y = "Count (log scale)") +
  theme_minimal() +
  annotate("text", x = max(citations_nonzero)/2, y = max(table(citations_nonzero)),
           label = paste("Mean =", round(mean_citations, 2), "\n",
                         "Median =", median_citations, "\n",
                         "Mode =", mode_citations),
           hjust = 0)
```



On a log-log scale, the distribution will appear more linear if it follows a power-law distribution. This is common in citation data. The mean is typically higher than both the median and mode due to the right-skewed nature of citation distributions. The mode being zero indicates that many papers have no citations. The mean is often far from the median and mode in skewed distributions. This spread indicates that using the mean alone might not accurately represent the typical citation count.