Group 22

PKQV7, LRZZ1, MDSZ1, PKVJ4, NFHJ5, MBXX0

2025-05-11

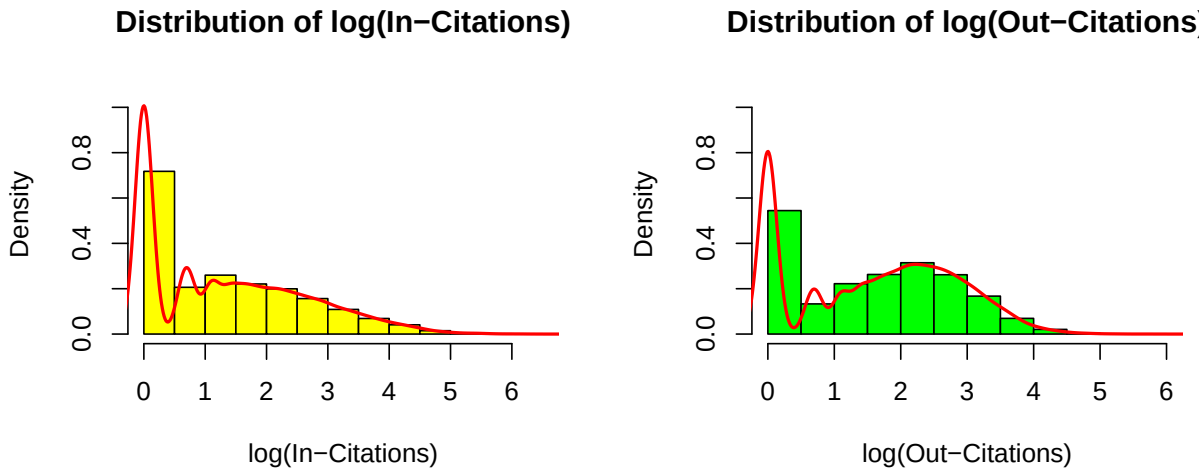
Task 1

```
# Read data files
papers <- read.csv("papers.csv", header = TRUE)
citations <- read.csv("citations.csv", header = TRUE)

# Calculate out-citations (papers cited by a paper)
out_citations <- table(citations$FromID)
out_citations_df <- data.frame(PaperID = as.numeric(names(out_citations)),
                               OutCitations = as.numeric(out_citations))
# Calculate in-citations (papers citing a paper)
in_citations <- table(citations$ToID)
in_citations_df <- data.frame(PaperID = as.numeric(names(in_citations)),
                              InCitations = as.numeric(in_citations))

# Combine citation data with the full list of papers
paper_citations <- merge(papers, out_citations_df, by = "PaperID", all.x = TRUE)
paper_citations <- merge(paper_citations, in_citations_df, by = "PaperID", all.x = TRUE)
paper_citations[is.na(paper_citations)] <- 0 # Replace NAs with 0 for zero citation

cat(capture.output({cat("Summary Statistics for In-Citations:\n")
 print(summary(paper_citations$InCitations))        # Printing in-citation statistics
 cat("\nSummary Statistics for Out-Citations:\n")
 print(summary(paper_citations$OutCitations))       # Printing out-citation statistics
 }), sep = "\n")
```

```
## Summary Statistics for In-Citations:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   8.688   8.000 620.000
##
## Summary Statistics for Out-Citations:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   5.000   8.752  12.000 369.000
```

The data reveals a striking disparity between in-citations (citations received) and out-citations (references made). For in-citations, the median paper receives just 2 citations, with 25% of papers never cited at all, demonstrating that most research has limited scholarly impact. However, the maximum of 620 in-citations shows how select papers achieve exceptional influence. In contrast, out-citations show a different pattern - while half of papers make 5 or fewer references, the upper quartile cites 12+ sources, with extreme cases reaching 369 references. This contrast highlights that extensive referencing (high out-citations) doesn't necessarily correlate with academic impact (high in-citations).

```r
par(mfrow = c(1,2))  # set plotting area to show 1 row and 2 columns
max_density <- max(density(log(1 + paper_citations$InCitations))$y,
  density(log(1 + paper_citations$OutCitations))$y)  # Get max y-value from density plots
hist(x = log(1 + paper_citations$InCitations),       # histogram for log(in-citations)
     main = "Distribution of log(In-Citations)",     # title of plot
     xlab = "log(In-Citations)", col = "yellow",     # x-axis label, color of bars
     freq = FALSE, ylim = c(0, max_density * 1.05))  # density on y-axis
lines(density(log(1 + paper_citations$InCitations)), col = "red", lwd = 2) # density line
hist(x = log(1 + paper_citations$OutCitations),      # histogram for log(out-citations)
     main = "Distribution of log(Out-Citations)",    # title of plot
     xlab = "log(Out-Citations)", col = "green",     # x-axis label, color of bars
     freq = FALSE, ylim = c(0, max_density * 1.05))  # density on y-axis
lines(density(log(1 + paper_citations$OutCitations)), col = "red", lwd = 2) # density line
```

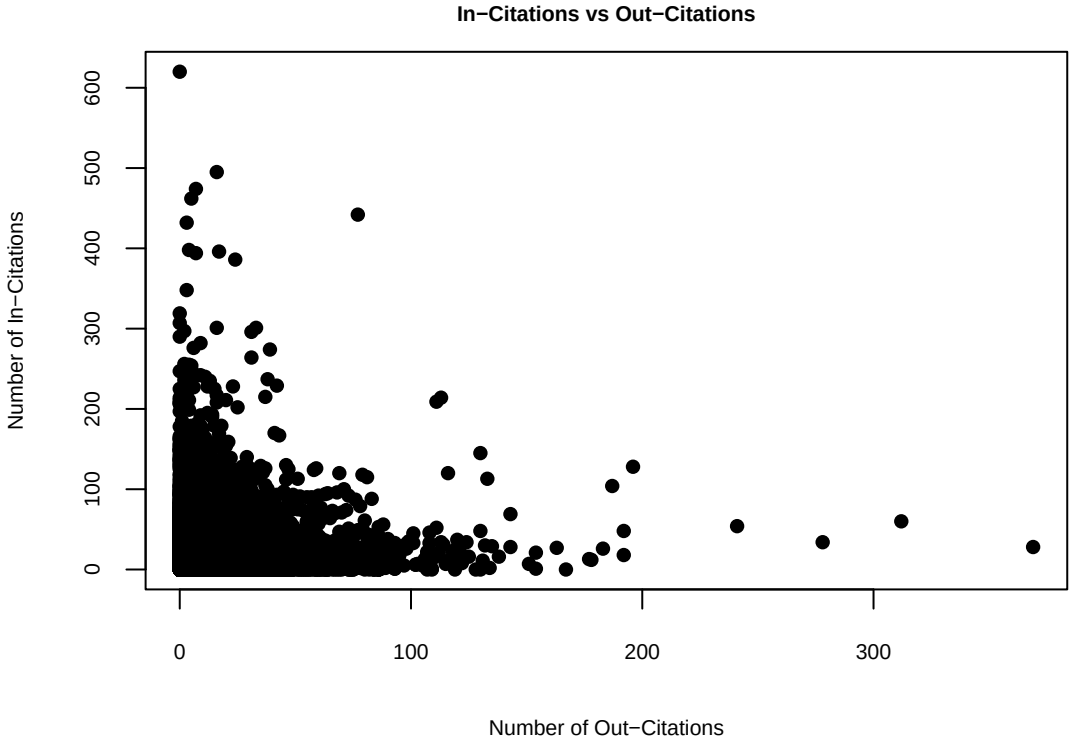


The log plot of in-citations reveals a highly right-skewed distribution, with most papers receiving few or no citations and a few achieving exceptional impact — highlighting extreme inequality in scholarly recognition. In contrast, out-citations, while still right-skewed, show a more balanced, mildly bi-modal distribution, suggesting two common referencing behaviors: minimal and moderate citation. The slower decline in out-citation density indicates more standardized referencing practices, showing that while academic impact is concentrated, referencing is more evenly distributed — aside from a subset of papers with no references.

```r
# Calculate correlation between In-Citations and Out-Citations
correlation <- cor.test(paper_citations$InCitations, paper_citations$OutCitations)
print(paste("Correlation between In- and Out- Citations:", round(correlation$estimate, 2)))
```

```
## [1] "Correlation between In- and Out- Citations: 0.16"
```

```r
# Scatter plot of in-citations vs out-citations
par(mfrow = c(1, 1), mar = c(4, 4, 2, 2), # set layout with margins(below, left, top, right)
    cex.lab = 0.7, cex.axis = 0.7)        # axis label and text size
plot(paper_citations$OutCitations, paper_citations$InCitations,
     main = "In-Citations vs Out-Citations",     # title of plot
     xlab = "Number of Out-Citations",           # x-axis Label
     ylab = "Number of In-Citations",            # y-axis Label
     pch = 16, cex.main = 0.7)                   # point type, title text-size
```



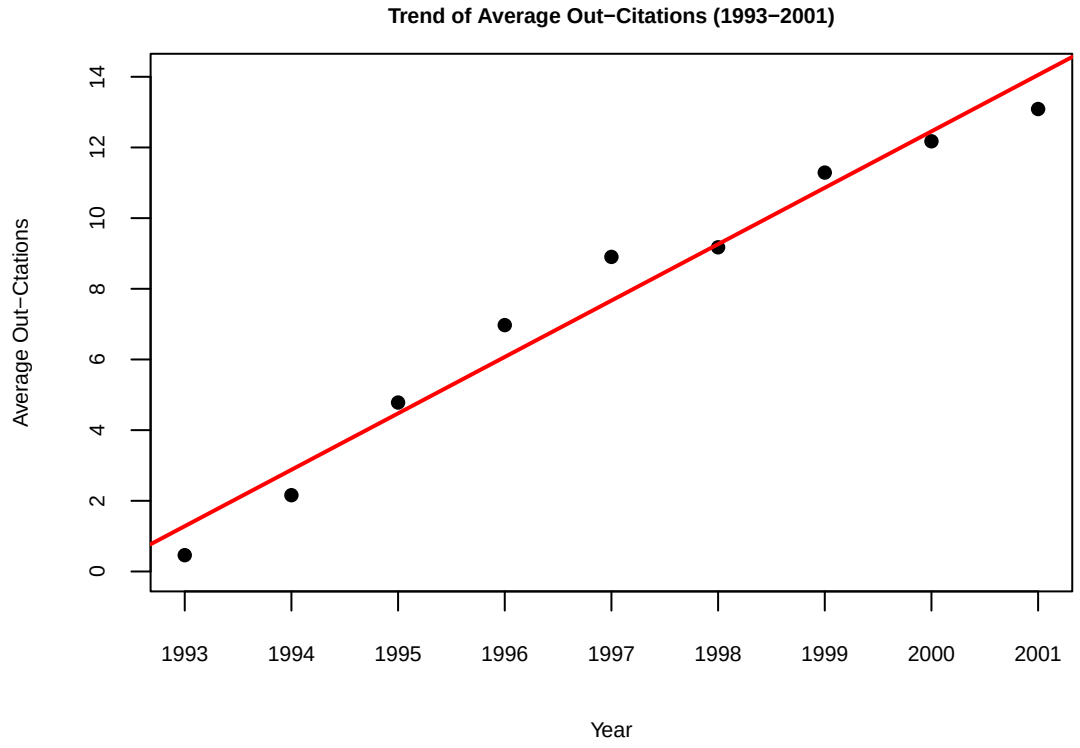
The scatter plot supports the weak correlation correlation coefficient between in-citations and out-citations, suggesting that the number of references a paper cites is not a good predictor of how often it will be cited. The scatter plot illustrates this clearly: most papers cluster in the lower citation ranges, with only a few outliers achieving high citation counts in either dimension. This spread in the data shows that while some papers with many references also receive many citations — possibly because they are well-researched — others become highly cited even though they reference only a few papers. This suggests that factors like originality, topic importance, and authorship may play a more significant role in a paper's impact than the number of references it includes.

Task 2

```r
# Ensure 'date' is in Date format
paper_citations$date <- as.Date(paper_citations$date, format="%m/%d/%Y")
# Extract the Year and store as separate column in the dataframe
paper_citations$Year <- as.numeric(format(paper_citations$date, "%Y"))

# Filter data for years between 1993 and 2001
filtered_citations <- subset(paper_citations, Year >= 1993 & Year <= 2001)

# Compute average out-citations per year
avg_out_citations_per_year <- aggregate(OutCitations~Year, data = filtered_citations, mean)

# Build a simple linear regression model
citation_trend_model <- lm(OutCitations ~ Year, data = avg_out_citations_per_year)
cat(capture.output({
  cat("Gradient (Yearly Change):", round(coef(citation_trend_model)[2], 4), "\n")
  cat("R-squared:", round(summary(citation_trend_model)$r.squared, 4), "\n")}), sep = "\n")
```

```
## Gradient (Yearly Change): 1.5961
## R-squared: 0.9694
```

```r
# Plot the trend
par(mar = c(4, 4, 2, 2),              # margins sizes(bottom, left, top, right)
    cex.lab = 0.7, cex.axis = 0.7)    # axis label and text size
plot(avg_out_citations_per_year$Year, avg_out_citations_per_year$OutCitations,
     main = "Trend of Average Out-Citations (1993-2001)",           # title of plot
     xlab = "Year", ylab = "Average Out-Ctations",                  # x and y axis labels
     pch = 16,                                                      # solid circle points
     cex.main = 0.7,                                                # title size
     xaxt = "n",                                                    # suppress x-axis ticks
     ylim = c(0, max(avg_out_citations_per_year$OutCitations) + 1)) # y-axis limits
axis(1, at = 1993:2001, labels = 1993:2001)                         # x-axis limits
abline(citation_trend_model, col = "red", lwd = 2)  # Add a red-colored regression line
```



The goal of this analysis is to determine whether high-energy physics papers have been citing more references (out-citations) over time. To investigate this, we calculate the average number of references cited by papers each year from 1993 to 2001 and visualize the trend using a simple line of best fit.

With a gradient of 1.5961, our model shows that the average number of references per paper increases by about 1.6 citations per year. This represents substantial growth — in practical terms, it means that every three years, papers cite nearly five more references on average. According to our model, the average paper in 1993 included roughly 0.5 references, increasing to around 13.0 references by 2001 — a dramatic over 20-fold increase in just nine years.

The model has an R-squared value of 0.96, indicating an excellent fit. This means that 96% of the variation in average out-citations can be explained by the passage of time (year). The steady progression of blue dots tracking upward along the red regression line highlights a consistent pattern of growth in referencing behavior. These findings support the claim that researchers are increasingly citing past studies, likely due to an expanding literature base, heightened academic expectations, or shifts in citation practices.

Task 3

```
# Extract the month (from the given dates) and assign numerical values (1 to 12)
paper_citations$Month <- as.numeric(format(paper_citations$date, "%m"))

# Categorize paper citations into two combined seasons
spring_summer_citations <- paper_citations$InCitations[
  paper_citations$Month %in% c(4, 5, 6, 7, 8, 9)]  # Spring and Summer months
autumn_winter_citations <- paper_citations$InCitations[
  paper_citations$Month %in% c(10, 11, 12, 1, 2, 3)]  # Autumn and Winter months
```

Our goal is to determine whether whether the quality of scientific papers—measured by the number of in-citations—varies by season of submission. We extracted publication month from dates and calculated the number of in-citations for papers submitted during spring-summer (April–September) and autumn-winter (October–March) periods. To assess whether the difference is statistically meaningful, we apply the Wilcoxon rank-sum test — a method suitable for comparing two groups when the data may not follow a normal distribution. We define the hypotheses as:

Null Hypothesis (H0): There is no difference in in-citation numbers between papers published in autumn/winter (October–March) and spring/summer (April–September).

Alternative Hypothesis (H1): Papers published in autumn/winter (October–March) have higher in-citation numbers than those published in spring/summer (April–September).

```
# Compute the mean in-citations for each combined season
mean_autumn_winter <- mean(autumn_winter_citations)
mean_spring_summer <- mean(spring_summer_citations)
cat("Average citations per paper:\n",
    "Autumn-Winter (Oct-Mar):", round(mean_autumn_winter, 2), "\n",
    "Spring-Summer (Apr-Sep):", round(mean_spring_summer, 2), "\n")
```

```
## Average citations per paper:
##  Autumn-Winter (Oct-Mar): 8.57
##  Spring-Summer (Apr-Sep): 8.8
```

The small difference in average citations between the Autumn-Winter and Spring-Summer seasons suggests there may be no statistically significant seasonal effect on citation counts. However, formal statistical testing is needed to confirm this observation.

```
# Perform the one-tailed Wilcoxon Test
wilcox_test_result <- wilcox.test(autumn_winter_citations, spring_summer_citations,
                                  alternative = "greater")

# Interpreting the result in context
if (wilcox_test_result$p.value < 0.05) {
  cat("There is a significant evidence that autumn-winter papers receive more citations")
} else {
  cat("There is no significant evidence that autumn-winter papers receive more citations")}
```

```
## There is no significant evidence that autumn-winter papers receive more citations
```

Our analysis finds no significant seasonal effect on citation counts. Papers published in autumn/winter months do not systematically outperform those published in spring/summer. Researchers need not prioritize submission timing based on citation potential. The findings suggest that factors like research novelty, methodology, and author reputation may dominate citation impact over timing of submission.