STAT0004 Take Home Assessment: Physics arXiv paper citations

Group 24

2025-05-01

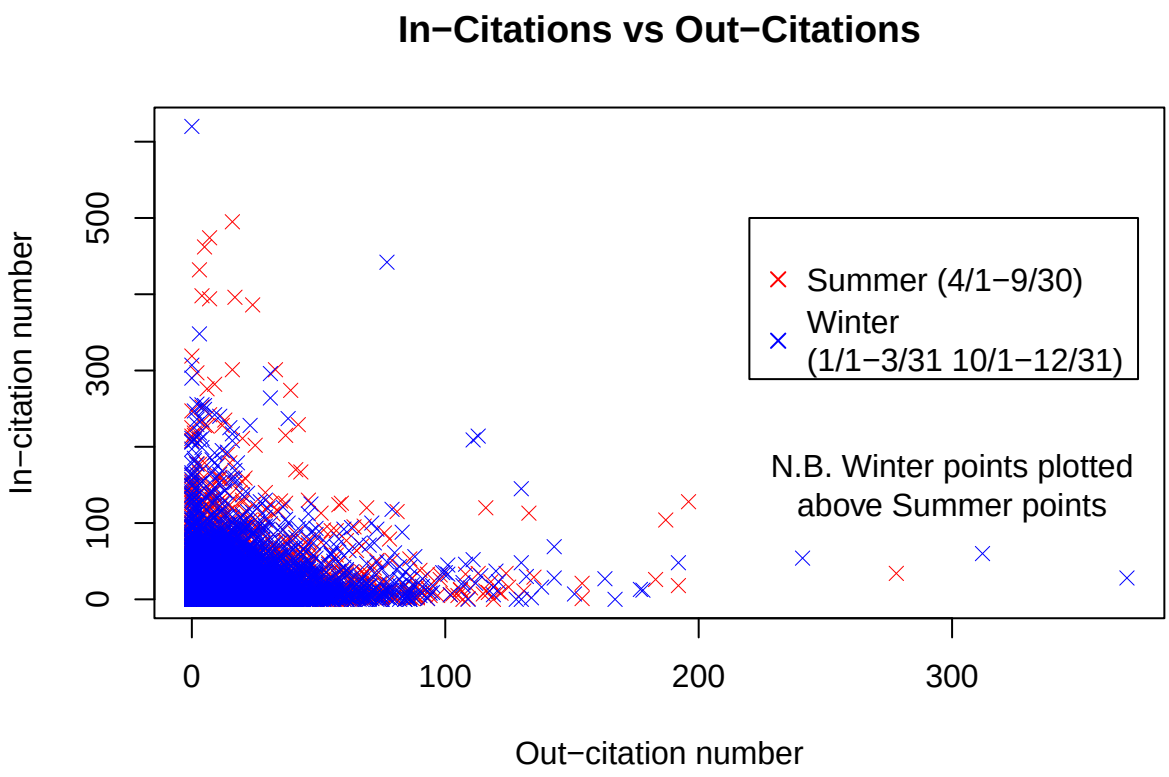
Task 1

Data wrangling and computing citation numbers

```
citations <- read.csv( "citations.csv" )
papers <- read.csv( "papers.csv" )
#calculating a decimal number representation of the dates
days_in_months <- c(31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
days_before_month <- c(0, cumsum(days_in_months)[-12])
find_year_decimal <- function(d){
  #iterates over the string to find the / dividers
  sep <- c(0, 0)
  for (i in 1:nchar(d)){
    if (substr(d, i, i) == "/"){
      if (sep[1] == 0){
        sep[1] <- i
      }else{
        sep[2] <- i
        break
      }
    }
  }
  #converts the sections between the /'s to number data types and appropriately sums them
  Month <- as.numeric( substr(d, 1, sep[1] - 1) )
  Day <- as.numeric( substr(d,(sep[1] + 1), (sep[2] - 1)) )
  Year <- as.numeric( substr(d, (sep[2] + 1), nchar(d)) )
  Year + (days_before_month[Month] + Day - 1)/365
}
papers["year"] <- apply(array(papers$date), MARGIN = 1, FUN = find_year_decimal)
#remove rows with duplicate paperID, Keeps the row with the earliest submission date.
papers <- sort_by.data.frame(papers, ~ list(PaperID, year))
papers <- papers[papers$PaperID != c(-1, papers$PaperID[ -length( papers$PaperID ) ]), ]
#efficiently calculating the in and out citation numbers expressed in "table" data type:
out_citations_table <- table(citations$FromID)
in_citations_table <- table(citations$ToID)
#converts citation number tables to vectors
papers["out_num"] <- as.vector( out_citations_table[ as.character( papers$PaperID ) ] )
papers["in_num"] <- as.vector( in_citations_table[ as.character( papers$PaperID ) ] )
#if a PaperID has no citations, it will give NA's. These need to be replaced with zeros:
papers[is.na(papers$out_num), "out_num"] <- 0
papers[is.na(papers$in_num), "in_num"] <- 0
```

Visualising the data

```
#cut-off dates for Summer/Winter as decimal
April1 <- find_year_decimal("4/1/0")
September30 <- find_year_decimal("9/30/0")
#separates the citation numbers data into 2 data frames for Winter months & Summer months.
#For the sake of brevity, "Winter" includes Autumn and "Summer" includes Spring
out_in_summer <- papers[(papers$year %% 1 >= April1) & (papers$year %% 1 <= September30),
                        c("out_num", "in_num")]
out_in_winter <- papers[!((papers$year %% 1 > April1) & (papers$year %% 1 < September30)),
                        c("out_num", "in_num")]
plot(out_in_summer, pch = 4, lwd = 0.5, col = 'red',
     xlim = c(0, max(papers$out_num)), ylim = c(0, max(papers$in_num)),
     main = "In-Citations vs Out-Citations",
     xlab = "Out-citation number", ylab = "In-citation number")
points(out_in_winter, pch = 4, col = 'blue', lwd = 0.5)
legend(x = 220, y= 500, legend = c("Summer (4/1-9/30)", "Winter\n(1/1-3/31 10/1-12/31)"),
       col = c('red', 'blue'), pch = c(4,4))
text(x = 300, y = 150, labels = "N.B. Winter points plotted\nabove Summer points")
```



According to the scatter plot, most data points are densely concentrated in the lower-left corner, showing that the majority of papers cite only a few others and get only a few citations. A small number of points extend horizontally and vertically, representing papers that either cite many works but receive few citations, or are heavily cited despite referencing few others. These extreme points might indicate a potential negative non-linear association but further analysis is needed to confirm the true relationship. Summer and winter

data points largely overlap in distribution, with only a small cluster of highly cited summer papers appearing in the upper-left area, which needs further analysis to determine significance.

Exploring the data with descriptive statistics and correlation

```
options(digits = 4)  #sets digit-width of display
out_summary <- summary(papers$out_num, quantile.type = 1)
out_upper_tail <- quantile(papers$out_num, quantile.type = 1, probs = c(0.9, 0.95, 0.99))
c( out_summary[-length(out_summary)], out_upper_tail, out_summary[length(out_summary)] )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.     90%     95%     99%    Max.
##   0.000   0.000   5.000   8.904  12.000  23.000  32.000  58.000 369.000
```

```
in_summary <- summary(papers$in_num, quantile.type = 1)
in_upper_tail <- quantile(papers$in_num, quantile.type = 1, probs = c(0.9, 0.95, 0.99))
c( in_summary[ -length(in_summary) ], in_upper_tail, in_summary[ length(in_summary) ] )
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.     90%     95%     99%    Max.
##   0.000   0.000   2.000   8.904   9.000  24.000  40.000  92.000 620.000
```

Both distributions display positive skew based on the quartiles and extremes. The quantiles up to the 90th percentile in the summaries are very similar which demonstrates similar distributions around the centres. The quantiles in the top 10% suggest a trend in the shapes of the 2 tails increasingly diverging, with in-citations concentrated into the very best papers.

```
in_out_spearman <- cor(papers$out_num, papers$in_num, method = "spearman")
in_out_spearman
```

```
## [1] 0.4297
```

```
#p-value for a 2 tailed test of the spearman rank correlation against the null hypothesis
#of independence:
2 * pnorm( atanh( in_out_spearman ), mean = 0, sd = 1/sqrt( length( papers$out_num ) - 3),
           lower.tail = FALSE)
```

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

```
#Recalculating correlation removing the (0, 0) data points
papers_no_00 <- papers[!((papers$out_num == 0) & (papers$in_num == 0)), ]
in_out_spearman_no_00 <- cor(papers_no_00$out_num, papers_no_00$in_num, method = "spear")
in_out_spearman_no_00
```

```
## [1] 0.1373
```

```
2 * pnorm( atanh( in_out_spearman_no_00 ), mean = 0,
           sd = 1/sqrt( length( papers_no_00$out_num ) - 3), lower.tail = FALSE)
```

```
## [1] 1.136e-122
```

The in and out citation numbers exhibit weak rank correlation (0.430), but with very high statistical confidence. After excluding the papers with no in or out citations from the analysis, the correlation is much lower (0.137) but still with very high statistical confidence. This demonstrates that the correlation was primarily due to the high density of data points at (0, 0) rather than a consistent trend.

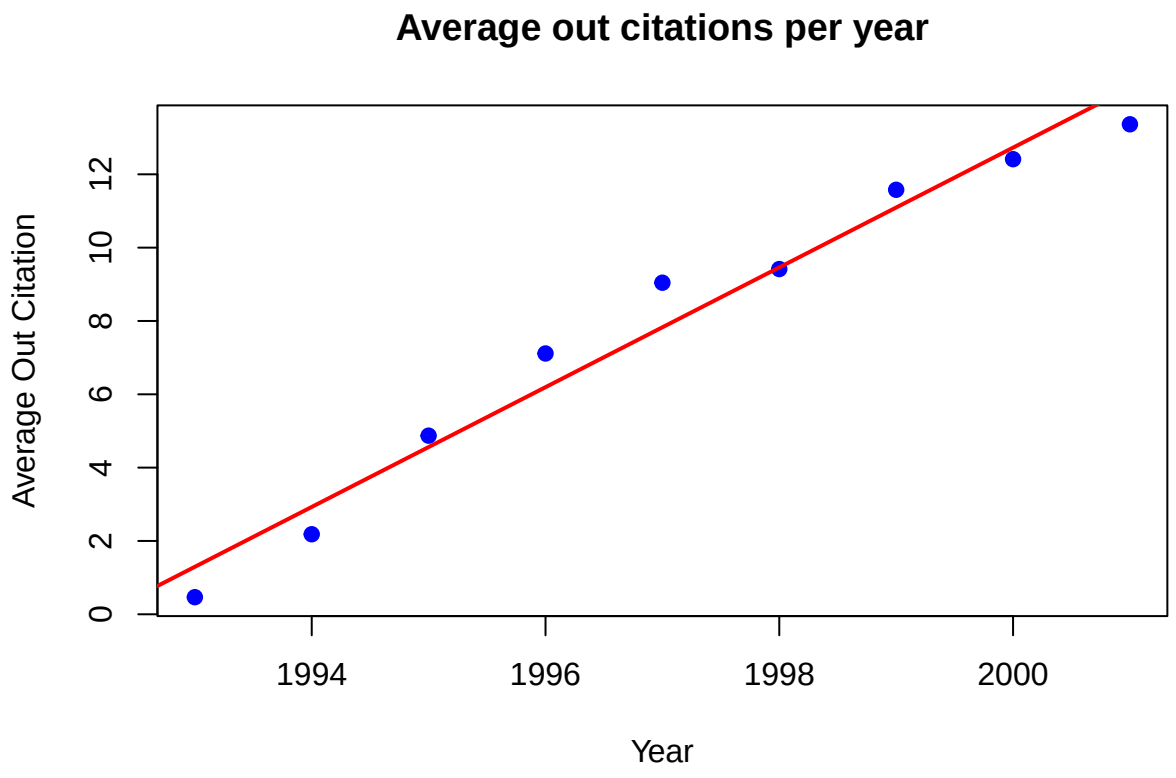
Task 2

Data Wrangling

```
# Range of years for the for loop
unique_years <- floor(min(papers$year)):floor(max(papers$year))
# Calculate the annual mean out citations for all years with papers
avg_out <- c()
for (y in unique_years) {
  avg_out <- c(avg_out, mean(papers[(papers$year >= y) & (papers$year < y+1), "out_num"]))
}
```

Linear Modelling

```
# Construct a simple linear model for annual mean out citations and time
mean_out_citation_lm <- lm(formula = avg_out ~ unique_years)
# Plot annual mean out citations with regression line
plot(unique_years, avg_out, pch = 19, col = "blue",
  main = "Average out citations per year", xlab = "Year", ylab = "Average Out Citation")
abline(mean_out_citation_lm, col = "red", lwd = 2)
```



A visual inspection of the data plotted with the linear model suggests a positive linear trend between mean annual out-citations and time, implying that bibliography length has increased. This assumes out-citations are a good proxy for length, though confounding factors like changes in citation styles could affect this.

Task 3

Statistical Tests for Seasonal Citation Difference

```
# Perform Welch's t-test (unequal variance t-test)
welch_test <- t.test(out_in_winter$in_num, out_in_summer$in_num,
                     alternative = "g",  # Tests winter > summer
                     var.equal = FALSE)        # Force Welch adjustment

# Perform Mann-Whitney U-test (non-parametric)
mw_test <- wilcox.test(out_in_winter$in_num, out_in_summer$in_num,
                       alternative = "g",  # Tests winter > summer
                       exact = FALSE,             # Required for large samples
                       correct = TRUE)           # Apply continuity correction

# Display results
print(welch_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  out_in_winter$in_num and out_in_summer$in_num
## t = -0.68, df = 35901, p-value = 0.8
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.5051     Inf
## sample estimates:
## mean of x mean of y
##     8.830     8.977
```

```
print(mw_test)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  out_in_winter$in_num and out_in_summer$in_num
## W = 1.6e+08, p-value = 1
## alternative hypothesis: true location shift is greater than 0
```

We conducted statistical analyses using both the Welch t-test and the Wilcoxon test at the 5% level, with null hypothesis that the sampling distributions of in-citations for Winter and Summer papers have the same distribution and alternative hypothesis that Winter is higher than Summer. The data has positive skew implying non-normality, so the non-parametric Mann-Whitney test is probably the most valid test to use. We still used the t-test to compare means since the test is robust against non-normality for large sample sizes despite the assumption of normality. The results of the Welch t-test indicate that contrary to the proposed alternative, the mean citation count in Summer (8.977) is higher than that in Winter (8.830). This gave a p-value of 0.8 which is not significant. This indicates that the test has failed to reject the null hypothesis. The equivalent Wilcoxon test gives a p-value of ~1 which gives the same conclusion. In fact, the Wilcoxon test statistic very strongly supports the notion that the Winter distribution position is significantly lower than that of Summer. Overall, neither the Welch t-test nor the Wilcoxon test supports the physicist's claim that the quality of papers written in Autumn or Winter is higher than in Spring or Summer.