Statistics 3.1.2 Exercise 4

Literary Style of Different Authors

Lucas Liona

Table of contents

Problem Description

The literary style of different authors differs widely from one to another, and it is possible to measure these differences statistically. By style we mean those aspects of writing that might be independent of the subject matter; for example, the lengths of words and sentences or the frequencies of different words. This exercise analyzes word frequencies in two books: *Pride and Prejudice* by Jane Austen and *Spy Hook* by Len Deighton.

Data Loading and Preparation

```
# Load required libraries
library(ggplot2)
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# Read the data
book_data <- read.table("book.txt", header = TRUE)

# Check the structure of our data
str(book_data)
```

```
'data.frame':   36 obs. of  3 variables:
 $ pcto : num  2.03 2.24 2.37 1.8 1.79 1.58 1.7 2.55 1.68 3.73 ...
 $ pcthe: num  5.29 3.06 3.96 3.51 2.95 2.05 3.39 3.9 4.11 2.55 ...
 $ book : chr  "pride" "pride" "pride" "pride" ...
```

```r
# View summary statistics
summary(book_data)
```

```
      pcto           pcthe           book
 Min.   :1.130   Min.   :2.050   Length:36
 1st Qu.:1.798   1st Qu.:3.042   Class :character
 Median :2.425   Median :3.905   Mode  :character
 Mean   :2.428   Mean   :4.268
 3rd Qu.:2.920   3rd Qu.:5.020
 Max.   :4.270   Max.   :8.390
```

```r
# Create a nicer looking table of the data
knitr::kable(book_data, caption = "Percentage of 'to' and 'the' in Each Book")
```

Table 1: Percentage of ‘to’ and ‘the’ in Each Book

pcto	pcthe	book
2.03	5.29	pride
2.24	3.06	pride
2.37	3.96	pride

pcto	pcthe	book
1.80	3.51	pride
1.79	2.95	pride
1.58	2.05	pride
1.70	3.39	pride
2.55	3.90	pride
1.68	4.11	pride
3.73	2.55	pride
1.24	2.66	pride
3.07	4.11	pride
2.77	8.39	pride
1.91	2.97	pride
2.08	2.54	pride
1.13	3.67	pride
2.07	2.99	pride
3.42	3.71	spyhook
3.78	4.56	spyhook
1.46	6.42	spyhook
1.81	4.16	spyhook
2.56	4.31	spyhook
4.27	4.51	spyhook
2.92	3.70	spyhook
3.85	3.59	spyhook
2.88	3.91	spyhook
1.95	5.62	spyhook
1.76	7.12	spyhook
2.92	6.90	spyhook
2.48	2.28	spyhook
2.75	3.87	spyhook
2.73	5.94	spyhook
2.53	2.86	spyhook
2.97	7.65	spyhook
1.36	5.51	spyhook
3.28	4.93	spyhook

(a) Scatter Plots

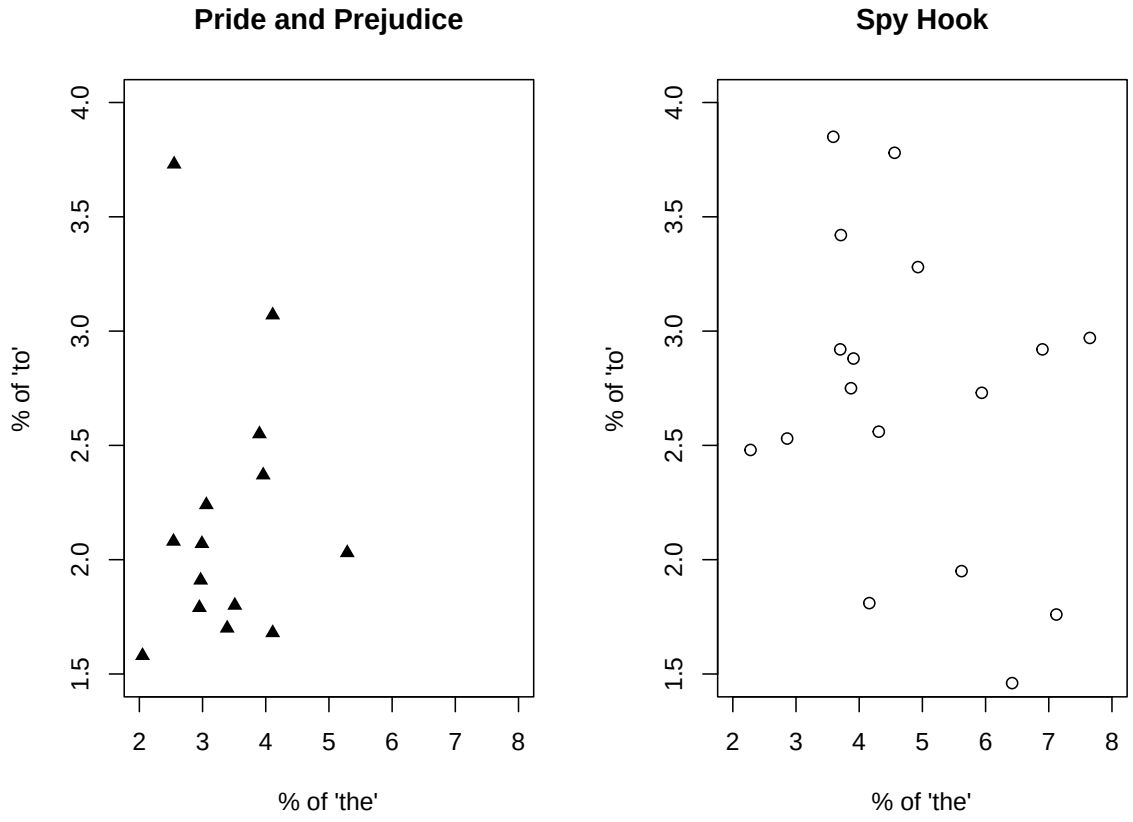
Let's create scatter plots of percentage of "to" versus percentage of "the" for each book separately:

```r
# Separate the data by book
pride_data <- book_data[book_data$book == "pride", ]
spyhook_data <- book_data[book_data$book == "spyhook", ]

# Create individual plots
par(mfrow = c(1, 2))

# Plot for Pride and Prejudice
plot(pride_data$pcthe, pride_data$pcto,
     main = "Pride and Prejudice",
     xlab = "% of 'the'",
     ylab = "% of 'to'",
     pch = 17,
     xlim = c(2, 8),
     ylim = c(1.5, 4.0))

# Plot for Spy Hook
plot(spyhook_data$pcthe, spyhook_data$pcto,
     main = "Spy Hook",
     xlab = "% of 'the'",
     ylab = "% of 'to'",
     pch = 1,
     xlim = c(2, 8),
     ylim = c(1.5, 4.0))
```



(b) Features of the Plots

Looking at the individual plots, several interesting features emerge:

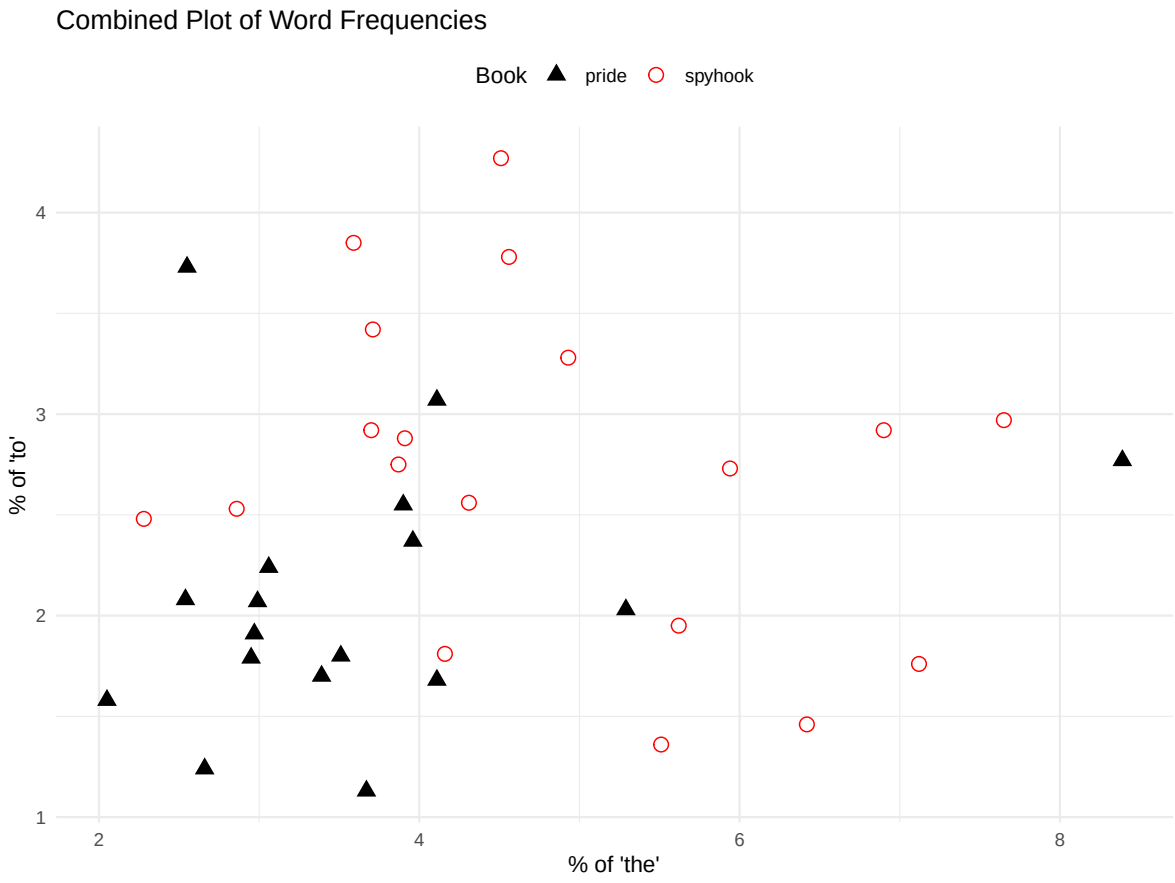
1. **Pride and Prejudice**: The plot shows a strong positive relationship between the percentage of "to" and the percentage of "the". Most points cluster together, suggesting consistent word usage patterns, with two points appearing as outliers (visible in the upper-left portion of the plot).
2. **Spy Hook**: The relationship appears weaker and possibly nonlinear, with an initial downward trend that curves upward. There's more scatter in the data points, suggesting greater variability in word usage patterns across different pages.

The outliers in the Pride and Prejudice plot warrant further investigation as they may represent pages with unusual characteristics.

(c) Combined Plot

Now let's create a combined scatter plot using different plotting symbols for each book:

```
# Create a combined plot
ggplot(book_data, aes(x = pcthe, y = pcto, shape = book, color = book)) +
  geom_point(size = 3) +
  scale_shape_manual(values = c("pride" = 17, "spyhook" = 1)) +
  scale_color_manual(values = c("pride" = "black", "spyhook" = "red")) +
  labs(title = "Combined Plot of Word Frequencies",
       x = "% of 'the'",
       y = "% of 'to'",
       color = "Book",
       shape = "Book") +
  theme_minimal() +
  theme(legend.position = "top")
```



The combined plot reveals a clear separation between the two books:

- **Pride and Prejudice** points tend to cluster in the lower-left region (fewer "to"s and "the"s).
- **Spy Hook** points generally appear in the upper-right region (more "to"s and "the"s).

This pattern suggests that these word frequencies could effectively discriminate between the two books.

(d) Differentiation Between Books

Yes, it is possible to differentiate between *Pride and Prejudice* and *Spy Hook* using these data. The combined plot shows that:

1. The two books occupy largely separate regions of the plot
2. Pride and Prejudice tends to use both "to" and "the" less frequently than Spy Hook
3. With only a few exceptions, you could draw a diagonal line that would separate most points from each book

This suggests that word frequency analysis can be an effective tool for authorship attribution. The differences likely reflect distinct writing styles - Jane Austen's more formal 19th-century prose versus Len Deighton's more contemporary thriller style.

Conclusions

The analysis demonstrates that even simple word frequency measurements can reveal significant differences in literary style. The distinct clustering patterns between the two books suggest that:

1. Authors have consistent word usage patterns within their works
2. These patterns differ measurably between authors
3. Such statistical approaches could be extended to broader authorship attribution problems

These findings support the use of computational stylistics in literary analysis and forensic linguistics.