Letters used in Pride and Prejudice*

Sindhu Priya Mallavarapu Jena Shah

March 11, 2024

This paper explores the distribution of the letter 'e' in Jane Austen's Pride and Prejudice to understand if its frequency of occurrence correlates with the amount of words in the text. The methodology involves counting the occurrences of the letter 'e' in the first ten lines of each chapter gathered from the dataset obtained from Project Gutenberg. We found that the frequency of the letter "e" increases as the amount of words in the text increases. This analysis can be of interest to linguists as it can help them to identify patterns and characteristics of the English language, better understand the writing style of Jane Austen, and understand the evolution of the English language over time.

1 Introduction

You can and should cross-reference sections and sub-sections.

The remainder of this paper is structured as follows. Section 2….

2 Data

Some of our data is of (**?@fig-bills**), from (**palmerpenguins?**).

Talk more about it.

And also (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Should this go in results or data?

Talk way more about it.

*Code and data are available at: https://github.com/MSindhuPriya/letters_jane_austen

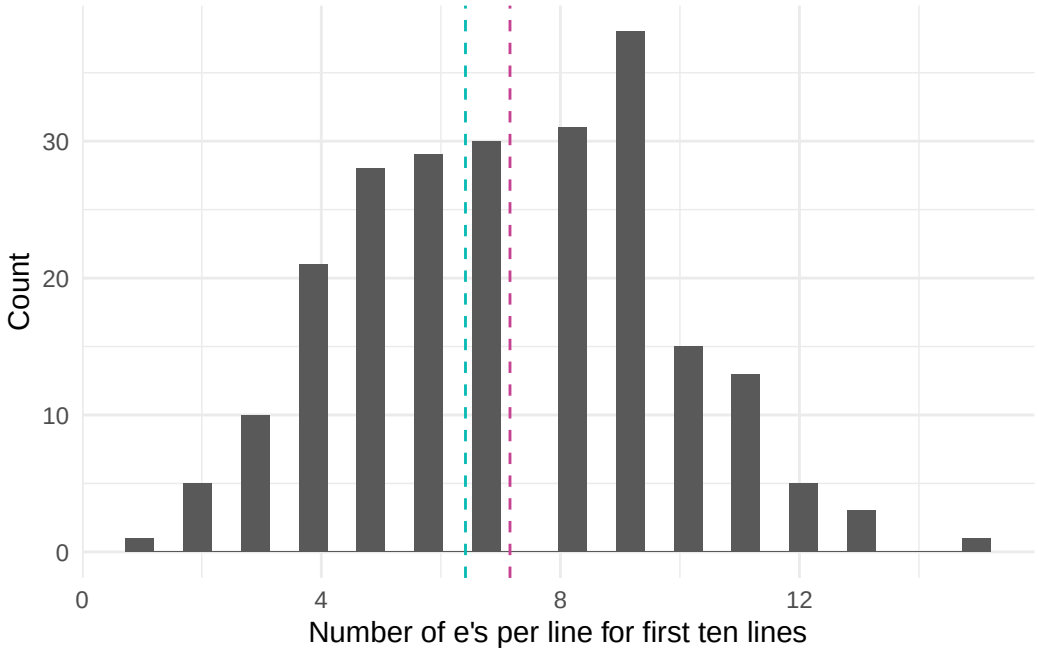


Figure 1: Frequency of the letter "e" in the first line of each chapter

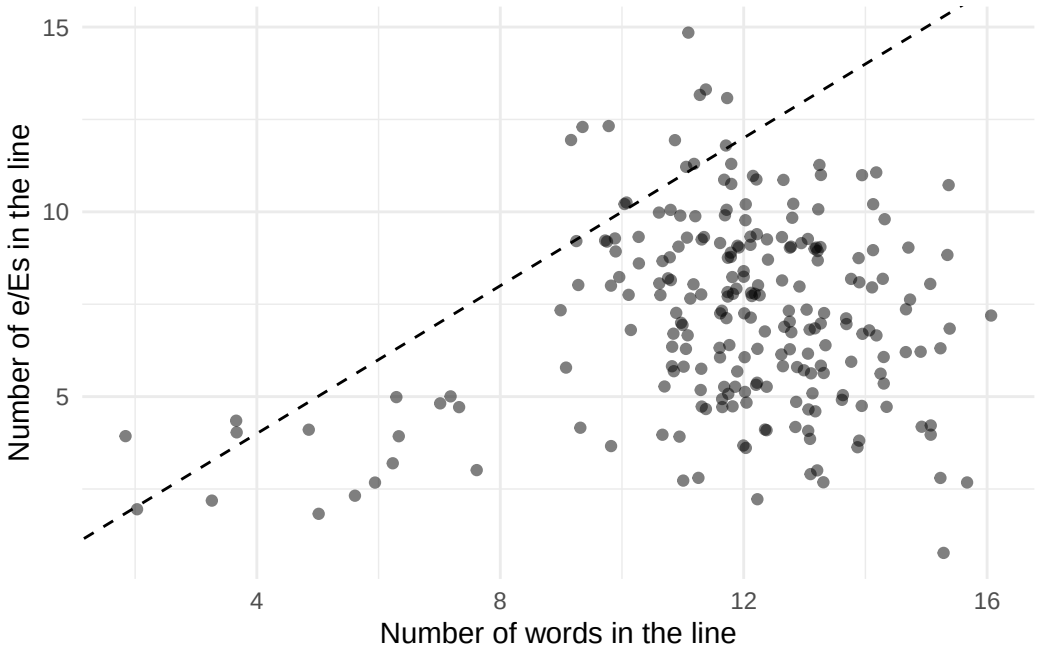


Figure 2: Frequency of the letter "e" in the first line of each chapter

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\lambda \sim \text{Poisson}(\lambda_i) \tag{1}$$
$$log(\lambda_i) = \beta_0 + \beta_i \times \text{Number of words}_i \tag{2}$$
$$\beta_0 \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta_1 \sim \text{Normal}(0, 2.5) \tag{4}$$

We run the model in R (R Core Team 2022) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in Table 1.

Table 1: Explanatory models of the number of 'E/e's seen based on the number of words on the line

	Model
(Intercept)	1.63
	(0.14)
word_count	0.03
	(0.01)
Num.Obs.	230
Log.Lik.	−536.945
ELPD	−538.8
ELPD s.e.	8.9
LOOIC	1077.7
LOOIC s.e.	17.8
WAIC	1077.7
RMSE	2.50

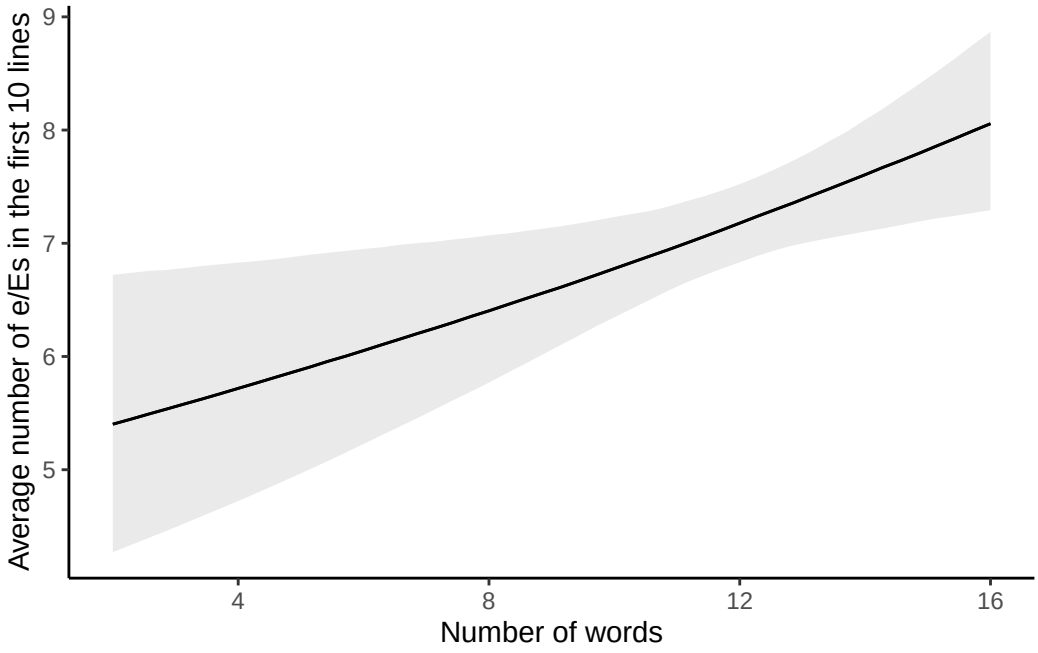


Figure 3: The number of 'E/e's predicted by the model for the first 10 lines

5 Discussion

5.1 First discussion point

In this study, we extracted the first ten lines of each chapter and counted the occurrences of "e" or "E" in each unit. Our goal was to investigate if there is a correlation between the frequency of "e" and the length of texts, which could provide insights into the distribution of this letter in English literature.

Our analysis revealed that the frequency of the letter "e" does/does not consistently increase as more words are used in the first ten lines of each chapter. This suggests that the distribution of "e" in Pride and Prejudice is/is not uniform and (may vary based on factors like word choice, sentence structure, and the literary style). Our findings add to the understanding of the linguistic characteristics of the novel as well as the distribution of letters in different texts.

Furthermore, we attempt (?? word choice) to highlight the importance of considering methodological approaches in linguistic research. By adapting a methodology from a study done on ancient poetry to analyze a work of fiction from the 19th century, we demonstrate the flexibility and applicability of quantitative methods in studying language and literature. This approach allows us to gain newer insights into familiar texts and opens up ways for future research in literature and linguistics.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

One weakness of our study is the limited scope of analysis, focusing only on the first ten lines of each chapter of Pride and Prejudice. While this approach provides us with a manageable dataset, it may not capture the full complexity of the entire novel's linguistic features. Additionally, our analysis does not consider contextual factors such as dialogue, character names, or punctuation, which could influence the distribution of "e" in the text.

5.5 Future

Future research could expand on our analysis by considering a broader range of textual features and linguistic elements. For example, studying the distribution of other letters or letter combinations could provide a more comprehensive understanding of the novel's language. Additionally, exploring the relationship between letter frequency and literary devices [cite something here] such as rhyme, meter, or alliteration could offer deeper insights into Austen's writing

style and the broader context of 19th-century English literature. Overall, our study lays the foundation for further investigations into literature, linguistics, and digital humanities, which highlights the great potential for interdisciplinary research in this field.

Appendix

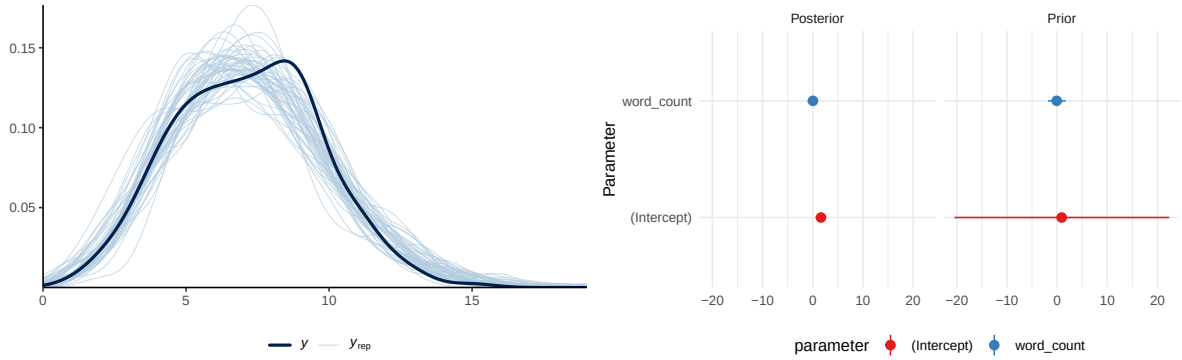
A Additional data details

B Model details

B.1 Posterior predictive check

In Figure 4a we implement a posterior predictive check. This shows...

In Figure 4b we compare the posterior with the prior. This shows...



(a) Posterior prediction check

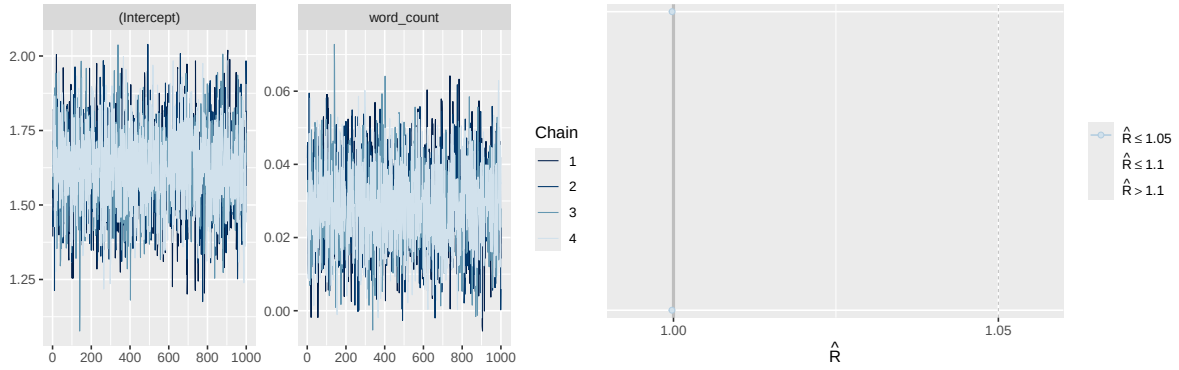
(b) Comparing the posterior with the prior

Figure 4: Examining how the model fits, and is affected by, the data

B.2 Diagnostics

Figure 5a is a trace plot. It shows... This suggests...

Figure 5b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.