Examining the Frequency of ‘E’: A Quantitative Analysis of Pride and Prejudice*

Sindhu Priya Mallavarapu Jena Shah

March 16, 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

In the 1830s, Samuel Morse, inventor of the telegraph and Morse code, conducted research on the frequency of each letter in printed materials. This revealed that certain letters, such as ‘e’, were more prevalent in written English than others. This formed the basis for Morse’s development of Morse code by assigning the most used letters to the shortest and simplest codes for the optimization of message transmissions of the telegraph (Notre Dame 2023). Other research has also shown that the letter ‘e’ is most used in English text (Robert L. Solso 1976). Rohan Alexander’s work in ‘Telling Stories with Data’ examined the frequency of ‘e’ was studied in the infamous ‘Jane Eyre’ by Charlotte Bronte, concluding that the letter ‘e’ increases with the words per line (Alexander 2023). His work showed that studying the frequencies of English letters is applicable in fields besides telecommunications as it can be used in analysing literature. This helps with understanding the linguistics of English used in various texts.

Influenced by Alexander’s work, we study the letter ‘e’ in yet another literary favourite – Jane Austen’s ‘Pride and Prejudice’ (Austen 1813). We do this by using data from the `gutenbergr`

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

(Johnston and Robinson 2023) package in R (R Core Team 2022) and using Poisson regression to draw conclusions. Through examining the first ten lines of each chapter, we use that as a representative sample of the entire text. The estimand for this study is the expected frequency of the letter 'e' in each line of text, with the expectation that it will increase with the number of words per line. This captures the relationship between line length (in terms of words) and the frequency of 'e', providing the distribution of that letter within the text. Through this study, we hope to contribute to the intersection of language, literature, and data analysis.

The structure of our paper ahead consists of three parts. We start with the discussion of the data used in our study in Section 2. This includes elements from data collection, data cleaning, and measurement. Further, we discuss our model in Section 3 Later, we share the results of our study in Section 4 and the process of analysis. We came to a conclusion that the frequency of the letter 'e' increases as the number of words per line increases. Finally, we discuss what the results indicate, limitations, and improvements we can make in Section 5. The entirety of this study is done through R (R Core Team 2022) and its packages – which, along with other studies we use in our work, are mentioned in the bibliography.

2 Data

2.1 Overview

The Pride and Prejudice text that was used in this paper was obtained from Project Gutenberg using the `gutenbergr` (Johnston and Robinson 2023) package. The analysis in this paper is done through R (R Core Team 2022) using numerous packages such as `tidyverse` (Wickham et al. 2019), `marginaleffects` (Arel-Bundock 2024), `arrow` (Richardson et al. 2024), `rstanarm` (Goodrich et al. 2022), and `ggplot2` (Wickham 2016). Moreover, in order to assist in this analysis we used some code found in the book 'Telling Stories with Data' (Alexander 2023) to create a model and to create figures in this paper.

Multiple versions of the text Pride and Prejudice were found in the `gutenbergr` package, however, most of them were either archived and therefore inaccessible via R or their access was blocked for other unknown reasons. There was only one version of this text that was available for download at the time of the analysis, and thus that was the version of the text that was used. This downloaded text also only had the first 23 chapters of Pride and Prejudice as opposed to the original 61.

2.2 Measurement

Several decisions were made with regards to the measurement of how many times we see 'e'. First, all the letters in the text were converted to lowercase so that all capital 'e's were counted along with the lowercase 'e's. This was done to ensure that we did not miss the counting of any

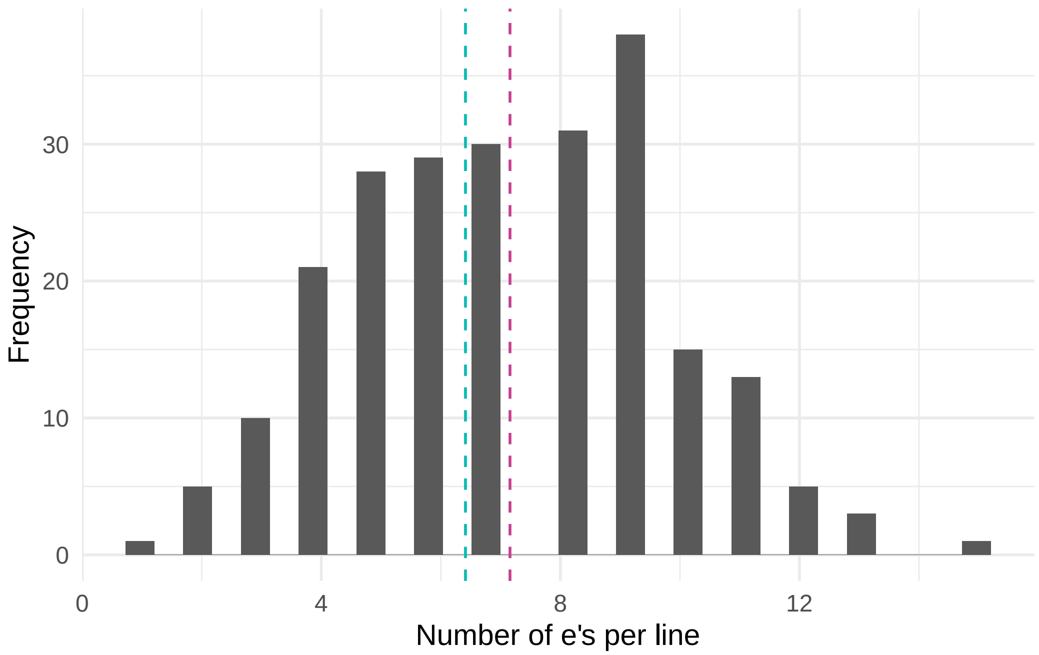


Figure 1: Frequency of the letter “e” in the first ten lines of each chapter

’e’s due to their case. The text was then processed to extract the first ten lines of each chapter, excluding any additional text such as chapter titles or introductory paragraphs. Moreover, special characters, punctuation, and whitespaces were removed. Further, we decided to only count the number of ’e’s seen in the first ten lines of each chapter, excluding the title of the chapter. These decisions were made to ensure that we could only count the frequency of the letter ’e’ from words that were written in the main body of the original story. The number of ‘e’s that occurred beyond the first ten lines of each chapter were not counted. This was done to ensure that the dataset was not too large so that the model could run on the dataset in a reasonable amount of time. It is also important to note that the data available on `gutenbergr` only included the first 23 chapters in Pride and Prejudice whereas the original text consisted of 61 chapters. Thus any occurances of the letter ’e’ outside of these first 23 chapters were not counted. This resulted in a total of 230 observations and the final cleaned analysis dataset consisted of 4 variables. Further discussion of the variables will be done in Section 2.3.

2.3 Description of Variables

Within our cleaned dataset, there are four main variables - `text`, `chapter`, `count_e`, and `word_count`.The actual string of text found in the lines is represented in the `text` variable, `chapter` indicates a string of the chapter in which the text is found. However, the variables that are used in our analysis are `count_e`, an integer that represents the occurrences of the letter ‘e’ per line, and `word_count` an integer variable representing the number of words in the line. As seen in Figure 1 the letter ‘e’ appears in a single line anywhere from 0 to 33 times and the number of words in a line range from 0 to 15.

3 Model

The goal of our modelling strategy is to use Poisson distribution to model the frequency of the letter ‘e’ in each line of “Pride and Prejudice”. Poisson distributions are commonly used to model the number of events occurring in a fixed interval of time or space, given the average rate of occurrence of the event.

In our study, we treated each line of text as a separate interval, with the number of ‘e’ occurrences in each line following a Poisson distribution. The Poisson distribution is characterized by a single parameter, , which represents the average rate of occurrence of the event. In our case, corresponds to the average frequency of ‘e’ in the text for the amount of words per line.

3.1 Assumptions

The Poisson distribution assumes that the events occur independently and at a constant average rate. In the context of our study, this assumption implies that the occurrence of the letter ‘e’ in one line is independent of its occurrence in other lines, and that the average frequency of ‘e’ remains constant across all lines. We see that these assumptions hold as the the number of occurances of ‘e’ in one line does not impact how many times ‘e’ occurs on another line.

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

3.2 Model set-up

We define y_i as the number of ‘e’s that are found in the line and is the mean number of ’e’s that occur on the line. To implement the Poisson distribution model, we calculated the average frequency of ’e’ in the text per line (depending on the amount of words per line). We then used this value as the parameter for the Poisson distribution to model the number of ‘e’ occurrences in each line as seen below

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

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`. Th

3.2.1 Model justification

We expect a positive relationship between the number of 'e's and the number of words in a line. To validate the Poisson distribution model, we compared the predicted frequency of 'e' for each line with the actual observed frequency. We found that the model accurately predicted the frequency of 'e' in the majority of lines, providing further support for its justification in modeling the letter frequency in the text.

Thus, the Poisson distribution model captures the frequency of the letter 'e' in "Pride and Prejudice", providing insights into the distribution of letters within the text. This model contributes to our understanding of the text and the English language in classic literature, highlighting the patterns of letter frequency and textual structure in the work of Jane Austen.

4 Results

The analysis focused on the frequency of the letter 'e' in the first ten lines of each chapter of "Pride and Prejudice". A total of 23 chapters were included in the analysis, resulting in a dataset of 230 lines.

The frequency of the letter 'e' varied across the lines, ranging from 0 to 33 occurrences per line. The average frequency of 'e' across all lines was 7 occurrences per line. To investigate the relationship between the number of words per line and the frequency of the letter 'e', a scatter plot was created Figure 2. Each point on the plot represents a line from the text, with the x-axis indicating the number of words in the line and the y-axis indicating the frequency of the letter 'e'.

Our results are summarized in Table 1.

The scatter plot Figure 2 shows a clear trend of increasing frequency of 'e' with a higher number of words per line. We created a best fit line for the scatter plot which indicated a positive relation between the number of words and the occurrences of the letter 'e'.

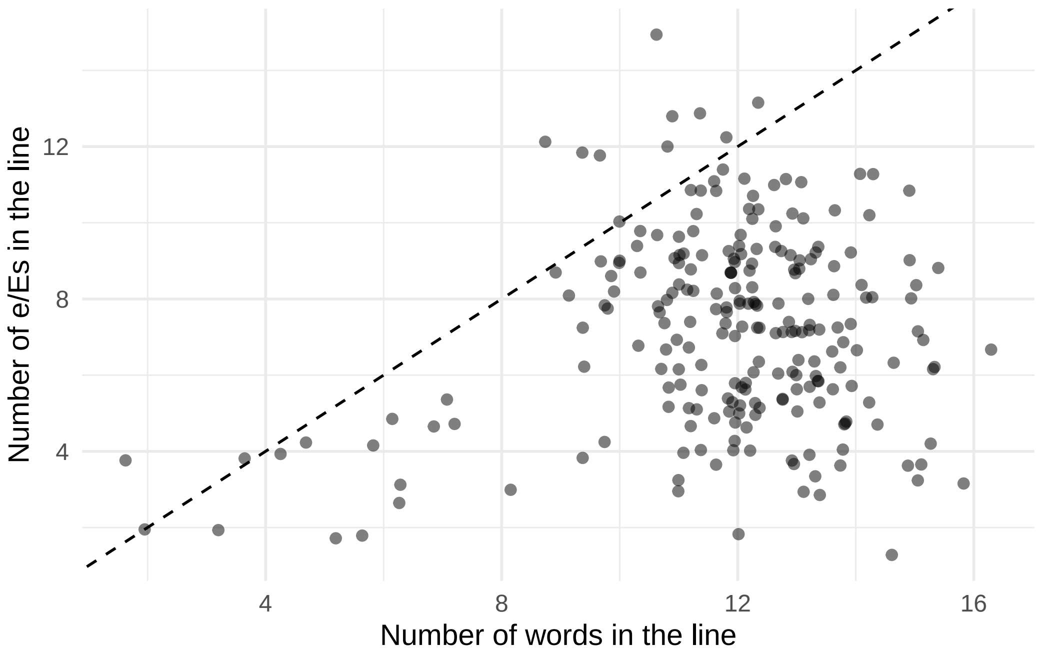


Figure 2: Frequency of the letter “e” depending on the number of words in a line

4.1 Poisson Distribution Analysis

The Poisson distribution was used to model the frequency of the letter ‘e’ in each line as shown in Figure 3. The model acts according to expectation that there is a positive relationship between the number of ‘e’s seen in a line and the number of words in the line. We see from Figure 3 also that the model most accurately predicts the number of ’e’s when there are 12 words in a line. We also see from Table 1 that the intercept is 1.63. This also represents a positive relationship between the count of ’e’ and the count of words in a line.

5 Discussion

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’ increases 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 uniform and (may vary based on factors like word choice, sentence structure, and

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

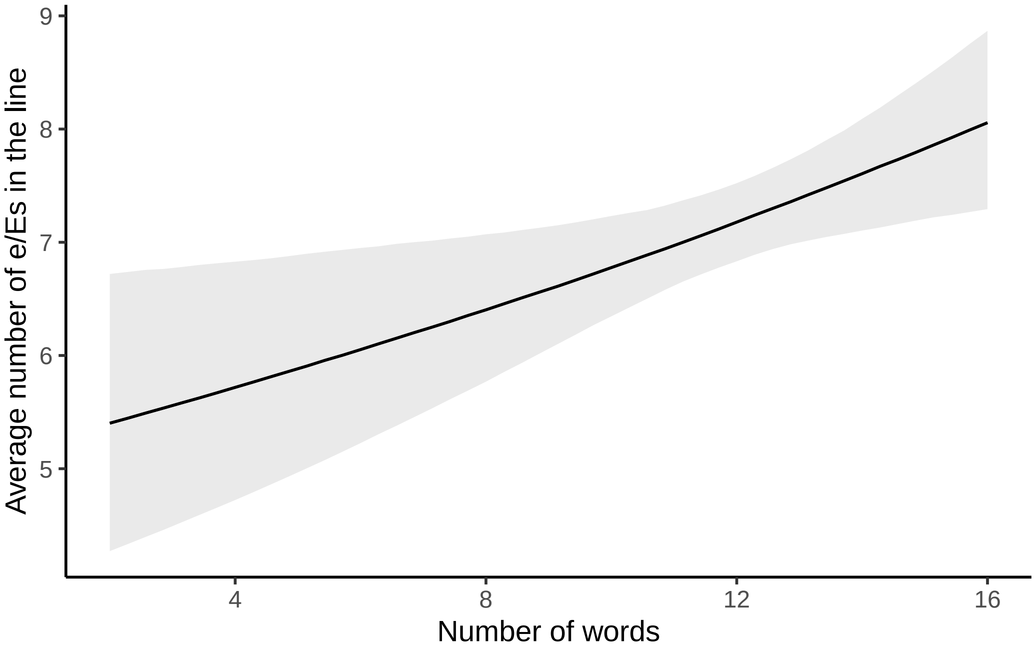


Figure 3: The number of ’E/e’s predicted by the model for the number of words in the line

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 to highlight the importance of considering methodological approaches in linguistic research. By adapting a methodology from a study done on ancient poetry to analyse 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.

5.1 Weaknesses

Our study has a 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.

5.2 Future Research

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 could offer deeper insights into Austen's writing style since literary devices are shown to play a big role in fiction literature (Yeung 2021). Overall, our study lays the foundation for further investigations into literature and linguistics which highlights the great potential for interdisciplinary research in this field.

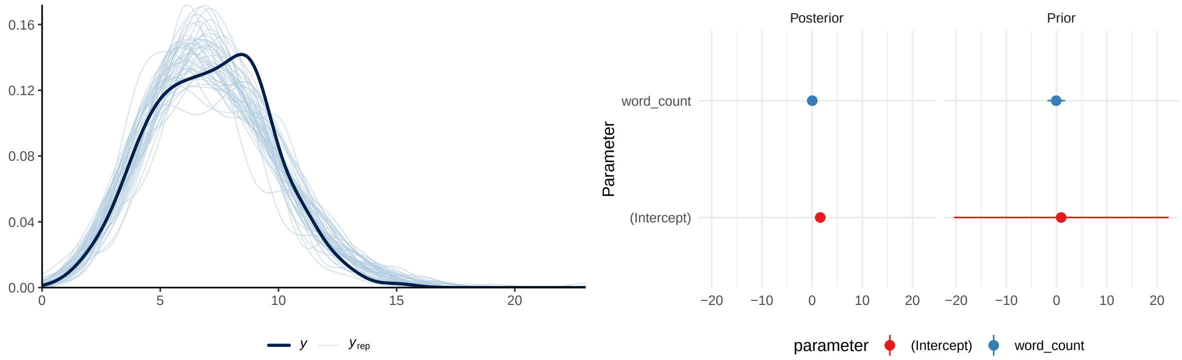
Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

The posterior predictive check assesses the Poisson distribution model's ability to accurately predict the frequency of the letter 'e' in each line of "Pride and Prejudice". The peak at around 7-8 indicates that the model aligns well with the observed data, supporting its validity in capturing the variability in letter frequency across the text.



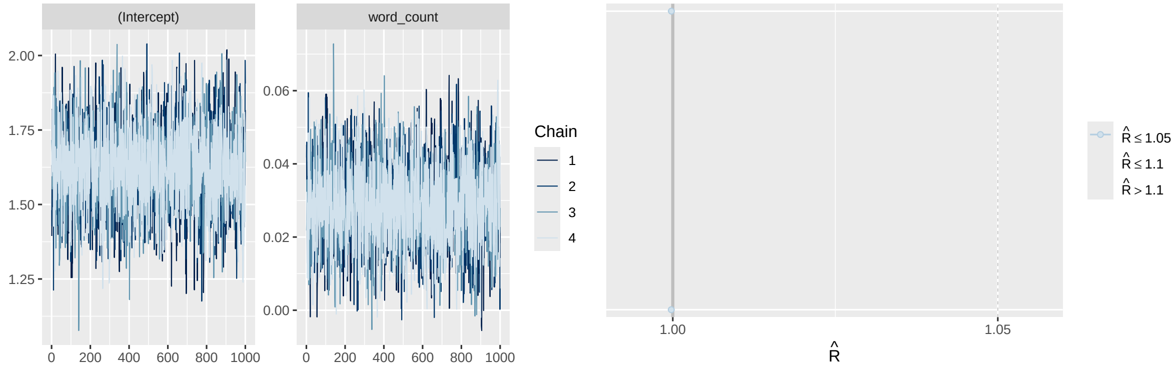
(a) Posterior prediction check

(b) Comparing the posterior with the prior

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

B.2 Diagnostics

In the trace plot Figure 5a, each line represents the sampled values of a parameter at each iteration of the MCMC algorithm, showing the convergence of the chain. The Rhat plot Figure 5b compares the variance within chains to the variance between chains, showing that they fall by 1 (indicating convergence).



(a) Trace plot

(b) Rhat plot

Figure 5: Checking the convergence of the MCMC algorithm

References

Alexander, Rohan. 2023. *Telling Stories with Data: With Applications in r and Python*. Chapman; Hall/CRC.

Arel-Bundock, Vincent. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects.

Austen, Jane. 1813. *Pride and Prejudice*. Penguin Classics.

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

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://CRAN.R-project.org/package=gutenbergr.

Notre Dame, University of. 2023. *Letter Frequencies in the English Language*. University of Notre Dame.

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

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Robert L. Solso, Joseph F. King. 1976. *Frequency and Versatility of Letters in the English Language*. Behavious Research Methods & Instrumentation.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Yeung, Lorriane K. C. 2021. *Why Literary Devices Matter*. The Polish Journal of Aesthetics.