An Analysis of the distribution of the letter ‘e’/‘E’ in Mary Shelley’s Frankenstein*

Irene Huynh

March 16, 2024

This paper analyzes the frequency and distribution of the letter ‘e’/‘E’ in the English literature classic, Frankenstein by Mary Shelley. Utilizing data from Project Gutenberg, the study examines the occurrences of the letter ‘e’/‘E’ in the first ten lines of each chapter to gauge whether the amount of ‘e’/‘E’s found in the text increases as more words are used. I find that the letter ‘e’/‘E’ usage does increase as the number of words increase. Our findings and methodologies may be employed by linguists to build on or corroborate existing data on letter frequency in the English language, by cryptographers to aid in the deciphering of classical or simple substitution ciphers, and by keyboard designers.

1 Introduction

Frankenstein by Mary Shelley combines Gothic horror story and science fiction elements to chronicle the life of Victor Frankenstein and his monstrous creation. The novel was first released in 1818 to unfavorable critiques, but has since solidified itself as a cult classic in literature and icon in pop culture (Gale 2006). This drastic change in reception is on account of Universal Studios’ 1931 horror film adaptation of the novel, bringing Frankenstein’s monster to life (Early and Malkowicz 2017). The film’s depiction of the creature with bolts protruding from its neck and a square-shaped head continues to be one of the most recognizable movie monsters (Early and Malkowicz 2017). According to Penguin Random House, the top publishing company as of 2024, Mary Shelley’s Frankenstein is one of the top twenty must-read classics, as chosen by their readers (McKenna 2022). Thus, analyzing this text can be of great use for studying the English language.

The English language has evolved a great deal over centuries and investigating its evolution can give insight into the people who use it. Linguists are concerned with studying languages

*Code and data are available at: LINK.

as a science; they observe and listen to speakers, conduct experiments, and perform analyses of languages to investigate the properties and characteristics of particular languages (Buffalo - Deparment of Linguistics, n.d.). Linguists may study the frequency of letter usage for a basic understanding of the particular writing system adopted by a language, whether it be syllabic, ideographic, or alphabetic (Buffalo - Deparment of Linguistics, n.d.). According to a study analyzing the occurrences of letters in words in the Concise Oxford Dictionary's (9th edition, 1995) main entries, the letter 'e'/'E' appears 11.1607% of the time in all words, the highest frequency out of the entire English alphabet (Notre Dame, n.d.). The frequency of letters may vary from language to language and such information is of interest to cryptographers, keyboard designers, and alphabet-based game creators (Grigas and Juškevičienė 2018). Cryptographers use letter frequency data to break classical or simple substitution ciphers (Guide, n.d.). By being cognizant of the frequency of letters in a particular language, cryptographers can correlate the distribution of letters in the substituted cipher with the actual distribution of letters in the language to ultimately match each letter with its substitute and break the cipher (Guide, n.d.). Moreover, keyboard designers rely on the frequency of letters to determine the location of keys on the keyboard (Grigas and Juškevičienė 2018).

In this paper, I analyze the frequency of the letter 'e'/'E' and the number of words in the first ten lines of each chapter in Mary Shelley's Frankenstein to produce a predictive model that determines whether the number of 'e'/'E's increase as more words are used. The estimand for this study is the effect of the number of words used on the number of times 'e'/'E' appears. This study finds that as the number of words used increases, the frequency of the letter 'e'/'E' also increases.

The remainder of this paper is structured as follows: the Data section details the collection and processing of Mary Shelley's Frankenstein text data. Additionally, the section includes discussion on the measurement of the original dataset. The following Results section displays and analyzes tables and figures that model the data. The Discussion section draws conclusions from those tables and figures and details the importance of my findings and weaknesses of the paper.

2 Data

2.1 Source

The data utilized in this paper was retrieved from Project Gutenberg (Johnston and Robinson 2023). The statistical programming language R was used to retrieve, clean, and process the data (R Core Team 2023). In particular, the following R packages were used: `gutenbergr` (Johnston and Robinson 2023) and `tidyverse` (Wickham et al. 2019) for data acquisition, cleaning, and processing; `rstanarm` (Goodrich et al. 2024) and `marginaleffects` (Arel-Bundock 2024) for model creations, and `ggplot2` (Wickham 2016) for creating figures.

2.2 Measurement

The original data set was retrieved from Project Gutenberg (Johnston and Robinson 2023), which is a reliable and reputable source for eBooks and other electronic texts. As the data I am concerned with was collected and cataloged from an existing novel of text, there is a low possibility for measurement error; however, possible avenues for the introduction of errors would be in the archiving of the novel by software bugs or technological malfunctions. Additionally, the digitized copy of Frankenstein was last updated in 2022 by Project Gutenberg. Considering that digital book scanning software as of recent years is considered highly sophisticated (Prainse 2022) and that Project Gutenberg has digitized over 70,000 books, the quality of the data used in this study is reliable.

2.3 Examined Variables

The original data set consists of all text in Mary Shelley's Frankenstein, including the title page, preface text, and inclusions such as letters. The main variable of interest is the frequency of the letter 'e'/'E' in the novel. An analysis using each line in the entire novel would be ideal; however, such a task would be cumbersome. Thus, I will only focus on the frequency of the letter 'e'/'E' in the first ten lines of each chapter.

Figure 1 verifies that the mean and variance of the number of 'e'/'E's are roughly similar by plotting all of the data. The mean, in pink, is 6.9, and the variance, in blue, is 5.8. They are not exactly the same; however, they are similar.

The diagonal in Figure 2 helps visualize the data. If the data were found on the y = x line, then on average, there would be one 'e'/'E' per word; however, Figure 2 demonstrates that most of the data points are found below the line. Thus, on average, there is less than one 'e'/'E' per word.

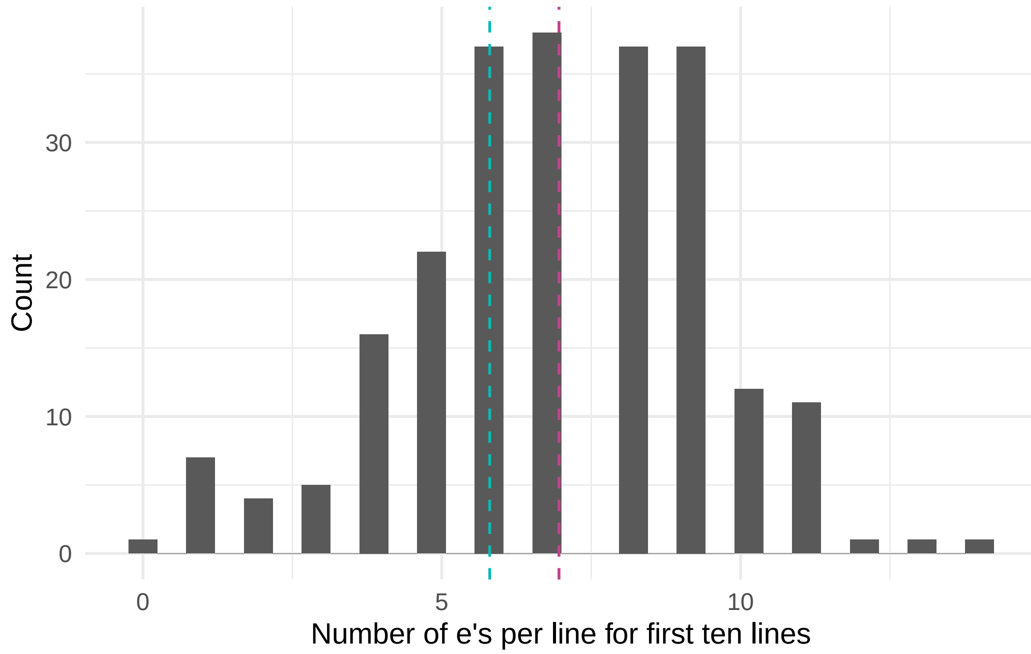


Figure 1: Distribution of the number of e/Es

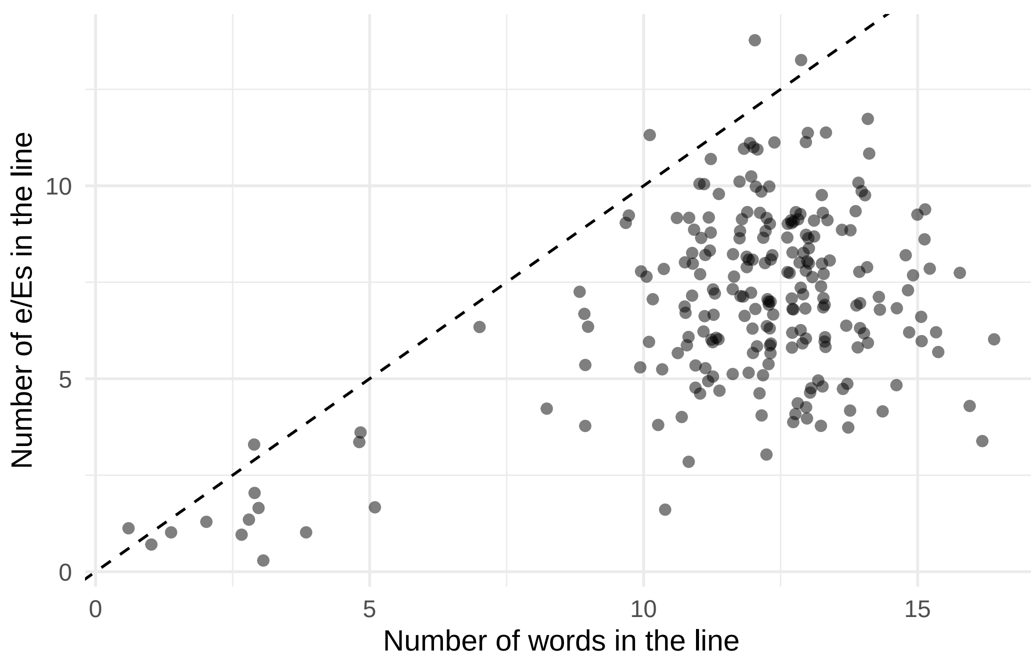


Figure 2: Comparison of the number of e/Es in the line and the number of words in the line

3 Model

The goal of our modelling strategy is to utilize the Poisson distribution to build a Poisson regression model as our data is concerned with count data and λ distributes probabilities over non-negative integers.

3.1 Model set-up

Define y_i as the number of 'e'/'E's in the line and the explanatory variable as the number of words in the line.

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (1)\\
\log(\lambda_i) &= \beta_0 + \beta_1 \times \text{Number of Words}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
& && (5)
\end{aligned}
$$

I run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2024). We use the default priors from `rstanarm`.

3.1.1 Model justification

I expect a positive relationship between the number of 'e'/'E's in the line and the number of words in the line. This draws from the

4 Results

Our results are summarized in Table 1.

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	0.98
	(0.14)
word_count	0.08
	(0.01)
Num.Obs.	230
Log.Lik.	−509.987
ELPD	−511.8
ELPD s.e.	7.1
LOOIC	1023.7
LOOIC s.e.	14.2
WAIC	1023.6
RMSE	2.17

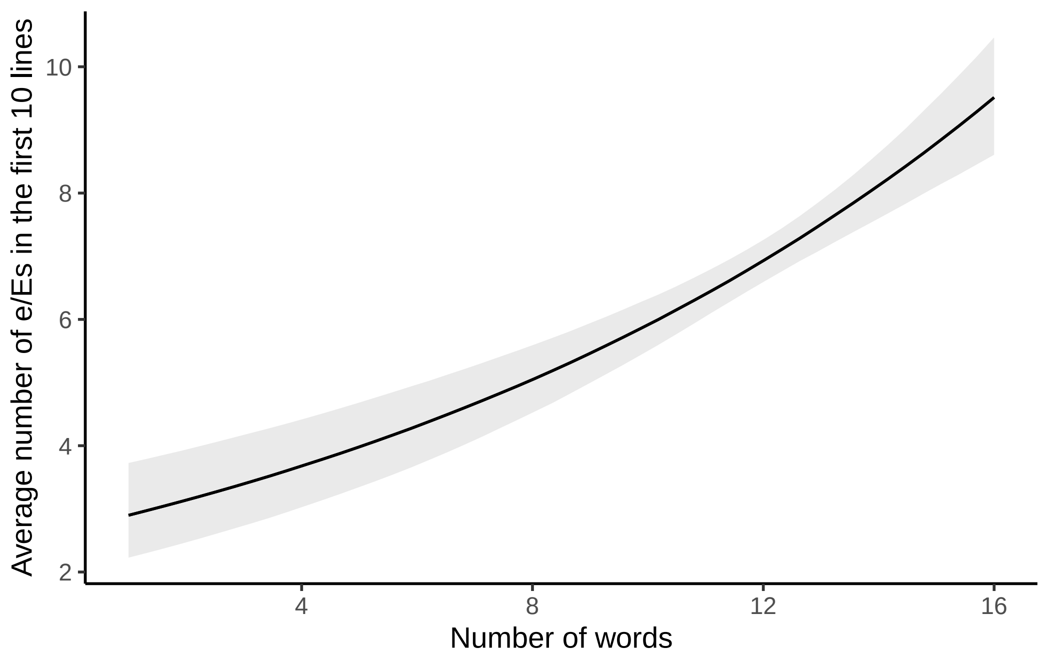


Figure 3: Explanatory models of flight time based on wing width and wing length

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

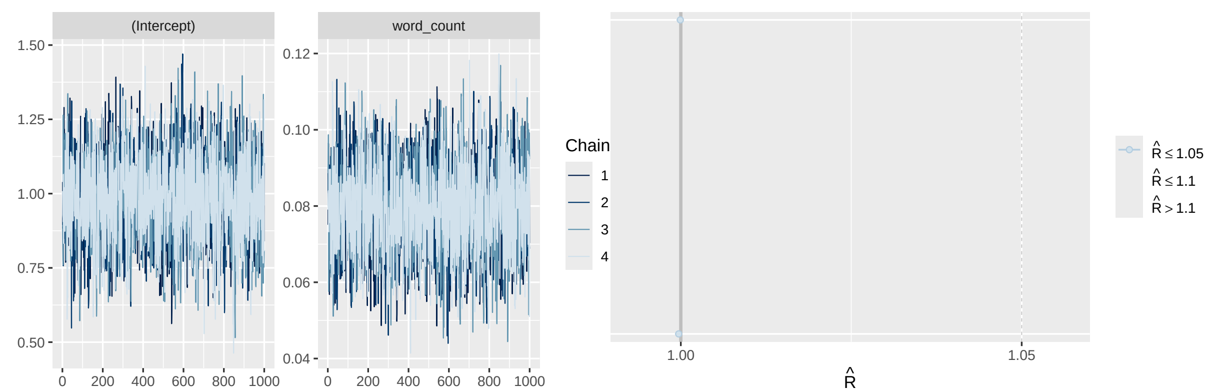
In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...

B.2 Diagnostics

Figure 4a is a trace plot. It shows... This suggests...

Figure 4b is a Rhat plot. It shows... This suggests...



(a) Trace plot

(b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://marginaleffects.com/.

Buffalo - Deparment of Linguistics, University at. n.d. *What Is Linguistics?* https://arts-sciences.buffalo.edu/linguistics/about/what-is-linguistics.html.

Early, Rosalind, and Tom Malkowicz. 2017. *Frankenstein's Monster in Popular Culture*. https://source.wustl.edu/2017/10/images-of-frankenstein/.

Gale. 2006. *An Introduction to Frankenstein*. https://www.gale.com/open-access/frankenstein#:~:text=Initial%20critical%20reaction%20to%20Frankenstein,was%20ahead%20of%20its%20time.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Grigas, Gintautas, and Anita Juškevičienė. 2018. *Letter Frequency Analysis of Languages Using Latin Alphabet*. *International Linguistics Research*. Vol. 1. https://doi.org/10.30560/ilr.v1n1p18.

Guide, Computer Science Field. n.d. *Substitution Ciphers - Problems with Substitution Ciphers*. https://www.csfieldguide.org.nz/en/chapters/coding-encryption/substitution-ciphers/#:~:text=A%20substitution%20cipher%20simply%20means,H%27s%2C%20E%27s%2C%20and%20L%27s.

Johnston, Myfanwy, and David Robinson. 2023. *Gutenbergr: Download and Process Public Domain Works from Project Gutenberg*. https://docs.ropensci.org/gutenbergr/.

McKenna, Sarah. 2022. *100 Must-Read Classics, as Chosen by Our Readers*. https://www.penguin.co.uk/articles/2022/05/100-must-read-classic-books.

Notre Dame, University of. n.d. *The Frequency of the Letters of the Alphabet in English*. https://www3.nd.edu/~busiforc/handouts/cryptography/letterfrequencies.html.

Prainse, George. 2022. *Book Scanners Speed: Why Is It Important?* https://www.libraryvision.org/digitization/book-scanner-speed/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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.