Homework Assignment 3: Applied Probabilistic Models

Word Distributions

5273

1 Introduction

For this work, data is collected on the free eBooks library Project Gutenberg [5]. The chosen book for the analysis is: “The Autobiography of Benjamin Franklin” [4]. Data obtained from the Project Gutenberg are in `txt` format.

For the analysis, it is used the R software in its version 4.0.2 [3], and the code used is available on the GitHub repository [6]. This work is run on a MacBook Air with an Intel Core i5 CPU @ 1.8 GHz and 8 GB RAM.

2 Data Distribution

The book is downloaded directly from the web and in order to develop the analysis, the following code is used.

```r
require(gutenbergr) #Download books from online library
require(tidytext) #Clean text
require(dplyr)  #Data Manipulation
require(textshape)
require(tokenizers)

library(fitdistrplus)

#Load the book: "The Autobiography of Benjamin Franklin"
book<-gutenberg_download(c(148))

#Variables used:
words <- book %>% unnest_tokens(word, text, "words")  #contains words
sentences <- book %>% unnest_tokens(sentence, text, "sentences")
paragraphs <- book %>% unnest_tokens(paragraph, text, "paragraphs")
```

a3.R

In this work, it is analyzed how frequencies of English words are distributed. The first discussed aspects are events that describe two known distributions. The second part corresponds to the other two events that can be worth finding which distribution best fits the data. For this last issue, it is recommended to see the paper of Delignette-Muller et al. [2].

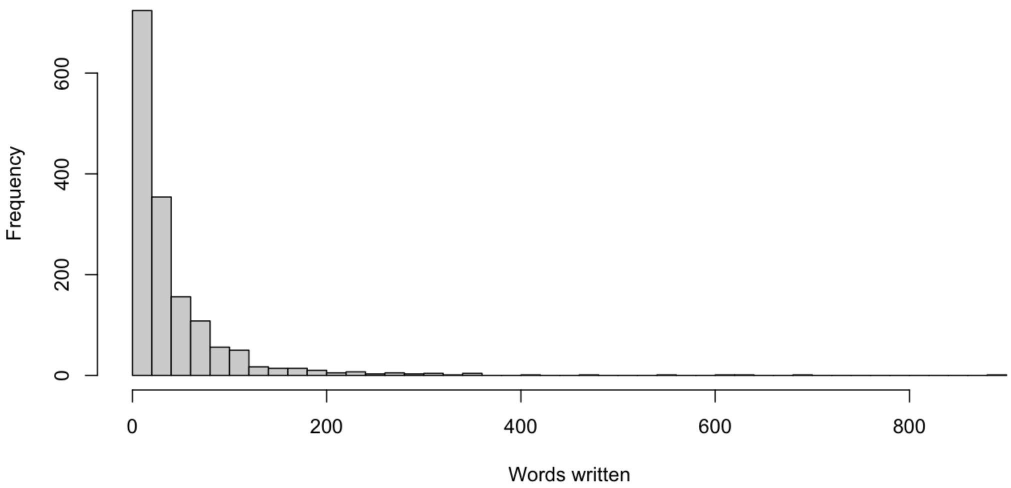


Figure 1: Histogram of the amount of words used before the pronoun "I"

2.1 Geometric Distribution

Since the book is an autobiography it is redacted in the first person, so it is expected the word "I" appears multiple times referring to facts corresponding to the author himself. Some statistics of this pronoun are studied. The first aspect to analyze is how many words have been used when the "I" appears. This may correspond to a geometric distribution, which can be described as the number of repetitions resulting in failures until the first success is achieved. In this case, all the words written until the pronoun is used can be considered as failures, and the use of "I" is the success. Figure 1 shows a histogram of how this aspect is represented in the book.

2.2 Binomial distribution

The number of times this pronoun is used in each sentence of the book is another event that could be analyzed. This case could be described as a binomial distribution, which describes the number of success obtained in n Bernoulli repetitions. For this example, those repetitions are the total number of sentences in the text whereas the number of success is given by the presence of the pronoun "I" in k times. This is shown in the histogram corresponding to Figure 2.

2.3 Other distributions

Length of English words and also its quantity when constructing paragraphs are the other considered aspects. Figure 3 shows a histogram of words length used throughout the book and Figure 4 shows a histogram of the amount of word used per paragraph in the document.

Last, Figure 5 show a skewness-kurtosis plot such as the one proposed by Cullen et al. [1] for the empirical distribution of both events. In this plot, values for common distributions are displayed in order to help the choice of distributions to fit data. The distribution is represented by a single point on the plot. According to Delignette-Muller et al. [2] skewness and kurtosis are known not to be

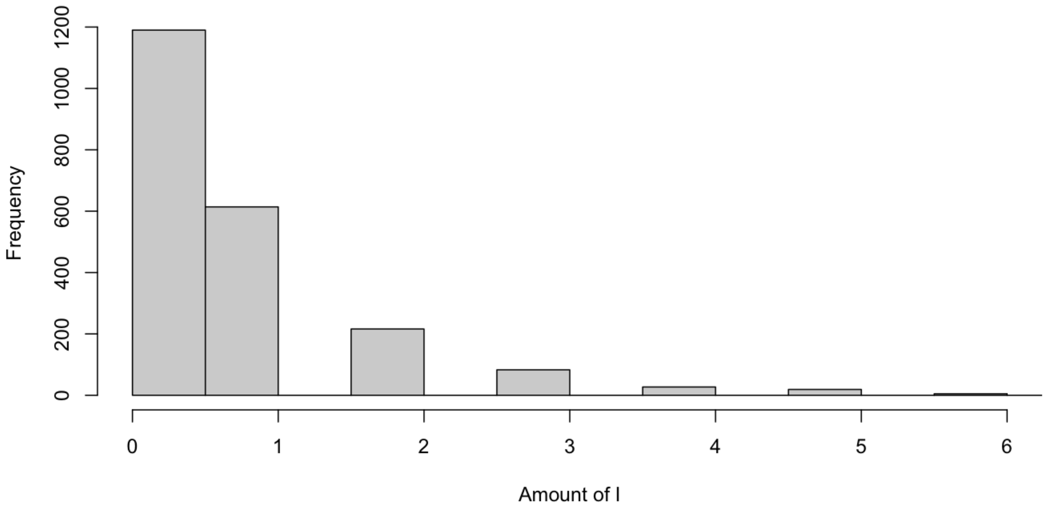


Figure 2: Histogram of the amount of times the pronoun "I" is mentioned in sentences

robust, thus the plot should then be regarded as indicative only. A non-zero skewness reveals a lack of symmetry of the empirical distribution. For words length and words per paragraph, the skewness has values of 0.996 and 3.128 respectively. The kurtosis value quantifies the weight of tails in comparison to the normal distribution for which the kurtosis equals 3. The values of kurtosis are 3.519 for word length and 27.017 for words per paragraph.

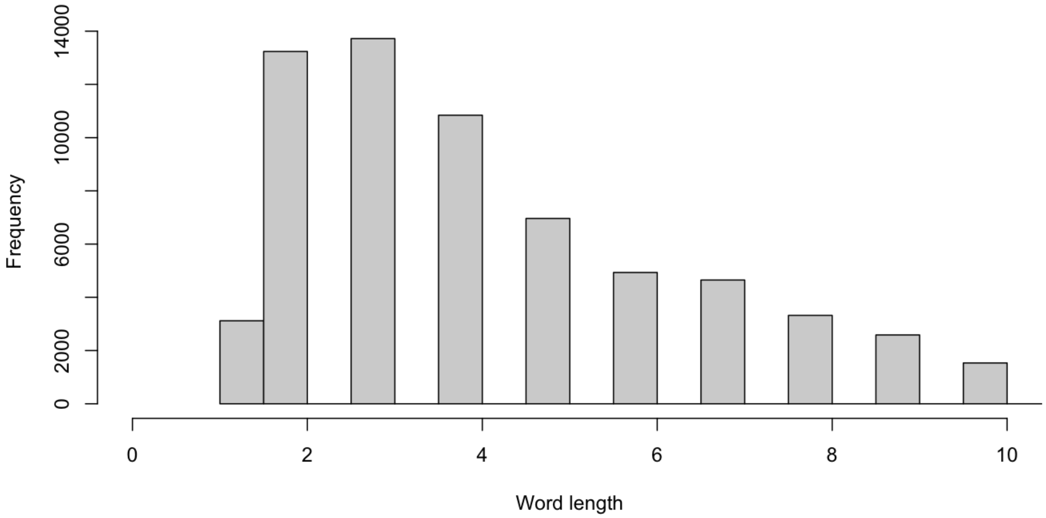


Figure 3: Histogram of lenght of words used in the book

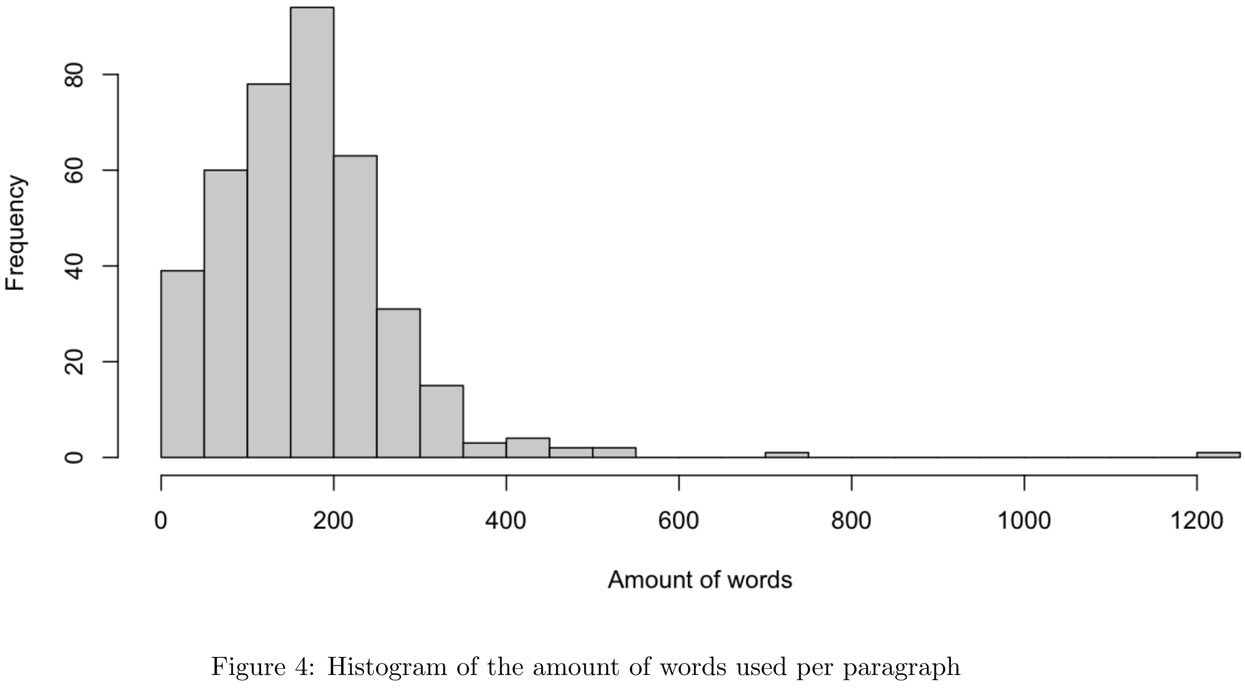
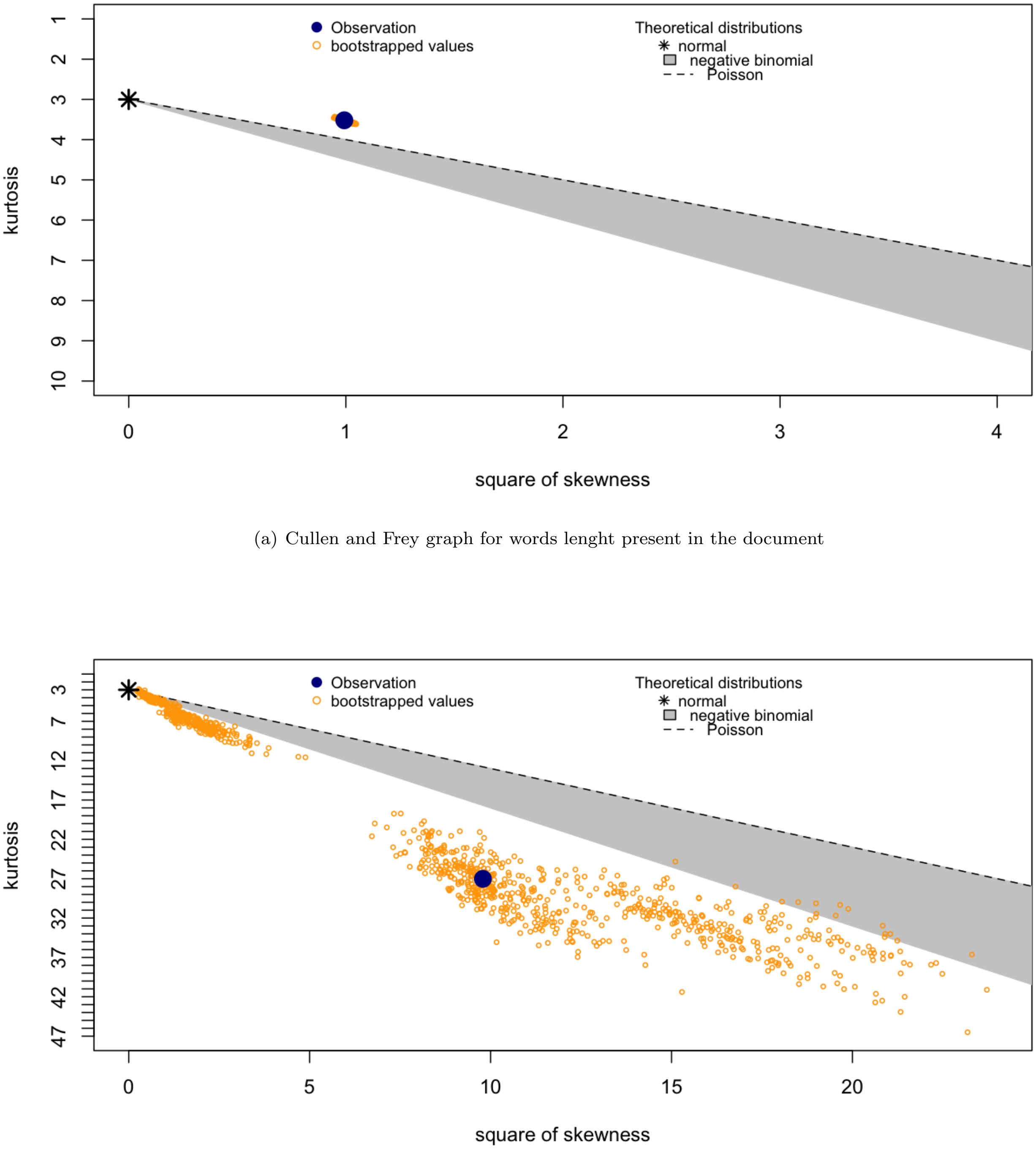


Figure 4: Histogram of the amount of words used per paragraph



(a) Cullen and Frey graph for words lenght present in the document

(b) Cullen and Frey graph for words per paragraphs

Figure 5: Cullen and Frey Graphs

References

[1] Alison C Cullen, H Christopher Frey, and Christopher H Frey. *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs*. Springer Science & Business Media, 1999.

[2] Marie Laure Delignette-Muller, Christophe Dutang, et al. fitdistrplus: An r package for fitting distributions. *Journal of statistical software*, 64(4):1–34, 2015.

[3] The R Foundation. The R Project for Statistical Computing. `https://www.r-project.org/`, 2020.

[4] Benjamin Franklin. *The Autobiography of Benjamin Franklin: 1706-1757*, volume 1. 2007.

[5] Michael Hart. Project Gutenberg, 1971. `http://www.gutenberg.org/ebooks/`, Last accessed on 2020-09-09.

[6] Oscar Hernandez. Probability in R. `https://github.com/oscaralejandro1907/probability-in-R/blob/master/assignment1/t1.R`, 2020.