Dustin Tran
STAT-301-01
Winter 2021

Is the percentage of the 100 most common English words in novels significantly less than 50%?

Table of Contents

Dustin Tran
STAT-301-01
Winter 2021

Introduction

I grew immediately fascinated in the direction of this question when I discovered a blog on basic literature statistics on the internet[1]. How much do the 100 most common English words make up English literature[2]? It raised questions on the diction and composition of English literature that is a subtle part of everyday life. In addition, the analytics of getting a program to process and filter out specific words intrigued me as a computer science major. Hence, I decided to chose this as my project.

There is no official study regarding this topic, but I calculated this '50%' statistic by averaging the seven values that was provided. On an intuitive level, I expected to find a much lower value between 30% and 40% in my study. Regardless, the results of this study should be interesting to those who are curious of how English literature may have an inherent pattern in composition.

[1] http://www.tylervigen.com/literature-statistics
[2] List of 100 most common English Words; https://web.archive.org/web/20130616200847/http://www.duboislc.org/EducationWatch/First100Words.html

Dustin Tran
STAT-301-01
Winter 2021

Data Gathering Method

Observational Unit

The observation units of this study are novels, more specifically, novels in the form of text files. Novels, as a majority, using everyday language and have a minimum word count of 40,000. This high word count is needed to dilute the effect of any individual number of the hundred most common words being counted. The usage of novels for this study is, in my opinion, better used to represent everyday language than other works of literature like poetry. 250 observational units are being used for this study.

Novels are randomly selected through systematic sampling on a random book generator[3]. The fourth novel title is selected in each iteration from the generator, with the filter set to have no other works of literature. Although the number of novel titles in the random generator is not given, it can be estimated to be at least over 10000 given that there are no duplicate titles during the entire study.

Some concern of sampling bias is warranted, given that the novel title generator is for the "Top Rated Books". In addition, I am only conducted the data analysis for the novels of text files that can be found. There are 38 recorded instances of being unable to find the generated novel's respective text file. With this possible bias in mind, the publishing years of the novels is recorded.

[3] https://www.bestrandoms.com/random-book-generator

Dustin Tran
STAT-301-01
Winter 2021

Variable of Interest

The variable of interest for this study is the “common word proportion”. This will be a measure of the number of the hundred most common words found in the text file divided by the total word count in the file. This proportion will be determined by running each text file through a program that does the following:

Open file →

Look at each line of the text file →

Remove special characters(“’\/-_[] , etc.) and punctuation →

Break up line into a list of ‘values’ based on white space →

Look at each value of the list→

If the ‘value’ has all alphabet characters, add to word count →

If it is in 100 common words, add to counter for 100 common words →

Repeat for each line until each of file →

Divide count of common words over word count →

Record data(Novel Title, Author, Publishing Year, Common Word Proportion, Word Count Error) →

Repeat for all text files.

Concerns can be raised regarding the functionality of the program since it is found to not perfectly measure the word count. With this is mind, a variable will be recorded called the ‘word count error’, a measurement of the program’s calculated word count divided over a more trusted word count.

Dustin Tran
STAT-301-01
Winter 2021

Analysis of Results

Hypotheses

The null hypothesis for my research question will be that the long-run common word proportion will be 0.5. The alternative hypothesis is that the long-run common word proportion will be less than 0.5.

H_0: $\pi = 0.5$

H_A: $\pi > 0.5$

Possible Sampling Bias

The issue of sampling bias was always going to be an issue with an unverified random generator, especially one that is aptly made for the "top rated books". Unfortunately, I had to the final decision to use the random generator regardless knowing that only more well-known novels would be found on the internet. I approached this novel title generator with the expectation that it may be skewed towards more 'classic' titles, which are normally older novels. An analysis of the publishing year shows that this belief is not unwarranted.

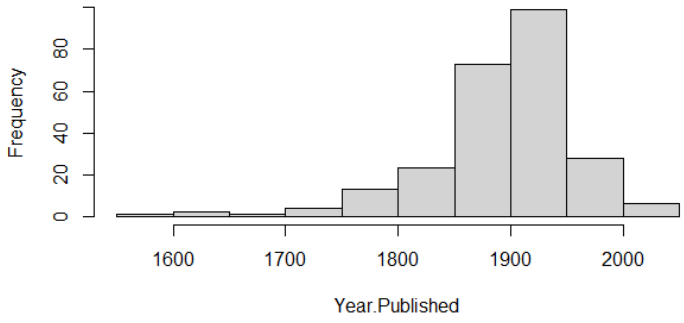

Figure 1: Histogram of Publishing Years

Dustin Tran
STAT-301-01
Winter 2021

The distribution of the novels' publishing years is skewed to the left with the median year being 1903 with the average year being 1891. The boxplot to the right will show that the middle 50% of the publishing years are between 1862 and 1927 with multiple outliers based on the 1.5IQR criterion. It is safe to say that the novel titles given by the random generator are heaviliy skewed toward being published between those years.

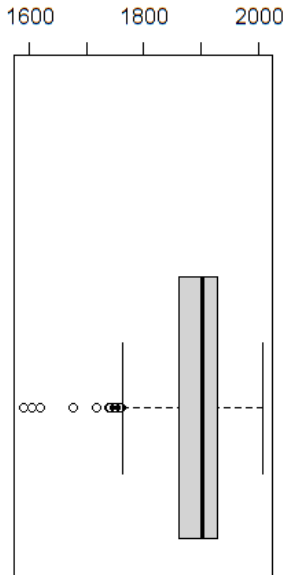

Figure 2: Boxplot of Publishing Years

In the context of study, I believe that a better random generator would have a more uniform distribution of the publishing years. However, it is hard to argue for what the range of the publishing years should be. The start of 'modern' English literature has no concrete year and earlier novels may have been translated into English decades, if not centuries, earlier. In my pure opinion, a lesss bias generator would have a uniform ditribution at least between 1750 and 2000 before leveling off in either direction.

Other possible biases, such as the novel's subgenre(s), are left unrecorded due to the subjectiveness of the issue and an underlying assumption that genre has overall little to no effect on whether it is put on a random generator for top-rated novels.

Dustin Tran
STAT-301-01
Winter 2021

Possible Programming Error

The program used to process text files is not perfect in its ability to count words. This problem could potentially cause notable errors when measuring the common word proportion. With this issue in mind, I also recorded a new variable, the "word.count.error", to measure the error in the program's word count for each text file. Below are the steps for getting this value.

Getting the "Word.Count.Error" for a Text File:

Run program and record the program's given word count for text file →

Determine text file's word count from another verified source →

Divide the program's word count value over the new word count →

Subtract previous value from 1 and round to 4 decimal places →

Record previous value as "Word.Count.Error" for the text file.

(A positive value indicates that the program's word count is too large while a negative value indicates the program's word count is too small)

An analysis of all the Word.Count.Error values will show that the middle 50% of values are between 0 and 0.017(1.7%). The lowest value is -0.015 with the largest being 0.048, the latter being the only outlier based on the 1.5IQR criterion. The average value is 0.009, meaning that on average the program's word count is 0.9% larger than it should be. I ultimately deem this as an acceptable margin of error and unlikely to change calculations of the common word proportion to a notable degree.

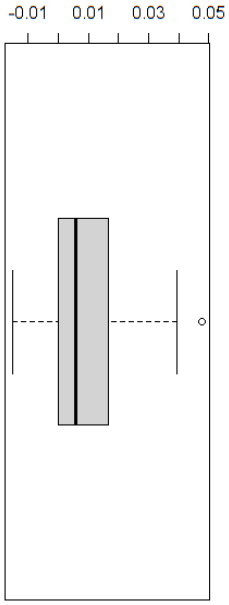

Figure 3:Boxplot of Word.Count.Error

Dustin Tran
STAT-301-01
Winter 2021

Common Word Proportion

An examination of the common word proportion distribution for this study will show that the distribution is centered at 0.456 with a standard deviation of 0.024. The middle 50% of Common.Word.Proportion values are between 0.442 and 0.470 with the minimum being 0.371 and the maximum being 0.538.

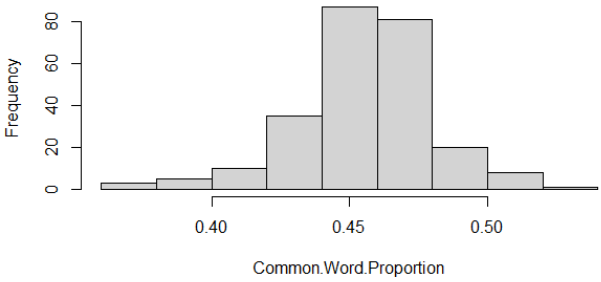

Figure 4: Histogram of Common Word Proportion

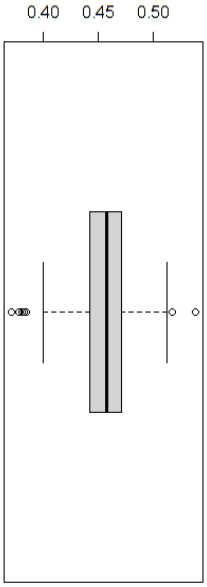

Figure 5: Boxplot of Common Word Proportion

The first fact to take away from this data is that obtaining a mean of 0.456 is unlikely to be by random chance. Under a theoretical binomial distribution where the probability of success is 0.5, the probability of obtaining a value of 0.456 is 0.092. It would be possible to reject the possibility of the common word proportion values being pure chance under a 10% test of significance.

Using the t-procedure would show that the probability of having a population mean of 0.5(50%) with my data is 2.2×10^{-16}(< 0.001). In addition, a constructed 95% confidence interval of the mean would be (0.453, 0.459). Under a 5% test of significance, we can confidently reject the idea that the actual common word proportion is 0.5. This is further supported by the fact that the constructed confidence interval does not contain 0.5.

Conclusions

All in all, the study provides strong evidence that the common word proportion is indeed less than 50, although not to the, in retrospect, extreme extent I was expecting. Where I was expected an average common word proportion between 30 to 40% from my data, I instead got a value of 45.6% from my study, which turned out to be a significant smaller value compared to 50% in context. At a 5% test of significance, I can reject that the hypothesis that the common word proportion is 0.5.

This study was initially intended to be representative of all English literature but scaled down due to resource and time constricts. As of now, I am only willing to generalize to mainstream English novels. However, if you accept the premise that novels can represent the entirety of English literature, you can also make a generalization that the results of this study can be generalized to all English works.

From this study, I hope for readers to gain both some insight, but also more questions regarding the composition of literature. Given more time and resource, I would have liked to expand beyond just novels and into other works of literature. This would have

most likely led to more variability in the data, but also more certainty in that the study is representative of all English literature. Based on the design of this study, comparative studies could be designed between different works of English literature or difference languages altogether.