

# Statistics in Text

Wesley Jenkins



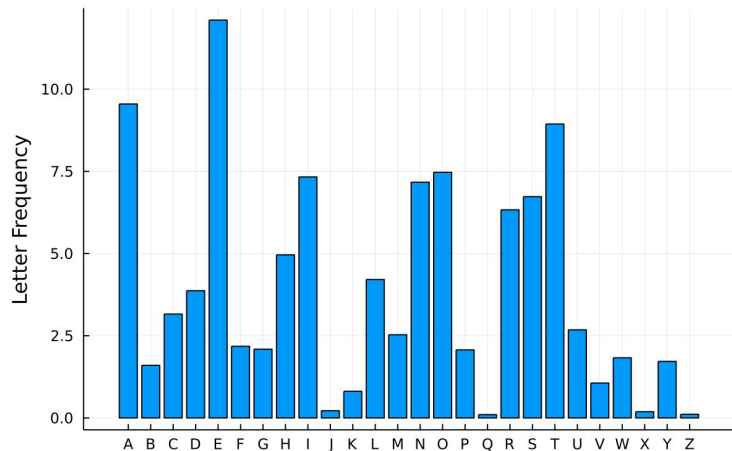
# What is Text Analysis?

- Natural text is not random. Completely random text would be meaningless. Thus, it must follow some prescribed pattern, which humans can understand.
- Some patterns are simple, while others are extremely complex and not yet understood.
- But even the most simple of patterns can be useful!

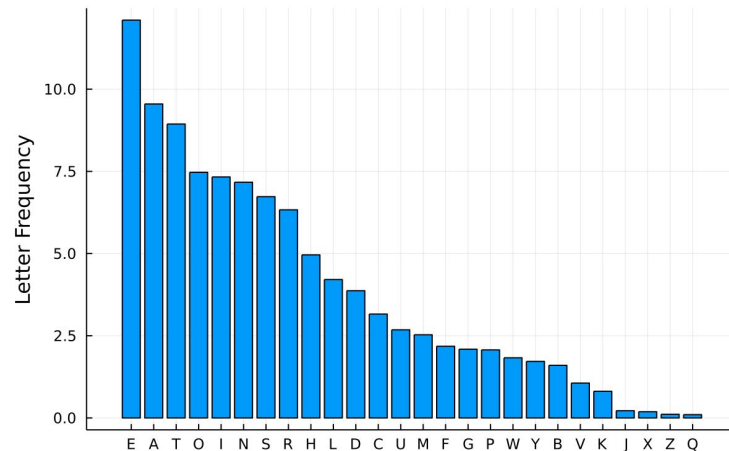
# Letter Frequency

- The most obvious characteristic of text is its letter frequency.
- Not all letters are used equally.
- In English, letters are distributed very unevenly:

Frequency of Letters in English Text

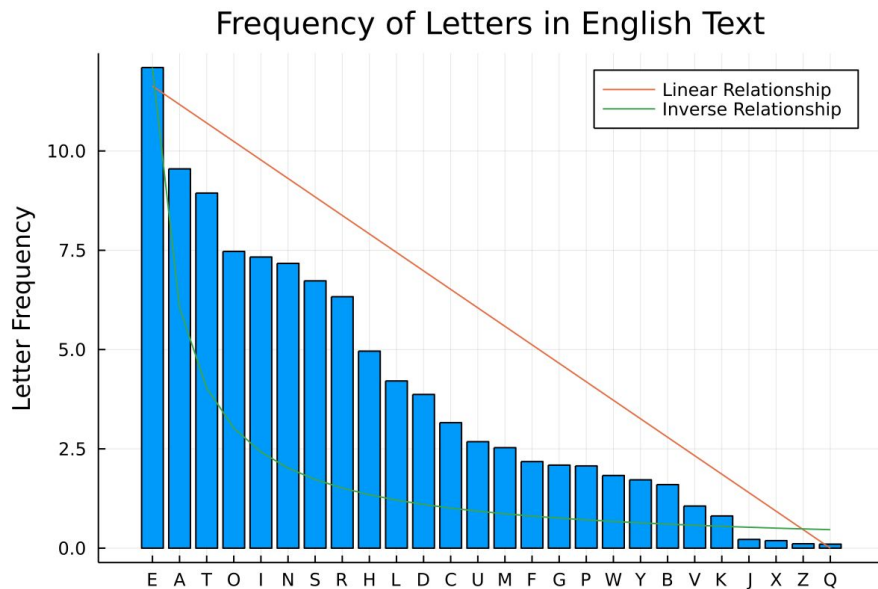


Frequency of Letters in English Text



# Letter Frequency

- Letter frequency doesn't follow any simple relationship.
- It seems to lie somewhere within a linear relationship and inverse relationship.
- However, given the small number of letters, there's too much variation for them to follow a neat relationship.
- Hopefully, something else will have a nicer relationship.

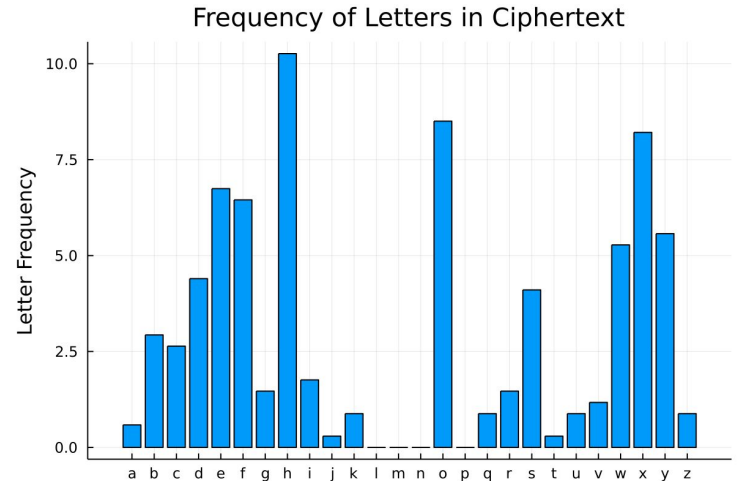


# Uses for Letter Frequencies

- Letter frequencies may be simple, but they have an important use case in cryptography.
- Given a simple example of a Caesar cipher or substitution cipher, letter frequencies can be used to trivially solve it.

Eo dwx oyh vhx0 fs oezhx, eo dwx oyh dfcx0 fs oezhx,  
eo dwx oyh wih fs dexgfz, eo dwx oyh wih fs sffuexybhxx,  
eo dwx oyh hrfqy fs vhuehs, eo dwx oyh hrfqy fs ebqchgkueoa,  
eo dwx oyh xhwxfb fs Leiyo, eo dwx oyh xhwxfb fs Dwcjhbxx,  
eo dwx oyh xrcebi fs yfrh, eo dwx oyh debohc fs ghxrvec,  
dh ywg hthcaoyebi vhsfch kx, dh ywg bfoeyebi vhsfch kx,

It was the best of times, it was the worst of times,  
it was the age of wisdom, it was the age of foolishness,  
it was the epoch of belief, it was the epoch of incredulity,  
it was the season of Light, it was the season of Darkness,  
it was the spring of hope, it was the winter of despair,  
we had everything before us, we had nothing before us,



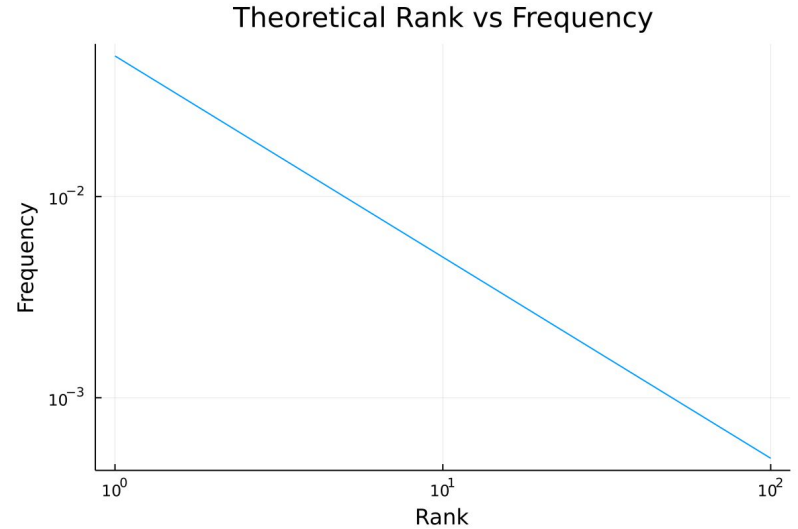
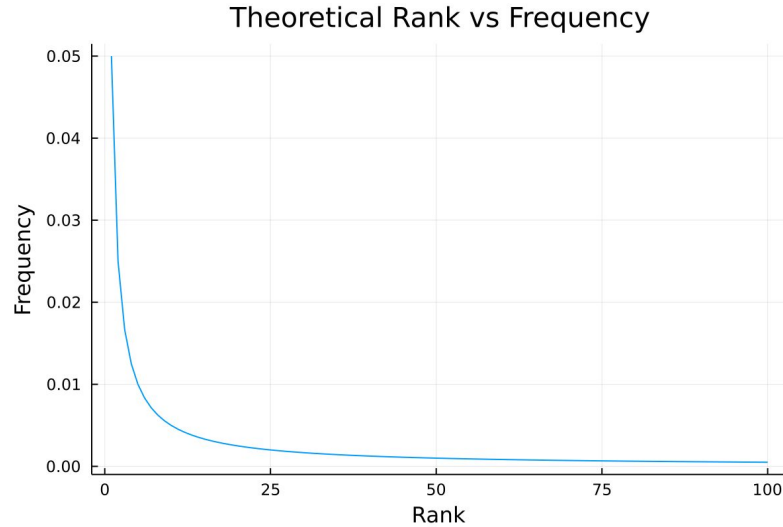
# Word Frequencies

- Not all letters are equal, but also, not all words are equal.
- The most common word in English is “the,” representing a massive percentage of all used words, closely followed by “of” and “and.”
- These three words alone account for over 12% of all written English words!
- And more importantly: There are a lot more words than letters!

# What is Zipf's Law?

- Zipf's Law is a law describing the relationship between the rank of words and frequency of words in natural languages.
- It states that the frequency of words are inversely proportional to the rank of said words.
- That is, the most common word is used twice as often as the second-most common word and three times as often as the third-most common word.
- Zipf's Law is named after George Kingsley Zipf, a linguist born in 1902. He first wrote about the law in 1935, though he did not claim to have found it himself.
- The first writings about Zipf's law are from a few decades earlier from a French stenographer named Jean-Baptiste Estoup.

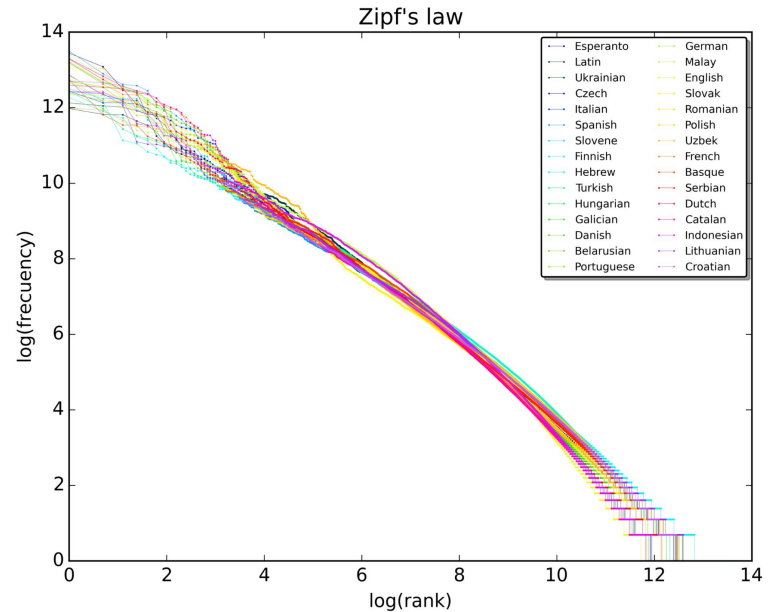
# Theoretical Plot and Log-Log plot



- In a log-log plot, Zipf's Law is observed when the line is linear with a slope of -1. This makes for a very easy visual check.

# Does Zipf's Law Work?

- Here is a plot of the first 10 million words in the Wikipedia pages of various languages.
- The graph is not perfect, especially on the edges. It has a clear curve on the extremes, where the relationship breaks down somewhat.
- Note the inclusion of Esperanto.
  - Esperanto is a completely constructed language (Or conlang).
  - Therefore, it is not only natural languages.



# Other Languages?

- How about other languages which don't use alphabets the same way?
- Korean uses system of syllabic blocks called Hangeul. Individual letters are formed into blocks, which represent a syllable. (Korean used to use Chinese characters called Hanja, which are still sometimes used, though rare).
- Chinese uses a system of logograms, where there are thousands of characters in use, though most are uncommon.
- Japanese uses a mixture of two alphabets (Hiragana and Katakana) and Chinese characters called Kanji.
- Modern Vietnamese uses the Latin alphabet, but adds a massive number of diacritics. How should a language like this be analyzed? Should every combination of letter and diacritic be considered its own letter? Or should they be analyzed all together?

# Other Languages?

- And some languages don't use words in the same way either. Highly synthetic languages construct long words that encompass a lot of meaning that would normally use multiple words in a less synthetic language like English.
- For example, a single word in Georgian:
  - Georgian: გადმოგვახტუნებინებდნენო
  - Translit: gadmogvakht'unebinebdneno
  - Meaning: 'They said that they would be forced by them (the others) to make someone to jump over in this direction'
- This of course makes analysis of words difficult, and would require them to be split up before any kind of analysis could be useful.

# Origin of Zipf's Law?

- There has been a lot of debate about Zipf's law, and its origins in specific.
- Is it mathematical in origin?
  - Some theorize that it can be explained when one considers randomly combining letters into probable word combinations.
- Is it psychological in origin?
  - Others theorize that it can be explained by the principle of least effort. That is: humans are lazy.
  - Humans like to use as few different words as possible. Zipf's law means that the vast majority of words people use are just a few hundred words.
- Currently: No one knows exactly why.

# Other uses for Zipf's Law?

- Some have also claimed that Zipf's law can be more generally applied to any ranking.
  - Correlating ranking of song to number of listens.
  - Correlating ranking of cities to their population.
  - Correlating ranking of income to income itself.
- However, others have questioned the validity of this, and it doesn't always work very well. It's unknown exactly which datasets work well, and which don't.