Letter Frequency

In the English Language

By Arvind Nagabhirava
MDM4UE - Mr. Heathfield
Tuesday June 19th, 2018

Abstract

In the popular board game Scrabble, there is a bag of 100 tiles, two of which are blank tiles. The number of each tiles for each letter is said to be based off of the letter frequencies the creator observed in the New York Times. Scrabble players have loved the game but a few of the reviews indicated that there were too many “I” tiles in the game. When the company Zynga created a similar game, they had a tile bag with slightly different tile distributions. For example, the number of “I” tiles was reduced and the number of “S” tiles was increased. This experiment was run to answer the question: *“Are the letter distributions of Scrabble and Words with Friends accurate representations of the letter distributions of today’s English language?”* Analysis of nearly 23 million characters indicates that Words With Friends is more accurate than Scrabble to today’s letter distribution.

Procedure

Due to the sheer volume of published text in the world, it would have been impossible to run this experiment by evaluating the population. For this experiment, only texts written after 1950 and originally published in English were taken into account. To ensure that the texts are an accurate representation of what the world is reading, only texts that sold over 10 million copies were taken into consideration. Furthermore, since the actual counting of letters was be done by a program, only texts that had been published online could be used for this experiment due to its programming limitations. Finally, texts under 100 pages were not taken into consideration as they would reduce the sample size.

A list of the most sold publications was used to choose the writing.[1] Of the 80 books that were on the list, 48 met all requirements for the experiment. Some books were invalidated because they were too short or written using a rhyming scheme which would have skewed letter frequencies in favor of the letters that made up the rhyming syllable. To ensure that there were no biases that a single author might have, only one book written by each author was taken from the list. For example, in the Lord of the Rings series, one of the most common words is a name, “Bilbo”, this is problematic because the letter “B” might be overcounted due to this discrepancy compared to a book where the main characters name isn’t used that often. If an author had multiple books on the top sold list, only their most sold book was used.

Once the books were found online, the table of contents, acknowledgements, author bibliography, and story summaries were deleted from the file.[2] Then, a python program was used to count the letters and post the recorded frequencies to a spreadsheet. The program was also modified to only count the first 45328 characters from each book to get rid of any issues that might have come with the differing lengths of each book. 45328 is the number of letters in the shortest book and this number was used to ensure the remaining sample would still be as large as possible.

---

[1] “List of Best-Selling Books.” *Wikipedia*, Wikimedia Foundation, 15 June 2018, en.wikipedia.org/wiki/List_of_best-selling_books.

[2] “OceanofPDF.” OceanofPDF, 18 June 2018, oceanofpdf.com/.

Python Code Used

http://bit.ly/LetterCounterCode

This code was written by Yash Dani and improved by Eric Chen.

Analysis

The initial data that was collected showed the total number of each letter in each of the 48 books and the total number of letters in each book. Instead of finding a percentage by comparing total “a” against the total characters from the 48 books, the individual percentage of a letter within each book was calculated. Calculating percentages the first way would have been an ineffective approach because the final distribution would be skewed towards the letter distributions of the longer books from the sample. The final percentages of each letter were instead calculated by taking the mean of the percentages of each letter within each book. Table 1 is the output yielded by the percentages for each letter as well as the standard deviation of each letter in the English alphabet.

TABLE 1 - Unrestricted Set

	Average %	Standard Dev		Average %	Standard Dev
a	8.11%	0.34%	n	6.74%	0.26%
b	1.57%	0.17%	o	7.60%	0.40%
c	2.36%	0.34%	p	1.68%	0.27%
d	4.55%	0.56%	q	0.08%	0.02%
e	12.39%	0.54%	r	5.52%	0.46%
f	2.04%	0.20%	s	6.12%	0.32%
g	2.24%	0.22%	t	9.11%	0.39%
h	6.31%	0.74%	u	2.94%	0.30%
i	6.92%	0.49%	v	0.91%	0.16%
j	0.18%	0.10%	w	2.49%	0.32%
k	1.08%	0.25%	x	0.13%	0.05%
l	4.17%	0.28%	y	2.16%	0.38%
m	2.50%	0.24%	z	0.10%	0.05%

To see if the length of passage affected the spread of the data or the standard deviations in *Table 1*, the shortest book of the 48 sampled was found. Since this book had 45328 letters, the first 45328 characters in each book were counted to produce an output similar to the unrestricted output previously seen. This second output only differed by taking the same number of letters from each book. The same process was used to generate the averages for the second set of outputs shown below.

TABLE 2 - Passage Length Restricted					
	Average %	Standard Dev		Average %	Standard Dev
a	8.14%	0.40%	n	6.75%	0.28%
b	1.62%	0.22%	o	7.54%	0.44%
c	2.39%	0.32%	p	1.68%	0.29%
d	4.56%	0.58%	q	0.08%	0.03%
e	12.45%	0.60%	r	5.61%	0.39%
f	2.09%	0.23%	s	6.18%	0.33%
g	2.23%	0.26%	t	9.05%	0.51%
h	6.35%	0.73%	u	2.89%	0.35%
i	6.78%	0.55%	v	0.89%	0.16%
j	0.17%	0.11%	w	2.46%	0.31%
k	1.06%	0.25%	x	0.14%	0.05%
l	4.18%	0.31%	y	2.14%	0.39%
m	2.49%	0.28%	z	0.09%	0.05%

As seen from comparing the two tables, the average standard deviation increased between the first and second set of data. This indicated that longer passages even out the letter frequency throughout the passage. This second set of data was still close to the first set of data which was produced. Looking closer, it was evident that the frequencies of “z”, “j”, “q”, and “x” were well below 1 in 100. This meant that if a distribution was made using the frequencies calculated here, there would simply be no “z”, “j”, “q”, or “x” tiles in the generated distribution. These letters are also worth more points in both games due to them being less common and harder to play. To accommodate for this, a tertiary set of outcomes was generated by fixing these letters as well as the letter “s”. The letter “s” was fixed because although the letter “s” is quite common in the English language (see *Table 1*), the number of “s” tiles in both Scrabble and Words With Friends is reduced because playing the letter is very easy due to the design of the game. The letter “s” can be utilised to pluralize most of the words that have already been played on the board. For playability reasons, the number of “s” tiles is also fixed for this next set.

Of the 98 letter tiles in Scrabble, there are 1 each of “z”, “j”, “q”, and “x” and 4 “s” tiles. Taking away these letters from the 98 tiles leaves behind 90 tiles. Similarly in Words With Friends there are also 1 each of “z”, “j”, “q”, and “x” but there are 5 “s” tiles. This leaves 93 letters since there are 102 letters tiles including those tiles. By “fixing” these letters, the remaining letters percentage is calculated as the percent they make up of the unfixed letters. For example in the word “exquisite” there are 9 letters, 3 fixed letters, and 6 unfixed letters. This means that the letter “e” makes up ⅓ of the unfixed letters in this passage. By using the first set of collected data and changing the total characters in each book to the total unfixed characters in each book, a new set of percentages was calculated. It is important to note that these percentages

are equal to the percentage a letter makes up of the unfixed letters and not the percentage they make up of all the letters. The bolded letters in the table below are the fixed letters who's percentages are not calculated for this set of data.

TABLE 3 - Fixed and Unfixed Letters					
	Average %	**Standard Dev**		**Average %**	**Standard Dev**
a	8.68%	0.36%	n	7.21%	0.28%
b	1.68%	0.19%	o	8.14%	0.43%
c	2.53%	0.37%	p	1.79%	0.29%
d	4.87%	0.60%	**q**		
e	13.26%	0.58%	r	5.91%	0.50%
f	2.18%	0.21%	**s**		
g	2.40%	0.23%	t	9.76%	0.41%
h	6.76%	0.78%	u	3.15%	0.33%
i	7.41%	0.54%	v	0.97%	0.17%
j			w	2.67%	0.33%
k	1.16%	0.27%	**x**		
l	4.47%	0.29%	y	2.32%	0.41%
m	2.68%	0.26%	**z**		

These three sets of processed data show the frequencies of each letter in a different way. To see whether Scrabble of Words With Friends is more accurate to today's language, the distributions of the two games needed to be calculated. The frequencies of letters in Scrabble and Words With Friends is shown above.

To evaluate how accurate the distributions of Scrabble and Words With Friends were, the percentages generated in *Table 1* and *Table 2* were multiplied by both 98 and by 102 to generate distributions equal to the size of Scrabble and to the size of Words With Friends. The percentages generated in *Table 3* were multiplied by 90 and 93 to generate the distributions of the two games accounting for fixed letters. Shown below are two of the six distributions generated by this process, the 98 tile distribution generated using the values in *Table 1* and the 90 tile distribution generated using the values in *Table 3*.

TABLE 4 - Unrestricted 98 Tile Distribution			
	Frequency		Frequency
a	8	n	7
b	2	o	7
c	2	p	2
d	4	q	0
e	12	r	5
f	2	s	6
g	2	t	9
h	6	u	3
i	7	v	1
j	0	w	2
k	1	x	0
l	4	y	2
m	2	z	0

TABLE 5 - Fixed Letter 90 Letter Distribution			
	Frequency		Frequency
a	8	n	6
b	2	o	7
c	2	p	2
d	4	**q**	
e	12	r	5
f	2	**s**	
g	2	t	9
h	6	u	3
i	7	v	1
j		w	2
k	1	**x**	
l	4	y	2
m	2	**z**	

To calculate how far off the generated distributions were from the distributions of the game, the sum of the absolute differences between the number of a letter in the distribution and the number of letters in the game was taken. Since the distributions generated using the values in *Table 3* would produce a difference of zero due to them being fixed, the sum of the absolute differences was not an adequate result. To further analyze this data, the sums were divided by 26 or by 21 based on whether the distribution produced values for all the letters or just the unfixed letters. The following results were found by using this process.

TABLE 6 - Average Deviation Per Letter Per Distribution			
	Unrestricted Set	Passage Length Restricted Set	Fixed and Unfixed Letters Set
Scrabble	0.9067	0.9159	0.8545
Words With Friends	0.7684	0.7728	0.7433

Conclusions

Based off the analysis of the raw data collected, it is evident that between Scrabble and Words With Friends, Words With Friends' letter distribution is more accurate to today's letter frequencies in the English language. Since this experiment revealed that both board games undercount the number of "h" and "t" tiles and overcount the number of "i" tiles, a 100 tile distribution that is actually representative of today's language was generated. This distribution was generated using a 99.99% confidence interval.

TABLE 7 - Representative 100 Tile Distribution					
	Frequency		**Frequency**		**Frequency**
a	8	j	1	s	5
b	2	k	1	t	9
c	2	l	4	u	3
d	4	m	2	v	1
e	12	n	7	w	2
f	2	o	8	x	1
g	2	p	2	y	2
h	6	q	1	z	1
i	7	r	5		

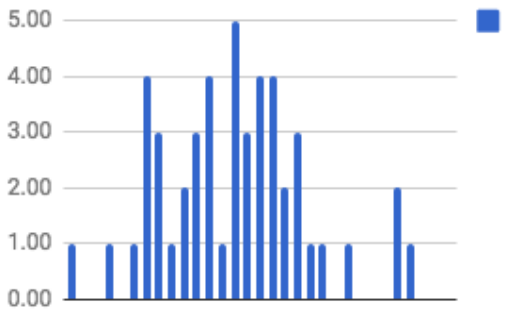
Sources of Error

Although the sample size used in this experiment was nearly 23 million characters from 48 of the most read books written by different authors since 1950, there is still room for error. Since the sample collected included only the most sold books, these letter frequencies cannot be assumed to be accurate for other forms of media. Furthermore, written language itself has changed since 1950 and so using book written in only the 21st century may have reduced any discrepancies between the generated frequencies and the actual frequencies. Furthermore, the books used were written in both first and third person, this results in discrepancies in the frequency of the letter “i”.

Critiques/Questions

Question: *“How consistent was your data?”*

Since the sample size was nearly 23 million characters, the standard deviation was incredibly low. When the frequency of a letter across the 48 books was published, most of the data showed up as a normal distribution which also indicates a good spread of data was collected. The histogram below shows the frequency of the letter “c” across the 48 books. The x-values on the graph range from 1.60% to 3.18%.



Question: *"Why was 1950 chosen as the cutoff year for the publishing date?"*

This was done because most of today's readers are reading books written since that time. The oldest generation is reading the books earlier in the range and the younger readers on the newer side of it. This was done to get an accurate representation of "today's English language" as stated in the question this report sought to answer.

Follow Up Questions

Would the frequency of the letter "i" decrease if books written in first person were omitted?
The frequency of the letter "i" would probably decrease since the word "I" is used far more often in first person novels as compared to third person novels.

Since one letter words cannot be played in either board game, would rooting out the words "a" and "I" result in more accurate results to the game?
The results of this experiment show that both the letters "a" and "i" are more frequent in the board game than in publications. This means that removing the letters "a" and "i" would likely change the probabilities to be even further off from the board game distribution than they already are.

Appendices

Link to Complete Data: http://bit.ly/2tgmD82

Appendix A - Scrabble Distribution			
	Frequency		**Frequency**
a	9	n	6
b	2	o	8
c	2	p	2
d	4	q	1
e	12	r	6
f	2	s	4
g	3	t	6
h	2	u	4
i	9	v	2
j	1	w	2
k	1	x	1
l	4	y	2
m	2	z	1

Appendix B - Words With Friends Distribution			
	Frequency		**Frequency**
a	9	n	5
b	2	o	8
c	2	p	2
d	5	q	1
e	12	r	6
f	2	s	5
g	3	t	7
h	4	u	4
i	8	v	2
j	1	w	2
k	1	x	1
l	4	y	2
m	2	z	1

Appendix C - Unrestricted Data Sample				
	7 Habits	**Angels and Demons**	**Brief History of Time**	**Catcher in the Rye**
a	36949	56913	22804	23752
b	6337	9639	4774	4707
c	15255	20227	9072	5593
d	17542	33462	9287	14178
e	63143	87234	37720	31869
f	10545	13238	6358	4625
g	9459	17248	4966	7063
h	23103	42696	15702	18157
i	38199	49986	20917	21202
j	507	656	270	574
k	2959	6321	1412	4084
l	19149	32321	12098	13255
m	12040	16450	6360	7035
n	36507	49762	18884	18500
o	38730	52318	20977	23034
p	11750	12076	5445	3626
q	592	568	409	192
r	28912	40393	17131	12985
s	29808	41925	17665	16657
t	45768	61894	29176	27739
u	15464	18146	7938	9079
v	5927	7440	4042	2569
w	10110	14277	5714	7825
x	881	1061	783	323
y	11080	10906	4817	8615
z	454	860	239	236
TOTAL	491170	698017	284960	287474

This is a segment of the raw data collected for 4 of the 48 books sampled. All the letters in the book are counted for this set of data and the results produced from the analysis of this data is equal to the values of *Table 1*.

Appendix D - Number of Characters Restricted Data Sample				
	7 Habits	**Angels and Demons**	**Brief History of Time**	**Catcher in the Rye**
a	3649	3720	3679	3841
b	543	652	742	713
c	1425	1264	1325	938
d	1699	2211	1547	2165
e	5840	5709	6180	4925
f	979	884	1040	754
g	866	1123	747	1182
h	2299	2660	2621	2868
i	3439	3201	3258	3313
j	36	52	51	90
k	269	443	163	643
l	1815	2296	1823	2015
m	1129	1117	953	1145
n	3215	3264	2935	2819
o	3297	3273	3269	3644
p	1155	789	912	631
q	71	47	53	37
r	2640	2617	2572	2147
s	2814	2819	2956	2591
t	4182	3688	4917	4279
u	1390	1210	1123	1462
v	515	402	665	400
w	999	942	922	1186
x	79	94	100	65
y	951	805	749	1440
z	32	46	26	35
TOTAL	45328	45328	45328	45328

This is a segment of the raw data collected for 4 of the 48 books sampled. In this set of data, only the first 45328 letters are counted as that is the number of letters in the shortest of the 38 books. The results produced from the analysis of this data is equal to the values of *Table 2*.

Appendix C - Number of Letters Restricted Data Sample				
	7 Habits	**Angels and Demons**	**Brief History of Time**	**Catcher in the Rye**
a	36949	56913	22804	23752
b	6337	9639	4774	4707
c	15255	20227	9072	5593
d	17542	33462	9287	14178
e	63143	87234	37720	31869
f	10545	13238	6358	4625
g	9459	17248	4966	7063
h	23103	42696	15702	18157
i	38199	49986	20917	21202
j	**507**	**656**	**270**	**574**
k	2959	6321	1412	4084
l	19149	32321	12098	13255
m	12040	16450	6360	7035
n	36507	49762	18884	18500
o	38730	52318	20977	23034
p	11750	12076	5445	3626
q	**592**	**568**	**409**	**192**
r	28912	40393	17131	12985
s	**29808**	**41925**	**17665**	**16657**
t	45768	61894	29176	27739
u	15464	18146	7938	9079
v	5927	7440	4042	2569
w	10110	14277	5714	7825
x	**881**	**1061**	**783**	**323**
y	11080	10906	4817	8615
z	**454**	**860**	**239**	**236**
TOTAL	458928	652947	265594	T

This is also a segment of the raw data collected for 4 of the 48 books sampled. The Totals presented here are the total unfixed letters which is equal to the total characters minus the sum of the fixed characters that are highlighted. The results produced from the analysis of this data is equal to the values of *Table 1*.

Bibliography

Works Cited

"List of Best-Selling Books." *Wikipedia*, Wikimedia Foundation, 15 June 2018, en.wikipedia.org/wiki/List_of_best-selling_books.

"OceanofPDF." *OceanofPDF*, 18 June 2018, oceanofpdf.com/.

Books Used

1. Harry Potter Philosopher's Stone
2. The Lord of the Rings
3. The Lion the Witch and the Wardrobe
4. The Catcher in the Rye
5. The Bridges of Madison County
6. Watership Down
7. Charlotte's Web
8. To Kill a Mockingbird
9. Angels and Demons
10. Kane and Abel
11. The Thorn Birds
12. The Kite Runner
13. The Lost Symbol
14. Who Moved My Cheese?
15. The 7 Habits of Highly Effective People
16. The Celestine Prophecy
17. The Hunger Games
18. The Fault in our Stars
19. The Shack
20. The Godfather
21. Love Story
22. Gone Girl
23. The Girl on the Train
24. Things Fall Apart
25. Jaws
26. The Power of Positive Thinking
27. The Secret
28. Dune
29. Charlie and the Chocolate Factory
30. The Naked Ape
31. The Outsiders
32. Shogun
33. The Pillars of the Earth
34. Interpreter of Maladies
35. The Hitchhiker's Guide to the Galaxy
36. Tuesdays with Morrie
37. Wrinkle in Time
38. Long Walk to Freedom
39. The Old Man and the Sea
40. Me Before You
41. The Exorcist
42. Eye of the Needle
43. Brief History of Time
44. The Lovely Bones
45. Wild Swans
46. What Color is your Parachute
47. Life of Pi
48. The Giver