The calculation model for average scores to analyze the variation of Wordle game

Gexuan Zhu[1, a], Haorong Sun[2, b]and Xuran Wang[2, c]

[1,] Southern University of Science and Technology, Shenzhen, 518000, China

[2,] The Ohio State University, Columbus, Ohio,43210, United States of America

[2,] University of Toronto, Toronto, Ontario, M5S 1A1, Canada

[a] 2363396816@qq.com; [b]13589260027@163.com, [c] 2859255125tank@gmail.com

Abstract. Wordle is a very popular puzzle game offered by The New York Times. Every day, the puzzle editor provides a five-letter word as a riddle. we regard the associated percentages of (1, 2, 3, 4, 5, 6, X) as statistical probabilities and propose a calculation model for average scores to analyze the variation of reported results. At the same time, we establish a probability model based on normal distribution to estimate the reported scores. However, the reported scores are mainly related to the difficulty of the solution word. To address this, we discuss the attributes related to the difficulty of the solution word and establish index variables. Based on the index variables we propose, we establish a linear regression model to predict the reported scores.

Keywords: associated percentages, probability model, linear regression model, reported scores prediction model.

1. Introduction

Predicting Wordle results is to predict the scores of all players. Wordle [1] is a very popular puzzle game software which is offered by the New York Times daily newspaper. It is meaningful to help the editor prepare for the puzzle in each day. As we know that the popularity of the game is mainly related to the difficulty. On the one hand, the puzzle game should be in a certain difficulty to keep it challenging and attractive to the players.

Word puzzle is an interesting game that attracts many people focus on this game as shown in Fig. 1. Crossword puzzle is first proposed by Arthur Wynne [3] in 1913. It was named as Word-Cross puzzle. Although more than 100 years have passed, we are enjoying various puzzles and clues spawned by that “fun”-filled grid [3]. The Wordle game is developed by software engineer Josh Wardle. His companion, Palak Shah, especially likes to play crossword puzzles. In order to let her spend time during the epidemic of COVID-19, Wardle developed this game. The game was then brought by the New York Times and became popular in various main social network platform. The popularity of Wordle keeps increasing. The game is now promoted to many countries in more than 60 languages of different versions. The editor of New York Times set a five-letter word as the solution of the puzzle of Wordle every day.

2. A Probability Model of Distribution of the Reported Scores

In this section, we study on the change of the number of reported results. The mathematical model for scores variation is to establish a function to describe the relationship between the scores and features of the words’ attributes, including the number of repeat letters, the number of vowel letters, frequency of using, the first letter, and the last letter. We establish a discrete normal distribution-based model to express the probability of different scores. In other word, we use the probability of different scores to predict the associated percentages of (1, 2, 3, 4, 5, 6, X). Besides, we also built a linear regression-based prediction model based on the features of solution words, which are related to the attributes effecting the reported scores.

The reported score of Wordle game is the number of times that a player cost to guess the solution word. As is well known, players' scores are related to their ability to solve word puzzles. Each player's

ability is different, with varying levels of skill. Most players have average ability, while the number of players with high or low ability is relatively small. In other words, there are relatively few players with high or low scores, and most players' scores are concentrated around a certain average level. Therefore, the reported scores have a clear normal distribution feature, as shown in Fig. 1. Fig. 1 shows the distribution of reported scores for December 31, 2022, on Twitter. We can clearly see that the reported scores on this day have a clear normal distribution feature. Therefore, we assume that reported scores of a specific solution word follow a normal distribution.

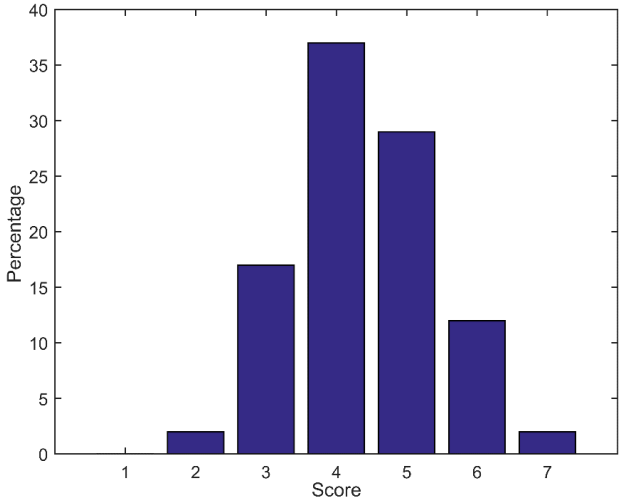

Fig. 1 Example of percentage of different scores for December 31, 2022 to Twitter [3]

Based on the above normal distribution assumption, we establish a probability model for the reported scores. We define x denotes the number of times a player guessed the solution word. We use a discrete normal distribution model to represent the distribution of scores, as shown in formula (1):

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Whereμdenote the mean value of reported score. σ^2denotes variance of reported scores.

Therefore, we can use sample data to estimate the sample mean and variance of the distribution for each solution word. It is noteworthy that score X represents the percentage of players that could not solve the puzzle. Therefore, the data of score X could not be used to estimate the mean and variance of score. The percentage of different score are treated as the probability of one player guessed solution word to the score. Therefore, we estimate the mean and variance of score by Equation (2) and Equation (3):

$$\bar{x} = \frac{\sum_{i=1}^{6} p_i x_i}{\sum_{i=1}^{6} p_i} \tag{2}$$

$$s^2 = \frac{\sum_{i=1}^{6} p_i (x_i - \bar{x}_i)^2}{\sum_{i=1}^{6} p_i} \tag{3}$$

New York Times has provided a dataset on scores results of Wordle players. This dataset is collected from the reported results that the players shared on Twitter. The dataset includes a total of 359 days of data, comprising date， contest number, word of the day, the number of players reporting their scores that day, the number of players on hard mode, and the percentage that players guessed the word in one try, two tries, three tries, four tries, five tries, six tries, fail solving the puzzle.

Based on the given data, we first calculated the sample's mean, as shown in Figure 2. The figure presents a sequence of the average score over time for 359 days, where the horizontal axis represents the number of days and the vertical axis represents the values of the average score.

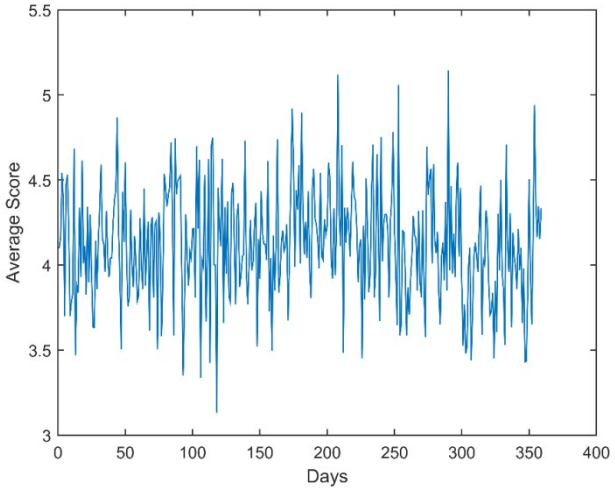

Fig. 2 Average scores from 2022/01/07 to 2022/12/31

As shown in Fig. 2, the reported scores exhibit significant fluctuations. In addition, there is no clear trend in the average scores over time. Therefore, we first assume that the average score of the solution word is random and follows a normal distribution, as shown in Equation (4):

$$p(\mu_i) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} e^{-\frac{(\mu_i-\tilde{\mu})^2}{2\tilde{\sigma}^2}} \tag{4}$$

Where μ_i denotes the mean value of the reported scores on the i-th day. $(\tilde{\mu}, \tilde{\sigma}^2)$is the kernel of the distribution.

Therefore, we can conclude that the variation of reported scores is random and independent of time. Then, we predict the number of reported result with the probability model in Equation (1)~(4). The prediction interval of mean value μ_i is as in Equation (5):

$$\mu_i \in [\tilde{\mu} - 3\tilde{\sigma}, \tilde{\mu} + 3\tilde{\sigma}] \tag{5}$$

Substitute the given dataset to calculate the mean value and variance of the average scores as shown in Table 1.

Table 1 Mean value and variance of the average scores

	Value
$\tilde{\mu}$	4.1203
$\tilde{\sigma}$	0.3389

Besides, we calculate the mean value of variance of reported scores $\bar{s}$ = 1.0522. Finally, we predict the number of reported results as shown in Table 2.

Table 2 Prediction of the number of reported results

	1 try	2 tries	3 tries	4 tries	5 tries	6 tries	X
Prediction Scores	0	5	22	32	25	10	7
Prediction interval	[0, 5]	[0,22]	[2,38]	[8,27]	[8,15]	[1,11]	[0,3]

3. Attributes of the Solution Word of Wordle Game

We analyzed the variation of reported scores in the previous section and used a probability model to predict the distribution of the percentage of reported scores. Through data analysis and observation, we can see that the variation of reported scores is independent of time. This is obvious because the solution words for each day are independent of each other, and reported scores are related to the difficulty of the solution words. Therefore, we will use the attributes of the solution words to predict their difficulty and thereby achieve the prediction of reported scores.

Generally, the words given by wordle each day are random and have no pattern. the distribution of people can guess the word in how many tries is not related to time. We consider that the distribution is related to the following factors:

Number of Repeat Letter. As the name suggests, Repeat Letter refers to the letters that appear two or more times in a word. The existence of repeated letters in a word affects its uniqueness, making it less likely to be guessed. According to the rules of the Wordle game, the fewer repeated letters in a word, the greater the probability that the letter elements in the solution word will be guessed. The more letters that are guessed, the greater the probability that the solution word will be guessed. Therefore, the number of repeat letters is an important factor that affects the scores of players. In this article, we use the number of repeated letters in each word as an important attribute feature of the word, obtained through statistical analysis.

Number of Vowels. The number of vowels, or what is also known as the number of vowel letters, is an important feature of a word in the context of Wordle. Vowel letters are the letters that make up the sounds of a language, namely the letters a, e, i, o, and u in the English language. Vowel letters play a significant role in the composition of English words, and most English words contain at least one or two vowel letters. The number of vowel letters in a five-letter word refers to how many a, e, i, o, and u are in the word. Due to the frequency of occurrence of vowel letters, words with more vowel letters may be more easily guessed. As shown in Fig. 3, we have counted the frequencies of all the letters in the given dataset. From the figure, it is clear that vowel letters appear more frequently than other letters. Therefore, the number of vowel letters is an important feature that we consider in this study.

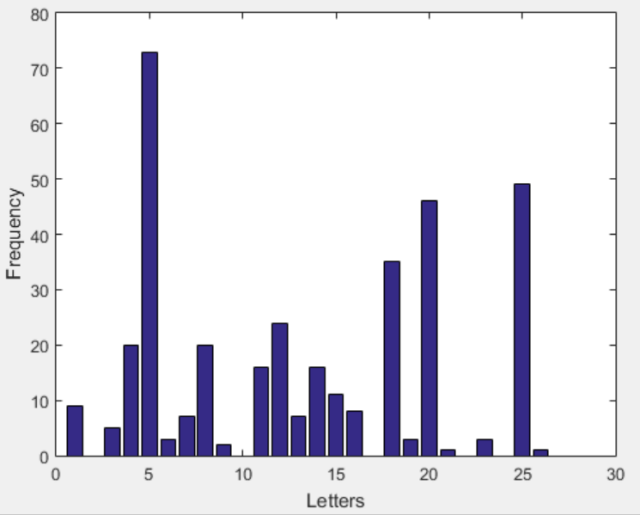

Fig. 3 Frequency of letters of given dataset

Frequency of word usage. We used the website Wordcount [4] to find the frequency of word usage. "Rank" represents the frequency of word usage of a word, and it indicates the position of the word's frequency ranking. For example, the rank of the word "other" is 71, which means that it is ranked 71st among all 86,800 words included on the website, and thus the lower the rank number, the higher the word frequency. Words that are not included on the website are represented by the number 0. In the data processing, since most of the words that are not included are rare and uncommon vocabulary, we replaced all of them with the maximum rank value in order to facilitate data analysis. To facilitate data analysis, we used a normalization method: we divided all the ranks by the maximum rank value and subtracted this value from 1 to obtain a new variable called "Frequency of word usage". The "Frequency of word usage" is a value between 0 and 1, and the closer the "Frequency of word usage" value is to 1, the higher the frequency of the corresponding word usage. In the process of guessing words, people are more likely to associate and guess the words they commonly use, so "Frequency of word usage" is considered an important attribute feature.

Frequency of First Letter. We calculated the number of times of each of the 26 letters appeared as the first letter in 359 words, as shown in Fig. 4 (1). We obtained the frequency of the first letter by dividing the number of occurrences of each letter by 359. We find out that the frequency of each letter appearing as the first letter in a word is highly variable, with letters that appear more frequently being

more likely to be discovered by people early on. Therefore, we consider the frequency of the first letter to be an important attribute feature of words.

Level of Difficulty. We used dictionary.com [5] to determine the difficulty levels of each word, and recorded them numerically, with 1 being elementary level, 2 being middle school level, 3 being high school level, 4 being college level, and 5 being post-college level. We believe that higher-level words may be a challenge for some younger players, so we consider the level of difficulty as a factor to be taken into consideration when selecting attribute features.

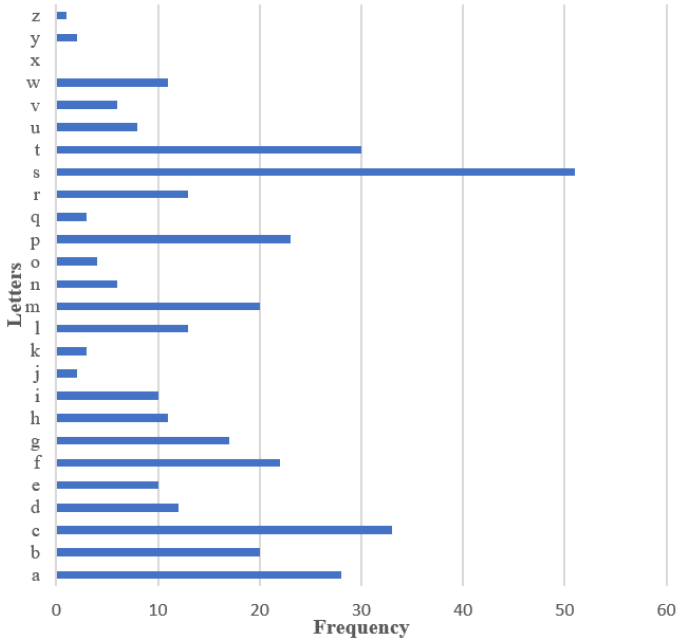

Fig. 4 Frequency of all letters in the solution words

3.1 One-Way ANOVA for Attributes of the Solution Word

3.3.1 One-Way ANOVA

Analysis of variance (ANOVA) is a statistical method used to test whether the means of multiple populations are equal, in order to determine whether a categorical independent variable has an impact on a numeric dependent variable. Here, we use ANOVA to perform statistical analysis on the variables of repeated word count and vowel count, in order to assess the level of significance of their impact on the difficulty of the solution word. The p-value in the analysis results is the main criterion for judgment. If the p-value is less than the significance level standard of 0.05, then the difference is significant; if the p-value is greater than the significance level standard of 0.05, then the difference is not significant.

3.3.2 Analysis of Variance for Number of Repeated Letters

According to the rules of the game Wordle, the fewer repeated letters a word has, the higher the probability that the letters in the solution word will be guessed. The more letters that are guessed correctly, the higher the probability of guessing the solution word. Therefore, the number of repeated letters is an important variable for measuring the difficulty level of the solution word. We conducted a statistical analysis on the number of words with different number of repeat letters in each word in the dataset:

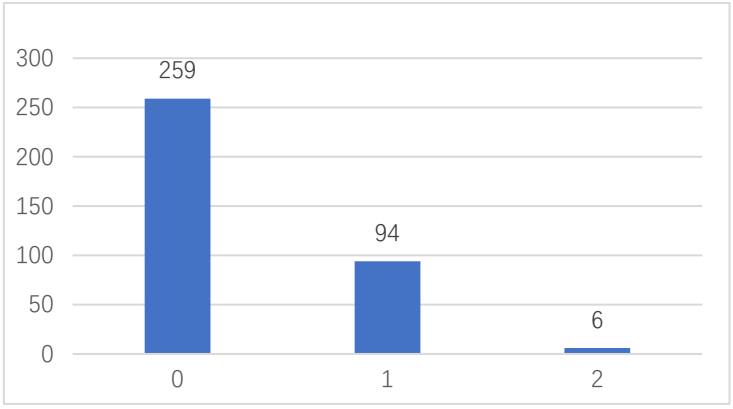

Fig. 5 Number of words with different number of repeat letters

The statistical results show that most of the words in the dataset contain 0 to 3 repeated letters. There are 259 words with no repeated letters, 94 words with one repeated letter, and 6 words with two repeated letters. The average number of repeated letters per word is 0.31, with a variance of 0.24. To further illustrate the differences in the number of repeated letters in different words, we conducted an ANOVA, the result of which is shown in Table 4.

Table 3 ANOVA of number of repeat letters

	SS	df	MS	F	P-value(F)
Group	38.3517	142	0.27008	1.87	1.66229e-05
Error	30.8976	214	0.14438		
Total	69.2493	356			

The result of the ANOVA shows that the p-value for the number of repeated letters is 1.66229×10^{-5}, which is much less than 0.05. This indicates that the difference in the number of repeated letters is significant and has a relatively large impact on the difficulty of the solution word.

3.3.3 Analysis of Variance for Number of Vowels

Due to the special nature of vowel letters, they appear more frequently in words. Therefore, according to the rules of Wordle, words that contain more vowel letters may be easier to guess. Thus, the number of vowel letters is an important variable for measuring the difficulty of the solution word. We conducted a count of the number of vowels in each word in the dataset

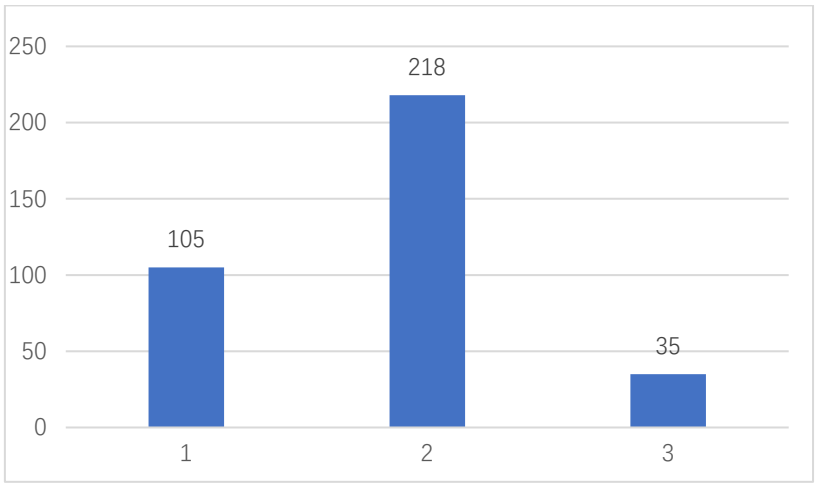

Fig. 6 Number of words with different number of vowel letters

The statistical results show that most words in the data set have 1 to 3 vowel letters. Specifically, there are 105 words with 1 vowel letter, 218 words with 2 vowel letters, and 35 words with 3 vowel letters. On average, each word contains 1.79 vowel letters with a variance of 0.35. To further

demonstrate the differences in the number of vowel letters in different words, we conducted a variance analysis, the results of which are shown in Table 4:

Table 4 ANOVA of number of vowel letters

	SS	df	MS	F	P-value(F)
Group	52.34	143	0.36601	1.05	0.3606
Error	73.935	213	0.34711		
Total	126.275	356			

The result of the ANOVA shows that the p-value for the number of vowel letters is 0.3606, which is greater than 0.05, indicating that the difference in the number of vowel letters is not significant and has a relatively small impact on the difficulty of the solution word.

3.2 Linear Regression Model for Average Score

Based on the given information, we believe that the average score as we calculated in section 3.1 is related to the factors we mentioned in section 3.2. To estimate the average score, we use an OLS model with the following equation.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon$$

y_i represents the average score of the i-th word. $\boldsymbol{x_i} = (x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$ are the variables that is going to affect the scores: x_{i1} is the number of repeated letters; x_{i2} is the number of vowels; x_{i3} is the frequency of word usage; x_{i4} is the frequency of the first letter; x_{i5} is the frequency of the last letter.

By conducting the OLS model, we get the following result:

Table 5 Coefficient result of linear function.

	coefficient	Standard error	p-value
Constant	4.472877	0.0313939	0.000
X_1	0.2776244	0.0254936	0.020
X_2	-0.0597996	0.0602675	0.000
X_3	-0.2756903	0.4067096	0.000
X_4	-1.838498	0.0745811	0.000

Table 6 ANOVA results of linear function

	Sum of square	df	Mean square	R square	F	Sig
SSR	11.2792769	4	2.81981922	0.2744	33.47	0.0000
SSE	29.8262458	354	0.084254932			
SST	41.1055226	358	0.114819896			

Based on the regression result in Table 6, we get the linear function of average score as the equation:

$$\mathrm{Y} = 4.472877 + 0.2776244\mathrm{X}_1 - 0.0597996\mathrm{X}_2 - 0.2756903\mathrm{X}_3 - 1.838498\mathrm{X}_4$$

The interpretation of coefficients for the linear model is holding other variables constant, when X_1 increase by 1 unit, the dependent variable y will increase by 0.2776244. Other dependent variable coefficients have the same meaning. Our coefficient result of our linear function is statistically significant, since all the P-value are smaller than 0.05.

To the word EERIE, we use out OLS model to estimate its average score:

In the word EERIE, there is two repeated E and four vowels. And it has a rank of 15351, that is, after transformation, its frequency of usage is $1 - \frac{15351}{77893} = 0.802922$. It also has a E as its first letter, with a frequency of occurrence of 0.027855. Therefore, its average score is calculated as:

$$4.47 + 0.278 \times 2 - 0.06 \times 4 - 0.28 \times 0.80 - 1.84 \times 0.028 = 4.52$$

With OLR model, we calculate that the average score of the word EERIE is 4.51635823. And based on the possibility model in section 3.1, we estimate the percentage of different scores as shown in Table 7.

Table 7 Prediction of the number of reported results by linear regression

	One try	Two tries	Three tries	Four tries	Five tries	Six tries	X
Prediction Scores	0	2	11	31	31	16	9

4. Conclusion

In this paper, we study the prediction of reported scores of solution word. Solution word is a non-numeric variable. Therefore, there are great difficulties in this prediction problem. In this paper, we have established a series of attribute features to effectively express the difficulty and feature of solution word. The factors that affect the reported scores of solution word are mainly related to its letter composition, order and familiarity. Although we have put forward some effective features, there are still some interesting features that are related to the diversity of solution word. Here we think there are still some other features to be considered:

References

[1] G. H. Chen, S. Nikolov, and D. Shah, A latent source model for nonparametric time series classification, in Advances in Neural Information Processing Systems, pp. 10881096, 2013.

[2] Krenker A, Bešter J, Kos A. Introduction to the Artificial Neural Networks. InTech; 2011.

[3] SHARMA S. Activation Functions in Neural Networks, 2017.

[4] Subasi A , Gursoy M I . EEG signal classification using PCA, ICA, LDA and support vector machines[J]. Expert Systems with Applications, 2010, 37(12):8659-8666.