APPLICATION OF WORD LENGTH USING THREE DISCRETE DISTRIBUTIONS (A Case Study of Students' Research Projects)

BY
OPARA JUDE
E-mail Address: judend88@yahoo.com
Department of Statistics, Imo State University
PMB 2000, Owerri Nigeria

IHEAGWARA ANDREW I.
Procurement Officer/Director Planning, Research & Statistics, Nigeria Erosion & Watershed Management Project (World Bank-Assisted), Ministry of Petroleum & Environment, Plot 36, chief Executive Quarters, Area "B", New Owerri, Imo State Nigeria

And
Onyegbuchulem B.O.
E-mail Address: besta75@yahoo.com
Department of Maths/Statistics, Imo State Polytechnic Umuagwo Ohaji

ABSTRACT
This paper examined the application of word length using three discrete distributions. The study tends to estimate word length frequency distributions of five randomly selected students' research project of the department of English and literally studies from the library project catalog of Imo State University Owerri. The five selected students' research projects were studied, and the sample sizes (number of pages) of each of the research project were computed via the Slovians' formula. Three discrete distributions such as 1-Displaced Singh-Poisson, 1-Displaced Hyper-Poisson, and 1-Displaced Geometric were clearly explained and their parameters estimated. The adequacy of the three models on the five selected students' research project was analyzed according to their goodness of fit properties. In order to test the goodness of fit of these probability models, we employed the standardized discrepancy coefficient, and the result of the analysis revealed that both the 1-Displaced Hyper-Poisson and 1-Displaced Singh-Poisson Distributions are good fit for the selected students' research project except for the case of the forth project, where it is not adequate for both distributions. It was concluded from the analysis that the 1-displaced geometric distribution model is not a good fit for all the students' research project data used in this study.

Key words: 1-Displaced Hyper-Poisson, 1-Displaced Geometric, 1-Displaced Singh-Poisson, Word Length.

Introduction

Word length has been studied by some past researchers, and it is defined by the number of letters per word according to Mendenhall (1901). Many contemporary approaches measure word length in the number of letters per word, not paying due attention to the arbitrariness of writing systems.

Thus, the least one would expect would be to count the number of sounds, or phonemes, per word; as a matter of fact, it would seem much more reasonable to measure word length in more immediate constituents of the word, such as syllables, or morphemes. Yet, even today, there are no reliable systematic studies on the influence of the measuring unit chosen or on possible interrelations between them (and if they exist, they are likely to be extremely language specific). However, we defined word length in this study as the number of syllables in a word.

Throughout the history the problem of modeling the distribution of word length was not only the interest of linguists, but also of scientists from other areas such as physics, mathematics and statistics. In 1851 the English mathematician and logician Augustus De Morgan was the first who already pointed out the relevance of the length of a linguistic unit. He mentioned word length as a possible style characteristic which may be helpful as an indicator in determining authorship (Lord, 1958). Several other scientists have dealt with the same topic counting even the frequency with which words of a given length occur in a text. Using mostly graphically methods to represent the results obtained, they noticed that the word length is not only a characteristic feature of the individual style of an author, as De Morgan stated, but also a characteristics of e.g. a certain genre.

Related Literature Review

Narisong et al (2014) worked on word length distribution in Mongolian. The study addressed the distribution features of word length and stem length in Mongolian, employing both dynamic (a corpus of 1 million Mongolian word tokens) and static (an orthographic Mongolian dictionary and a Mongolian stem dictionary) language resources. The results showed that the Mongolian words and stems abide by the Poisson distribution. Concretely, the word lengths from the dynamic corpus abide by the Dacey-Poisson distribution, and all the others abide by the Conway-Maxwell-Poisson distribution. In addition, the Mongolian word lengths are influenced by word frequencies, basically abiding by Zipf's Principle of Least Effort. The fitting experiments of power functions relationship between Mongolian word lengths and word frequencies using individual short texts, continuous long texts, and fixed-length texts indicate that the individual texts with fixed length (about 2000 words) yield the best fitting results.

Pande and Dhami (2013) researched on Analysis for the significance of statistical word-length features in genre discrimination of Hindi texts. In automatic text categorization procedure, quantifiable features' information is extracted from a text and on the basis of the information the text is sorted as a category. This information consists of values of set of one or more measurements, where the measurements can be considered as frequencies or function of frequencies of linguistic elements. In the process of text classification and genre discrimination, the role of the systematic study of word length and the analyses of word-length statistics of different texts has been established by researchers for various languages. In their work, an attempt was made to test the contribution of quantitative word length features in classification of written texts of Hindi Language by extracting quantitative measures with the help of word length profiles and frequencies. Classificatory tasks in computational linguistics are mainly performed on the basis of text genre detection and authorship attribution. Genre determination of text usually refers to identification of the kind of the text. Research articles, news articles, court

decisions, home pages, poems, novels are some examples of genres of texts. Genre discrimination practices have applications in many natural language processing tasks.

Kalimeri et al (2014) researched on Entropy analysis of word-length series of natural language texts: Effects of text language and genre. They estimated the n-gram entropies of natural language texts in word-length representation and found that these were sensitive to text language and genre. They attributed the sensitivity to changes in the probability distribution of the lengths of single words and emphasized the crucial role of the uniformity of probabilities of having words with length between five and ten. Furthermore, comparison with the entropies of shuffled data revealed the impact of word length correlations on the estimated n-gram entropies.

It is important in this paper to look at these three discrete distributions namely; 1-Displaced Hyper-Poisson, 1-Displaced Geometric and 1-Displaced Singh-Poisson distributions to know the nature of the result having reviewed past researchers' work.

Material and Methods

A set of data was collected from the five randomly selected students' research project of the department of English and literally studies from the library project catalog of Imo State University Owerri. The total pages of each research project were recorded (taken N as, the population size). The researchers determined the sample sizes by using the Slovian's formula accordingly. The Slovian's formula written as;

$$n = \frac{N}{1 + Ne^2} \qquad (1)$$

Model Detection

In determining for an appropriate model for word length frequency distributions, an ideal solution for future interpretations of the model parameters would be the existence of a unique model, appropriate for all analyzed materials of the text basis under study. Since we are concerned with words that have at least one syllable, these models will be considered to be 1-displaced. In order to test the goodness of fit of these probability models, we employ the standardized discrepancy coefficient C, where N is the text length (number of words in the text material). As an empirical rule of thumb we consider the fit of the model (a) as not appropriate in case of $C > 0.02$, (b) as sufficient if $0.01 < C \leq 0.02$, and (c) as extremely good if $C \leq 0.01$

The 1-Displaced Hyper-Poisson Distribution

The Hyper-Poisson Distribution has its pdf as;

$$p_x = \frac{a^x}{{}_1F_1(1;b;a)b^x}, x = 0,1,2,... \qquad (2)$$

Here, ${}_1F_1(1;b;a)$ is the confluent hyper-geometric function

$$ {}_1F_1(1;b;a) = \sum_{j=0}^{\infty} \frac{a^j}{b^{(j)}} = 1 + \frac{a^1}{b^{(1)}} + \frac{a^2}{b^{(2)}} + \ldots \quad (3) $$

And

$$ b^{(0)} = 1 $$

$$ b^{(j)} = b(b+1)(b+2)\ldots(b+j-1) \quad (4) $$

Since the support of Equation (2) is x = 0,1,2,… $a > 0, b > 0$, and since there are no zero-syllable words in our study data, we then concentrate with the **1-Displaced Hyper-Poisson Distribution**, which consequently takes the following dimension:

$$ p_x = \frac{a^{x-1}}{{}_1F_1(1;b;a)b^{(x-1)}}, x = 1,2,\ldots \qquad a > 0, b > 0 \quad (5) $$

Where ${}_1F_1(1;b;a)$ is defined in Equation (3) and $b^{(x-1)}$ is defined as;

$$ b^{(x-1)} = b(b+1)\ldots(b+x-2) \quad (6) $$

The mean and the variance of the 1-displaced Hyper-Poisson distribution are;

$$ \mu = E(X) = a + (1-b)(1-P_1) + 1 \quad (7) $$

$$ Var(X) = (a+1)\mu + \mu(2-\mu-b) + b - 2 \quad (8) $$

From equation (7),

$$ \hat{a} = \overline{X} - (1-\hat{b})(1-\hat{P}_1) - 1 \quad (9) $$

But $Var(X) = E(X^2) - [E(X)]^2$

From equation (8),

$$ \hat{b} = \frac{\overline{X}^2 - m_2' + \overline{X}(1+\hat{P}_1) - 2}{\hat{P}_1\overline{X} - 1} \quad (10) $$

1-Displaced Geometric Distribution

The Geometric distribution can be described as

$$ p_x = p(1-p)^x, x = 0,1,2,\ldots \quad (11) $$

Since the support of (11) is x = 0,1,2,… with $0 \le p \le 1$, $q = 1-p$, and since there are no zero-syllable words in the study data, we are concerned with the 1-displaced geometric distribution, which consequently takes the following dimension:

$$ p_x = p(1-p)^{x-1}, x = 1,2,\ldots \quad (12) $$

To get the parameter p, we can obtain it through the Maximum Likelihood Method (MLE) of (12) as;

$$ L(x_1, x_2, \ldots, x_n; p) = p(1-p)^{x_1-1}.p(1-p)^{x_2-1}\ldots p(1-p)^{x_n-1} = \prod_{i=1}^{n} p(1-p)^{x_i-1} \quad (13) $$

Taking the natural logarithm of Equation (13) to get;

$$\ln L = n\ln p + \sum_{i=1}^{n} x_i - n\ln(1-p) \quad (14)$$

Differentiating (14) w.r.t. p to get;

$$\therefore \hat{p} = \frac{1}{\bar{x}} \quad (15)$$

$$\hat{q} = 1 - \hat{p} \quad (16)$$

The 1-Displaced Singh-Poisson Distribution

The 1-Displaced Singh-Poisson Distribution introduces a new parameter α changing the relationship between the probability of the first class and the probabilities of the other classes. It is given as;

$$p_x = \begin{cases} 1-\alpha+\alpha e^{-a}, x=1 \\ \dfrac{\alpha a^{(x-1)} e^{-a}}{(x-1)!}, x = 2,3,... \end{cases} \quad (17)$$

Where $a > 0$ and $0 \le \alpha \le 1/(1-e^{-a})$.

The mean and variance can be gotten as;

$$\therefore E(X) = \bar{X} = 1 + \alpha a \quad (18)$$

$$Var(X) = \alpha a(1 + a - \alpha a) \quad (19)$$

$$\text{But } Var(X) = E\left[X(X-1)\right] + E(X) - \left[E(X)\right]^2 \quad (20)$$

Where $E[X(X-1)] = \sum_{x=2}^{\infty} X(X-1)P_x$

To obtain the parameters $\hat{a}$ and $\hat{\alpha}$, we use Equations (18) and (19), thus;

From second factorial moment, we get;

$$M_2 = E[X(X-1)]$$

$$\text{But } Var(X) = E[X(X-1)] + E(\bar{X}) - [E(\bar{X})]^2$$

From Equation (20), we get

$$Var(\bar{X}) = m_2 + \bar{X} - \bar{X}^2 \quad (21)$$

Put Equation (21) into Equation (20) to get

$$m_2 + \bar{X} - \bar{X}^2 = a\alpha(1 + a - \alpha a) \quad (22)$$

From Equation (18)

$$\hat{\alpha} = \frac{\bar{X} - 1}{a} \quad (23)$$

Put Equation (23) into Equation (22) to obtain

$$m_2 + \bar{X} - \bar{X}^2 = \frac{a(\bar{X}-1)}{a}\left[1 + a - \frac{(\bar{X}-1)}{a}a\right]$$

$$\hat{a} = \frac{m_{(2)}}{\bar{X}-1} - 2 \quad (24)$$

Put Equation (24) into Equation (23) to have;

$$\hat{\alpha} = = \frac{\left(\overline{X} - 1\right)^2}{m_{(2)} - 2\overline{X} + 2} \tag{25}$$

Where $m_{(2)}$ is an estimation of the second factorial moment $\mu_{(2)}$ where $\mu_{(2)}$ is given by;

$$\mu_{(2)} = E[(X(X-1)] = E(X^2) - E(X) \tag{26}$$

Test for Goodness-of-fit

This involves goodness-of-fit using χ^2 test. In goodness-of-test, we seek to measure how well an observed data supports an assumption about the distribution of a population or random variable of interest, i.e. how well an assumed distribution fits the data. To test the hypothesis regarding a population distribution, the chi-square test determines if the sample observations fit our expectation based on the hypothesized distribution. The test statistic is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \tag{27}$$

$p(\chi^2) =$ probability of the chi-square

$$C = \frac{\chi^2}{N} \text{ discrepancy coefficient} \tag{28}$$

where N is the text length (number of words in the text material)

Data Analysis

The three discrete distributions discussed in this paper were used to analyze the data generated for this study. For instance, the 69 selected pages studied in the first project material were used to obtain the number of syllables in each word and their respective frequencies. The collected data are presented in Table 1.

Table 1: Fitting the 1-Displaced Singh-Poisson Distribution to the first Project Material

X_i	f_i	$f_i X_i$	$f_i X_i^2$	p_x	$e_i = Np_x$	$(f_i - e_i)^2 / e_i$
1	6998	6998	6998	0.54906	6854.46504	3.005674
2	3348	6696	13392	0.29070	3629.0988	21.77305
3	1525	4575	13725	0.11949	1491.71316	0.742779
4	608	2432	9728	0.03274	408.72616	97.15567
5	8	40	200	0.00801	99.99684	84.63686
Total	12484	20741	44043			207.314

$$\overline{X} = \frac{\sum_{i=1}^{n} f_i X_i}{\sum_{i=1}^{n} f_i} = \frac{20741}{12484} = 1.66141$$

$$E(X^2) = \frac{\sum_{i=1}^{n} fX_i^2}{\sum_{i=1}^{n} f_i} = \frac{44043}{12484} = 3.52796$$

Using Equation (26);
$\therefore \mu_{(2)} = 3.52796 - 1.66141 = 1.86655$

Using Equation (24), $\hat{a} = \frac{1.86655}{1.66141 - 1} - 2 = 0.82208$

Using Equation (25), $\hat{\alpha} = \frac{(1.66141 - 1)^2}{1.86655 - 2(1.66141) + 2} = 0.80456$

Using Equation (17), we get;
$p_1 = 1 - 0.80456 + 0.80456e^{-0.82208} = 0.54906$

$p_2 = \frac{0.80456(0.82208)^1 e^{-0.82208}}{1!} = 0.29070$, $p_3 = \frac{0.80456(0.82208)^2 e^{-0.82208}}{2!} = 0.11949$, and so on.

The calculated chi-value is computed using Equation (27) as;
$\chi^2_{cal} = 207.314$

The standardized discrepancy coefficient, using Equation (27) is given by
$C = \frac{207.314}{12484} \approx 0.02$
The result suggests that the 1-displaced Singh-Poisson distribution is an adequate fit probability distribution for word lengths in the first selected Project Material.

Table 2: Fitting the 1-Displaced Hyper-Poisson Distribution for the first Project Material

X_i	f_i	$b^{(x-1)}$	a^{x-1}	${}_1F_1(1;b;a)b^{(x-1)}$	p_x	$e_i = Np_x$	$(f_i - e_i)^2/e_i$
1	6998	1	1	1.79863	0.55598	6940.85432	0.47049
2	3348	2.54727	1.34134	4.58160	0.29277	3654.94068	25.77677
3	1525	9.03585	1.79919	16.25215	0.11070	1381.9788	14.80129
4	608	41.08845	2.41333	73.90292	0.03266	407.72744	98.37233
5	8	227.92873	3.23710	409.95945	0.00789	98.49876	83.14851
Total	12484						222.5694

Using Equation (10), we get;
$\hat{b} = \frac{(1.66141)^2 - 3.52796 + 1.66141(1 + 0.56056) - 2}{1.66141(0.56056) - 1} = 2.54727$
Applying Equation (9), we obtain;
$\hat{a} = 1.66141 - (1 - 2.54727)(1 - 0.56056) - 1 = 1.34134$

It should be noted that $b^{(0)} = 1$
We shall obtain the following; using Equation (6):

$b^{(1)} = b = 2.54727$

$b^{(2)} = b(b+1) = 2.54727(3.54727) = 9.03585$

$b^{(3)} = b^{(2)}(b+2) = 9.03585(4.54727) = 41.08845$

$b^{(4)} = b^{(3)}(b+3) = 41.08845(5.54727) = 227.92873$

Applying Equation (3), we get;

$${}_1F_1(1;b;a) = 1 + \frac{1.34134}{2.54727} + \frac{1.79919}{9.03585} + \frac{2.41333}{41.08845} + \frac{3.23710}{227.92873} = 1.79863$$

Applying Equation (5),

$p_1 = 0.55598,\ p_2 = 0.29277,$

And so on.

$\chi^2_{cal} = 222.5694$

Applying Equation (27), we get;

$$C = \frac{222.5694}{12484} \approx 0.02$$

Since C = 0.02, we conclude that 1-Displaced Hyper-Poisson distribution is adequate for the first selected Project Material.

Table 3: Fitting the 1-Displaced Geometric Distribution for the first Project Material

X_i	f_i	p_x	$e_i = Np_x$	$(f_i - e_i)^2 / f_i$
1	6998	0.60190	7514.1196	35.450519
2	3348	0.23962	2991.41608	42.505652
3	1525	0.09539	1190.84876	93.762579
4	608	0.03798	474.14232	37.790085
5	8	0.02511	313.47324	297.67740
Total	12484			507.18624

Applying Equation (15), we get;

$$\hat{p} = \frac{1}{1.66141} = 0.60190$$

Applying Equation (16), w obtain

$\hat{q} = 0.39810$

Applying Equation (12), we get;

$p_1 = 0.6019(0.3981)^0 = 0.6019,\ p_2 = 0.6019(0.3981)^1 = 0.23962$

$\chi^2_{cal} = 27.32101$

Applying Equation (27), we get;

$$C = \frac{507.18624}{12484} \approx 0.04$$

Since $C > 0.02$, we conclude that 1-Displaced Hyper-Geometric distribution is not adequate for the first selected students' Project.

The same procedure was adopted for the remaining four selected students' research project and they are summarized in Tables 4, 5 and 6 respectively for 1-Displaced Singh-Poisson Distribution, 1-Displaced Hyper-Poisson Distribution, and 1-Displaced Geometric Distribution.

Table 4: Fitting the 1-Displaced Singh-Poisson Distribution to Students' Projects

Students' Projects	$\hat{\alpha}$	$\hat{a}$	C
2	0.812416	0.813214	0.02
3	0.823456	0.843264	0.01
4	0.823255	0.834543	0.03
5	0.812345	0.821233	0.02

Table 5: Fitting the 1-Displaced Hyper-Poisson Distribution to Students' Projects

Students' Projects	$\hat{b}$	$\hat{a}$	C
2	2.576565	1.43244	0.02
3	2.654322	1.54422	0.02
4	2.53432	1.23253	0.04
5	2.45768	1.35643	0.02

Table 6: Fitting the 1-Displaced Geometric Distribution to Students' Projects

Students' Projects	$\hat{p}$	$\hat{q}$	C
2	0.63245	0.36755	0.05
3	0.58997	0.41003	0.04
4	0.64869	0.35131	0.06
5	0.61343	0.38657	0.05

Conclusion

It can be seen from the analysis of data collected for this study that both the 1-Displaced Hyper-Poisson and 1-Displaced Singh-Poisson Distributions are good for the selected Students' Projects except for the case of the forth material, where it is not adequate for both distributions (See Tables 4 and 5). It can be concluded from the analysis that the 1-displaced geometric distribution model do not fit for all the project materials data used in this study.

References

Grzybek, P., Stadlober, E., Kelih, E., and Antic, G. (2005): Quantitative Text Typology: The Impact of Word Length. In: C. Weihs and W. GAUL (Eds.), Classification – The Ubiquitous Challenge. Springer, Heidelberg; 53-64.

Kalimeri M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakanos, F.K. and Papageorgiou, H. (2014). Entropy analysis of word-length series of natural language texts: Effects of text language and genre. Journal of Quantitative Linguistics. http://www.researchgate.net/publication/259893423

Lord, R. (1958). Studies in the history of probability and statistics. VIII: De Morgan and the statistical study of literary style. *Biometrika*, *45*, 282.

Mendenhall, Thomas C.(1901). "A mechanical solution of a literary problem", in: Popular Science Monthly, vol. 60, pt. 7; 97–105.

Narisong, H., Jingyang, J. and Haitao, L. (2014). Word Length Distribution in Mongolian .Journal of Quantitative Linguistics Volume 21, Issue 2, 2014.

Pande, H. and Dhami, H. S. (2013). Mathematical Modelling of the Pattern of Occurrence of Words in Different Corpora of the Hindi Language. *Journal of Quantitative Linguistics, 20:1*, 1-12