Procreation of Distribution for words

Raj Kishore Bisht & H.S.Dhami*
Dept. of Mathematics,
University of Kumaun
S.S.J.Campus Almora
(Uttaranchal), India 263601

Abstract

In the present paper an attempt has been made to find the probability distributions based on the frequency of words in a small sample of text. The test of goodness of fit of the distributions has been also worked out on exemplary basis for some selected words.

1.Introduction

The aim of Linguistic science is to be able to characterize and explain the multitude of Linguistic observation circling around us, in conversation, writings and other media. Part of it has to do with understanding the linguistic structures by which language communicates. While analyzing a text mathematically, the first question is about the distribution of words and the probability of occurrences of a word in a particular number of times in a document.

* To whom all correspondence be addressed.

E mail – drhsdhami@yahoo.com

A different source of information about terms that can be exploited in information retrieval is co-occurrence, the fact that two or more terms occur in the same document more often than chance. The models of term distribution can be embedded in information retrieval frameworks such as term weighting. In addition to information retrieval, a good understanding of distribution pattern is useful wherever we want to assess the likelihood of a certain number of occurrences of a specific word in a unit of text. The frequency of words and their behavior in a text can have applications in cryptography or to give sort of indication of style or author-ship.

A better fit to the frequency distribution of content words is provided by the two Poisson model, a mixture of two Poissons by Bookstein et al [1]. Katz et al [3] defined a K mixture for empirical word distribution which was more accurate than the Poisson distribution and the two Poisson models. Residual inverse document frequency was introduced by church et al [2]. Wong [6] has used information theory to motivate inverse document frequency. An application of residual inverse document frequency to the characterization of index terms is described by Yamamoto et al [7]. Naranan & Balasubrahmanyan [4-5] have developed information theoretical models in structural linguistics based on word frequencies and rank frequency distribution of phonems and alphabet.

We are making attempts to investigate discrete distributions based on the frequency of the words in the sample of text. While going through all existing distributions, it was observed that occurrence of the specific words in the text has been at such low level that even Poisson distribution could not be applied. The need to have a distribution which can represent a random variable which has very low frequency in a given data, lead to the search of new distributions. It was further examined that the words for whom the difference in frequencies at $x = 0$ and at $x = 1$ is less than 80 percent

(approximately) of the sample size satisfy distribution first and otherwise distribution second.

2.Methodology:-

We have compiled a corpus of 25600 words from the book of noble laureate Amertya Sen and other prominent books on poverty alleviation programmes. For the sake of simplification the corpus was divided in 256 units, each unit containing 100 words.

Notations used:-

$X = x_i$, Number of times the word 'w' appears in a unit

F_0 = Number of units possessing x_i^{th} attribute (observed frequency)

F_e = Number of units possessing x_i^{th} attribute theoretically (theoretical frequency)

3.Distributions of words:-

In this section we are presenting two type of distributions generated from power series distribution and inflated power series distribution which we think are appropriate to delineate the distribution of a random variable having very low frequency in the data. The test of goodness of fit of the distributions for different words and corresponding collocations in a text has been also worked out.

Distribution I :-

We have the generalized power series distribution defined as

$$P(X=x) = \begin{cases} \frac{a_x \theta^x}{f(\theta)}; x = 0,1,2,3.....; a_x \geq 0 \\ 0, elsewhere \end{cases}$$

where $f(\theta) = \sum_{x=0}^{\infty} a_x \theta^x, \theta \geq 0$

For distribution I, we shall set $f(\theta) = 1$ on the basis of deductive reasoning, which in turn implies that $a_x = 1 - \theta$

Thus it can be asseverated that a random variable X assuming non-negative values shall have distribution I if it has probability mass function of the form

$$P(X=x) = (1-\theta)\,\theta^x; \quad x = 0,1,2,3,\ldots\ldots\ldots$$

where 'θ' is the parameter of the distribution and $0 \leq \theta \leq 1$. We observe that this distribution is applicable when n, the number of trials is infinitely large and the probability of success for each trial is indefinitely small.

The corresponding Distribution function is

$$F(X) = P(X \leq x) = \sum_{r=0}^{x} P(r)$$

$$= (1-\theta) \sum_{r=0}^{x} \theta^r \quad ; x = 0,1,2,3,\ldots\ldots\ldots\ldots$$

We shall also have

Mean of the distribution $:= \dfrac{\theta}{(1-\theta)}$

Variance of the distribution $:= \dfrac{\theta}{(1-\theta)^2}$

Moment Generating Function $:= \dfrac{1-\theta}{1-\theta.e^t}$

Characteristic Function $:= \dfrac{1-\theta}{1-\theta.e^{it}}$

Test of goodness of fit of the distribution I to the observed data for words.

For the word 'poverty', we have following table

X	F_0	$F_0 . X$	P(X)	F_e
0	130	0	.498	127.5 = 128
1	56	56	.249	63.7 = 64
2	29	58	.125	32
3	25	75	.063	16.1 = 16
4	11	44	.032	8.19 = 8
5	3 } 16	15	.016	4.09 = 4 } 14
6	2	12	.008	1.94 = 2
	$\sum F_0$ =256	$\sum F_0.X$=260		

For which we can easily obtain

$\overline{X}$ = 1.01, θ = .502 and $\sum (F_0 - F_e)^2 / F_e$ =6.655

Since $\chi_{.05}^2$ = 7.815 (for three degree of freedom), this depicts that the distribution is of a good fit of the observed data at 5% level of significance.

For the word 'line' we observe that

X	F_0	$F_0 . X$	P(X)	F_e
0	213	0	.813	208.1 = 208
1	32	32	.152	38.9 = 39
2	8	16	.028	7.17 = 7
3	1 } 11	3	.005	1.28 = 1 } 8
4	1	4	.001	.256 = 0
5	1	5	.0002	.051 = 0
	$\sum F_0$ =256	$\sum F_0.X$=60		

For which we can easily obtain

$\overline{X}$ = 0.23, θ = 0.187 and $\sum$ $(F_0 - F_e)^2 / F_e$ =2.501

Since $\chi_{.05}^2$ = 3.841 (for one degree of freedom), hence the distribution is again a good fit for the observed data at 5% level of significance.

For the collocation ' poverty- line' we have

X	F_0	$F_0 . X$	P(X)	F_e
0	210	0	.848	217.08 =217
1	28	28	.128	32.77 = 33
2	7	14	.020	5.12 = 5
3	0 } 8	0	.003	0.78 = 1 } 6
4	1	4	.0004	0.11 = 0
	$\sum F_0$ =256	$\sum F_0.X$=46		

For which we have

$\overline{X}$ = 0.18, θ = 0.152 and $\sum$ $(F_0 - F_e)^2 / F_e$ =1.651

Since $\chi_{.05}^2$ = 3.841(for one degree of freedom), therefore this distribution is a good fit of the observed data at 5% level of significance.

Distribution II :-

An inflated power series distribution inflated at zero is given by P (X=x) = $\begin{cases} 1-\alpha+\dfrac{\alpha . a_0}{f(\theta)}; x=0 \\ \alpha \dfrac{a_x \theta^x}{f(\theta)}; x=1,2,3,.... \end{cases}$

where α (0 <α ≤ 1) is the inflation parameter.

From where we have $f(\theta) = a_0 + \sum_{x=1}^{\infty} a_x \theta^x$

In order that it could fit our conditions, let us assume $f(\theta)$=1 and

$a_x \theta^x = \dfrac{1}{(x-1)!} + \dfrac{1}{x!}$

which yields a_0= 2 (1- e), and thus we have the following definition.

A random variable X shall have distribution II if it assume non negative values and has probability mass function as

$$P(X=x) = \begin{cases} 1+\alpha-2e\alpha; x=0 \\ \alpha\left[\frac{1}{(x-1)!}+\frac{1}{x!}\right], x=1,2,3.... \end{cases}$$

here $0 < \alpha \leq 0.1$

so that the corresponding distribution function is

$$F(X) = 1+\alpha - 2e\alpha + \sum_{r=0}^{x} \alpha\left[\frac{1}{(r-1)!}+\frac{1}{r!}\right], \qquad x = 1,2,3,\ldots$$

Here we shall have

Mean of the distribution : = $3e\alpha$

Variance of the distribution : = $e\alpha\ (7 - 9e\alpha)$

Moment Generating Function:=

$$(1+\alpha-2e\alpha) + \alpha[(e^t+1)\exp e^t - 1]$$

Characteristic Function :=

$$(1+\alpha-2e\alpha) + \alpha[(e^{it}+1)\exp e^{it} - 1]$$

Test of goodness of fit of the distribution II to the observed data for words.

If we consider the distribution of the word 'entitlement' then we have

X	F_0	$F_0 . X$	P(X)	F_e
0	218	0	.824	210.9=211
1	14	14	.080	20.48= 20
2	12	24	.060	15.36 =15
3	8	24	.027	6.91 = 7
4	4 } 13	16	.008	2.13 = 2 } 10
5	1	5	.002	0.51 = 1
	$\sum F_0$ =256	$\sum F_0.X$=83		

So that $\overline{X}$ = 0.324, α = .04 and $\sum (F_0 - F_e)^2 / F_e$ = 3.53

Since $\chi_{.05}^2$ = 5.99 (2 degree of freedom), therefore this distribution is a good fit of the observed data at 5% level of significance.

If we consider the word 'exchange' then the distribution of the word in the text is

X	F_0	$F_0 . X$	P(X)	F_e
0	233	0	.912	233.4=233
1	14	14	.04	10.24= 10
2	5	10	.03	7.68 = 8
3	2 } 10	6	.013	3.14 = 3 } 12
4	1	4	.004	1.06 = 1
5	2	10	.001	0.256 = 0
	$\sum F_0$ =256	$\sum F_0.X$=44		

We have $\overline{X}$ = 0.171, α = .02 and $\sum (F_0 - F_e)^2 / F_e$ = 1.93

so that $\chi_{.05}^2$ = 3.841(for one degree of freedom) shall exhibits that the distribution is a good fit of the observed data at 5% level of significance.

And for the collocation 'exchange -entitlement', we have

X	F_0	$F_0 . X$	P(X)	F_e
0	239	0	.94	240.64 =241
1	10	10	.026	6.65 = 7
2	4	8	.019	4.99 = 5
3	2 } 7	6	.0087	2.22 = 2 } 8
4	11	4	.002	0.69 = 1
	$\sum F_0$ =256	$\sum F_0.X$=28		

$\overline{X}$ = 0.109, α = .013 and $\sum (F_0 - F_e)^2 / F_e$ = 1.426

Since $\chi_{.05}^2$ =3.841 (for one degree of freedom), therefore this distribution is a good fit of the observed data at 5% level of significance.

4.Conclusion:

In order to exhibit the test of goodness of fit we have taken a few words only on exemplary basis, though there are many other words which satisfy these distributions. Therefore we conclude that the distribution of words in a sample of text follows a geometric type of distribution and in some particular situation where the frequency is very low it follows a linear type of distribution. Since the corpus is divided in units and the distributions are independent of the number of words in a unit, thus it can be said that these distributions shall be well defined for a large corpus.

5.Acknowledgement:-

Authors are grateful to Dr. Rajiv Pandey, Head, Dept. of Statistics, K.U. S.S.J. Campus Almora for providing assistance in the preparing of the paper and also to Council of Scientific & Industrial Research (CSIR), New Delhi for providing financial assistance to carry out the research work in the form of junior research fellowship to first author.

6. REFERENCES:-

1. Bookstein, Abraham and Don R. Swanson (1975), A decision theoretic foundation for indexing, Journal of the American Society for Information Science, 26: 45-50.

2. Church, Kenneth W. and William A. Gale (1995), Poisson Mixture, Natural Language Engineering, 1: 163-190.

3. Katz, Salva M. (1996), Distribution of Content words and Phrases in text and language modelling, Natural Language Engineering, 2: 15-59.

4. Naranan & Balasubrahmanyan, V.K. (1992), Information theoretic models in statistical linguistic, Part II, Word frequencies and hierarchical structure in language-Statistical tests; Current Science, Vol 63, No. 6, 297-305.

5. Naranan & Balasubrahmanyan, V.K. (1993), Information theoretic models for frequency distribution of words and speech sound (phonems) in language, Journal of Scientific and Industrial Research, Vol 52, 728-738.

6. Wong, S.K.M. and Yao Y.Y. (1992), An information- theoretic measure of term specificity, Journal of the American Society for Information Science, 43: 54-61.

7. Yamamoto, Mikio and Kenneth W. Church (1998), Using Suffix array to compute term frequency and document frequency for all substrings in a corpus. In proceeding of the 6th workshop on very large corpora, pp. 28-237.