© 2020 | International Journal of Scientific Research in Science and Technology
IJSRST | Volume 7 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
DOI :https://doi.org/10.32628/IJSRST

# Use of Statistical tools for Style by Marker Word

**Dr. Ashok Y. Tayade**

Assistant Professor, Department of Statistics, Dr. B. A. M. University, Aurangabad-431004, Maharashtra State, India

## ABSTRACT

In this artical our contribution is towards statistical stylistics, a discipline established in the recent years. We have analysed samples from the book entitled, 'Glimpses of World History' (1931) by Pandit Jawaharlal Nehru. Our attempt is to study author's writing style with the help of marker words. A method of selecting statistical model for the literary style has been described, and different methods of estimation of parameters appearing in the models have been discussed.

Keywords :Marker Words, Statistical models, Kolmogoroff-Smirnoff test, Chi-square test.

## I. INTRODUCTION

An author has his / her own style of writing. A person who has examined some work of an author is able to describe author's style as reflected in features of his writing. These features which do not depend on the subject matter of writing but on the habit and personality of author are parts of author's style. Moreover these features persist from one work of an author to his / her another work.

A statistical finds in study of texts, an interesting material to use for testing and refining the distributional formulae. The statistical investigations of texts are most directly concerned with the description and explanation of the features inherent in the text, their organization and variability. We have formulated a statistical model which reflects some of the features inherent in an author's style. This statistical model is employed to describe style.

## II. METHODS AND MATERIAL

### 2.1 Marker words

Mosteller and Wallace (1964) have considered the problem of stylistics from many view points. In their book they have pointed out the existence of marker words, which were brought to their noticed by Douglass Adair. Generally marker words for English Language, Deway (1923) collected near about one lakh words. The percentages in descending order of frequencies of the first few words noted by Deway (1923) are as folloes:

**TABLE NO: 2.1.1**

Deway's Table

| Word | Deway'sValue(%) |
|---|---|
| the | 7.3 |
| of | 4.9 |
| and | 3.3 |
| to | 2.9 |
| a | 2.1 |
| in | 2.1 |

With the help of above table, we have selected the most frequent words, namely 'the' and 'of'. We have made an attempt of describing author's style on the basis of distribution of marker words. Shende and PrabhuAjgaonkar (1988), Muthe and Prabhu-Ajgaonkar (1999) have made attempts in this direction.

## 2.2. Methodology and Procedure:

For studying the writing style of the former Prime Minister Pandit Jawaharlal Nehru, we selected his book entitled, "Glimpses of World History" (1931).

This book contains letters from a father to his daughter. The volume was written in prison, which contained a rambling account of History for the young people. The volume contains a survey of world history from the earliest times to our own day, or to be more exact, till the summer of 1933. The book is primarily meant for the young, but it has plenty of interest and instruction even to the grown up people. Those who are eager to understand the tangled web of present day world affairs, and especially those Indians who desire to see their own national struggle in proper perspective with the world as its background, will find the book eminently useful. The book is not the work of a historian, but that of a modern thinker of history and world events.

For the sake of comparison we have selected two samples from this book. For both the samples we consider 30,000 consecutive words, being size of each sample. For collection of data it is noted that each block contains 100 consecutive words. In this way we constructed 300 blocks. Further we counted for each block frequencies of the marker words 'the' as well as 'of'. For the first sample we considered the first hundred and one pages of this book, for the second sample we considered the material between the pages 521, fifth line, third word and the end of the page number 621. Consider the first block of 100 consecutive words starting from the words "Do" on the first page of the sample I. In this block 'the' occurs three times, and the word 'of' occurs six times. This is the first observation. Next consider the next consecutive block of 100 words. The word 'the' occurs 'seven' times in this block, similarly the word 'of' occurs 'twenty six' times. This is the second observation. In this way we obtained the first and the second sample. The frequencies of marker words among these blocks were counted and these are shown in table 2.2.1 to 2.2.4

**Table showing frequencies of Markers words**

**TABLE NO : 2.2.1**

| Sample-I : 'the' | |
|---|---|
| $x_i$ | $F_i$ |
| 0 | 3 |
| 1 | 7 |
| 2 | 25 |
| 3 | 30 |
| 4 | 39 |
| 5 | 49 |
| 6 | 42 |
| 7 | 30 |
| 8 | 29 |
| 9 | 18 |
| 10 | 13 |
| 11 | 08 |
| 12 | 02 |
| 13 | 03 |
| 14 | 01 |
| 15 | 00 |
| 16 | 01 |
| 17 | 00 |
| 18 | 00 |
| Total | 300 |

TABLE NO: 2.2.2

| Sample-II : 'the' | |
|---|---|
| $x_i$ | $F_i$ |
| 0 | 1 |
| 1 | 2 |
| 2 | 5 |
| 3 | 15 |
| 4 | 21 |
| 5 | 26 |
| 6 | 35 |
| 7 | 46 |
| 8 | 31 |
| 9 | 28 |
| 10 | 21 |
| 11 | 20 |
| 12 | 18 |
| 13 | 11 |
| 14 | 09 |
| 15 | 03 |
| 16 | 04 |
| 17 | 01 |
| 18 | 01 |
| Total | 298 |

TABLE NO: 2.2.3

| Sample-I : 'of ' | |
|---|---|
| $x_i$ | $F_i$ |
| 0 | 06 |
| 1 | 26 |
| 2 | 37 |
| 3 | 38 |
| 4 | 61 |
| 5 | 58 |
| 6 | 35 |
| 7 | 22 |
| 8 | 10 |
| 9 | 04 |
| 10 | 01 |
| 11 | 00 |
| 12 | 00 |
| 13 | 00 |
| 14 | 00 |
| 15 | 00 |
| 16 | 01 |
| 17 | 01 |
| 18 | 00 |
| Total | 300 |

TABLE NO: 2.2.4

| Sample-II : 'of ' | |
|---|---|
| $x_i$ | $F_i$ |
| 0 | 00 |
| 1 | 12 |
| 2 | 42 |
| 3 | 52 |
| 4 | 55 |
| 5 | 68 |
| 6 | 28 |
| 7 | 19 |
| 8 | 15 |
| 9 | 06 |
| 10 | 02 |
| 11 | 00 |
| 12 | 00 |
| 13 | 00 |
| 14 | 00 |
| 15 | 00 |
| 16 | 00 |
| 17 | 00 |
| 18 | 00 |
| Total | 299 |

Next we have examined whether these samples are homogeneous.

**2.3 K-S test for Homogeneity:**

If tow random samples are drawn from the same population, then with high probability their distribution functions will be the same. It is describe to test the hypothesis that the distribution functions of the characteristic of interest in the two given populations are identical. And this has to be carried out without specifying form of the distribution.

The non-parametric test termed Kolmogoroff (1933) (Sach 1984) Smirnoff (1939) (Sach 1984) is applied to test whether the two samples are drawn from the same population. This test is distribution free test and depends on the sample size only. It is a sharpest homogeneity test and covers all sorts of differences in shapes of the distributions.

The test is, $\widehat{D} = Max.\left|\frac{F_1}{n_1} - \frac{F_2}{n_2}\right|$

Here the cumulative frequencies $F_1$ and $F_2$ (with equal class limit for both the samples) are divided by the corresponding sample size $n_1$ and $n_2$. Then the differences $(F_1/n_1 - F_2/n_2)$ are computed at regular intervals. The maximum of the absolute differences furnishes the test statistics D.

The critical value of D can be approximated for medium to large samples of size ( $n_1 + n_2$ ) > 35, by

$$\widehat{D_\alpha} = K_{(\alpha)}.\sqrt{\frac{n_1 + n_2}{n_1 n_2}}$$

Where $K_{(\alpha)}$ is a constant depending on the level of significance.

If the values of $\widehat{D}$ determined from the two samples equals or exceeds the critical value of $D_{(\alpha)}$ then a significance difference exists between the distributions of the two populations.

**'The' marker word:**

First examine the marker word 'The' only. Observed frequencies of both samples of 'The' marker word with equal class limits are shown in the following table

**TABLE NO: 2.3.1**

| $X_i$ | $F_1$ | $F_2$ | Cumu. $F_1$ | Cumu. $F_2$ | $\left\|\frac{F_1}{n_1} - \frac{F_2}{n_2}\right\|$ |
|---|---|---|---|---|---|
| 0 | 03 | 01 | 03 | 01 | 0.0066 |
| 1 | 07 | 02 | 10 | 03 | 0.0233 |
| 2 | 25 | 05 | 35 | 08 | 0.0898 |
| 3 | 30 | 15 | 65 | 23 | 0.1395 |
| 4 | 39 | 21 | 104 | 44 | 0.1990 |
| 5 | 49 | 26 | 153 | 70 | 0.2751 |
| 6 | 42 | 35 | 195 | 105 | **0.2976** |
| 7 | 30 | 46 | 225 | 151 | 0.2433 |
| 8 | 29 | 31 | 254 | 182 | 0.2359 |
| 9 | 18 | 28 | 272 | 210 | 0.2020 |
| 10 | 13 | 21 | 285 | 231 | 0.1748 |

| | | | | | |
|---|---|---|---|---|---|
| 11 | 08 | 20 | 293 | 251 | 0.1344 |
| 12 | 02 | 18 | 295 | 269 | 0.0806 |
| 13 | 03 | 11 | 298 | 280 | 0.0537 |
| 14 | 01 | 09 | 299 | 289 | 0.0302 |
| 15 | 00 | 03 | 299 | 292 | 0.0201 |
| 16 | 01 | 04 | 300 | 296 | 0.0100 |
| 17 | 00 | 01 | 300 | 297 | 0.0067 |
| 18 | 00 | 01 | 300 | 298 | 0.0333 |
| Total | 300 | 298 | -- | -- | -- |

$$\therefore \widehat{D_{cal.}} = Max.\left|\frac{F_1}{n_1} - \frac{F_2}{n_2}\right| = 0.2976$$

$$\widehat{D_{tab.}} = K_{(\alpha)}.\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.26\sqrt{\frac{300 + 298}{300 \times 298}}$$

= 0.103051 at 5% level of significance

$$\therefore \widehat{D_{cal.}} > \widehat{D_{tab.}}$$

Therefore we reject the hypothesis that the samples are drawn from the same population.

Distribution of the marker word 'the' in the blocks of consecutive 100 words in the first hundred pages differs from that observed in the last hundred pages of the book.

Next we carry out the homogeneity test for the marker word 'of' **'OF' marker word:**

Here we examine the marker word 'OF' only. Observed frequencies of both samples of 'of' marker word, with equal class limit are shown in the following table:

**TABLE NO: 2.3.2**

| $X_i$ | $F_1$ | $F_2$ | Cumu. $F_1$ | Cumu. $F_2$ | $\lvert F_1/n_1 - F_2/n_2\rvert$ |
|---|---|---|---|---|---|
| 0 | 6 | 0 | 6 | 0 | 0.0200 |
| 1 | 26 | 12 | 32 | 12 | **0.0665** |
| 2 | 37 | 42 | 69 | 54 | 0.0494 |
| 3 | 38 | 52 | 107 | 106 | 0.00215 |
| 4 | 61 | 55 | 168 | 161 | 0.0215 |
| 5 | 58 | 68 | 226 | 229 | 0.0215 |
| 6 | 35 | 28 | 261 | 257 | 0.0105 |
| 7 | 22 | 19 | 283 | 276 | 0.0202 |
| 8 | 10 | 15 | 293 | 291 | 0.0034 |
| 9 | 04 | 06 | 297 | 297 | 0.0033 |
| 10 | 01 | 02 | 298 | 299 | 0.0067 |
| 11 | 0 | 0 | 298 | 299 | 0.0067 |
| 12 | 0 | 0 | 298 | 299 | 0.0067 |
| 13 | 0 | 0 | 298 | 299 | 0.0067 |
| 14 | 0 | 0 | 298 | 299 | 0.0067 |
| 15 | 0 | 0 | 298 | 299 | 0.0067 |
| 16 | 1 | 0 | 299 | 299 | 0.0033 |
| 17 | 1 | 0 | 300 | 299 | 0.0000 |
| 18 | 0 | 0 | 300 | 299 | 0.0000 |
| Total | 300 | 299 | | | |

$$\therefore \widehat{D_{cal.}} = Max. \left|\frac{F_1}{n_1} - \frac{F_2}{n_2}\right|$$

$$\widehat{D_{tab.}} = 1.26 \sqrt{\frac{300 + 299}{300 \times 299}}$$

$$= 0.0665$$

$$\therefore \widehat{D_{cal.}} < \widehat{D_{tab.}}$$

Therefore we cannot reject the hypothesis that the samples are drawn from the same population at 5% level of significance.

## 2.4 Models

The statistical models which we consider for fitting to the data are unimodal. Therefore we considered the data where in the random variable x assumes values 0,1,2,…….. up to 12. Because when x=12, $F_{12}$ is 2, however when x=13, $F_{13}$ is 3. The frequency function goes on increasing. Therefore we consider values of the random variable x up to 12.

On the same lines other frequency functions are considered and frequencies are as follows:

TABLE NO: 2.4.1

| Sample-I | |
|---|---|
| 'The' marker word | |
| $X_i$ | $F_i$ |
| 0 | 3 |
| 1 | 7 |
| 2 | 25 |
| 3 | 30 |
| 4 | 39 |
| 5 | 49 |
| 6 | 42 |
| 7 | 30 |
| 8 | 29 |
| 9 | 13 |
| 10 | 18 |
| 11 | 8 |
| 12 | 2 |
| Total | 295 |

TABLE No: 2.4.2

| Sample-II | |
|---|---|
| 'The' marker word | |
| $X_i$ | $F_i$ |
| 0 | 1 |
| 1 | 2 |
| 2 | 5 |
| 3 | 15 |
| 4 | 21 |
| 5 | 26 |
| 6 | 35 |
| 7 | 46 |
| 8 | 31 |
| 9 | 28 |
| 10 | 21 |
| 11 | 20 |
| 12 | 18 |
| 13 | 11 |
| 14 | 9 |
| 15 | 3 |
| Total | 292 |

TABLE NO: 2.4.3

| Sample-I | |
|---|---|
| 'OF' marker word | |
| $X_i$ | $F_i$ |
| 0 | 6 |
| 1 | 26 |
| 2 | 37 |
| 3 | 38 |
| 4 | 61 |
| 5 | 58 |
| 6 | 35 |
| 7 | 22 |
| 8 | 10 |
| 9 | 4 |
| 10 | 1 |
| Total | 298 |

TABLE NO: 2.4.4

| Sample-II | |
|---|---|
| 'OF' marker word | |
| $X_i$ | $F_i$ |
| 0 | 0 |
| 1 | 12 |
| 2 | 42 |
| 3 | 52 |
| 4 | 55 |
| 5 | 68 |
| 6 | 28 |
| 7 | 19 |
| 8 | 15 |
| 9 | 6 |
| 10 | 2 |
| Total | 299 |

## 2.5. Marker word: 'The' :

**Statistical model for data of 'the' marker word:**

Consider 'the' marker word from both the samples. An important problem here is to formulate a statistical model. For literary style of an author, there exist several models. The problem of formulating an appropriate model is an important one. Generally, it is noted that for,

(1) Poisson distribution, $\mu_1' = \mu_2$

(2) Binomial distribution, $\mu_1' > \mu_2$

(3) Negative Binomial distribution, $\mu_1' < \mu_2$.

The terms $\mu_1'$ and $\mu_2$ are respectively the mean and variance of the population, these are not known. We have samples values. Based on these samples values we obtained estimates of $\mu_1'$ and $\mu_2$. These are $\overline{x}$ and $s^2$ respectively. Since these are based on sample values, $\overline{x}$ and $s^2$ are random variables assuming different values for different samples. Based on these sample values we apply a test to infer regarding ratio of population values $\mu_1'$ and $\mu_2$. This is due to Katz (1963).

## 2.6. Katz's criterion:

This criterion shows whether the data follows the Poisson probability law or the Negative Binomial Probability law or the Binomial Probability Law.

Let

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

And $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2$

Consider the statistic,

$$z = \frac{s^2 - \overline{x}}{\overline{x}}$$

We have,

$E(Z) = \frac{\mu_2}{\mu_1'} - 1$, approximately

And $V(Z) = \frac{2}{n}$, approximately

Futher for large n,

The statistics,

$$y = \frac{Z - E(Z)}{\sqrt{\frac{2}{n}}}$$

is distributed as the standard normal random variable.

**Fitting of Poisson distribution where $\mu_1' = \mu_2$:**

**Case: (A) When $s^2 > \overline{x}$,**

From the theory of testing of hypothesis, we have

$H_0: \frac{\mu_2}{\mu_1'} = 1$ against $H_1: \frac{\mu_2}{\mu_1'} > 1$ as $s^2 > \overline{x}$.

under $H_0, E(z) \div 0$ and the statistic y reduce to

$$y = \frac{Z}{\sqrt{\frac{2}{n}}} \div N(0,1) \quad \text{for large n.}$$

Let $y_0$ be the calculated value of y.

$H_0$ is rejected at 5% level of significane,

If $y_0 > 1.645$.

$H_0$ is accepted at 5% level of significance,

If $y_0 < 1.645$.

This is one sided test.

**Case: (B) When $s^2 < \overline{x}$**

From the theory of testing hypothesis, we have,

$H_0: \frac{\mu_2}{\mu_1'} = 1$ Against $H_1: \frac{\mu_2}{\mu_1'} < 1$ as $s^2 < \overline{x}$.

Under $H_0, E(z) = 0$ and the statistic

$y = \frac{Z}{\sqrt{\frac{2}{n}}} \div N(0,1)$for large n.

Let $y_0$ be the calculated value of y.

$H_0$is rejected at 5% level of significance,

If $y_0 < -1.645$.

$H_0$is accepted at 5% level of significance,

If $y_0 > -1.645$.

This is one sided test.

### 2.7. Sample – I :

$s^2 = 6.41992505$

$\overline{x} = 5.579661016$

$z = \frac{s^2 - \overline{x}}{\overline{x}} = 0.1506$

$V(z) = \frac{2}{n} = \frac{2}{295} = 0.0068$

$\therefore\ y_0 = \frac{z}{\sqrt{\frac{2}{n}}} = \frac{0.1505941}{0.0823359} = 1.8289$

$\therefore\ y_0 = 1.8289$

Here $s^2 > \overline{x}$,therefore we refer to case A.

We have,

$H_0: \frac{\mu_2}{\mu_1'} = 1$ Against $H_1: \frac{\mu_2}{\mu_1'} > 1$,

Here $y_0 > 1.645$.

So, it is significant.

$\therefore\ H_0$ is rejected at 5% level of significance.

Therefore $H_1$ is accepted.

i.e. A Negative Binomial Distribution is to be considered.

### 2.8. Sample – II :

$s^2 = 9.523116$

$\overline{x} = 7.75$

$z = \frac{s^2 - \overline{x}}{\overline{x}} = 0.228789$

$V(z) = \frac{2}{n} = \frac{2}{292} = 0.006849315$

$\therefore\ y_0 = \frac{z}{\sqrt{\frac{2}{n}}} = \frac{0.228789}{0.0827605} = 2.76447$

Here, $s^2 > \overline{x}$,therefore we refer to case A.

We have,

$H_0: \frac{\mu_2}{\mu_1'} = 1$ Against $H_1: \frac{\mu_2}{\mu_1'} > 1$,

Here $y_0 > 1.645$.

So, it is significant.

$\therefore H_0$ is rejected and alternately $H_1$ is accepted at 5% level of significance

i.e. A Negative Binomial Distribution is to be considered.

From the above tests it is noticed that for both the samples of 'the' marker word, a

Negative Binomial distribution should be considered as a statistical model.

### 2.9. Negative Binomial model:

The statistical model of Negative Binomial distribution has two parameters. If we examine this model it is noted that frequencies in this model go on increasing and reach a maximum value and thereafter frequencies go on decreasing.

The model of Negative Binomial distribution is as follows:

$p(x) = \binom{x+r-1}{r-1} p^r (1-p)^x$ , x = 0, 1, 2, ……

$= 0$ otherwise

Where 'p' and 'r' are the parameters of the model

### 2.10. Estimation of Parameters:

There are several methods suggested in the literature to estimate parameters of the model. From among these we have considered the method of moments for estimating the parameters.

It is noted that,

$\mu_1' = \frac{r(1-p)}{p}$and$\mu_2 = \frac{r(1-p)}{p^2}$

$\therefore \frac{\mu_1'}{\mu_2} = \frac{r(1-p)}{p} \times \frac{p^2}{r(1-p)}$

Which gives,

$\hat{p} = \frac{\mu_1'}{\mu_2}$and $\hat{r} = \frac{\mu_1' p}{q} = \frac{\mu_1' p}{(1-p)}$.

The estimates of $\hat{p}$ and $\hat{r}$ in the models for both samples of 'the' marker word are presented below:

**TABLE NO: 2.10.1 :** Estimates of parameters

| Marker word | $\hat{p}$ estimate | $\hat{r}$ estimate |
|---|---|---|

| | Sample- I | Sample - II | Sample- I | Sample - II |
|---|---|---|---|---|
| the | 0.8691161 | 0.8138092 | 37.050952 | 33.873968 |

Thus we obtain the statistical model of the negative Binomial distribution incorporating the estimates of parameters. With the help of this model we calculate expected frequencies.

The expected frequencies for both the samples of 'the' marker word are presented below:

**TABLE NO: 2.10.1**

| Exp. freq. Sample -I | Exp. freq. Sample -II |
|---|---|
| 1.65613 | 0.274448 |
| 8.031788 | 1.7374 |
| 19.99867 | 5.64144 |
| 34.073828 | 12.55892 |
| 44.654121 | 30.40888 |
| 47.984228 | 36.68104 |
| 42.219427 | 38.90316 |
| 35.430415 | 37.00808 |
| 25.535879 | 32.05868 |
| 16.7258565 | 25.5938 |
| 10.083159 | 19.00336 |
| 5.64512 | 13.23052 |
| 2.9587615 | 8.69284 |
| | 5.41952 |
| **295.00009** | 3.21784 |
| Total | **291.99124** |

## 2.11. K-S test for testing goodness of fit :

The validity of the model is tested with the help of Kolmogoroff (1941) (Sach,1984) and Smirnoff (1948) (Sach, 1984) test of goodness of fit. The given test is distribution free test. It corresponds to the Chi-square goodness of fit test.

The K-S test is more likely to detect the deviations from the normal distribution. The K-S test is more sensitive to departures from the shape of the distribution functions. The K-S test values and critical values at 5% level of significance of the samples of 'the' marker word are presented in the following table:

**TABEL NO: 2.11.1 :** K-S test values and critical values.

| Marker word | K-S values | | Critical values | |
|---|---|---|---|---|
| | Sample- I | Sample -II | Sample- I | Sample -II |
| the | 0.0306732 | 0.0391341 | 0.0790653186 | 0.0794709 |

The table values of the test, determined by Liliefor's (1967) and presented in the book by Sach (1984), are used for testing of significance.

Here it is noted that the observed K-S values are less than the corresponding critical values at 5% level of significance. Therefore the statistical model of Negative Binomial probability cannot be rejected.

## 2.12. Chi-square test:

Karl Pearson (1900) (Gupta and Kapoor, 1992) formulated a test for goodness of fit called 'CHI-SQUARE' test of goodness of fit. Chi-square test of fit is generally used for testing significance of the discrepancy between theory and experiment. It enables us to find if the deviation of the experiment from the theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

The test is given by,

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-k-1}$$

Where $O_i$ – observed frequency in $i^{th}$ class
$E_i$ – Exp. Frq. in $i^{th}$ class.
n – number of classes after adjusting extreme classes to have theoretical frequency 5 or more k- the number of parameter estimated.

There are also inadequacies of $\chi^2$test. Sach (1984) while describing K-S test, pointed out that the $\chi^2$test is better only for detecting irregularities in the distribution. Also Pearson and Hartley (1958) noted that the $\chi^2$test is good only when the number of observations are large and test is powerful only if he most of classes have high expected frequencies. The $\chi^2$ test is not considered to be very sensitive in measuring deviations from normality.
Sach (1984) while describing the K-S test (P-330) pointed out that the $\chi^2$ test and K-S test require infinitely many classes (k). Still both the tests can be employed even for small samples with few classes.

The $\chi^2$ values for both the samples are presented in table below:

**TABEL NO.: 2.12.1**

| Marker word | Observed $\chi^2$ values | | Tabulated $\chi^2$values | |
|---|---|---|---|---|
| | Sample- I | Sample -II | Sample- I | Sample -II |
| the | 4.9556411 | 8.5225232 | 15.5073 | 18.307 |

Here it is noted that all the observed $\chi^2$ values are less than tabulated values at 5% level of significance. This shows that the proposed models of negative Binomial distribution cannot be rejected.

## 2.13. Combined Samples data for 'the' marker word :

It is observed that the model of Negative Binomial distribution fits well to the data of individual samples of 'the' marker word. Now the problem is to find out the statistical model for the combined sample. Here also the problem of choosing a statistical model is solved by computing moments from the combined data. The observed combined frequencies of 'the' marker word are shown below:

**TABLE NO: 2.13.1**
Frequency of combined sample

| $X_i$ | $F_i$ |
|---|---|
| 0 | 4 |
| 1 | 9 |
| 2 | 30 |
| 3 | 45 |
| 4 | 60 |
| 5 | 75 |
| 6 | 77 |
| 7 | 76 |
| 8 | 60 |
| 9 | 46 |
| 10 | 34 |
| 11 | 28 |
| 12 | 20 |
| Total | 564 |

From the above combined data of 'the' marker word, the sample values of moments of 'the' marker word are presented below:

**TABLE NO. 2.13.2**
Moments

| Marker word | $\overline{X}$ | $S^2$ |
|---|---|---|
| The | 6.374113 | 7.4185497 |

We consider Katz's criterion. The calculations of this criterion are shown below:
We have,

$$S^2 = 7.418550,$$
$$\overline{X} = 6.374113.$$

$Z = 0{,}163856.$

$V(z) = \frac{2}{n} = \frac{2}{564} = 0.003546$

Form (4.7) we have,

Here, $H_0 = \mu_1' = \mu_2$ Against $H_1 = \mu_1' < \mu_2$,

Since $\overline{X} < S^2$.

Now,

$$y_0 = \frac{z}{\sqrt{2/n}} = \frac{0{,}1638559}{0.059491} = 2.751608.$$

We consider the case (A), as $S^2 < \overline{X}$.

Here, $y_0 > 1.645$, at 5% level of significance.

We reject $H_0$ hypothesis and accept $H_1$. A negative Binomial distribution is to be considered for the combined data.

The statistical model and estimates of parameters are given in the section 2.9. Estimated values of parameters $\hat{p}$ and $\hat{r}$ of the above model for 'the' marker word of combined data are presented in the table below:

**TABLE NO : 2.13.2**

Parameters of combined data

| Marker word | $\hat{p}$ estimate | $\hat{r}$ estimate |
|---|---|---|
| the | 0.859213 | 38.900720 |

Thus we have the statistical model of Negative Binomial distribution incorporating estimates of parameters. The expected frequencies according to this model for the combined data are presented below:

**TABLE NO : 2.13.4**

Expected frequencies of 'the' marker word for the combined data,

| Expected frequency |
|---|
| 1.57356 |
| 8.62356 |
| 24.2238 |
| 46.49616 |
| 68.57676 |
| 82.83468 |
| 85.3332 |
| 77.05932 |
| 62.24304 |
| 45.66708 |
| 30.80004 |
| 19.27752 |
| 11.28 |
| Total = 563.9887 |

**Goodness of fit:**

By employing Kolmogoroff-Smirnoff test and also Chi-square test of goodness of fit, the validity of model is tested. The test is described in detail in the section No. 2.11, 2.12. The K-S and Chi-square values are calculated. The corresponding critical and tabulated values are presented below:

**TABLE NO : 4.15.5**

| Marker word | K-S values calculated | K-S Critical values | $\chi 2_{cal}$ | $\chi 2_{tab,5\%\ l.s.}$ |
|---|---|---|---|---|
| the | 0.0371703 | 0.0571821 | 15.952594 | 16.9190 |

In case of K-S test, the calculated value is less than the critical value at 5% level of significance. Similarly, in case of the $\chi^2$test goodness of fit, the calculate value is less than the tabulated value of $\chi^2$, at 5% level of significance. This implies that the proposed model of Negative Binomial distribution to the combined data of two samples cannot be rejected.

## 2.14. Marker word: 'Of' Statistical model for the data of 'of' marker word:

In the formulation of statistical model the constants, which are determined from the data, play very important role. The values of constant (moments) for both the samples data of 'of' marker word are presented in the table below:

**TABLE NO : 2.14.1**

Moments

| Marker word | moments | Sample-I | Sample-II |
|---|---|---|---|
| OF | $\overline{X}$ | 4.154362416 | 4.371237458 |
| | $s^2$ | 4.029863518 | 3.67823626 |

We use the above sample moments, for testing the statistical model with help of Katz's criterion. The results of this criterion are given below:

## 2.15. Katz's criterion

**Sample – I**

$$S^2 = 4.029863518$$
$$\overline{X} = 4.154362416$$
$$Z = -0.0299682$$
$$V(z) = \frac{2}{n} = \frac{2}{298} = 0.006711$$

We have,

$H_0 = \mu_1' = \mu_2$ against $H_1 = \mu_1' > \mu_2$, since $\overline{X} > S^2$

$\therefore$ consider case (B).

Now,

$$y_0 = \frac{z}{\sqrt{2/n}} = \frac{-0.0299682}{0.0819231}$$

$y_0 = -36580.101.$

Here, $y_0 > -1.645.$

Therefore$H_0$ is accepted at 5% level of significance. A Poisson distribution is to be considered for statistical model.

A Poisson distribution is as follows:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \text{ x= 0, 1, 2, ……., } \infty.$$

Where, λ is a parameter.

**Estimation of parameter:**

The formula for estimation of parameter of Poisson distribution is as follows:

$$\overline{x} = \sum_{i=1}^{n} (f_i x_i / f_i)$$

$$\therefore \hat{\lambda} = \overline{x}$$

**TABLE NO :2.15.1 : Parameters**

| Marker word | $\hat{\lambda}$ |
|---|---|
| OF | 4.15436 |

The statistical model of Poisson distribution incorporating the estimate of parameter helps to derive the expected frequencies. The expected frequencies are shown below:

**TABLE :2.15.2**

| Expected frequencies |
|---|
| Sample – I |
| 4.6961224 |
| 19.509494 |
| 40.524841 |
| 56.043817 |
| 58.283972 |
| 48.42652 |
| 33.530185 |
| 19.899516 |
| 10.333686 |
| 4.769966 |
| 1.9816404 |
| 297.99979 |

**Validity of model**

The validity of model can be tested with the help of Chi-square test and K-S test of goodness of fit. The details of this test are given section 2.11 and 2.12. The number of degree of freedom for the $\chi^2$ test is 8 d.f. . The calculated and tabulated values are given below:

**TABLE NO: 2.15.3**

| Test | Of marker word sample I |
|---|---|
| $\chi 2_{cal}$ | 11.407793 |
| $\chi 2_{tab}$ | 15.5073 |
| **K-S values** | |
| Calculated | 0.0462223 |
| Tabulated | 0.0786668 |

From the above values it is noted that, the $\chi^2$ and K-S calculated values are less than the corresponding tabulated values at 5% level of significance. This implies that the proposed model is Poisson distribution fits well to the data.

## 2.16. Sample –II:

$$S^2 = 3.67823626$$

$$\overline{X} = 4.371237458$$

$$Z = \frac{-0.6930012}{4.371237458} = -0.1585366$$

$$V(z) \div \frac{2}{n} = \frac{2}{299} = 0.00668896$$

We have,

$H_0 = \mu_1' = \mu_2$against $H_1 = \mu_1' > \mu_2$, since $\overline{X} > S^2$

Now,

$$y_0 = \frac{z}{\sqrt{2/n}} = -1.938432$$

$$y_0 < -1.645.$$

Therefore $H_0$ is accepted at 5% level of significance. A Binomial distribution is to be considered for statistical model.

$f(x) = \binom{n}{x} p^x q^{n-x}$ , x= 0, 1, 2, ……n

$= 0$ otherwise.

Where, p is parameter.

## 2.17. Estimation of parameters:

The formulae of parameters of Binomial model are follows:

$$\mu_1' = np,$$

$$\therefore p = \frac{\mu_1'}{n}$$

and, $q = 1 - p$

**TABLE NO : 2.17.1 :** Parameter

| Marker Word | Sample II | |
|---|---|---|
| | p | q |
| OF | 0.4371237 | 0.5628762 |

The expected frequencies according to this model are shown below:

**TABLE NO: 2.17.2 :** Expected frequencies

| Sample II | |
|---|---|
| Obs. Freq | Exp. Freq |
| 0 | 0.958893 |
| 12 | 7.4466548 |
| 42 | 26.023525 |
| 52 | 53.892298 |
| 55 | 72.771607 |
| 68 | 67.816399 |
| 28 | 43.887878 |
| 19 | 19.475903 |
| 15 | 5.6717908 |
| 6 | 0.978926 |
| 2 | 0.0760058 |
| 299 | 298.99988 |

**Validity of model:**

The validity of model can be tested with the help of $\chi^2$ test and K-S test of goodness of fit. The tabulated values are at 5% level of significance. The calculated and tabulated values are given below:

**TABLE NO : 2.17. 3 :** Marker word 'OF '

| Test | Of marker word sample I |
|---|---|
| $\chi^2$ calculated | 60.883893 |
| $\chi^2$ tabulated | 12.6 |
| **K-S values** | |
| Calculated | 0.0654546 |
| Tabulated | 0.0785351 |

From the above values it is noted that, the $\chi^2$calculated is greater than tabulated. According to the $\chi^2$ test a binomial distribution does not fit.
The K-S calculated values are less than the tabulated values. This shows that the fit is good.
When we fit a Poisson distribution to the first sample, it shows that the fit is good.
Therefore for the second sample also we attempt a Poisson model.

A Poisson distribution is as follows:
$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$, x= 0, 1, 2, ……., ∞.
Where, λ is a parameter.

**Estimation of parameter:**
The formulae of parameter of Poisson model is as follows:

$$\hat{\lambda} = \overline{x}$$

**TABLE NO: 2.17.4 : Parameter**

| Marker word | $\hat{\lambda}$ |
|---|---|
| of | 4.371237 |

The expected frequencies based on the above model are presented as below:

**TABLE NO : 2.17.5**

| Expected frequencies |
|---|
| Sample II |
| 3.7986455 |
| 16.604816 |
| 36.291873 |
| 52.880123 |
| 57.787915 |
| 50.520954 |
| 36.806481 |
| 22.98425 |
| 12.558688 |
| 6.0996299 |
| 2.6663325 |
| 298.99971 |

**Validity of model:**

The validity of model is tested with help of Chi-square and K-S test of goodness of fit. The details of this test are given in section 2.11 and 2.12.

The Chi-square test and K-S test calculated and tabulated values are shown below:

**TABLE NO. : 2.17.6**
Marker word 'of '

| Test | Of marker word sample I |
|---|---|
| $\chi 2_{cal.}$ | 13.894602 |
| $\chi 2_{tab.}$ | 14.0671 |
| **K-S values** | |
| Calculated | 0.0371761 |
| Tabulated | 0.0785351 |

From these above it is noted that, the $\chi^2$ and K-S calculated values are less than tabulated values of 5% level of significance. This implies that the proposed model of Poisson distribution fits well to the collected data.

## 2.18 Combined samples data of 'of ' marker word :

It is observed that the proposed model of Poisson distribution fits well to the collected data of individual samples of 'OF' marker word. Now the problem is to find out the statistical model for the combined sample. The observed frequencies for the combined samples are given below:

**TABLE NO : 2.18.1**

Observed frequency of combined data

| $X_i$ | $F_i$ |
|---|---|
| 0 | 06 |
| 1 | 38 |
| 2 | 79 |
| 3 | 90 |
| 4 | 116 |
| 5 | 126 |
| 6 | 63 |
| 7 | 41 |
| 8 | 25 |
| 9 | 10 |
| 10 | 03 |
| Total | 597 |

From above combined data of 'OF" marker word, the values of moments are presented
below:

**TABLE NO : 2.18.2**

Moments

| Marker word | $\overline{X}$ | $S^2$ |
|---|---|---|
| OF | 4.262981574 | 3.865514057 |

For the formulation of statistical model, we use here Katz's test.

## 2.19. Katz's Criterion:

The details of this test are discussed in section 4.8.

$$S^2 = 3.865514057$$

$\overline{X} = 4.262981574$

$$Z = \frac{S^2 - \overline{X}}{\overline{X}} = -0.0932369$$

$$V(z) = \frac{2}{n} = \frac{2}{597} = 0.0033500838$$

We have,

$H_0 = \mu_1' = \mu_2$ Against $H_1 = \mu_1' > \mu_2$, since $\overline{X} > S^2$

$\therefore$ Consider case (B).

Now,

$$y_0 \div \frac{z}{\sqrt{2/n}} = \frac{-0.0932369}{0.0033500838} = -1.6108681$$

Here, $y_0 > -1.645$.

Therefore $H_0$ is accepted at 5% level of significance. A Poisson distribution is to be considered for statistical model.

A Poisson distribution is as follows:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad , \quad x = 0, 1, 2, \ldots, \infty.$$

Where, $\lambda$ is a parameter.

Base on this model the expected frequencies are given below:

**TABLE NO : 2.19.1** Expected frequencies

| Obs. Freq | Exp. Freq |
|---|---|
| 06 | 8.4441471 |
| 38 | 35.997428 |
| 79 | 76.728231 |
| 90 | 109.03029 |
| 116 | 116.19859 |
| 126 | 99.070478 |
| 63 | 70.389225 |
| 41 | 42.866809 |
| 25 | 22.842533 |
| 10 | 10.81967 |
| 03 | 4.6123974 |
| 597 | 596.9998 |

## 2.20. Validity of model:

The validity of model can be tested with the help of Chi-square test and K-S test of goodness of fit. The details of this test are given section 2.11 and 2.12.

**TABEL NO : 2.20.1**

| Marker word | $\chi 2_{cal}$ | $\chi 2_{tab}$ | K-S values | |
|---|---|---|---|---|
| | | | calculated | Critical values |
| OF | 12.972122 | 15.5073 | 0.0291435 | 0.555792 |

## III.CONCLUSION

Every author has his/her style of writing. This style depends on use of words, or some words in certain ways. Recently statistician have considered this problem and developed a branch of statistics called “Stylistics”.

Marker words which are the words with maximum frequencies have been utilised to describe the style. For this purpose samples have been selected and probability distribution has been fitted. It was noticed that in some cases Negative Binomial and Poisson distribution gives good fits.

## IV. REFERENCES


[1]. Mosteller Frederick and Wallace David L.(1964): Inference and Disputed Authorship: The Federalist, Reading Mass, Addison-Wesley Publishing Company, Inc.

[2]. Mosteller Frederick and David L. Wallace(1963): “ Inference in an Authorship Problem,” Journal of the American Statistical Association, LVIII, pp.-275-309.

[3]. Dewey G.(1923): Relative Frequency of English Speech Sounds., Harvard University Press, Rev. Edition, 1950.

[4]. Shende P.S. and Prabhu-Ajgawonkar S.G.(1988): On a statistical measure of style. Journal of the Linguistic Soceity of India, Vol. 49, pp-1-10.

[5]. Muthe S.P. and Prabhu-Ajgaonkar S.G.: Statistical Parameter Representing style of Marathi Novelist Phalke, Indian Linguistics, Journal of the Linguistic society of India, Vol. 60{1999), pp.-111-119.

[6]. Katz L.(1963): Unified Treatment of a broad class of discrete probability distributions, Proceeding of the International Symposium on Discrete Distributions, Montreal, pp.-175-182.

[7]. Nehru Pandit Jawaharlal(1931): “ Glimses of word History” , First edition, Kitabstan Publication, Allahabad.

[8]. Prabhu-Ajgaonkar S.G(1969): Determination of phonemic and Graphemic frequencies by sampling Techniques, Deccan College, deccan College Post-graduate and Research Institute, Poona.

[9]. Prabhu-Ajgaonkar S.G(1973): Frequency count and sampling method, Journal of Ganganatha, The Kendriya Sanskrit Vidyapeetha, Allhabad, Vol. XXIX, Parts, pp.-1-4.

[10]. Prabhu-Ajgaonkar S.G(1975): On determining Average Number of phonemes per word, Natural Science Journal, Marathwada University, Vol. XIV, Science 7.

[11]. Prabhu-Ajgaonkar S.G. and Somaya P.C.(1972): Fitting of a Mixture of Poisson Distribution to the linguistic data, Journal of Science, Marathwada University, Vol.XI Science 4.

[12]. Willam C.B.(1970): Style and vocabulary: Numerical studies, Ist edition Griffin Company, London.

[13]. Williams C,B.(1956): Studies in the history of Probability and statistics, IV : A note on an early statistical study of literary style, Biometrika, Vol. 43, pp.-248-256.

[14]. Good I.J.(1957): Distribution of word frequencies, Nature, London, Vol. 179, pp-595.

[15]. Gore A.P., Mrs.Gokhale M.K. and Mrs.Joshi S.B.: On disputed authorship of editorials in Kesari. Maharashtra Association for Caltivation of Science, Pune.

**Cite this article as :**

Dr. Ashok Y. Tayade, "Use of Statistical tools for Style by Marker word", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 7 Issue 2, pp. 574-590, March-April 2020. Journal URL : http://ijsrst.com/IJSRST207330