Advances in Computational Research, ISSN: 0975–3273, Volume 1, Issue 2, 2009, pp-34-38

# Characteristic of style: sentence– length

**Tayade A.Y.* and Prabhu-Ajgaonkar S.G.**
*Department of Statistics, Dr.B.A.M. University Aurangabad-431004(M.S.)

**Abstract**- This paper is a contribution to stylostatistics which has, as a discipline, established itself in the recent years. We have considered probability distribution to the data of sentence-length collected from the book of late Pandit Jawaharlal Nehru.
**Keywords**- Statistical Methods, Probability distributions, Non-parametric test

## Introduction

It is noted that generally authors use different lengths of sentences. Sentence is a mixture of words to signify complete sense of thought. It depends on a number of words. The length of a sentence is measured by the number of words appearing in it. Yule [10] was the first to propose use of sentence - length as a statistical characteristic of style. The shape of a typical sentence-length distribution is positively skewed that is, it exhibits high frequency of short sentences and declining frequency of longer ones. It has also been argued that on log scale the distribution of sentence length resembles to that of the normal distribution. Suppose that style of author A is characterized, in general by used of short sentences. On the other hand the style of another B is distinguished generally by use of long sentences. Nevertheless this stylistic regularity does not prohibit author A from use of longer sentences in certain cases, author B is not deprived of the possibility of using short sentences. Thus it is noted that sentence –length is a tendency that can descried by probability distributions of the individual sentence lengths. In this way, a probabilistic conception enables us to describe style not as a fixed habit, but rather as a preference for one or another mode of expression. A probabilistic approach revels the flexible character of stylistic features, which resist any description in terms of necessity or in terms of strict rules. Yule [10] referred to the English Dictionary for definitions of sentence and word. He considered that the number of words between successive full stops is a sentence and in a sentence whatever mixture of letters of A to Z is used for the proper meaning of word is called word so thus words form a sentence.

Williams [9] and Wake [8] indicated that the distribution of sentence-length in terms of words follows the log-normal distribution. Rao Subba [7] studied sentence-length in eight works in Kannada Prose by three different authors. Bhattacharya [1] used the lognormal distribution for his collected data and also in his Ph.D. thesis entitled "Some Statistical Studies on languages" (1965).

## Methodology

For examining the writing style of the first Prime Minister of India, Pandit Jawaharlal Nehru, we selected the book entitled "India's freedom “[5].

The total number of pages of this book is 89. This complete book is divided into two equal parts, each part is considered to be a sample. Accordingly, the first sample contains 895 sentences and the second sample contains 760 sentences. The frequency table of number of words occurred, that is the number of size one sentences occurred how many times and so on prepared. For the classification of this data we have used Sturge's rule.

Sturge's rule is,

$$k = 1 + 3.32 \, \mathrm{Log}\, n$$

Where k is the number of classes and n is total of frequencies.

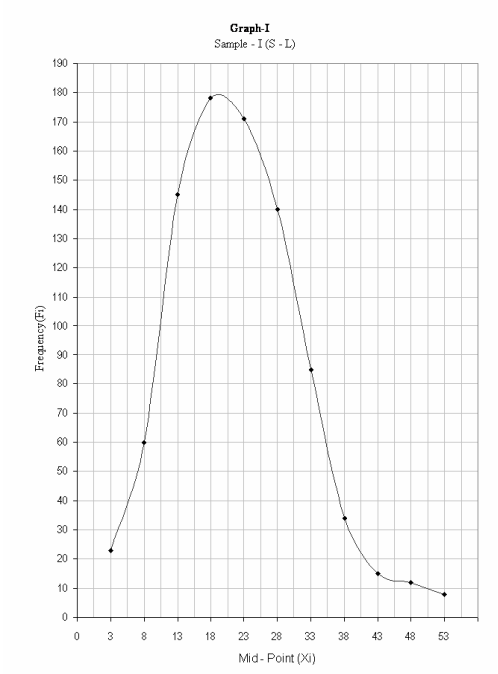
The values of k are 11 for both samples.

Nearly eleven classes should be considered for the statistical investigation. Each class is of size five. The data is classified as follows:

*Table 1-Sample – I*

| C - I | Mid-Points | Frequency |
|---|---|---|
| | (xi) | (fi) |
| 1-5 | 3 | 23 |
| 6-10 | 8 | 60 |
| 11-15 | 13 | 145 |
| 16-20 | 18 | 178 |
| 21-25 | 23 | 171 |
| 26-30 | 28 | 140 |
| 31-35 | 33 | 85 |
| 36-40 | 38 | 34 |
| 41-45 | 43 | 26 |
| 46-50 | 48 | 12 |
| 51-55 | 53 | 5 |
| Total | | 879 |

The graphical representations are as follows:



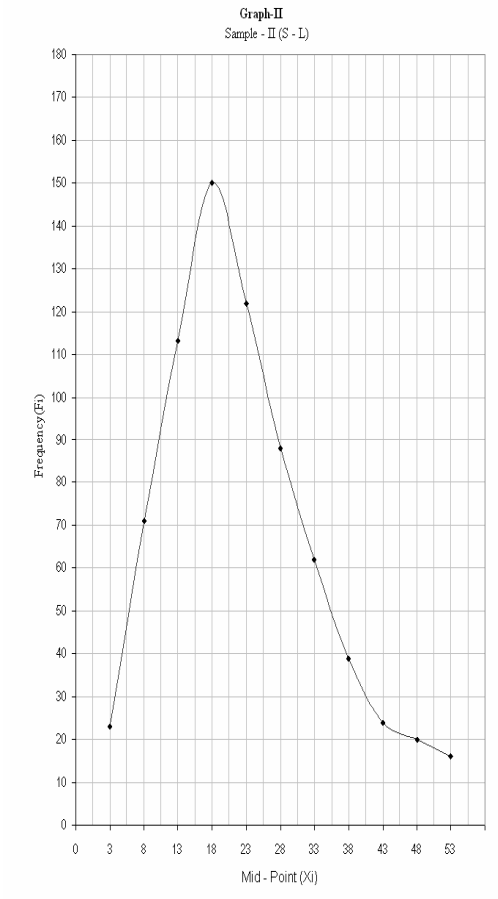




*Table 2-Sample – II*

| C - I | Mid-Points | Frequency |
|---|---|---|
| | (xi) | (fi) |
| 1-5 | 3 | 22 |
| 6-10 | 8 | 71 |
| 11-15 | 13 | 113 |
| 16-20 | 18 | 150 |
| 21-25 | 23 | 122 |
| 26-30 | 28 | 88 |
| 31-35 | 33 | 62 |
| 36-40 | 38 | 39 |
| 41-45 | 43 | 24 |
| 46-50 | 48 | 20 |
| 51-55 | 53 | 16 |
| Total | | 727 |

It is noted that the data is not unimodel, whereas the probability models generally are unimodel. Using the above data, we plotted the frequency curves for both samples. It is noted that the curves are bell-shaped. So we consider first fitting of the Normal distribution to the data.

The Normal distribution is,

$$F(x) = \frac{1}{\sigma\sqrt{2\Pi}}\, e^{-\frac{(\chi-\mu)^2}{2\sigma^2}}, \qquad -\infty \le \chi \le \infty$$

where, μ and σ are parameters.

*Table 3-Estimates of Parameters*

| Samples | Parameters | |
|---|---|---|
| | $\bar{\chi}$ | s2 |
| Sample I | 22.23777 | 9.728353 |
| Sample II | 22.61486 | 11.33413 |

Next we fitted the normal distribution .The expected frequencies of both samples are shown below:

*Table 4-Expected frequency*

| Sample - I | | Sample - II | |
|---|---|---|---|
| Observed frequency | Expected frequency | Observed frequency | Expected frequency |
| 23 | 37.797 | 22 | 43.62 |
| 60 | 61.53 | 71 | 58.16 |
| 145 | 116.028 | 113 | 87.24 |
| 178 | 161.736 | 150 | 116.32 |
| 171 | 175.8 | 122 | 130.86 |
| 140 | 152.067 | 88 | 109.05 |
| 85 | 97.569 | 62 | 87.24 |
| 34 | 50.103 | 39 | 50.89 |
| 26 | 19.338 | 24 | 29.08 |
| 12 | 5.274 | 20 | 7.27 |
| 5 | 1.758 | 16 | 7.27 |
| Total =879 | 879 | 727 | 727 |

**Validity of Model**

Generally the $\chi^2$ - test is used for testing goodness of fit. It should be noted that there are inadequacies of $\chi^2$ – test. Sach [4] while describing Kolmogoroff-Smirnoff test pointed out that the $\chi^2$ – test is better only for detecting irregularities in the distribution. Also Pearson and Hartley [6] noted that the $\chi^2$test is good only when the number of observations is large and the test is powerful only if the most of the classes have high expected frequencies. The $\chi^2$ test is not considered to be very sensitive in measuring deviations from normality. We therefore consider non-parametric Kolmogoroff -1941, Sach [4] and Smirnoff -1948, Sach [4] test called K-S test of goodness of fit.

**K-S test**

This test of goodness of fit is how well an observed distribution fits to a theoretically expected one. Sach [4] noted that the K-S test is more likely to detect deviations from normal distribution particularly when the sample sizes are small. The K-S test is also more sensitive to departure from the shape of the distribution function. The test is given as follows:

$$D = \frac{\text{Max.}\left|F_{(O_i)} - F_{(E_i)}\right|}{n}$$

Where, F (Oi) =The observed cumulative frequency of ith class

F (Ei) = The expected cumulative frequency of the i th class.

n = Sample size.

The table values for $\Lambda D$, given by Liliefor's -1967, presented in the book by Sach [4], are used for testing of significance level. K-S values of both samples and their corresponding tabulated and critical values at 5% level of significance are presented below:

*Table 5-*

| Samples | K-S values | Critical values |
|---|---|---|
| First | 0.032889 | 0.045804 |
| Second | 0.069684 | 0.050365 |

The K-S value of first sample is insignificant and for the second sample it is significant at 5% level of significance.

Next we consider the log-normal distribution,

Fitting of a log-normal distribution to the data. The log-normal distribution is,

$$f(x) = f(x, \mu, \sigma^2) = \frac{1}{x\sqrt{2\pi\sigma}} \exp\left[-\left(\log_e x - \mu\right)^2 / 2\sigma^2\right] I_{(0,\infty)}(x)$$

where, $\infty < x < \infty$ and $\sigma > 0.$, and $\mu$ and $\sigma$ are parameters.

Estimation of parameters are as follows:

*Table 6-*

| Samples | Parameters | |
|---|---|---|
| | $\bar{x}$ | s2 |
| Sample I | 3.014236 | 0.418464 |
| Sample II | 3.006563 | 0.47338 |

Thus we fitted the Log-normal distribution, and the expected frequencies are worked out.

*Table 7-Observed and Expected frequency*

| Sample - I | | Sample - II | |
|---|---|---|---|
| Observed frequency | Expected frequency | Observed frequency | Expected frequency |
| 23 | 0.797802 | 22 | 2.217876 |
| 60 | 49.8182 | 71 | 59.73316 |
| 145 | 177.9097 | 113 | 150.7375 |
| 178 | 218.243 | 150 | 165.8187 |
| 171 | 178.5302 | 122 | 130.0379 |
| 140 | 111.7809 | 88 | 88.63868 |
| 85 | 67.99039 | 62 | 55.59318 |
| 34 | 36.61017 | 39 | 34.15432 |
| 26 | 20.47683 | 24 | 20.03423 |
| 12 | 10.9919 | 20 | 12.49364 |
| 5 | 5.850545 | 16 | 7.540517 |
| Total=879 | 878.9996 | 727 | 726.9996 |

**Validity of model**

The K-S test of goodness of fit is applied to test the goodness of fit. The K-S values of both samples and their corresponding tabulated and critical values at 5% level of significance are presented below:

*Table 8-*

| Samples | K-S values | Critical values |
|---|---|---|
| First | 0.054948 | 0.045804 |
| Second | 0.042894 | 0.050365 |

From the above table it is noted that for the first sample K-S test is significant. For the second sample the K-S test is insignificant at 5% level of significance. Since the normal and log-normal probability distributions do not fit both the samples, we examine the suitability of the Pearsonian system of curves. Next we apply Pearson's criterion, For applying pearsons criteria used book by Elderton W. P. and Johnson N. L. [2]. The moments and constants of data are determined and presented in Table 9.

The Pearson K criterion to be applied is given below :

$$K = \frac{\beta_1 (\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}$$

as shown in equation (4) by Elderton and Johnson (1969). The values of K are shown in Table 10 for both samples.

*Table 9-*

| Moment | Sample - I | Sample - II |
|---|---|---|
| $\Lambda$ $\mu'_1$ | 22.2377 | 22.6222 |
| $\Lambda$ $\mu_2$ | 3.7856 | 5.1251 |
| $\Lambda$ $\mu_3$ | 3.5005 | 7.7896 |
| $\Lambda$ $\mu_4$ | 44.9918 | 79.9518 |
| Co-efficients | | |
| $\Lambda$ $\beta_1$ | 0.2259 | 0.4507 |
| $\Lambda$ $\beta_2$ | 3.1395 | 3.0438 |

*Table 10-*

| Samples | K values |
|---|---|
| First | – 0.44939 |
| Second | – 0.30222 |

From the above results it is noted that the Pearson type-I probability distribution is suitable to the data. The equation to Pearson type-I and constants of curve are given below:

$$Y = Y_o \left(1 + \frac{x}{a_1}\right)^{m_1} \left(1 - \frac{x}{a_2}\right)^{m_2}$$

Where,

$$\frac{m_1}{a_1} = \frac{m_2}{a_2} \quad \text{and} \quad -a_1 < x < a_2$$

Here the origin is at the mode. The constants of the curve to be determined are given by,

$$r = 6(\beta_2 - \beta_1 - 1) / (6 + 3\beta_1 - 2\beta_2)$$

And,

$$a_1 + a_2 = \frac{1}{2}\sqrt{\mu_2}\sqrt{\beta_1 (r+2)^2 + 16(r+1)}$$

The values of m1 and m2 appearing in the expression are calculated as follows:

$$\frac{1}{2}\left(r - 2 \pm r(r+2)\sqrt{\frac{\beta_1}{\beta_1 (r+2)^2 + 16(r+1)}}\right)$$

Where $m_2$ is the positive root when $\mu_3$ is positive. Further

$$Y_0 = \frac{N}{a_1 + a_2} \frac{m_1^{m_1} m_2^{m_2}}{(m_1 + m_2)^{m_1 + m_2}} \frac{\Gamma(m_1 + m_2 + 2)}{\Gamma(m_1 + 1)\Gamma(m_2 + 1)}$$

If the curve is expressed with origin at the mean then,

$$\text{Mode} = \text{Mean} - \frac{1}{2} \frac{\mu_3}{\mu_2} \frac{r+2}{r-2}$$

$$A_1 + A_2 = a_1 + a_2$$

$$\frac{m_1 + 1}{A_1} = \frac{m_2 + 1}{A_2}, \text{ i.e. } A_1 = \frac{A_1 + A_2}{(m_1 + m_2 + 1)}(m_1 + 1) \text{ and } A_2 = \frac{A_1 + A_2}{(m_1 + m_2 + 1)}(m_2 + 1)$$

$$Y_e = \frac{N}{A_1 + A_2} \frac{(m_1 + 1)^{m_1}(m_2 + 1)^{m_2}}{(m_1 + m_2 + 2)^{m_1 + m_2}} \frac{(m_1 + m_2 + 2)}{(m_1 + 1)(m_2 + 1)}$$

Subsequently the equations to the curve with origin at the mean reduces to,

$$Y = Y_e \left(1 + \frac{x}{A_1}\right)^{m_1} \left(1 - \frac{x}{A_2}\right)^{m_2}$$

The constants are shown below for the data of both samples:

*Table 11-Value of different constants*

| Contents | Sample –I | Sample – II |
|---|---|---|
| r | 28.80102 | 7.558948 |
| m1 | 5.381941 | 0.962129 |
| m2 | 21.41908 | 4.596818 |
| A1 | 5.135646 | 3.921525 |
| A2 | 20.43887 | 11.18584 |
| A1+A2 | 25.57452 | 15.10737 |

The values of A1 and A2 are to be calculated, when origin is shifted at the mode.

For the second sample the values of A1 and A2 are to be calculated, when origin is shifted at mean. These values are used at the time of calculations of transformation. The values are given below:

*Table 12-*

| | |
|---|---|
| Mode (Sample - I) | 21.70643 |
| Mean (Sample- II) | 22.62253 |

The values of gamma function are calculated from the gamma table. The values of this function are taken from 'TABLE OF INCOMPLETE BETA FUNCTION' edited by Karl Pearson (1968), and published for the Biometrika Trustees at 'Cambridge University Press' . The expected frequencies for both samples are presented below:

*Table 13-*

| Sample - I | | Sample - II | |
|---|---|---|---|
| Observed Frequency | Expected Frequency | Observed Frequency | Expected Frequency |
| 23 | 15.85364 | 22 | 18.58474 |
| 60 | 69.22784 | 71 | 77.85101 |
| 145 | 140.3052 | 113 | 127.1664 |
| 178 | 180.9965 | 150 | 132.8736 |
| 171 | 173.3338 | 122 | 102.5819 |
| 140 | 104.4583 | 88 | 97.75278 |
| 85 | 96.88373 | 62 | 71.77802 |
| 34 | 55.13308 | 39 | 42.65709 |
| 26 | 27.03936 | 24 | 31.31596 |
| 12 | 11.53593 | 20 | 15.36529 |
| 5 | 4.227655 | 16 | 9.072815 |
| Total=879 | 878.997 | 727 | 726.9996 |

**Validity of model**

The $\chi^2$ and K–S tests of goodness of fit are applied for both samples. The tabulated and critical values at 5% level of significance are as follows:

*Table 14-*

| Samples | K-S values | Critical values |
|---|---|---|
| First | 0.037343 | 0.045804 |
| Second | 0.026056 | 0.050365 |

From the above table it is noted that the K–S test is insignificant for both samples at 5% level of significance.

**Conclusion**

This paper considers writing style of authors. Yule (1939) suggested that sentence-length could be used as criterion to describe style. In this contribution we have analyzed samples from the book 'India's Freedom' (1936) written by Pandit Jawaharlal Nehru. The method of selecting statistical models for literary style has been described. Estimation of parameters appearing in the models has discussed. The Normal, Log-normal and Pearson type-I probability models are considered for the collection data. It is noted that for the data the normal as well as the log-normal distribution do not give good fit to the data. However, the Pearsonian Type-I probability distribution gives good fit.

**References**

[1] Bhattacharya N. (1974) *Sankhya*, 36,series B,Pt.4,323-347 .

[2] Elderton W.P. and Johnson N.L. (1969) *System of frequency curves. Cambridge University Press.*

[3] Karl Person (1968) *'Table of incomplete beta function' edited by and published for the Biometrika Trustees at Cambridge University Press.*

[4] Lothar Sach (1984) *Applied Statistics: A Handbook of Techniques, Second edition, Springer-Verlag, New York,Berlin Heidelberg Tokyo.*

[5] Nehru Pandit Jawaharlal (1936) *'India's freedom' Unwin Book, George Allen and Unwin Ltd. Ruskin House, Museum Street,London W.C.I.*

[6] Pearson E.S. and Hartly (1958) *Biometrika Tables for Statisticians, Cambridge University Press.*

[7] Rao Subba(1960) T*he Half-Yearly Journal of the Mysore University,New Series , Section A-Art*, Vol. XX,No.1,pp. 1-12.

[8] Wake C. Williams (1948) *Hibbert Journal*, 47, 50-55.

[9] Williams C.B.(1939-40) *Biometrika*, 31, 356-361.

[10] Yule G. Udny (1938-39) *Biometrika*, 30, 363-390.