International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
DOI :https://doi.org/10.32628/IJSRSET

Statistical Measures of Writing style by using K-Characteristics Criteria

Dr. Ashok Y. Tayade
Assistant Professor, Department of Statistics, Dr. B. A. M. University, Aurangabad, Maharashtra, India

ABSTRACT

This research article is contributes to the writing style which has, as a discipline, recognized itself in the recent years. We have considered some statistical parameters and large sample test is taken in consideration to the data for K-Characteristics. The data is collected from the book of former Priminister Pandit Jawaharlal Nehru.
Keywords : Statistical Methods, K-Characteristics Criteria, Large Sample Test.

I. INTRODUCTION

In this chapter a criterion is examined for measuring writing style of Pandit Jawaharlal Nehru. Nehru's book, entitled "The Discovery of India" (1946) is used for collection of data.
Yule (1944) derived a "Characteristics" to measure writing style which is independent of sample size. K – Characteristics is discussed in detail in the book entitled "The statistical study of literary Vocabulary (1944) by Yule. This book contains eleven chapters. Chapter first, second and third discussed theoretical part of the K- characteristics. In chapter four practical examples are presented.

A number of characteristics have been suggested in the literature to describe writing style of an author. Some others have fancy for writing long sentences while others use small sentences. Apparently sentence-length would be a criterion to distinguish writing style of authors. If we examine passages of writing of an author, we notice certain words are repeated in this writing. Moreover some authors have tendency to use certain particular words. Naturally words with their frequencies vary from author to author. Thus a criterion can be formulated based on words with their frequencies.

This characteristic is independent of size of sample within the limits of fluctuations of sampling. This seemed to be an important result. Yule (1944) noticed this by taking a series of samples spread over one and same work.

It is noted that the characteristics remained the same within the limits of fluctuation of sampling, whether the distribution was based on either one or two or three or four or five thousand occurrences. Therefore one was able to compare two distributions without regard to the number of occurrences.

The results gave one confidence in the general notions on which the theory was based. Next, the characteristics having been obtained, it was obviously desirable to find out the extent to which it would be likely to vary in the data drawn from different but similar work of one and the same author.

Yule (1944) considered this point and formulated a criterion based on facts. For this purpose Yule (1944) imagined the frequency of words to be similar to the

frequency of accidents within a given period. In the distribution of accidents we have the intervals without accident. However in the word distribution such a thing is not possible.

II. K- Characteristics and large sample test

It is noted that Yule (1944) was a Cambridge statistician who pioneered several impotent stylostatistical measures. His main concern was to devise a criterion which would apply largely independently of sample size. This work is based on tables showing the number of words used once, twice, thrice,...etc. , by an author in his / her writing . Commonsense suggests that there ought to be something the same or rather approximately the same within the limits of fluctuations of sampling. The equivalent for the word distribution would be the (unknown) total of words at risk. The characteristics of the accident-distribution are independent of the period of exposure to risk. The characteristics are also independent of S_1. If we express this characteristics in terms of the two sums S_1 and S_2, we get the expression $(S_2\text{-}S_1) / (S_1)^2$ must also be independent of S_1, where S_1 and S_2 are respectively the first and second moments (i.e. $S_1 = \sum_{i=1}^{n} fx$, $S_2 = \sum_{i=1}^{n} fx^2$). The above expression is therefore constant for the decapitated distribution. In actual example it is found that, since S_1^2 greatly exceeds S_2, the above expression gives a very small decimal. It is inconvenient working with small decimals and for practice handier to multiply the expression by 10,000 thus the characteristics is given by

$$K = 10000\frac{S_2 - S_1}{S_1^2}$$

where, S_1 and S_2 are respectively the first and second moments.

It has been noted that for large samples an estimate of a parameter may be obtained by calculating from the sample values, the value of parameter. For sample of size n, the standard error gives valid measure of precision, provided that the sampling distribution of the statistic under discussion approaches normality and that n is large.

III. Methodology

In this present chapter, Jawaharlal Nehru's work entitled "The Discovery of India" (1946) is considered. The book was written by him in the prison of Ahmadnagar fort during the five months, April to September 1944.

The principal merit of the Discovery is that it let us sees the mind of its author and helps us to forget the links of our racial memory and firmly turns our face to the culture. The Discovery of India is a happy blending of the past with the present. Jawaharlal Nehru projected India's illustrious past in comparison with the present. And through this he tried to present both the aspects in front of the minds of his countrymen to let them decide, to retrieve what they had lost, what they had forsaken and what exactly they needed to retrieve. Nehru took the solid material of history to bring forth yet another image of India into focus. This work exposed Nehru to a cross-section of world opinion.

The Discovery of India is a work of perennial value. Written in 1944 in the confines of the Ahmadnagar fort, it has an edge over his two other major works, "Glimpses of World History" and "Autobiography" in so far as the writer stands mellowed. It shows a much more balanced mind, a mind which always tried to look beyond the narrow confinements around him. The book shows a vision which never ceased to work and which ultimately made the man one of the greater visionaries of his time. For the sake of

comparison we have selected five samples from his book. Each sample contains one thousand words. Further count of words occurring once, twice, thrice has been determined for each sample.

The following data are complete in the sense that each sample contains one thousand words. The number of different words occurred how many times and the total number of words occurred i.e. the total frequencies of each sample are shown below:

TABLE NO. 3.1

NO. Of different Words, (X)	Number of words occurring (per 1000 words)				
	Sample-I A	Sample-II B	Sample-III C	Sample-IV D	Sample-V E
1	283	267	277	240	293
2	61	64	50	51	68
3	32	32	23	20	18
4	10	13	14	10	13
5	07	05	06	05	08
6	05	05	01	07	02
7	02	07	06	06	03
8	02	02	04	03	04
9	--	03	01	03	--
10	01	02	01	01	03
11	02	--	01	01	--
12	01	02	01	02	01
13	02	01	--	--	01
14	02	--	01	01	01
15	01	01	--	--	--
16	01	01	01	--	01
17	--	--	01	--	--
18	--	--	--	--	01
20	--	--	--	01	--
22	--	--	01	01	--
23	01	--	01	01	--
24	--	--	01	01	02
27	--	01	--	--	--
29	--	--	--	01	01
30	01	01	--	01	--
32	--	--	--	--	01
34	01	--	--	--	--
39	--	--	--	01	--
40	--	--	01	--	--
41	--	01	01	--	--
44	--	--	--	--	01

45	--	01	--	01	--
47	--	--	--	--	01
50	--	--	--	01	--
54	01	--	--	--	--
55	--	--	--	01	01
61	--	--	01	--	--
70	--	--	01	--	--
72	01	--	--	--	--
79	--	01	--	--	--
Total	417	410	395	360	424

IV. Statistical Measures

In table no. 3.1 the data are presented. It was considered to one thousand occurrences, so as to give a fairly substantial basis for the characteristics. We formed five of such samples say A, B, C, D and E, and by adding together the number of times the words occurred in any two of them A and B, A and C, B and C and so on, we could get a sample based on 2000 occurrences. By adding together the number of times the words occurrences, by adding together the number of times the word occurred in any three we could get a sample based on 3000 occurrences. By adding together the number of times these occurred in all four we could get a sample based on 4000 occurrences. And finally by adding together the number of times the words occurred in all the five, we could get a sample based on 5000 occurrences. This combined data are shown in table numbers 4.1, 4.2, 4.3 and 4.4.

The work was continued in precisely the same way. The combined samples are AB, AC, AD, AE, BC, BD, BE, CD, CE, and DE; ABC, ABD, ABE, ACD, ADE, ACE, BCD, BCE, BDE and CDE ; ABCD, ABCE, ACDE, BCDE, and ABDE and finally ABCDE. The addition of two samples per 2000, of three samples per 3000, of four samples per 4000 and of all the five samples per 5000 respectively is compiled. The distributions are shown in table 3.1 and 4.1 to 4.4. The table 3.1 gives the five distribution for the initial samples A, B, C, D and E for 1000 occurrences each : the actual number of occurrences S_1 is calculated in col.(3) of table no. 4.5, and their range is from 978 to 1002. Table no. 4.1 gives the ten distributions based on the possible pairs of A, B, C, D and E, each distribution being therefore based on 2000 occurrences each. Table no. 4.2 shows the ten distributions based on the possible triplets of A, B, C, D and E, each distribution being therefore consisted of 3000 occurrences each. Table no. 4.3 presents the five distribution based on the possible four distributions of A, B, C, D and E , each distribution being therefore based on 4000 occurrences each. Table no. 4.4 gives a distribution based on 5000 occurrences formed by all five samples taken together. The combined samples data are presented in the following tables:

TABLE NO. 4.1 : Ten Samples (per 2000 words)

NO. Of different Words, (X)	Number of words occurring (per 2000 words)									
	1	2	3	4	5	6	7	8	9	10
	AB	AC	AD	AE	BC	BD	BE	CD	CE	DE
1	550	560	523	576	544	507	560	517	570	533
2	125	111	112	129	114	115	132	101	118	119
3	64	55	52	50	55	52	50	43	41	38
4	23	24	20	23	27	23	26	24	27	23
5	12	13	12	15	11	10	13	11	14	13
6	10	06	12	07	06	12	07	08	03	09
7	09	08	08	05	13	13	10	12	09	09
8	04	06	05	06	06	05	06	07	08	07
9	03	01	03	--	04	06	03	04	01	03
10	03	02	02	04	03	03	05	02	04	04
11	02	03	03	02	01	01	--	02	01	01
12	03	02	03	02	03	04	03	03	02	03
13	03	02	02	03	01	01	02	--	01	01
14	02	03	03	03	01	01	01	02	02	02
15	02	01	01	01	01	01	01	--	--	--
16	02	02	01	02	02	01	02	01	02	01
17	--	01	--	--	01	--	--	01	01	--
18	--	--	--	01	--	--	01	--	01	01
20	--	--	01	--	--	01	--	01	--	01
22	--	01	01	--	01	01	--	02	01	01
23	01	02	02	01	01	01	--	02	01	01
24	--	01	01	01	01	01	02	02	03	03
27	01	--	--	--	01	01	01	--	--	--
29	--	--	01	01	--	01	01	01	01	02
30	02	01	02	01	01	02	01	01	--	01
32	--	--	--	01	--	--	01	--	01	01
34	01	01	01	01	--	--	--	--	--	--
39	--	--	01	--	--	01	--	01	--	01
40	--	01	--	--	01	--	--	01	01	--
41	01	01	--	--	02	01	01	01	01	--
44	--	--	--	01	--	--	01	--	01	01
45	01	--	01	--	01	02	01	01	--	01
47	--	--	--	01	--	--	01	--	01	01
50	--	--	01	--	--	01	--	01	--	01
54	01	01	01	01	--	--	--	--	--	--
55	--	--	01	01	--	01	01	01	01	02
61	--	01	--	--	01	--	--	01	01	--

70	--	01	--	01	01	--	--	01	01	--
72	01	01	01	--	--	--	--	--	--	--
79	01	--	--	--	01	01	01	--	--	--
Total	827	812	777	841	805	770	834	755	819	784

TABLE NO. 4.2 : Ten Samples (per 3000 words)

NO. Of different Words, (X)	Number of words occurring (per 3000 words)									
	1	2	3	4	5	6	7	8	9	10
	ABC	ABD	ABE	ACD	ADE	ACE	BCD	BCE	BDE	CDE
1	827	790	843	800	816	853	784	837	800	810
2	175	176	193	162	180	179	165	182	183	169
3	87	84	82	75	70	73	75	73	70	61
4	37	33	36	34	33	37	37	40	36	37
5	18	17	20	18	20	21	16	19	18	19
6	11	17	12	13	14	08	13	08	14	10
7	15	15	12	14	11	11	19	16	16	15
8	08	07	08	09	09	12	09	10	09	11
9	04	06	03	04	03	01	07	04	06	04
10	04	04	06	03	05	05	04	06	06	05
11	03	03	02	04	03	03	02	01	01	02
12	04	05	04	04	04	03	05	04	05	04
13	03	03	04	02	03	03	01	02	02	01
14	03	03	03	04	04	04	02	02	02	03
15	02	02	02	01	01	01	01	01	01	--
16	03	02	03	02	02	03	02	03	02	02
17	01	--	--	01	--	01	01	01	--	01
18	--	--	01	--	01	01	--	01	01	01
20	--	01	--	01	01	--	01	--	01	01
22	01	01	--	02	01	01	02	01	01	02
23	02	02	01	03	02	02	02	01	01	02
24	01	01	02	02	03	03	02	03	03	04
27	01	01	01	--	--	--	01	01	01	--
29	--	01	01	01	02	01	01	01	02	02
30	02	03	02	02	02	01	02	01	02	01
32	--	--	01	--	01	01	--	01	01	01
34	01	01	01	01	01	01	--	--	--	--
39	--	01	--	01	01	--	01	--	01	01
40	01	--	--	01	--	01	01	01	--	01
41	02	01	01	01	--	01	02	02	01	01
44	--	--	01	--	01	01	--	01	01	01
45	01	02	01	01	01	--	02	01	02	01

47	--	--	01	01	01	01	--	01	01	01
50	--	01	--	01	01	--	01	--	01	01
54	01	01	01	01	01	01	--	--	--	--
55	--	01	01	01	02	01	01	01	02	02
61	01	--	--	01	--	01	01	01	--	01
70	01	--	--	01	--	01	01	01	--	01
72	01	01	01	01	01	01	--	--	--	--
79	01	01	01	--	--	--	01	01	01	--
Total	1222	1187	1251	1172	1201	1236	1165	1229	1194	1179

TABLE NO. 4.3: Five Samples (per 4000 words)

NO. Of different Words, (X)	Number of words occurring (per 4000 words)				
	1	2	3	4	5
	ABCD	ABCE	ACDE	BCDE	ABDE
1	1067	1120	1093	1077	1083
2	226	243	230	233	244
3	107	105	93	93	102
4	47	50	47	50	46
5	23	26	26	24	25
6	18	13	15	15	19
7	21	18	17	22	18
8	11	12	13	13	11
9	07	04	04	07	06
10	05	07	06	07	07
11	04	03	04	02	03
12	06	05	05	06	06
13	03	04	03	02	04
14	04	04	05	03	04
15	02	02	01	01	02
16	03	04	03	03	03
17	01	01	01	01	--
18	--	01	01	01	01
20	01	--	01	01	01
22	02	01	02	02	01
23	03	02	03	02	02
24	02	03	04	04	03
27	01	01	--	01	01
29	01	01	02	02	02
30	03	02	02	02	03
32	--	01	01	01	01
34	01	01	01	--	01

39	01	--	01	01	01
40	01	01	01	01	--
41	02	02	01	02	01
44	--	01	01	01	01
45	02	01	01	02	02
47	--	01	01	01	01
50	01	--	01	01	01
54	01	01	01	--	01
55	01	01	02	02	02
61	01	01	01	01	--
70	01	01	01	01	--
72	01	01	01	--	01
79	01	01	--	01	--
Total	1582	1646	1596	1589	1611

TABLE NO. 4.4: Five Samples (per 5000 words)

NO. Of different Words,	Number of words occurring (per 5000 words)
x	ABCDE
1	1360
2	294
3	125
4	60
5	31
6	20
7	24
8	15
9	07
10	08
11	04
12	07
13	04
14	05
15	02
16	04
17	01
18	01
20	01
22	02
23	03
24	04
27	01

29	02
30	03
32	01
34	01
39	01
40	01
41	02
44	01
45	02
47	01
50	01
54	01
55	02
61	01
70	01
72	01
79	01
Total	2006

TABLE NO. 4.5: STATISTICAL CONSTANTS

1	2	3	4	5	6	7	8
Samples	**S_0**	**S_1**	**S_2**	**M**	**σ^2**	**σ**	**K**
A	417	978	13938	2.3453	27.9240	5.2843	135.497
B	410	1002	14820	2.4439	30.1737	5.4930	137.6289
C	395	997	16322	2.5240	34.9509	5.9119	154.1729
D	360	998	15396	2.7722	35.0816	5.9230	144.5570
E	424	998	13186	2.3538	25.5587	5.0555	122.3690
Mean				**2.4878**	**30.7378**	**5.5335**	**138.8450**
AB	827	1978	28690	2.3918	28.9709	5.3825	68.2738
AC	812	1960	30260	2.4138	31.4396	5.6071	73.6672
AD	777	1976	29534	2.5431	31.5429	5.6163	70.5787
AE	841	1976	27124	2.3496	26.7315	5.1702	64.4065

BC	805	1982	31074	2.4621	32.5393	5.7343	74.0570
BD	770	1998	30148	2.5948	32.4203	5.6939	70.5160
BE	834	1998	27938	2.3957	28.2486	5.3159	64.9799
CD	755	1980	31718	2.6225	35.1331	5.9273	75.8545
CE	819	1980	29508	2.4176	30.1845	5.4940	70.2173
DE	784	1996	28582	2.5459	29.9750	5.4749	66.7317
Mean				**2.4737**	**30.7186**	**5.5385**	**69.9283**
ABC	122	2960	45012	2.4223	30.9676	5.5648	47.9958
ABD	1187	2976	44086	2.2072	30.8546	5.5547	46.4175
ABE	1251	2976	41876	2.3789	27.8148	5.2740	43.9222
ACD	1172	2958	45656	2.5239	32.5856	5.7084	48.7990
ADE	1201	2974	42520	2.4763	29.2718	5.4103	44.7116
ACE	1236	2958	43446	2.3932	29.4231	5.4243	46.2732
BCD	1165	2980	46470	2.5579	33.3456	5.7746	48.9730
BCE	1229	2980	44260	2.4247	30.1338	5.4894	46.4844
BDE	1194	2996	43334	2.5092	29.9970	5.4769	44.9397
CDE	1179	2978	44904	2.5259	31.7063	5.6308	47.2753
Mean				**2.4719**	**30.6100**	**5.5308**	**46.5792**
ABCD	1582	3958	60408	2.5019	31.9251	5.6502	36.0340
ABCE	1646	3958	58198	2.4046	29.5751	5.4383	34.6233
ACDE	1596	3956	58842	2.4787	30.7245	5.5430	35.0711
BCDE	1589	3978	59656	2.5035	31.2756	5.5924	35.1847
ABDE	1611	3974	57272	2.4668	29.4655	5.4282	33.7485

Mean				2.4711	30.5932	5.5304	34.9323
ABCDE	2006	4956	73594	2.4706	30.5831	5.5302	27.9449

Now let us examine the characteristics K. The values of K for each distribution are given in table no. 4.5 , column 8. For the initial sample A, B, C, D and E of 1000 occurrences each, the characteristics range from 135.497 to 154.1729, a range of 18.6759 units. The mean value of this first group is 138.8450. For the second group of distributions, based on 2000 occurrences each, the range of K values is 64.4065 for AE to 75.8545 for CD. The mean value of this group is 69.9283. For the third group, based on 3000 occurrences each, the range of values is from 43.9222 for ABE to 48.9730 for BCD. The mean value K of this group is 46.5792 . For the fourth group the mean value of K is 34.9323. Finally , for the total distribution based on the whole 4956 occurrences, the value of K is 27.9449. All the values of group are around their mean values, because of the fluctuations of sampling. The mean values of K-characteristics for all groups show a steady continuous decrease with increasing size of samples.

The effect sample size on values of characteristics K is shown below:

TABLE NO. 4.6

No of words (Per)	K- Characteristics
1000	138.8450
2000	69.9283
3000	46.5792
4000	34.9323
5000	27.9449

An examination of the above values of K-characteristics shows that characteristics are decreasing as the size of sample increases.

Brief inspection shows how greatly this behaviour differs from that of the mean and variance; etc. The means are given in table no. 4.5 of column 5, for A, B, C, D and E. Range is from 2.3453 to 2.7722, with a general average of 2.4878. For the second group the general average of mean is 2.4737, for the third it is 2.47195, and for the fourth group it is 2.4711 and for the final table the mean is 2.4706. The values of characteristics are at first large then the values get smaller as sample size increases. For the complete distribution the mean would of course be directly proportional to S_1 column 6 and 7 of table no. 4.5 show that mean values are nearly constant. The characteristics appear to be constant in each group.

V. SUMMARY

Five independent samples each of one thousand words were selected from the book entitled “The Discovery of India” (1946) by Pandit Jawaharlal Nehru. Values of characteristics K formulated by Yule (1944) and modified by Herdan (1964) signifying the style were calculated for each of the samples.

We determine the standard error of the characteristics K with the help of large sample theory. Five values of the characteristics K were tested for their equality with the help of normal distribution. It was noted that differences between the values of characteristics K considered two by two, were not significant at 5% level of significance.

The number of different words which constitute the working vocabulary of a writer is necessarily limited. As such this must result in particular words being drawn upon oftener as the number of occurrences increase. Consequently, the number of occurrences per word, i.e. the ratio between the number of occurrences and vocabulary, increases with number of occurrences. Similarly, the standard deviation must increase with number of occurrences because the number of different words, the vocabulary, does not argument so fast as does the range of the frequency with which words are used. This suggests that a statistic which is independent of the vocabulary N might satisfy the fundamental relation between sample statistics. Such a statistics is $V/\sqrt{N}$, that is the coefficient of variation divided by $\sqrt{N}$. $V/\sqrt{N}$ is independent of N and should therefore represent a parameter of the word count which satisfies the basic requirements for sample statistics that remains sensibly constant irrespective of sample size, and thus characterizes the population.

So far calculating the moments, i.e. mean, variance, covariance and standard error etc., the following derivation is considered. Suppose that the two following conditions are satisfied: H is continuous function of m_v and m_ρ, where m_v and m_ρ are respectively $v^{th}$and ρ^{th} sample moments about the sample mean.

$m_1 = 0$, $m_2 = s^2$

$H(m_v , m_\rho)$

The theorem is given on the page 353 of Harald Cramer's book entailed 'Mathematical Methods of Statistics' (1958).

$E(H) = H_0 + O(1/n)$

$D^2(H) = \mu_2(m_v) H_1^2 + \mu_1(m_v , m_\rho) H_1H_2 + \mu_2(m_\rho) H_2^2 + 0(1/n^{3/2})$

Where, $H_0 + H(\mu_v , \mu_\rho)$,

$H_1 = \frac{\partial}{\partial m_v} H(m_v, m_\rho) \mid m_v = \mu_v , m_\rho = \mu_\rho$,

$H_2 = \frac{\partial}{\partial m_\rho} H(m_v, m_\rho) \mid m_v = \mu_v , m_\rho = \mu_\rho$

In the present case,

$$H(m_v, m_\rho) = \frac{S}{\bar{X}} \quad , H(\mu_v, \mu_\rho) = \frac{\sigma}{m}$$

Where m is the population mean and $\mu_2 = \sigma^2$

$$H_1 = \frac{\partial}{\partial s}\left(\frac{s}{\bar{x}}\right)_{\bar{x}=m} = \frac{1}{m}^2$$

$$V(s) = \frac{\mu_4 - \mu_2^2}{4\mu_2 n}$$

$$H_2 = \frac{\partial}{\partial \bar{x}}\left(\frac{s}{\bar{x}}\right) = -\frac{\sigma}{m^2}$$

$$V(\bar{x}) = \frac{\mu_2}{n}$$

Appling the theorem,

$$D^2\left(\frac{s}{\bar{x}}\right) = V(s)\left[\frac{\partial}{\partial s}\left(\frac{s}{\bar{x}}\right)\right]^2 + 2Cov(s,\bar{x})H_1H_2 + V(\bar{x})\left[\frac{\partial}{\partial \bar{x}}\left(\frac{s}{\bar{x}}\right)\right]^2 \text{-------- (*)}$$

$$\because Cov(s,\bar{x}) = E[(s-\sigma)(\bar{x}-m)] \ , \ s - \sigma = \sqrt{m_2} - \sqrt{\mu_2}$$

We have,

$$\sqrt{m_2} - \sqrt{\mu_2} = \frac{m_2 - \mu_2}{2\sqrt{\mu_2}} - \frac{(m_2-\mu_2)^2}{2\sqrt{\mu_2}\left(\sqrt{m_2} + \sqrt{(\mu_2)^2}\right)}$$

$$\therefore Cov(s,\bar{x}) = E\left[\frac{m_2-\mu_2}{2\sqrt{\mu_2}}(\bar{x}-m)\right] - E\left[\frac{(m_2-\mu_2)^2}{2\sqrt{\mu_2}\left(\sqrt{m_2}+\sqrt{(\mu_2)^2}\right)}\right]$$

$$\therefore Cov(s,\bar{x}) = \frac{Cov(\bar{x}-m_2)}{2\sqrt{\mu_2}} = \frac{1}{2\sqrt{\mu_2}}\frac{(n-1)}{n^2}\mu_3 = \frac{1}{2\sqrt{\mu_2}}\frac{\mu_3}{n}$$

By neglecting the second term which is small, the equations (*) reduces to,

$$D^2\left(\frac{S}{\bar{x}}\right) = \frac{\mu_4 - {\mu_2}^2}{4\mu_2 n}\left[\frac{1}{m}\right]^2 + 2\frac{1}{2\sqrt{\mu_2}}\frac{\mu_3}{n}\frac{1}{m}\left(-\frac{\sqrt{\mu_2}}{m^2}\right) + \frac{\mu_2}{n}\left(-\frac{\sqrt{\mu_2}}{m^2}\right)^2$$

$$= \frac{{\mu_4 - \mu_2}^2}{4\mu_2 n m^2} - \frac{\mu_3}{nm^3} + \frac{{\mu_2}^2}{nm^4}$$

$$D^2\left(\frac{s}{\bar{x}}\right) = \frac{(\mu_4 - {\mu_2}^2)m^2 - 4m\mu_2\mu_3 + 4{\mu_2}^3}{4\mu_2 n m^4} \text{ ---------- (1)}$$

According to Herdan (1964) , the characteristic K should be (book page-70),

$$V_H = \frac{V_x}{\sqrt{N}} = \frac{\sigma}{m\sqrt{N}}$$

Where N is vocabulary,

$$V_H = \frac{S}{\bar{X}}\frac{1}{\sqrt{N}} = K' \quad \text{-----Suppose------ (2)}$$

Subsequently,

$$E(V_H) = \frac{\sigma}{m}\frac{1}{\sqrt{N}}$$

$$D^2(V_H) = \frac{(\mu_4 - {\mu_2}^2)m^2 - 4m\mu_2\mu_3 + 4{\mu_2}^3}{4\mu_2 n m^4 N} \quad \text{---------- (3)}$$

With the help of this formula we calculate the moments of distribution A, B, C, D and E. First we determine the raw moments, and then the central moments of distributions. We use the equation (2) for calculating K' characteristic, and equation (3) for the values of standard errors. The values of above constants are presented below :

TABLE NO. 5.1. : The values of constants

Samples	K'	Standard Error
A	0.110336	0.005915
B	0.1110039	0.007375
C	01178504	0.003665
D	0.1126054	0.001302
E	0.1043092	0.002293
AB	0.0782543	0.002622
AC	0.0815193	0.001752
AD	0.0792271	0.001157
AE	0.0758792	0.001495
BC	0.0816576	0.001948
BD	0.0790784	0.001396
BE	0.0768222	0.001829
CD	0.0822556	0.001
CE	0.0794089	0.0012
DE	0.0768027	0.001

For estimating $D^2(V_H)$ from sample values, we have

$$D^2(V_H) = \frac{\left(m_4 - m_2^{2)}\right)\bar{X}^2 - 4\bar{X}(m_2 m_3) + 4m_2^3}{4\bar{X}^4 m_2 (nN)}$$

Where,

$$m_2 = s^2 \quad , \quad m_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\sum_{i=1}^{n}(x_i - \bar{x})^r \ ,$$

$$r = 2,3, \ldots$$

Next we consider the problem whether the characteristics derived from different samples, are the same. For this purpose let V_H and V'_H be the two characteristics based on two different but independent samples.

We have ,

$$E(V_H) = K$$

And $(V'_H) = K'$

Let $Z = V_H - V'_H$

$E(Z) = K - K'$

and $V(Z) = D^2(V_H) + D^2(V'_H)$ since the samples are independent.

Consider the hypothesis,

$H_0 : K = K'$ against $H_1 : K \neq K'$

We consider the normal test based on large samples,

$$y = \frac{z}{\sqrt{V(Z)}} \sim N(0,1) \quad , \quad for\ large\ n$$

Making use of the above equations for calculating the values of y , the calculated values of pairs A-B, C-D, A-D, A-E, B-C, B-D, B-E, C-D, C-E, D-E, AB-CD, AD-CE, BC-DE are shown below:

TABLE NO. 5.2 : Values of y

Sample Pairs	y calculated
A-B	-0.005793
C-D	0.0744148
A-D	-0.0267124
A-E	0.0665231
B-C	-0.0651586
B-D	-0.0172128
B-E	0.0681615
C-D	0.0744162

C-E	0.1754327
D-E	0.1383622
AB-CD	0.0667309
AD-CE	0.00308357
BC-DE	0.0945367

We observe that the values of calculated y lie between -1.96 to 1.96, which are critical points of a standardized normal variate at 5% level of significance. So these are not significant. The difference of K - K' is just due to the fluctuations of sampling. The sample data do not provide sufficient evidence against the null hypothesis which may therefore be accepted. Thus the characteristicsK' remains the same.

The value of K-Characteristics:

We combine all the five independent values of K'-characteristics and determine unique value of K-characteristics i.e. $\overline{\dot{K}'}$,

$$\overline{\dot{K}'} = \frac{(K_1 + K_2 + K_3 + K_4 + K_5)}{5}$$

= 138.84496

$$V(\bar{k}') = \frac{V(K_1 + K_2 + K_3 + K_4 + K_5)}{n^2}$$

= 6.147556

S.E.$(\bar{k}') = 2.4794265$

∴ The Confidence – Interval for $'$,

$K' \pm \sqrt{V(\overline{K'})}$

138.84496 ± 2.4794265

VI. CONCLUSION

Five independent samples each of one thousand words were selected from the book entitled, " The Discovery of India" (1946) by Pandit Jawaharlal Nehru. Values of K characteristic formulated by Yule(1944) and modified by Herdan (1964) signifying the style were calculated for each of the sample.

It is noted that difference among the values of characteristics K' signifying the author's style, were statistically insignificant. Further the value of characteristics K', signifying the author's style, dcreases as the sample size increases.

VII. REFERENCES

[1]. Yule G. Udny(1944): The Statistical Study of Literary Vocabulary, Cambridge University Press.

[2]. Yule Udny G. And Kendall M.G.(1968): An Introduction to the theory of statistics, Fourteenth edition, Revised and Enlarged, Universal Book Stall, New Delhi.

[3]. Williams C.B.(1946) : Yule's Characteristic" and the " Index of Diversity", Nature, Vol. 157, p.-482.

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

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

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

[7]. Herdan Gustav(1956): Languqge asChoice and Chance., P. Noordhoff Ltd., Groninggen, Holland.

[8]. Herdan G.(1964): Quantitative Linguistics, Butter Worth and Company(Publishers)Ltd.

[9]. Herdan Gustav(1966 b): " How can Quantitative Methods Contribute to Our Undersanding of Language Mixture and Language Borrowing?" in Statistique et Analyse Linguistique., Paris, pp.-17-36.

[10]. Herdan Gustav(1953): Language in the light of the theory of information, part-I, Metron., Vol. XVII, Nos.1-2, pp.-89-125.
[11]. Herdan Gustav(1955): " A New Derivation and Interpretation of Yule's Characteristics-K," Journal of Applied Mathematics and Physics(ZAMP), VI, pp.-332-334.
[12]. Herdan Gustav(1955): Language in the light of the theory of information, Part-II, Metron, Vol. XVII, Nos. 3-4,pp.-93-121.
[13]. Herdan Gustav(1958 a): The relation between the dictionary distribution and the occurrence distribution of word-length and its importance for the study of quantitative linguistics, Biometrika, Vol. 45, pp.-222-228.
[14]. Herdan Gustav(1958 b): The mathematical relation between Greenberg's index of linguistic diversity and Yule's Characteristics, Biometrika, Vol.45,pp.-268-270.
[15]. Herdan Gustav(1960): Type-taken Mathematics: A Text-Book of Mathematical Linguistics, Mouton and Co., The Hague.
[16]. Herdan Gustave(1961): A critical examination of Simon's model of certain distribution function in linguistics, Applied Statistics, Vol. X, No. 2, pp.-65-76.
[17]. Herdan Gustav(1962): Calculus of Linguistic Observations, Mouton and Co., The Hague.
[18]. Herdan Gustav(1966 a): The Advanced Theory of Language as Choice and Chance, New York.
[19]. Nehru Pandit Jawaharlal(1946): " The Discovery of India", Second Edition, Meriden Books Ltd, London.

Cite this article as :

Dr. Ashok Y. Tayade, "Statistical Measures of Writing style by using K-Characteristics Criteria", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 1, pp. 312-326, January-February 2020.

Journal URL : http://ijsrset.com/IJSRSET207350