International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 06, June 2017
www.ijates.com
ijates
ISSN 2348 - 7550

HERDAN'S K*: A CRITERION FOR WRITING STYLE

Dr. Kumud Gore Kherdekar

Associate Professor and Head, Dept. of Statistics, Govt. College of Arts and Science, Aurangabad

ABSTRACT

Word distribution is a typical type of distribution, wherein the complete form ofthe distribution changes as the sample size increases. Hence Yule (1944) has suggested his well-known characteristic K, which characterizes word distribution and yetis independent of size of sample. Yule's K is based on the assumption that the worddistribution (complete) is of Poisson type. This assumption was attacked by linguistsand Ross (1950) and hence laid to reformulation of K. The new characteristic is givenby Herdan (1955) as Herdan's K. Present paper is study of Herdan's K*. Yule (1944)has posed a problem of standard errors of functions of word distribution to theoreticalstatisticians. Here we will try to establish standard error of Herdan's K*. Further wetry to analyze three novels of famous English author Thomas Hardy, which are"Two On A Tower", "Desperate Remedies" and "Tess Of D'Urbervillies".*

I. INTRODUCTION

Due to peculier characteristic of word distribution the comparison of the works of two authors and so also comparison of the works of the same author becomes very difficult. Hence something which characterises the word distribution and yet is independent of size of sample is necessary. This type of characteristic was provided by Yule(1944), and Herdan(1955), termed as Yule's Characteristic K or Herdan's Characteristic K*. According to Yule and Herdan the Characteristic charecterises the word distribution and yet it is independent of size of sample.

1.1 Yule's Characteristic K

Yule(1944), compared word distribution with the distribution of number of persons meeting with said number of accidents in a stipulated time, with resemblancesin two respects.First, as the length of the period of exposure to risk increases, the form of thedistribution changes. Similarly in case of word distribution, as the sample size isincreased the complete form of the distribution changes. And second, in case ofaccident distribution, number of persons met with a few accidents may drop of, whereas in word distribution number of words used a few times (once, twice,…)may get deleted.Similarly, these two distributions differ in two respects. One of which is that,in case of word distribution the total number of words at risk of being used (total number of words in author's treasure) are unknown, that is the population size is unknown. In this sense the word distribution is termed as “Incomplete or Decapitated distribution." The total number of words used in the text under consideration is called sample. Every word in the population is not at the same risk of getting used. The other difference is that, the word distribution does not involve time. Occurrence of said number of accidents in a stipulated time seem to be reasonable. But occurrence of a particular word so many times in a stipulated time period does not carry any meaning. An author may write a text in a few days or may take several years to write.

With these resemblances mentioned, Yule (1944), approximated the word distribution to a Poisson probability distribution. The random variable x is defined as,

$$X = x : Number\ of\ times\ a\ word\ ocurs.$$
$$x = 1, 2, 3, \dots, n$$
$$and\ f_x is\ the\ frequency\ of\ X = x$$

Then X is a Poisson variable with mean λ which is, average number of times a word occurs.

$$Thus\ E(x) = \lambda\ and$$
$$V(x) = \lambda$$

Yule introduced an arbitrary frequency $f_0$for the words not used in the sample, to complete the decapitated distribution. Since it is the frequency of$X = x$, defined as the number of times a word is used and $X = x = 0$ corresponding to f_0 for the sums $\sum x f_x , \sum x^2 f_x$ remain unaltered even after introduction of arbitrary frequencyf_0. By additive property of independent Poisson variates, λ defined above is also a Poisson random variable.

$$E(\lambda_i) = \bar{\lambda}$$
$$= mean\ of\ \lambda_i\ s$$

and standard error of $\lambda_i = \sigma_\lambda$.

Subsequently Yule (1944), has defined Characteristic K as,

$$\nu_\lambda = K = Coefficient\ of\ variation\ of\ \lambda$$
$$= \frac{\sigma_\lambda}{\bar{\lambda}}$$

Different components together for which is are computed will form a complete worddistribution.We define,

$$\left.\begin{array}{c} S_0 = \sum f_x \\ S_1 = \sum x f_x \\ and\ S_1 = \sum x^2 f_x \end{array}\right\} \quad (1)$$

The Characteristic K , independent of size of sample is for practical calculations is defined by Yule (1944), as,

$$K = 10{,}000 \left(\frac{S_2 - S_1}{S_1^2}\right)$$
$$= 10{,}000 \left(\frac{S_2}{S_1^2} - \frac{1}{S_1}\right)$$

The figure 10,000 is introduced to avoid too many decimals. As the sample size increases, the term$\frac{1}{S_1}$tends to be negligible. Hence, for large sample,

$$K = 10^4 \frac{S_2}{S_1^2} (2)$$

Yule's characteristic provides an objective measurement of one significant aspect of literary style. Thus if the samples are random and are chosen by correct sampling procedure and if they are representative of one and the same text, the characteristic K is expected to be constant. In spite of random sampling procedure, if K is exhibiting change then it reflects change in word distribution meaning thereby change in the style of the author with respect to the characteristic K.

1.2 Herdan'Sk

Derivation of Yule's K was possible only under the assumption that occurrenceof a word is governed by Poisson law. This assumption and hence the use of Yule'sK was attacked by linguists.The notion that words come out of a "treasure chest" is something like the sameway that coloured balls emerge from a container in a classical

instance of randomnesswas attacked by Ross,(1950) and led to a reformulation of K.Herdan (1955), analysed the situation without any assumption about stochastic process of the statistic.According to him, the number of different words which constitute the workingvocabulary of a writer is limited and it results in occurrence (frequency) of somewords. Consequently mean number of occurrences per word will increase with theincrease in sample size. Similarly the standard deviation will also increase. Since therate of increase of vocabulary N, that is number of different words decreases as thesample size$\sum xf_x$ increases. The statistic he formulated is Characteristic K*, known as Herdan's K* and,

$$K^* = \frac{v_x^2}{N}$$
$$= \frac{\sigma_x^2/N}{\bar{x}^2}$$
$$= \frac{\frac{1}{N}\left[\frac{1}{N}\sum x^2 f_x - \bar{x}^2\right]}{\bar{x}^2}$$
$$= \frac{\frac{1}{N}\sum x^2 f_x}{\bar{x}^2} - \frac{1}{N}$$
$$= \frac{\sum x^2 f_x}{(\sum x f_x)^2} - \frac{1}{N}$$
$$= \frac{S_2}{S_1^2} - \frac{1}{N}$$

$$For\ large\ N, \frac{1}{N} \to 0\ and\ can\ be\ neglected. Thus$$

$$K^* = \frac{v_x^2}{N}$$
$$= \frac{S_2}{S_1^2}$$

If K* is also multiplied by $10^4$to avoid too many decimals we get Yule's Characteristic K, for large samples. Thus it is very important to note that for large samples as$\frac{1}{N}$ and $\frac{1}{S_1}$tend to be negligible.

Yule's K = Herdan'sK* (without assumption of stochastic process)

Hence K*represents a parameter of the word count which satisfies the basic requirement for sample statistic, that it characterizes the population.It is independent of N, the vocabulary. Herdan did not assume the stochastic process of K*,as was implied in Yule's derivation of K (Poisson law). Now we use this Characteristic for the analysis of writing style of the author Thomas Hardy.

II. METHODOLOGY AND PROCEDURE

The three novels were considered for this study. When different texts of an author are to be compared, it is necessary to test if the texts are consistent within themselves. If so, then two or more texts of the same author can be considered for comparison. Hence first we are going to study the consistency pertaining to stylistic parameter K* for three novels of the author Thomas Hardy within themselves. If the consistency is confirmed, then we can compare two or more novels for the consistency of parameter, Characteristic K*.It is decided to divide each novel in two parts. Part I is starting of the novel. PartI ends at the page where twenty samples of predetermined size are drawn. SimilarlyPart II is end of the novel which consists of twenty samples of same size as in Part I. Two novels “Desperate Remedies" and “Tess OfD'Urbervillies" contain approximately 1, 50,000

words. Hence it is decided to take 20 samples of 3000 words fromeach Part. The third novel, "Two On A Tower " consists of approximately 95,000 words. If the same sample size is taken (3000), we would not have got non overlapping samples from two Parts. Hence it is decided to have 20 samples from each part of size 2000 words in each sample.

Use of computer made sampling easy, correct and quick.

2.1 Statistical Analysis

For the statistical analysis of the three novels we have used the Herdan's K* and not the Yule's K, because

i. There is no stochastic assumption for the Herdan's K*

ii. Herdan used all words and not "only nouns".

The random variable defined as X is,

$$X = x \qquad : Number\ of\ times\ a\ word\ occurs$$
$$f_x : frequency\ of\ X = x$$

Also we denote $S_0, S_1$and S_2as in (01).

Here we define Herdan's Characteristic as,

$$K^*_{ijp} = 10{,}000\left(\frac{S_2}{S_1^2} - \frac{1}{N}\right), \qquad N = S_0$$

It is the value of Herdan'sK* for j^{th}sample of i^{th}Part of p^{th}novel.

$$i = 1, 2 \quad indicates\ Part\ I\ and\ Part\ II\ respectively$$
$$j = 1, 2, \ldots, 20 \quad indicates\ 20\ samples\ from\ each\ part$$
$$and\ p = 1, 2, 3 \quad indicates\ the\ three\ novels\ defined\ in\ section.$$

$p = 1$ The novel Desperate Remedies

$p = 2$ The novel Two on a Tower

$p = 3$ The novel Tess OfD'Urbervillies

Then we define,

$\bar{K}^*_{ip} = \frac{1}{20}\sum K^*_{ijp}$ is the mean of Herdan's K^*_{ijp} for the i^{th}Part of p^{th}novel.

Hence from the Central Limit Theorem, we get,

$$\bar{K}^*_{ip} \to N\{E(K^*_{ijp}), V(K^*_{ijp})\}$$
for$i = 1, 2$ and $p = 1, 2, 3$

$$E(K^*_{ijp}) = E\left[\frac{1}{20}\sum_{j=1}^{20} K^*_{ijp}\right]$$
$$= \frac{1}{20}\sum_{1}^{20} E(K^*_{ijp})$$
$$where\ E(K^*_{ijp}) = E\left(\frac{vx^2}{N}\right)$$
$$= \frac{1}{N}E\left(\frac{m_2}{m_1'^2}\right)$$

$$= \frac{1}{N}\frac{\mu_2}{\mu_1'^2} \qquad (03)$$

$$V(\bar{K}_{ip}^*) = V\left[\frac{1}{20}\sum_{j=1}^{20} K_{ijp}^*\right]$$
$$= \frac{1}{400}\sum_{j=1}^{20} V(K_{ijp}^*) \qquad (\because K_{ijp}^* \ s \ are \ independent)$$
$$= \frac{1}{20}V(K_{ijp}^*)$$

To find,

$$V(K_{ijp}^*) = V\left(\frac{\nu x^2}{N}\right)$$
$$= \frac{1}{N^2}(2\nu_x)^2 \,.\, V(\nu_x) : (from\ 03) \qquad (04)$$
$$\nu_x = \frac{\sigma_x}{\bar{x}}$$
$$= \frac{\sqrt{m_2}}{m_1'}, \qquad m_1' > 0$$

where m_2 is the second sample moment about the mean and m_1' is the first sample moment about origin.

We denote Q for population coefficient,

$$Q = \frac{\sqrt{\mu_2}}{\mu_1'}, \qquad \mu_1' > 0$$

If $m_r \ and \ m_r' \ are \ r^{th}$ sample moments about mean and origin respectively and $\mu_r \ and \ \mu_r' \ are \ r^{th}$ population moments about mean and origin respectively then we have,

$$E(m_r) = \mu_r\,, \quad E(m_r') = \mu_r',$$
$$V(m_r) = \frac{1}{N}[\mu_{2r} - \mu_r^2 + r^2\mu_2\mu_{r-1}^2 - 2r\mu_{r-1}\mu_{r+1}]$$
$$V(m_r') = \frac{1}{N}(\mu_{2r}' - \mu_r'^2)$$

and

$$Con(m_r, m_1') = \frac{1}{N}(\mu_{r+1} - r\mu_2\mu_{r-1})$$

we get

$$V(\nu_x) = \frac{Q^2}{N}\left\{\frac{\mu_4 - \mu_2^2}{4\mu_2^2} + \frac{\mu_2}{\mu_1'^2} - \frac{\mu_3}{\mu_2\mu_1'}\right\}$$

Hence from (04) we get,

$$V(K_{ijp}^*) = \frac{4}{N^3}.Q^4\left\{\frac{\mu_4 - \mu_2^2}{4\mu_2^2} + \frac{\mu_2}{\mu_1'^2} - \frac{\mu_3}{\mu_2\mu_1'}\right\} \qquad (05)$$

(Ref: Kendall Maurice and Stuart Alan Vol I, (1977), pages 244 – 248) Estimate ofmean and variance of $\widehat{K_{ijp}^*}$ can be obtained from sample as,

$$Estimate\ of\ E(K_{ijp}^*) = \frac{1}{n}\sum K_{ijp}^*$$
$$Estimate\ of\ V(K_{ijp}^*) = \frac{1}{n}\sum K_{ijp}^{*\,2} - \bar{K}_{ip}^{*\,2} \qquad (06)$$

With this exercise, we can use large sample tests to test the significance of difference between $\bar{K}_{ip}^*$s.

2.2 Application of Large Sample Tests

To test the hypotheses

$$H_{0p} \quad : \quad E(\bar{K}_{1p}^*) = E(\bar{K}_{2p}^*), \qquad p = 1,2,3$$
$$against\ H_{1p} \quad : \quad E(\bar{K}_{1p}^*) \neq E(\bar{K}_{2p}^*)$$

The test statistic used is:

$$Z_{p_{cal}} = \frac{\bar{K}_{1p}^* - \bar{K}_{2p}^*}{\sqrt{V(\bar{K}_{1p}^*) + V(\bar{K}_{2p}^*)}} \rightarrow N(0,1)$$
$$= \frac{\bar{K}_{1p}^* - \bar{K}_{2p}^*}{\sqrt{\frac{1}{20}V(K_{1jp}^*) + \frac{1}{20}V(K_{2jp}^*)}} \rightarrow N(0,1)$$

To test the hypotheses $H_{0p}, p = 1,2,3$ at level of significance α, the tabulated value Z_{tab} is obtained at $Z_{\alpha/2}$ for two tailed tests. Hence, for $\alpha = 0.05$

$$Z_{tab} = 1.96$$

If $|Z|_{pcal} > Z_{tab}\, H_{0p}$ is rejected at α% level of significance, otherwise H_0 may be accepted.

The data for testing of hypotheses procedure for novels Desperate Remedies,Two On A Tower and Tess Of D'Urbervillies is obtained by the procedure describedearlier and the results are depicted in the following table.

TABLE

Parts of Novel	Sample Size, Mean and variance	Name of the Novel		
		Desperate Remedies $p = 1$	Two On A Tower $p = 2$	Tess Of D'Urbervillies $p = 3$
I.	Sample Size Mean Variance	3000 98.2780 111.9280	2000 95.5577 202.1120	3000 99.2282 158.3670
II.	Sample Size Mean Variance	3000 106.5070 232.2880	2000 95.2523 63.7262	3000 97.6296 113.7980
$To\ test \quad H_{0p} \quad : \quad E(\bar{K}_{1p}^*) = E(\bar{K}_{2p}^*)$ $against \quad H_{1p} \quad : \quad E(\bar{K}_{1p}^*) \neq E(\bar{K}_{2p}^*)$ $p = 1,2,3$				
$\lvert Z\rvert_{pcal}$		1.9835	0.08376	0.433349
Z_{ptab}		1.96	1.96	1.96
Conclusion		$Reject \quad H_{01}$	$Accept \quad H_{02}$	$Accept \quad H_{03}$

III. RESULT

The Table gives results of analysis of data of Herdan'sK_ by application of largesample test at a glance.

$Z_{1cal} = 1.9835$for the novel “Desperate Remedies" results in rejection of nullhypothesis. It clearly shows that the writing style of Thomas Hardy does not showconsistency regarding style Characteristic Herdan'sK*, between two parts of thenovel. That is from starting to end of the novel the Characteristic is not consistent.Whereas $Z_{2cal} = 0.08376$and$Z_{3cal} = 0.33349$result in acceptance of thenull hypotheses. This implies that the author has maintained uniformity pertaining to style trait Herdan'sK*within the novels “Two On A Tower" and “Tess OfD'Urbervillies"respectively.

Due to the consistency shown in above analysis for the two novels, within themselves, let us test the significance of difference between two novels regarding Characteristic K*.

IV. COMPARISON OF TWO NOVELS

Consider the two novels “Two On A Tower" and “Tess Of D'Urbervillies", to test consistency regarding K*, the style parameter.To test this consistency we obtain combined means for both the parts to representthe novel by using

$$Meanofcombinedseries\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

Also we obtain variance for combined samples of both the parts for a novel by using,

$$\sigma^2 = \frac{1}{n_1 + n_2}[n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)]$$

Where$n_1, n_2$are the sizes, $\bar{x}_1, \bar{x}_2$ are the means and $\sigma_1^2\ and\ \sigma_2^2$ are the variances for the two parts of a novel and $d_i = \bar{x}_i - \bar{x}, i = 1, 2$.Thus we get,

$\bar{x}$ = Mean of two combined parts of the novel, “Two On A Tower”

σ_x^2 =Variance of two combined parts of the novel, “Two On A Tower”

Similarly

$\bar{y}$andσ_y^2represent the mean and variance for the novel, “Tess Of D'Urbervillies”. We have,

$$\bar{x} = 95.4050, \quad \sigma_x^2 = 132.954067$$
$$\bar{y} = 98.4289, \quad \sigma_y^2 = 136.721380$$

To test the hypothesis,

$$H_0: \text{□}_x = \text{□}_y$$

$$against H_1: \text{□}_x \neq \text{□}_y$$

where□$_x$ and □$_y$ are population means corresponding to $\bar{x}$and $\bar{y}$respectively. The test statistic used is,

where□$_x$ and □$_y$ are population means corresponding to $\bar{x}$and $\bar{y}$respectively. The test statistic used is,

$$Z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{V(\bar{x})}{40} + \frac{V(\bar{y})}{40}}} \rightarrow N\ (0,1)$$

Hence,

$$Z_{cal} = 1.164599\text{and for two tailed}$$
$$test\ Z_{tab} = 1.96\text{ at 5\% level of significance}$$

Hence we accept $H_0 : \square_x = \square_y$ That is the two novels do not differ significantly with respect to Characteristic K*of writing style of Thomas Hardy.

V. CONCLUSION

i. $E(K^*) = \frac{1}{N}\frac{\square_2}{\mu'^2_1}$

ii. $V(K^*) = \frac{4}{N^3}Q^4\left\{\frac{\square_4-\square_2^2}{4\square_2^2} + \frac{\square_2}{\mu'^2_1} - \frac{\square_3}{\square_2\mu'_1}\right\}$

Where $Q = \frac{\sqrt{\square_2}}{\mu'_1},\ \mu'_1 > 0$

iii. Style analysis of three novels pertaining to Stylistic Criterion Herdan's K*:

a. The novel "Desperate Remedies" does not show consistency within itself.
b. The novels "Two OnA Tower" and "Tess Of D'Urbervillies" are consistentwithin themselves.
c. The novels "Two On A Tower" and "Tess Of D'Urbervillies" are consistentwith each other.

REFERENCES

[1.] Gupta S. C. and Kapoor V. K., Fundamentals of Mathematical Statistics, Eleventh Edition,(2002), Sultan Chand and Sons.

[2.] Hardy Thomas, Desperate Remedies, Edition (1896), Macmillan & Co. Ltd, New York.

[3.] Hardy Thomas, Tess OfD'Urbervillies, Library Edition (1952), Macmillan & Co. Ltd.

[4.] Hardy Thomas, Two On A Tower,Edition (1949), Macmillan & Co. Ltd.

[5.] Herdan Gustav, Quantitative Linguistic, Butterworth & Co.(Publishers) Ltd.,(1964)

[6.] Hogg Robert V. and Tanis Elliot A., Probability and Statistical Inference, Third Edition, Maxwell Macmillan International Editors,(1989).

[7.] Sir Kendall Maurice and Stuart Alan, The Advanced Theory of Statistics Vol I, Distribution Theory, Fourth Edition, Charles Grifin and Company United,(1977).

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