A NOTE ON THE STATISTICAL ANALYSIS OF SENTENCE-LENGTH AS A CRITERION OF LITERARY STYLE

By C. B. WILLIAMS, Sc.D.
Department of Entomology, Rothamsted Experimental Station

Some years ago I made a number of calculations of the frequency distribution of words of different length in different books to see to what extent authors kept to a definite distribution and so perhaps might be identified by such a method. The results obtained, however, were not striking and the work was put at one side.

Mr Udny Yule (1939), however, has attacked the problem of authorship from the angle of the variation in sentence length, and this appears to be a much more fertile method of approach.

Mr Yule shows that the frequency distribution of sentence length (i.e. number of words between successive full stops) is of the skew type and by comparing in two different manuscripts, the mean, the median, quartiles and deciles he is able to produce convincing mathematical evidence on the identity or otherwise of their authorship.

Mr Yule does not comment on the skew distribution further than to state (p. 371) "they are not of the Poisson type, but of the type in which the square of the standard deviation largely exceeds the mean".

When I converted some of Yule's tables into diagrams I was struck by their general resemblance to certain skew distributions with which I have recently been dealing in some Entomological problems, and which distributions, I found, became normal and symmetrical if the logarithm of the number was taken as a basis for subdivision into groups instead of the number itself (see Williams, 1927).

I was unable to test this transformation on Yule's figures as he unfortunately does not give the original data, but only the word length of sentences in groups of five; so it was necessary to obtain some new data.

These I obtained by counting the number of words in each of 600 sentences from the following three books:

(1) G. K. Chesterton, *A Short History of England*, 1917.

(2) H. G. Wells, *The Work, Wealth and Happiness of Mankind.*

(3) G. Bernard Shaw, *An Intelligent Woman's Guide to Socialism.*

All three works deal with the exposition of somewhat similar sociological subjects and none of them are in the "conversational" style.

The selection of the sentences was randomized as follows. Each of the books is divided up into chapters, sections or both. In Chesterton's book the first 30 sentences were counted in each of the first 20 chapters. In Wells's book the first 10 sentences were counted in each chapter subdivision up to chapter VII, division 11. In Shaw's book the first 15 sentences in each of sections 1–40 were taken. In each case the greater part of the book was covered.

The original data thus obtained are shown diagrammatically in Fig. 1. Each of the distributions is of the typical skew type obtained by Yule: Shaw is the most extreme and varies from 3 words to 143; Wells is less skew and ranges from 3 to 91; while the Chesterton curve is the least skew and varies from 5 to 91 with only two values over 60.

From Table I it will be seen that the arithmetic mean number of words per sentence is 25·87 for Chesterton, 24·11 for Wells and 31·23 for Shaw. The medians are also different and presumably the quartiles and deciles, but these latter were not calculated.

TABLE I

Frequency constants of the distributions of sentence length

	Chesterton	Wells	Shaw
Number of sentences	600	600	600
Number of words	15,521	14,463	18,735
Arithmetic mean no. of words	25·87	24·11	31·23
Median no. of words	25·3	20·8	26·0
Mean log no. of words	1·37	1·31	1·39
Geometrical mean no. of words	23·5	20·5	24·5
Standard deviation of mean log	0·200	0·237	0·290
Standard error of mean log	0·0080	0·0095	0·0112

If, however, instead of taking the frequency distribution of the actual number of words per sentence we take that of the logarithm of the number we get the distributions shown in Figs. 2–4. They undoubtedly show a very close resemblance to the "normal distribution". The mean log and standard deviation for Chesterton is $1·37 \pm 0·20$; for Wells $1·31 \pm 0·24$ and for Shaw $1·39 \pm 0·29$. The standard error of the mean is, owing to the large number of observations, in all cases very small and approximately $\pm 0·01$.

On each of the three figures is superimposed a normal curve of the same area, mean and standard deviation and it will be seen how closely it fits the observed values.

The following comments may, however, be made:

(1) The greater irregularity of the observed values in the lower portion of the distribution is due to the irregular distribution of the logarithms of integers when grouped in small artificial divisions as in the present case. Thus there is no logarithm of an integer between 0·01 and 0·25; none between 0·61 and 0·65 and between 0·71 and 0·75. On the other hand, there are two between 1·11

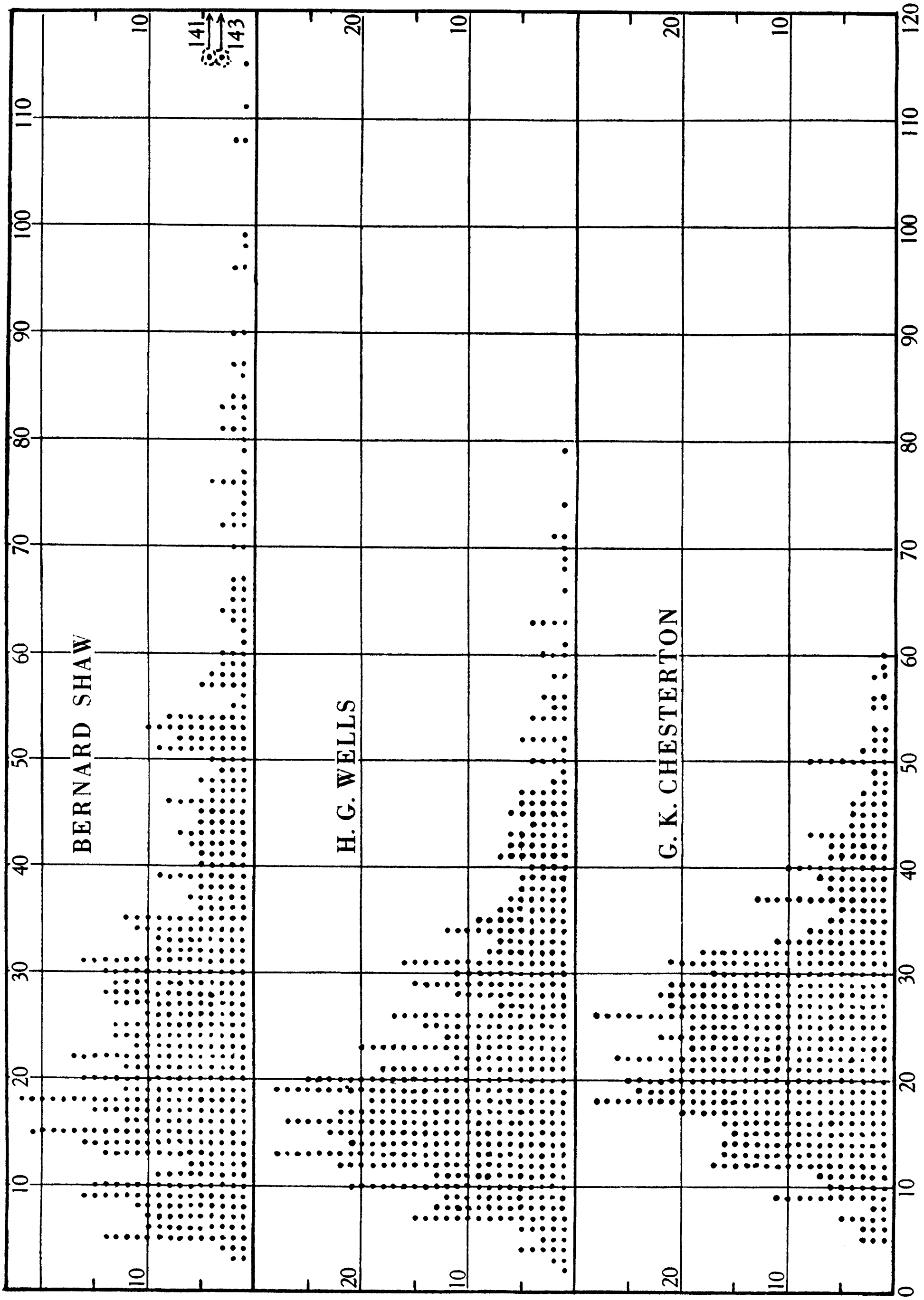
BERNARD SHAW
H. G. WELLS
G. K. CHESTERTON
141
143
0
10
20
30
40
50
60
70
80
90
100
110
120
10
20
10
20
10

Fig. 1.

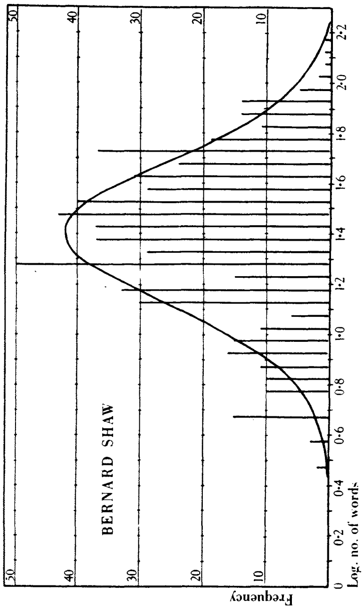

Fig. 2.

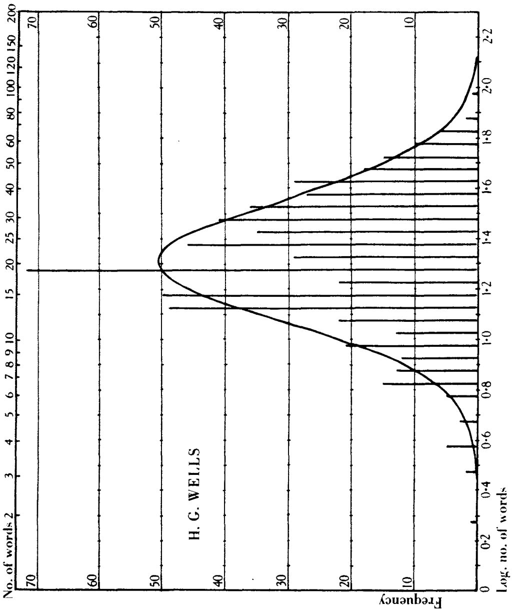

Fig. 3.

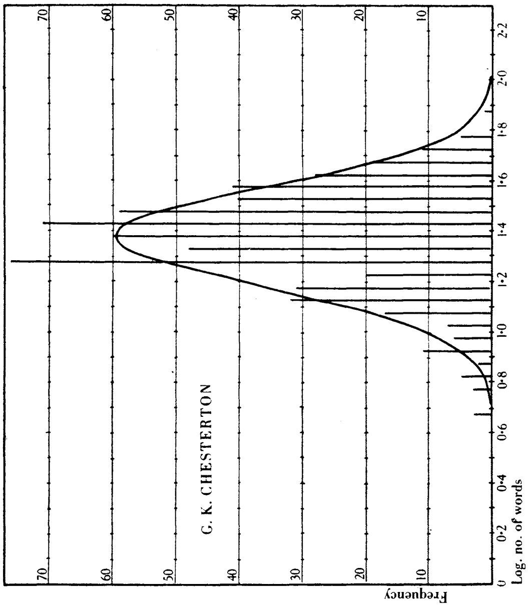

Fig. 4.

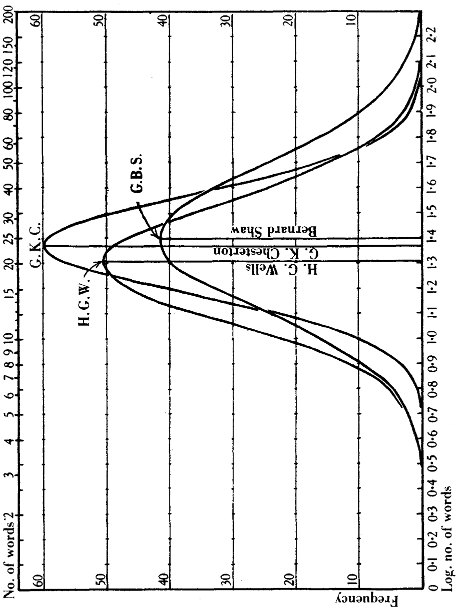

Fig. 5.

and 1·15; two between 1·16 and 1·20; only one between 1·21 and 1·25 and three between 1·26 and 1·30. Thus in all three diagrams the frequency of 1·21–1·25 is well down and that of 1·26–1·30 is far up. Some process of smoothing would undoubtedly eliminate these irregularities, but it was thought better to leave the data in their original form and draw attention to the sources of irregularity.

(2) There appears to be on all three diagrams a slight shortage of high values and a slight excess of low ones. On the latter I have no comment but it appears possible that the small deficit in the longer sentences might easily be due to a biased effect introduced by the habit that many writers have of cutting up unusually long sentences into their component parts when reading over their manuscript or proofs.

On the assumption that the normal curve is a sound representation of the frequency distribution of the log number of words in sentences, Fig. 5 has been prepared which shows the means and normal curves for the three books superimposed on one another. The means are close together but the distributions are very different.

The difference in means between Shaw and Wells is 0·09 and the standard error of the difference is only 0·015. Thus the difference is six times the standard error and hence certainly significant. Between Shaw and Chesterton the difference of the means is barely significant but that between the standard deviations is quite striking.

If the above reasoning is correct, it is unnecessary for the comparison between two documents to compare arithmetic means, medians, quartiles and deciles, but only the log mean and the log standard deviation; all other comparisons are included in these.

It follows also that Mr Bernard Shaw, while undoubtedly under the impression that he was punctuating at his own free will, was for this particular book hide-bound within the limits of

$$Z = \frac{1}{0{\cdot}29\sqrt{(2\pi)}}\exp\left[\frac{(1{\cdot}4-x)^2}{2(0{\cdot}29)^2}\right],$$

while similarly Mr Wells was writing under the restricting influence of

$$Z = \frac{1}{0{\cdot}24\sqrt{(2\pi)}}\exp\left[\frac{(1{\cdot}3-x)^2}{2(0{\cdot}24)^2}\right],$$

where Z is the frequency and x the logarithm of the number of words per sentence.

It is also perhaps worthy of passing comment that the curve representing Mr Shaw is short and broad while that representing Mr Chesterton is tall and slender; which shows how necessary it is to use these curves only for the purpose for which they were originally designed.

Perhaps something might be added on the meaning in words of the above mathematical transformation. If the log distribution is normal we can infer that the extent to which a sentence in the process of writing is likely to vary is at any level proportional to the length of the sentence. Thus when he is thinking in short sentences of about 10 words an author is likely to vary say from 8 to 12 words; when he is thinking in longer sentences of say 100 words he will vary from 80 to 120. In other words the variations are proportional or geometric and do not merely involve the addition or subtraction of x words at all levels. Further, if the geometric mean is taken as a basis, sentences between this and half its length are as frequent as those between it and twice its length; sentences down to one-quarter its length are as likely to occur as sentences up to four times its length; and so on.

If the arithmetic mean were the true basis then sentences of 10 words more than the arithmetic mean would be as likely to occur as sentences of 10 words less, and it is easy to see that this is not the case.

Before the whole theory of the use of such distributions for separating works of different authorships can be fully accepted it will of course be necessary to study the results obtained from many different works by the same author, in different styles, on different subjects and at different periods of his life. From these it may be possible to find what variation can occur "within authors" as compared with "between authors". This note is not meant to deal with this basic problem but only to draw attention to the simplification of the method of approach to such a problem by the use of a transformation which produces a normal instead of a skew distribution.

REFERENCES

WILLIAMS, C. B. (1927). *Ann. Appl. Ent.* **24**, 404.

→ YULE, G. UDNY (1939). *Biometrika*, **30**, 363–90.