J. R. Statist. Soc. A,
(1974), **137**, *Part* 1, p. 25

On a Distribution Representing Sentence-length in written Prose

By H. S. SICHEL

University of the Witwatersrand, Johannesburg, South Africa

SUMMARY

A new model for representing sentence-length distributions is suggested in equation (8) which is a special case of equation (2), with parameter $\gamma = -\frac{1}{2}$ known *a priori*.

Eight known sentence-length frequency counts taken from English, Greek and Latin prose were all satisfactorily described by distribution (8). For these eight fits, the average probability $P(\chi^2)$ was 0·50.

A ninth observed distribution, taken from a Latin text of unknown authorship failed the χ^2 test applied to the fit of the data to the model in equation (8). This corroborates Yule's (1939) conclusion that it is highly unlikely that de Gerson could have written *De Imitatione Christi*. It is further conjectured that the last-mentioned observed frequency distribution could be well represented by the more general model in equation (2), with a parameter γ much smaller than $-\frac{1}{2}$.

Keywords: SENTENCE-LENGTH; COMPOUND POISSON DISTRIBUTION; CLASSICAL PROSE

1. INTRODUCTION

THE first substantial investigation on sentence-length as a statistical tool to be used in deciding disputed authorship was published by Yule in 1939. Simple statistical indices such as the average number of words per sentence and the standard deviation of sentence-lengths were employed. Yule did not suggest a particular mathematical distribution model. Later (Yule, 1944) he explored word-frequency of an author in addition to sentence-length. Although Yule mentions in that book the negative binomial, he discards this distribution model as totally inadequate for representation of word frequencies *and* sentence-lengths.

Williams (1940, 1970) suggests and uses the lognormal distribution as a model for sentence-length. To verify lognormality, Williams plots the observed cumulative percentage frequencies of sentence-lengths on log-probability paper in the hope that these plots will approach a straight line. No χ^2 tests are given for any of Williams's examples.

Wake (1957), who discusses sentence-lengths in works of Greek authors, also makes use of the lognormal distribution by superimposing the observed histograms of the logarithms of sentence-lengths over the "expected" normal distributions. No χ^2 tests are given.

The authorship of Greek prose is again investigated by Morton (1965) who works with distribution-free statistics such as the mean, the median, the quartiles and the deciles.

Mosteller and Wallace (1963), in their study of the authorship of the Federalist papers, came to the conclusion that the mean and standard deviation of sentence-length was of no help in solving disputed authorship. In their particular research Mosteller and Wallace found the mean and standard deviations of sentence-length to be virtually identical for Madison and Hamilton. It can be shown, however, that two

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

discrete distribution models, with the same first two moments, may have entirely different shapes. For example, the negative binomial may be J-shaped whereas the new distribution discussed in this paper may be unimodal with a mode far away from zero, although the same mean and standard deviation are common to both models. Furthermore, the other investigators mentioned previously, have shown that some authors differ decisively in mean sentence-lengths.

It would be of great help to have a reasonable mathematical distribution model for sentence-length in order to sharpen our statistical tools, not only with respect to the enhanced power in significance testing but also to investigate the shape of the sentence-length distribution. In addition, a few pertinent statistical indices could be used to express sentence-lengths instead of showing massive tables of frequencies of the number of words in sentences.

The lognormal model suggested by Williams and used by Wake must be rejected on several grounds: In the first place the number of words in a sentence constitutes a discrete variable whereas the lognormal distribution is continuous. Wake (1957) has pointed out that most observed log-sentence-length distributions display upper tails which tend towards zero much faster than the corresponding normal distribution. This is also evident in most of the cumulative percentage frequency distributions of sentence-lengths plotted on log-probability paper by Williams (1970). The sweep of the curves drawn through the plotted observations is concave upwards which means that we deal with sub-lognormal populations. In other words, most of the observed sentence-length distributions, after logarithmic transformation, are negatively skew. Finally, a mathematical distribution model which cannot fit real data—as shown up by the conventional χ^2 test—cannot claim serious attention.

2. The Model

It has been pointed out by some of the writers mentioned previously that sentence-lengths are not randomly distributed throughout a given text written by a certain author. A tendency of some serial correlation between the lengths of successive sentences has been observed. This points to "clustering" and one immediately thinks of some compound Poisson process seeing that the underlying distributions must be discrete.

Recently, Sichel (1971) proposed a family of discrete distributions which arises from mixing Poisson distributions with parameter λ. The mixing distribution is given by

$$f(\lambda) = \frac{1}{2}\frac{\{2\sqrt{(1-\theta)/\alpha\theta}\}^{\gamma}}{K_{\gamma}\{\alpha\sqrt{(1-\theta)}\}}\lambda^{\gamma-1}\exp\left\{-\left(\frac{1}{\theta}-1\right)\lambda-\frac{\alpha^2\theta}{4\lambda}\right\}. \tag{1}$$

Here $-\infty<\gamma<\infty$, $0<\theta<1$ and $\alpha>0$ are the three parameters and $K_{\gamma}(.)$ is the modified Bessel function of the second kind of order γ. The resulting compound Poisson distribution is

$$\phi(r) = \frac{\{\sqrt{(1-\theta)}\}^{\gamma}}{K_{\gamma}\{\alpha\sqrt{(1-\theta)}\}}\frac{(\alpha\theta/2)^r}{r!}K_{r+\gamma}(\alpha), \tag{2}$$

where $r = 0, 1, 2, \ldots, \infty$.

A number of known discrete distribution functions such as the Poisson, negative binomial, geometric, Fisher's logarithmic series in its original and modified forms, Yule, Good, Waring and Riemann distributions are special or limiting forms of (2).

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

If parameter γ is made negative in (2), an entirely new set of discrete distribution is generated.

Mean and variance of the d.f. in (2) are, respectively,

$$E(r) = \frac{\alpha\theta}{2\sqrt{(1-\theta)}} \frac{K_{\gamma+1}\{\alpha\sqrt{(1-\theta)}\}}{K_{\gamma}\{\alpha\sqrt{(1-\theta)}\}} \tag{3}$$

and

$$\text{var}(r) = \frac{(\alpha\theta)^2}{4(1-\theta)} \frac{K_{\gamma+2}\{\alpha\sqrt{(1-\theta)}\}}{K_{\gamma}\{\alpha\sqrt{(1-\theta)}\}} + E(r)\{1-E(r)\}. \tag{4}$$

In general, all moments exist as long as $\theta < 1$.

If γ is known *a priori*, maximum likelihood estimators for parameters α and θ are available (Sichel, 1971). They are not required for the purpose of this investigation as sentence-length distributions are not excessively skew.

The first two probabilities (for $r = 0$ and $r = 1$) are derived from equation (2) as

$$\phi(0) = \{\sqrt{(1-\theta)}\}^{\gamma} \frac{K_{\gamma}(\alpha)}{K_{\gamma}\{\alpha\sqrt{(1-\theta)}\}} \tag{5}$$

and

$$\phi(1) = \{\sqrt{(1-\theta)}\}^{\gamma}(\alpha\theta/2) \frac{K_{\gamma+1}(\alpha)}{K_{\gamma}\{\alpha\sqrt{(1-\theta)}\}}. \tag{6}$$

All other probabilities are easily calculated from the recurrence formula

$$\phi(r) = \theta\left(\frac{r+\gamma-1}{r}\right)\phi(r-1) + \frac{(\alpha\theta)^2}{4r(r-1)}\phi(r-2). \tag{7}$$

A particularly interesting and simple case arises if we make $\gamma = -\frac{1}{2}$ in (2). This is the distribution which will be used to represent sentence-lengths. We have, from (2),

$$\phi(r) = \sqrt{(2\alpha/\pi)} \exp\{\alpha\sqrt{(1-\theta)}\} \frac{(\alpha\theta/2)^r}{r!} K_{r-\frac{1}{2}}(\alpha) \tag{8}$$

with a mean of

$$E(r) = \alpha\theta/\{2\sqrt{(1-\theta)}\} = \mu_1' \tag{9}$$

and a variance of

$$\text{var}(r) = \alpha\theta(2-\theta)/\{4(1-\theta)^{\frac{3}{2}}\} = \mu_2. \tag{10}$$

The population index of dispersion is defined as

$$\omega = \text{var}(r)/E(r) = (2-\theta)/\{2(1-\theta)\} \tag{11}$$

whence

$$\hat{\theta} = 1-(2\hat{\omega}-1)^{-1}. \tag{12}$$

From (9) we obtain

$$\hat{\alpha} = 2\bar{r}\sqrt{(1-\hat{\theta})}/\hat{\theta}. \tag{13}$$

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

In (12) and (13) $\hat{\theta}$ and $\hat{\alpha}$ are moment estimators of population parameters θ and α, $\bar{r}$ is the average sentence-length in the sample and $\hat{\omega}$ is the index of dispersion in the sample. For the type of skewness encountered in sentence-length distributions, estimators $\hat{\theta}$ and $\hat{\alpha}$ are reasonably efficient.

Strictly speaking, the sentence-length model in equation (8) should be truncated at zero as the minimum number of words per sentence is one. However, it has been found that the expected frequencies for $r = 0$, in the case where distribution (8) is fitted to real data, are very small. For ease of calculation zero-truncation does not appear necessary.

The first two proportionate frequencies are obtained from (5) and (6), with $\gamma = -\frac{1}{2}$:

$$\phi(0) = \exp\left[-\alpha\{1-\sqrt{(1-\theta)}\}\right] \tag{14}$$

and

$$\phi(1) = (\alpha\theta/2)\,\phi(0). \tag{15}$$

Any further probabilities are calculated from the recurrence relationship in (7) with $\gamma = -\frac{1}{2}$. It was shown by Sichel (1971) that the characteristic function of distribution (8) is

$$g(t) = \exp\left[\left[\alpha\left[\sqrt{(1-\theta)}-\sqrt{\{1-\theta\exp(it)\}}\right]\right]\right] \tag{16}$$

and hence we obtain the characteristic function of the arithmetic mean of samples of n, drawn from a population (8), as

$$[g(t/n)]^n = \exp\left[\left[n\alpha\left[\sqrt{(1-\theta)}-\sqrt{\{1-\theta\exp(it/n)\}}\right]\right]\right]. \tag{17}$$

It follows that the sampling distribution of the mean has the same form as the original population (8) but with parameter α replaced by $n\alpha$ and the arithmetic mean $\bar{r}$ advancing in steps of $1/n$. This property of population (8) is of considerable help in hypothesis-testing concerning the population mean.

3. Application

Several observed sentence-length distributions reported in the literature and taken from Greek, Latin and English texts were fitted to the distribution shown in equation (8). As mentioned before, zero-truncation was unnecessary as the expected frequencies at $r = 0$ were small. For the purpose of the χ^2 test, the expected frequencies were included in the first cell, i.e. the class containing 1–5 words.

Starting with examples from English authors the sentence-lengths from Macaulay's writings (Yule, 1939) are fitted to distribution (8) in Table 1. The fit is satisfactory. In contrast, the negative binomial does not represent these data as indicated in the last column of Table 2. The total χ^2 is 79·927 as compared to 16·846 for the new distribution model. The same tail-end grouping was used for both models. The deviations of the negative binomial from the data (and from distribution (8)) follow a systematic pattern although both distributions were fitted with the identical sample means and variances. At the start of the curve the negative binomial yields much larger frequencies. The position is reversed for the occurrences, at $6 \leqslant r \leqslant 25$. Once again the negative binomial frequencies exceed those of the new distribution in the range $26 \leqslant r \leqslant 65$. Finally, in the upper tail for $r \geqslant 66$, the negative binomial tends more rapidly to zero, that is the new distribution has the longer tail.

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

TABLE 1

Sentence-length distribution from Macaulay, fitted to the new model and also to the negative binomial (data from Yule, 1939)

	New distribution		*Negative binomial*
No. of words	*Observed no. of sentences* f_O	*Expected no. of sentences* f_E	*Expected no. of sentences* f_E
1– 5	46	57·8	110·3
6–10	204	201·0	185·7
11–15	252	244·6	202·6
16–20	200	209·1	184·2
21–25	186	157·5	152·3
26–30	108	113·0	118·6
31–35	61	79·5	88·6
36–40	68	55·6	64·4
41–45	38	38·9	45·7
46–50	24	27·3	31·9
51–55	20	19·2	22·0
56–60	12	13·6	14·9
61–65	8	9·6	10·1
66–70	2	6·8	6·7
71–75	4 } 14	4·9 } 15·2	4·5 } 14·1
76–80	8	3·5	2·9
81–85	2 } 4	2·5 } 4·3	1·9 } 3·2
86–90	2	1·8	1·3
91 and over	6	4·8	2·4
Total	1,251	1,251·0	1,251·0
Mean	22·07	—	—
Variance	230·22	—	—
χ^2	—	16·846	79·927
d.f.	—	13	13
$P(\chi^2)$	—	0·21	0·00
$\hat{\alpha}$	—	10·43117	$\hat{\gamma}$ = 2·34068†
$\hat{\theta}$	—	0·94965	0·90412

† The general distribution in equation (2) becomes the negative binomial with parameters γ and θ as $\alpha \to 0$.

In Table 2 sentence-length distributions from works of Wells and Chesterton as given by Williams (1940), are excellently represented by the new model. The large differences in the parameter estimates $\hat{\alpha}$ and $\hat{\theta}$ for these two authors are of interest.

Morton (1965) shows sentence-length distributions taken from eight works of Thucydides and from nine works of Herodotus. The examples from ancient Greek texts in Table 3, once again indicate the success of distribution (8) as a model for sentence-length distributions.

Negative binomials were also fitted to the two observed frequency counts in Table 3 making use of the same means and variances as derived from the samples. The respective total χ^2 values were 82·216 for Thucydides and 42·823 for Herodotus.

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

TABLE 2

Sentence-length distributions from H. G. Wells and G. K. Chesterton fitted to the new model (data from Williams, 1940)

	H. G. Wells		*G. K. Chesterton*	
No. of words	*Observed no. of sentences* f_O	*Expected no. of sentences* f_E	*Observed no. of sentences* f_O	*Expected no. of sentences* f_E
1– 5	11	11·0	3 } 30	1·7 } 25·6
6–10	66	63·3	27	23·9
11–15	107	106·8	71	76·1
16–20	121	109·6	112	114·8
21–25	75	90·6	108	116·2
26–30	61	67·7	109	94·0
31–35	52	48·1	64	66·5
36–40	27	33·2	41	43·3
41–45	29	22·7	28	26·6
46–50	17	15·3	19	15·8
51–55	12	10·3	9	9·2
56–60	8	6·9	6	5·2
61–65	5 } 9	4·7 } 7·8	1	2·9
66–70	4	3·1	—	1·6
71–75	3	2·1	—	0·9
76–80	1	1·4	1 } 9	0·5 } 11·9
81–85	— } 5	1·0 } 6·7	—	0·3
86–90	—	0·7	—	0·2
91 and over	1	1·5	1	0·3
Total	600	600·0	600	600·0
Mean	24·08	—	25·91	—
Variance	199·38	—	131·05	—
χ^2	—	9·132	—	5·788
d.f.	—	11	—	8
$P(\chi^2)$	—	0·61	—	0·67
$\hat{\alpha}$	—	13·04312	—	19·27635
$\hat{\theta}$	—	0·93575	—	0·89030

The systematic deviations of the negative binomials from the data and the new distribution were very similar to those described in the discussion on Table 1. In short, the negative binomial distribution cannot take on the shape of observed sentence-length frequency counts.

The data discussed so far display a concave upward curvature if plotted as a c.d.f. on log-probability paper. Some sentence-length distributions do approach a straight line on log-probability paper. To check whether such cases can be represented satisfactorily by distribution (8), two frequency counts given by Wake (1957) were fitted to (8) and they are shown in Table 4. The first example refers to sentence-lengths from *Timaeus* by Plato and the second comes from the *Hippocratic Corpus, Regimen in Acute Diseases.* As shown in Table 4, both observed frequency counts are well fitted by the new model.

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

TABLE 3

Sentence-length distributions from Thucydides and Herodotus, fitted to the new model (data from Morton, 1965)

No. of words	*Thucydides (Works 1–8)*		*Herodotus (Works 1–9)*	
	Observed no. of sentences f_O	*Expected no. of sentences* f_E	*Observed no. of sentences* f_O	*Expected no. of sentences* f_E
1– 5	48	49·8	90	93·1
6–10	201	203·4	354	338·4
11–15	274	279·2	388	406·0
16–20	257	260·3	322	328·1
21–25	232	209·6	223	228·3
26–30	144	158·9	171	149·3
31–35	135	117·4	96	95·1
36–40	73	85·8	56	59·9
41–45	70	62·6	39	37·6
46–50	45	45·6	26	23·6
51–55	32	33·3	15	14·8
56–60	15	24·4	7 } 10	9·3 } 15·2
61–65	22	17·9	3	5·9
66–70	17	13·2	1	3·7
71–75	9	9·7	2	2·4
76–80	5 } 10	7·2 } 12·6	4	1·5
81–85	5	5·4	1 } 10	1·0 } 10·6
86–90	2	4·0	1	0·6
91–95	5 } 9	3·0 } 9·2	—	0·4
96–100	2	2·2	—	0·3
101 and over	7	7·1	1	0·7
Total	1,600	1,600·0	1,800	1,800·0
Mean	24·98	—	19·04	—
Variance	293·73	—	140·45	—
χ^2	—	15·750	—	7·384
d.f.	—	15	—	10
$P(\chi^2)$	—	0·40	—	0·69
$\hat{\alpha}$	—	11·02074	—	11·07245
$\hat{\theta}$	—	0·95558	—	0·92729

Yule (1939) discussed the authorship of the Latin essay *De Imitatione Christi* whose author is unknown. He came to the conclusion that Jean Charlier de Gerson is unlikely to have written this work. In Table 5 the sentence-length distribution for the combined two samples from de Gerson's works, as quoted by Yule (1939), is fitted to the distribution (8). Bearing in mind that the sample size is $n = 2{,}417$, the fit is fair [$P(\chi^2) = 0{\cdot}11$ for 17 degrees of freedom]. Most of the contributions to total χ^2 come from two cells only, that is 1–5 words and 46–50 words per sentence. But for these two deviations from theory, amounting to a χ^2 contribution of 13·888, the fit would have been excellent.

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

TABLE 4

Sentence-length distributions from Plato and the Hippocratic Corpus, fitted to the new model (data from Wake, 1957)

No. of words	*Plato (Timaeus)*		*Hippocratic Corpus, Regimen in Acute Diseases*	
	Observed no. of sentences f_O	*Expected no. of sentences* f_E	*Observed no. of sentences* f_O	*Expected no. of sentences* f_E
1– 5	15	15·4	23	32·8
6–10	70	68·9	84	83·8
11–15	104	100·7	91	80·5
16–20	102	98·4	59	57·4
21–25	73	82·2	45	37·3
26–30	68	64·2	22	23·5
31–35	45	48·7	9	14·7
36–40	41	36·5	7	9·2
41–45	23	27·2	6	5·8
46–50	18	20·2	3 } 5	3·6 } 5·9
51–55	13	15·0	2	2·3
56–60	14	11·2	—	1·5
61–65	10	8·4	1	0·9
66–70	6 } 14	6·3 } 11·0	—	0·6
71–75	8	4·7	1 } 4	0·4 } 4·1
76–80	2 } 5	3·6 } 6·3	—	0·3
81–85	3	2·7	1	0·2
86–90	1 } 6	2·0 } 5·7	—	0·1
91–95	3	1·6	1	0·1
96–100	2	1·2		
101–105	—	0·9		
106 and over	4	5·0		
Total	625	625·0	355	355·0
Mean	26·77	—	16·96	—
Variance	337·24	—	133·56	—
χ^2	—	5·821	—	8·909
d.f.	—	14	—	8
$P(\chi^2)$	—	0·97	—	0·35
$\hat{\alpha}$	—	11·35126	—	9·47232
$\hat{\theta}$	—	0·95868	—	0·93221

In contrast, the sentence-length distribution of *De Imitatione Christi* cannot be represented by the distribution model of equation (8) as shown in the last two columns of Table 5. The χ^2 is 66·788 for nine degrees of freedom and the deviations of the data from the model are systematic suggesting that in the *general* model of equation (2) $\gamma \ll -\frac{1}{2}$. Consequently α and θ should be larger.

This is a very good example illustrating that, in addition to differences in means and variances, the shape of the distribution, as measured at least in part by parameter γ, is most useful in detecting statistically significant differences of sentence-length distributions.

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

TABLE 5

Sentence-length distributions from de Gerson and from De Imitatione Christi fitted to the new model (data from Yule, 1939)

	de Gerson		*De Imitatione Christi*	
No. of words	*Observed no. of sentences* f_O	*Expected no. of sentences* f_E	*Observed no. of sentences* f_O	*Expected no. of sentences* f_E
1– 5	116	93·1	39	91·8
6–10	362	353·1	302	290·6
11–15	431	456·2	376	305·0
16–20	394	405·3	237	217·7
21–25	301	313·5	119	134·6
26–30	213	229·3	52	78·5
31–35	167	163·9	42	44·7
36–40	125	116·2	20	25·2
41–45	94	82·1	8	14·2
46–50	80	58·1	11	8·0
51–55	37	41·2	7	4·5
56–60	26	29·4	2	2·5
61–65	22	21·0	2	1·5
66–70	18	15·0	2	0·8
71–75	7	10·8	—	0·5
76–80	8	7·8	1	0·3
81–85	3	5·6	—	0·2
86–90	3	4·1	—	0·1
91–95	5	3·0	—	0·1
96–100	1	2·2	—	0·1
101–105	—	1·6	—	0·1
106 and over	4	4·5	1	—
			8·0	6·2
	6	6·8		
Total	2,417	2,417·0	1,221	1,221·0
Mean	23·07	—	16·24	—
Variance	244·98	—	96·36	—
χ^2	—	24·389	—	66·788
d.f.	—	17	—	9
$P(\chi^2)$	—	0·11	—	0·00
$\hat{\alpha}$	—	10·78827	—	10·85495
$\hat{\theta}$	—	0·95059	—	0·90796

REFERENCES

MORTON, A. Q. (1965). The authorship of Greek prose. *J. R. Statist. Soc.* A, **128**, 169–224.

MOSTELLER, F. and WALLACE, D. L. (1963). Inference in an authorship problem. *J. Amer. Statist. Ass.*, **58**, 275–309.

SICHEL, H. S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In *Proceedings of the Third Symposium on Mathematical Statistics* (N. F. Laubscher, ed.), S.A. C.S.I.R., Pretoria, pp. 51–97.

WAKE, W. C. (1957). Sentence-length distributions of Greek authors. *J. R. Statist. Soc.* A, **120**, 331–346.

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions

WILLIAMS, C. B. (1940). A note on the statistical analysis of sentence-length as a criterion of literary style. *Biometrika*, **31**, 356–361.
—— (1970). *Style and Vocabulary: Numerical Studies*. London: Griffin.
YULE, G. U. (1939). On sentence-length as a statistical characteristic of style in prose: with applications to two cases of disputed authorship. *Biometrika*, **30**, 363–390.
—— (1944). *The Statistical Study of Literary Vocabulary*. Cambridge: University Press.

This content downloaded from 195.251.235.181 on Wed, 8 Oct 2014 07:42:29 AM
All use subject to JSTOR Terms and Conditions