

LITERATURE AND STATISTICS—II

NEW POEMS BY SHAKESPEARE?

The years 1981–1986 have witnessed the appearance, among other works, of Pol-latschek and Radday's further analysis of Biblical literature [13], Brainerd's extensions of type-token models [2,3], Ellegard's discussion on the identification of authorship [6], Kenny's expository monograph [9], the revised edition of Mosteller and Wallace's book on *The Federalist Papers* [12], Lanke's comments on Ellegard's work [11], Holmes' review of the analysis of literary style [8], and Sichel's paper on type-token characteristics [15]. All these are indicative of a new awareness of the value of statistical analysis in literature.

Most recently, popular interest has been aroused by Taylor's discovery on November 14, 1985, in a folio volume at the Bodleian Library, Oxford, of a nine-stanza poem beginning with the lines

Shall I die? Shall I fly
Lovers' baits and deceits,
sorrow breeding?

This was attributed by Taylor [17] to Shakespeare on the basis of a literary analysis, but several critics, including Robbins [14], have expressed dissenting views.

In their 1976 paper, Efron and Thisted [5] had already studied Shakespeare's vocabulary using Spevack's [16] concordance, and estimated the number of words Shakespeare might have known, but not used. It is not often that statisticians can test their results on new data, but with Taylor's discovery of the new poem, this has in fact proved possible (see Kolata [10], and the ensuing correspondence between Driver and Kolata [4] and Birkes [1]). To help resolve the question of its authorship, Thisted and Efron [18] decided to carry out a statistical analysis of the new poem based on their earlier work. Their conclusion was that "On balance, the poem is found to fit previous Shakespearean usage well, lending credence to belief that it was actually written by Shakespeare."

The new poem contains 429 words, 258 of them distinct, 9 of which had not appeared in

Shakespeare's previous work. These unusual words were "admiration," "besots," "exiles," "inflection," "joying," "scanty," "speck," "tormentor," and "explain" (see Driver and Kolata [4] for a clarification of types, or different words). On the basis of Shakespeare's known writings, the number of distinct words, $\hat{\nu}_x$, expected to occur $x = 0, 1, 2, 3, 4, \dots, 99$ times, respectively, in a poem of 429 words can be estimated (see the summary in Table 1); what Thisted and Efron [18] have done in their study is to compare these counts with the actual numbers m_x of such words appearing in the new poem.

To broaden the scope of their study, seven additional poems were considered: one by Ben Jonson, a second by Christopher Marlowe, and a third by John Donne, as well as four already included in the Shakespeare canon from *Cymbeline*, *A Mid-summer Night's Dream*, *The Phoenix and the Turtle*, and Sonnets 12–15.

Three different tests were used on the collected data, one of which (the slope test) proved to be the best discriminator for detecting non-Shakespearean authorship. All the tests relied on a regression model in which the observed numbers of words $\{m_x\}$ for the particular poem under study follow the Poisson distribution* with means $\{\mu_x\}$

Table 1. Expected and Observed Counts of Words Appearing x Times in a 429-Word Poem

Number of Occurrences x	Expected Number of Words $\hat{\nu}_x$	Observed Number of Words m_x
0	6.97	9
1	4.21	7
2	3.33	5
3	2.84	4
4	2.53	4
5	2.43	2
10	1.62	1
30	0.96	4
50	0.68	0
70	0.49	0
90	0.37	0
95	0.34	1
99	0.32	0

Table 2. Estimated Slope Values $\hat{\beta}_1$, Standard Errors $\hat{\sigma}$, and z-Values for 8 Poems

Poems	$\hat{\beta}_1$	$\hat{\sigma}$	z-Value $\hat{\beta}_1/\hat{\sigma}$
Jonson	0.229	0.11	2.08**
Marlowe	-0.323	0.08	-4.04****
Donne	-0.138	0.09	-1.53*
Cymbeline	-0.047	0.10	-0.47
Midsummer	-0.050	0.12	-0.42
Phoenix	-0.127	0.09	-1.41
Sonnets	-0.034	0.09	-0.38
New Poem	-0.075	0.09	-0.83

Table 3. Summary of Significant z-Values for Tests 1, 2, and 3

Poems	Test 1	Test 2	Test 3
Jonson			**
Marlowe	***		****
Donne		***	*
Cymbeline	***		
Midsummer		*	
Phoenix	****	**	
Sonnets			
New Poem	**		

Asterisks indicate significant values as follows: * $1.5 \leq |z| < 2$; ** $2 \leq |z| < 2.5$; *** $2.5 \leq |z| < 3$; and **** $3 \leq |z|$.

independently for $x = 0, 1, \dots, 99$, where

$$\mu_x = \hat{\nu}_x e^{\beta_0} (x+1)^{\beta_1}.$$

The first test was based on the total count of words occurring 99 times or less in each of the poems. For the new poem, for example, the actual number was $m_+ = \sum_{x=0}^{99} m_x = 118$, while its expectation was $\nu_+ = \sum_{x=0}^{99} \hat{\nu}_x = 94.95$, and $\mu_+ = \sum_{x=0}^{99} \mu_x$. The hypothesis tested was $H_1: \mu_+ = \hat{\nu}_+$; this proved the least reliable test for discriminating between Shakespearean and non-Shakespearean authorship.

The second test was concerned with the simple null hypothesis $H_2: \pi_0 = \hat{\nu}_0/\hat{\nu}_+$, where the zero count m_0 conditional on the total count m_+ in each of the poems follows a binomial distribution $B(m_+, \pi_0)$. This test proved only moderately useful in discerning Shakespearean authorship.

The third test (the slope test) of the hypothesis $H_3: \beta_1 = 0$ relied on the data (m_1, \dots, m_{99}) for each poem. This is equivalent to testing H_3 conditional on (m_+, m_0)

when the (m_1, \dots, m_{99}) follow a multinomial distribution depending on β_1 . The maximum likelihood estimates $\hat{\beta}_1$ and their standard errors $\hat{\sigma}$ were obtained (see Table 2); this test can be seen to provide the most promising method of discriminating Shakespearean authorship.

When the significant z-values for all three tests are summarized as in Table 3, the conclusion that test 3 is the most discriminating is strengthened. On the basis of these statistical tests, Thisted and Efron [18] reached the conclusion that the new poem "fits Shakespearean usage about as well as do the four Shakespeare poems." It is only fair to mention that many, including Foster [7] and Robbins [14], remain unconvinced that Shakespeare was its author.

Most recently, Foster [7] has analyzed the word frequency, frequency of subordinating conjunctions, and 15 other statistical measures of the Peter "Funeral Elegy" signed W. S. He has concluded from these 17 different tests that W. S. is very likely to be Shakespeare. Once again statistical analysis has assisted in the identification of authorship, and justified its value in the humanities.

REFERENCES

1. Birkes, D. (1986). Sly statistics. *Science*, **232**, 698.
2. Brainerd, B. (1981). Some elaborations upon Gani's model for the type-token relationship. *J. Appl. Prob.*, **18**, 452-460.
3. Brainerd, B. (1982). On the relation between the type-token and the species-area problems. *J. Appl. Probab.*, **19**, 785-793.
4. Driver, O. and Kolata, G. (1986). Shakespeare and statistics. *Science*, **231**, 1355.
5. Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**, 435-447.
6. Ellegard, A. (1982). Genre styles, individual styles, and authorship identification. In Text Processing, S. Allen, ed. *Proc. Nobel Symp.*, **51**, 519-537.
7. Foster, D. W. (1986). Elegy by W. S.—A Study in Attribution. Ph. D. thesis, University of California, Santa Barbara, CA.
8. Holmes, D. I. (1985). The analysis of literary style—A review. *J. R. Statist. Soc. A*, **148**, 328-341.

9. Kenny, A. (1982). *The Computation of Style*. Pergamon, Oxford, England.
10. Kolata, G. (1986). Shakespeare's new poem: An ode to statistics. *Science*, **231**, 335–336.
11. Lanke, J. (1985). On the art of conditioning on the right event. In *Contributions to Statistics in Honour of Gunnar Blom*, pp. 215–221.
12. Mosteller, F. and Wallace, D. L. (1984). *Applied Bayesian and Classical Inference—The Case of the Federalist Papers*. Springer, New York.
13. Pollatschek, M. and Radday, Y. T. (1981). Vocabulary richness and concentration in Hebrew Biblical literature. *Ass. Lit. Linguist. Comp. Bull.*, **8**, 217–231.
14. Robbins, R. (1985). ... and the counter-arguments. *TLS*, **4316**, December 20, 1985, 1449–1450.
15. Sichel, H. S. (1986). Word frequency distribution and type-token characteristics. *Math. Scientist*, **11**, 45–72.
16. Spevack, M. (1968). *A Complete and Systematic Concordance to the Works of Shakespeare*, 6 volumes. George Olms, Hildesheim, West Germany.
17. Taylor, G. (1985). A new Shakespeare poem? The evidence ... *TLS*, **4316**, December 20, 1985, 1447–1448.
18. Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, **74**, 445–455.

See also LINGUISTICS, STATISTICS IN; LITERATURE AND STATISTICS—I; and STYLOMETRY.

J. GANI