Biometrika (1987), **74**, 3, *pp.* 445–55
Printed in Great Britain

Did Shakespeare write a newly-discovered poem?

BY RONALD THISTED
Department of Statistics, University of Chicago, Chicago, Illinois 60637, U.S.A.
AND BRADLEY EFRON
Department of Statistics, Stanford University, Stanford, California 94305, U.S.A.

SUMMARY

The consistency of the word usage in a previously unknown nine-stanza poem attributed to Shakespeare with that of the Shakespearean canon is examined using a nonparametric empirical Bayes model. We consider also poems by Jonson, Marlowe and Donne, as well as four poems definitely attributed to Shakespeare. On balance, the poem is found to fit previous Shakespearean usage reasonably well.

Some key words: Authorship; Empirical Bayes; Poetry; Poisson regression; Shakespeare; Species.

1. INTRODUCTION

In 1976, we examined the problem of estimating the number of unseen species in ecological studies and applied our results to estimating the number of words that Shakespeare knew but had not used in his existing body of work (Efron & Thisted, 1976). Our results were based on the frequencies of the 31534 distinct words found in the body of 884647 total words in the Shakespearean canon, and we made our question operational by writing,

> ..., suppose another large quantity of work by Shakespeare were discovered, say $884647t$ total words. How many new words in addition to the original 31534 would we expect to find?

As no new works by Shakespeare had been discovered since the 17th century, we did not expect to be able to test our theory directly. On November 14, 1985, however, Shakespearean scholar Gary Taylor discovered a nine-stanza poem attributed to Shakespeare, hereinafter called the 'Taylor poem', in a bound folio volume that had been in the collection of the Bodleian Library since 1755 (Lelyveld, 1985; Taylor, 1985). The size of the newly discovered poem is small relative to the size of Shakespeare's total work, only 429 total words.

Can we prove that the poem was not actually written by Shakespeare? Our paper develops simple tests for this question based on the frequency of occurrence of unusual words. In fact 'unusual words' are very common in Shakespeare: two-thirds of the 31534 distinct words occur three or fewer times in the entire Shakespearean canon. As a result our tests have good power for detecting non-Shakespearean usage, even for new works as short as the Taylor poem.

The Taylor poem contains 429 words, of which 258 are distinct. By 'word' here and throughout we mean 'word type': any distinguishable arrangement of letters. Thus 'tormentor' and 'tormentors' are different word types. In keeping with the spelling practice adopted by the concordance on which our word counts are based, 'oh' and 'o', each of which appears once in the Taylor poem, are considered to be the same word type.

Table 1. *Number, m_x, of distinct words in the Taylor poem that appeared exactly x times in the Shakespearean canon*

x	0	1	2	3	4	5	6	7	8	9	Row total
0+	9	7	5	4	4	2	4	0	2	3	40
10+	1	0	3	0	1	1	1	2	1	0	10
20+	2	2	1	5	3	1	0	2	2	3	21
30+	4	1	1	1	2	1	0	0	3	3	16
40+	1	2	0	0	2	1	1	2	1	1	11
50+	0	1	1	1	1	0	0	1	0	2	7
60+	0	1	0	0	1	1	0	0	1	0	4
70+	0	0	1	0	0	1	0	0	1	1	4
80+	0	0	1	1	0	0	0	0	0	0	2
90+	0	0	0	1	0	1	1	0	0	0	3

For example, 9 distinct words in the poem appeared zero times in the canon, 7 appeared once each, etc.

Our analysis begins by ranking each of the 258 distinct words in the Taylor poem according to its rarity of usage in the Shakespearean canon; see Table 1. Let m_x denote the number of distinct words in the Taylor poem which occurred exactly x times in the 884647 total words of the Shakespearean canon. Table 1 shows for example that $m_{23} = 5$, corresponding to 5 distinct words in the Taylor poem that occurred exactly 23 times each in the canon. The table covers x from 0 to 99, as our analysis considers only those words. There are 118 such distinct words in the poem, leaving 140 distinct words that each appeared 100 or more times in the canon.

Of particular note is $m_0 = 9$, which is the number of distinct words in the poem that never appeared in the Shakespearean canon. This is the quantity estimated by $\hat{\Delta}(t)$ in the notation of Efron & Thisted (1976). Both the nonparametric empirical Bayes estimate due to Good, Toulmin and Turing (Good & Toulmin, 1956), expression (2.3) of Efron & Thisted (1976), and Fisher's negative binomial model (Efron & Thisted, 1976, eqn (3·4)), give estimates $\hat{\Delta}(t) = 6{\cdot}97$ in this case. Allowing for Poisson variation, this agrees reasonably with the observed value $m_0 = 9$.

The nine new words are 'admirations', 'besots', 'exiles', 'inflection', 'joying', 'scanty', 'speck', 'tormentor' and 'explain'. A feature of our analysis is that it is based on distinct word types and not linguistically equivalent words. Thus, for instance, 'admiration' appears 14 times in Shakespeare's works, but never as a plural. 'Besotted' appears once in the literature, but never 'besot', or 'besots'.

Efron & Thisted (1976) give a prediction only for m_0. It is easy to extend the theory to predictions for $m_1, \ldots, m_{99}$, which is done in § 2. Given the predictions and the observed data of Table 1, our analysis proceeds by a standard Poisson regression analysis, comparing observed with predicted counts. Seven other Elizabethan love poems, four definitely attributed to Shakespeare and three definitely attributed to other authors, are also analysed. This gives us an empirical basis for evaluating the theoretical model behind our predictions and tests.

2. NONPARAMETRIC ESTIMATES FOR THE EXPECTED COUNTS

This section describes simple nonparametric estimates for the expectations of the counts m_x appearing in Table 1, assuming Shakespearean authorship. Let $\hat{\nu}_x$ be the estimate of

the expected value of m_x; see Table 2. We now derive $\hat{\nu}_x$, under the same assumptions as in § 2 of Efron & Thisted (1976).

Suppose that there is a universe of S 'species', in our application distinct words, and that after trapping species, observing words, for one unit of time, say, $[-1, 0]$, we have captured X_s members of species s. Assume that the trapping process for species s is a Poisson process with rate parameter λ_s, and let $G(\lambda)$ denote the empirical cumulative distribution function of the numbers $\lambda_1, \ldots, \lambda_S$. Let n_x denote the number of species observed exactly x times in $[-1, 0]$, and let η_x denote the expected value of n_x. As in our previous paper, we can then write

$$\eta_x = E(n_x) = S \int_0^\infty e^{-\lambda} \frac{\lambda^x}{x!}\, dG(\lambda). \tag{2·1}$$

In our application, 'trapping for the interval $[-1, 0]$', corresponds to reading all 884647 total words of the Shakespearean canon. The values of n_x for $x = 1$ to 100 appear in Table 1 of Efron & Thisted (1976).

Now suppose we extend the trapping period another t time units, say to the interval $(0, t]$, which corresponds to reading $884647t$ new words of Shakespeare. Denote by $\nu_x = E(m_x)$ the expected number of species found in $(0, t]$ which were seen exactly x times in the initial interval $[-1, 0]$. Then

$$\nu_x = S \int_0^\infty e^{-\lambda} \frac{\lambda^x}{x!}(1 - e^{-\lambda t})\, dG(\lambda), \tag{2·2}$$

giving the formal equality

$$\nu_x = \sum_{k=1}^{\infty} (-1)^{k+1} \binom{x+k}{k} t^k \eta_{x+k}. \tag{2·3}$$

The right-hand side of (2·3) will converge provided that $t < 1$. In our case, $t = 4{\cdot}849 \times 10^{-4}$, and convergence is quite rapid. Expression (2·3) suggests the unbiased estimator $\hat{\nu}_x$ for ν_x obtained by substituting the observed values n_x for the η_x in (2·3), which gives

$$\hat{\nu}_x = \sum_{k=1}^{\infty} (-1)^{k+1} \binom{x+k}{k} t^k n_{x+k}. \tag{2·4}$$

We refer to (2·4) as the empirical Bayes estimate of ν_x. Entries 0–14 of Table 2 were obtained directly from (2·4). The higher entries of Table 2 were also obtained from (2·4),

Table 2. *Estimated expectation, $\hat{\nu}_x$, for the corresponding count m_x in Table* 1, *assuming Shakespearean authorship for the Taylor poem*

x	0	1	2	3	4	5	6	7	8	9
0–9	6·97	4·21	3·33	2·84	2·53	2·43	2·16	2·01	1·87	1·76
10–19	1·62	1·50	1·52	1·51	1·36	1·38	1·33	1·28	1·25	1·22
20–29	1·18	1·16	1·13	1·11	1·09	1·06	1·04	1·02	1·00	0·98
30–39	0·96	0·94	0·93	0·91	0·90	0·88	0·86	0·85	0·83	0·82
40–49	0·80	0·79	0·77	0·76	0·75	0·74	0·73	0·72	0·70	0·69
50–59	0·68	0·67	0·66	0·65	0·64	0·63	0·62	0·61	0·60	0·59
60–69	0·58	0·57	0·56	0·55	0·54	0·53	0·52	0·51	0·50	0·50
70–79	0·49	0·48	0·48	0·47	0·47	0·46	0·45	0·45	0·44	0·44
80–89	0·43	0·42	0·42	0·41	0·41	0·40	0·39	0·39	0·38	0·38
90–99	0·37	0·36	0·36	0·35	0·35	0·34	0·34	0·33	0·32	0·32

but a local linear smoother was applied to damp the statistical variability of the $\hat{\nu}_x$ arising from the smaller, less stable, values of n_x for x large.

The entries of Table 2 have small standard errors, of the order 0·06–0·10 for $x = 0$–14. These standard errors are based on the conservative assumption that n_x has approximate standard deviation $n_x^{\frac{1}{2}}$. The standard errors for the higher entries of Table 2 are smaller still because of the smoothing, roughly 0·030 according to a bootstrap analysis.

3. SEVEN MORE POEMS

To broaden the empirical base of our results, we consider seven more Elizabethan poems; three poems attributed to Ben Jonson, Christopher Marlowe and John Donne respectively, as well as four poems definitely attributed to Shakespeare. Table 3 summarizes the eight sources; full citations appear in the Appendix.

Table 4 gives the Shakespearean frequency counts for all eight poems. These are defined in Table 1, except that the counts have been summed over 11 categories. For example, 10 distinct words in the Jonson poem appeared between 60 and 79 times each in the Shakespearean canon.

The expected values for the counts in Table 4, assuming Shakespearean authorship, are given in Table 5. These were derived from (2·4), with smoothing for values with $x \geqslant 15$ as in Table 2.

Table 3. *The eight poems analysed in this paper*

Abbreviation	Description	Total length	Number distinct words
1. JON	Ben Jonson; 'An Elegy'	411	243
2. MAR	C. Marlowe: four poems	495	272
3. DON	J. Donne; 'The Ecstacy'	487	252
4. CYM	Shakespeare: from 'Cymbeline'	323	215
5. PUC	from 'A Midsummer Night's Dream'	234	156
6. PHO	'The Phoenix and Turtle'	352	216
7. SON	Sonnets, Nos. 12–15	448	264
8. TAY	Taylor poem	429	258

Table 4. *Words in the eight poems categorized according to their Shakespearean frequencies*

	Category of x										
Poem	0	1	2	3–4	5–9	10–19	20–29	30–39	40–59	60–79	80–99
1. JON	8	2	1	6	9	9	12	12	13	10	13
2. MAR	10	8	8	16	22	20	13	9	14	9	5
3. DON	17	5	6	5	12	17	14	6	12	3	10
4. CYM	7	4	3	5	13	17	9	12	17	4	4
5. PUC	1	4	0	3	9	6	9	4	5	9	3
6. PHO	14	5	5	9	8	18	13	7	13	8	5
7. SON	7	8	1	5	16	14	12	13	12	13	8
8. TAY	9	7	5	8	11	10	21	16	18	8	5

As in Table 1, except that the counts have been summed over 11 categories.

Table 5. *Estimated expected values for the counts in Table* 4, *assuming Shakespearean authorship*

	Category of x										
Poem	0	1	2	3–4	5–9	10–19	20–29	30–39	40–59	60–79	80–99
1. JON	6·68	4·03	3·19	5·14	9·81	13·16	9·94	8·18	12·68	9·17	6·83
2. MAR	8·04	4·86	3·85	6·19	11·81	15·91	12·03	9·92	14·92	10·72	8·26
3. DON	7·91	4·78	3·78	6·09	11·62	15·59	11·77	9·68	14·99	10·83	8·06
4. CYM	5·25	3·17	2·51	4·04	7·71	10·35	7·82	6·44	9·99	7·23	5·39
5. PUC	3·79	2·29	1·81	2·91	5·57	7·47	5·65	4·66	7·22	5·23	3·91
6. PHO	5·72	3·46	2·73	4·40	8·40	11·28	8·52	7·02	10·87	7·87	5·87
7. SON	7·28	4·40	3·48	5·60	10·69	14·52	11·10	9·06	13·71	10·02	7·96
8. TAY	6·97	4·21	3·33	5·36	10·24	13·96	10·77	8·87	13·77	9·99	7·48

As in Table 2, except that the expected values have been summed over 11 categories.

4. Modelling and testing

We wish to test whether the observed counts m_x for each of our eight poems fit the predicted values $\hat{\nu}_x$ based on the assumption of Shakespearean authorship. Our tests will rely upon the following regression model, that for $x = 0, 1, \ldots, 99$, the m_x have independent Poisson distributions with means μ_x, where

$$\log \mu_x = \log \hat{\nu}_x + \beta_0 + \beta_1 \log (x+1). \tag{4·1}$$

The quantities $\hat{\nu}_x$ are considered to be constants in what follows, for example having the values shown in Table 2 for the Taylor poem. The null hypothesis

$$H_0\colon \beta_0 = \beta_1 = 0 \tag{4·2}$$

corresponds to $\mu_x = \hat{\nu}_x$, that is, to perfect agreement with Shakespearean usage.

Model (4·1) is motivated in § 6. It is a generalized linear model of the simplest sort, as described by McCullagh & Nelder (1983, Ch. 2). Expression (4·1) can be rewritten as

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

We will be particularly interested in the slope parameter β_1. Notice that if β_1 is negative then the ratio $\mu_x/\hat{\nu}_x$ increases as $x \to 0$. In other words we are in a situation where relative usage is increasing as we move toward the rare end of the Shakespearean vocabulary. Conversely, β_1 greater than zero means that the relative usage $\mu_x/\hat{\nu}_x$ decreases as we move toward rarer Shakespearean vocabulary.

Three different tests will be applied to each poem: Test 1, total count; Test 2, new words; Test 3, slope.

Here is a brief description of each test. Let $m_+ = \Sigma\, m_x$ be the total count for categories 0–99; $m_+ = 118$ for the Taylor poem. Similarly let $\mu_+ = \Sigma\, \mu_x$, so that m_+ has a Poisson distribution of mean μ_+ according to (4·1).

Under the null hypothesis (4·2), $\mu_+ = \hat{\nu}_+ \equiv \Sigma\, \hat{\nu}_x$, with $\hat{\nu}_+ = 94{\cdot}95$ for the Taylor poem. Test 1 is just the usual test of the simple null hypothesis $H_1\colon \mu_+ = \hat{\nu}_+$.

The zero count m_0, considered conditional on the total count m_+, has a binomial distribution of index μ_+ and parameter $\pi_0 = \mu_0/\mu_+$. Test 2 is the usual test of the simple null hypothesis $H_2\colon \pi_0 = \hat{\nu}_0/\hat{\nu}_+$.

Test 3 is the usual test, using large-sample maximum likelihood approximations, of the null hypothesis $H_3\colon \beta_1 = 0$ in the generalized linear model (4·1), based, however, only on the data $(m_1, \ldots, m_{99})$. As discussed in § 6 this is equivalent to testing H_3 conditional

on (m_+, m_0), in which case $(m_1, \ldots, m_{99})$ has a multinomial distribution depending only upon the slope parameter β_1.

5. RESULTS AND COMMENTS

The results of Tests 1, 2 and 3 appear in Tables 6, 7 and 8 respectively. A z-value is given for each test and each poem, this equalling the equivalent normal deviate corresponding to the attained significance level of the one-tailed test result. For example Poem

Table 6. *Test* 1, *total count, for the eight poems*

Poem	Total count m_+	Expectation $\hat{\nu}_+$	z
1. JON	95	88·8	0·67
2. MAR	134	106·5	2·57***
3. DON	107	105·1	0·20
4. CYM	95	69·9	2·86***
5. PUC	53	50·5	0·37
6. PHO	105	76·1	3·13****
7. SON	109	96·7	1·24
8. TAY	118	95·0	2·29**

Asterisks indicate deviations from null hypothesis. * $1{\cdot}5 \leq |z| < 2{\cdot}0$; ** $2{\cdot}0 \leq |z| < 2{\cdot}5$; *** $2{\cdot}5 \leq |z| < 3{\cdot}0$; **** $3{\cdot}0 \leq |z|$.

Table 7. *Test* 2, *new words, for the eight poems*

Poem	New words m_0	Expected new words $\mu_+(\hat{\nu}_0/\hat{\nu}_+)$	z
1. JON	8	7·14	0·37
2. MAR	10	10·12	0·01
3. DON	17	8·06	2·90***
4. CYM	7	7·13	0·00
5. PUC	1	3·98	−1·64*
6. PHO	14	7·89	2·08**
7. SON	7	8·21	−0·39
8. TAY	9	8·66	0·16

The total counts m_+ appear in Table 6; $\hat{\nu}_0/\hat{\nu}_+ \doteq 0{\cdot}075$ for all eight poems. For asterisks, see Table 6.

Table 8. *Test* 3, *slope, for the eight poems*

Poem	Est. slope $\hat{\beta}_1$	Est. standard error $\hat{\sigma}$	$z = \hat{\beta}_1/\hat{\sigma}$
1. JON	0·229	0·11	2·08**
2. MAR	−0·323	0·08	−4·04****
3. DON	−0·138	0·09	−1·53*
4. CYM	−0·047	0·10	−0·47
5. PUC	−0·050	0·12	−0·42
6. PHO	−0·127	0·09	−1·41
7. SON	−0·034	0·09	−0·38
8. TAY	−0·075	0·09	−0·83

For asterisks, see Table 6.

1, Test 1 has an attained significance level pr {Poisson $(88{\cdot}8) \leqslant 95\} = 0{\cdot}748$, including a continuity correction which subtracts half of the atom of probability at $m_+ = 95$, so that $z = \Phi^{-1}(0{\cdot}748) = 0{\cdot}67$, where Φ is the standardized normal integral.

The z-values for Tests 1, 2 and 3 are independent and have nearly a standard normal distribution under null hypotheses 1, 2 and 3 respectively. Asterisks in Tables 6–9 indicate significant values.

Table 9 summarizes all the results using this coding scheme. Here are a few comments.

Test 1 is the least reliable for discriminating between Shakespearean and non-Shakespearean authorship. All eight of the z-values are positive, four of them including two for the known Shakespeare poems CYM and PHO greater than 2·0. It may be that the highly structured rhyming patterns of the poems induces a bias away from common words and toward unusual ones. The excess of m_+ over $\hat{\nu}_+$ is not huge, about 24% averaged over poems 4–7 in Table 6. About 78% of the 884687 total words in the Shakespearean canon consist of the 851 common words, those appearing 100 times or more each. A decrease in common words of just a few percent, say to 73% from 78%, would account for the observed effect.

Table 9. *Summary of significant z for three tests categorized as Table* 6

Poem	Total counts Test 1	New words Test 2	Slope Test 3
1. JON			**
2. MAR	***		****
3. DON		***	*
4. CYM	***		
5. PUC		*	
6. PHO	****	**	
7. SON			
8. TAY	**		

Two of the poems, the Shakespearean poem PHO and, especially, the non-Shakespearean poem DON, have a significant excess of new words; the Shakespearean poem PUC has a not so significant deficiency of new words. It is easy to believe that new word appearance might be sensitive to context. In any case, Test 2 seems only moderately useful for discerning Shakespearean authorship. Comparing columns 1 and 2 of Table 7, we see that there is no consistent trend toward an excess or deficiency of new words in the eight poems. It should be remembered that this comparison is conditional on the total counts m_+, and so includes an adjustment for their generally too-large values.

Test 3, the slope test, is definitely promising as a discriminator between Shakespearean versus non-Shakespearean authorship. None of the four Shakespearean poems gives a significant z. Two of the three non-Shakespearean poems give strong significance, while the Donne poem is borderline. On this evidence, the slope test has good power for detecting non-Shakespearean authorship, even for short poems. The theoretical $N(0, 1)$ distribution for $z = \hat{\beta}_1/\hat{\sigma}$ assuming Shakespearean authorship is roughly confirmed in Table 8, though it is mildly worrisome that all the values of z for poems 4–7 were negative. This may relate to the structure of highly-patterned poems, as mentioned above. It would obviously be informative to obtain counts for more poems. This is tedious work,

and we have only one other fragmentary result, three short poems of Jonson mentioned below.

An advantage of our approach is that all words in the poem are ordered on a one-dimensional scale, their rarity of Shakespearean occurrence, and therefore can be compared with Shakespeare using a one-degree-of-freedom test. This tactic, which is the basis of the slope test, is usually a good one for increasing power in high-dimensional testing situations.

On the basis of Table 9, the Taylor poem appears consistent with the hypothesis of Shakespearean authorship. In particular it passes the slope test, which is our best discriminator. It fails the total count test, but less dramatically than do two of the four Shakespearean poems. Overall it seems fair to say that the Taylor poem fits Shakespearean usage about as well as do the four Shakespeare poems.

The positive estimated slope for the Jonson poem stands out in Table 8. It seems to indicate a smaller working vocabulary than Shakespeare's, or at least a preference for more common words. The contrast with Marlowe is particularly striking. Three short Jonson poems, together only 231 total words in length, also showed a positive slope, though this value was not significantly different from zero, perhaps because of the short length.

Our tests are based on the counts of unusual words. Traditional authorship tests such as those of Mosteller & Wallace (1984) depend on common words. If a common-word test appropriate to this situation, i.e. for short, highly-structured poems, were derived, it would be essentially independent of our tests.

In fact we did try one common-word test on the eight poems, comparing the number of distinct words in each poem from Table 3 with its expectation assuming Shakespearean usage. Without going into details, the expectation depends mainly on the commonly-occurring words in the Shakespearean canon. The Taylor poem passes this test very nicely, as do all the Shakespearean poems. Only the Donne poem shows a significant discrepancy between actual and expected number of distinct words ($z = -2{\cdot}64$).

How well does model (4·1) fit the count data for the eight poems? Table 10 shows the deviance residuals from the maximum likelihood fit $\hat{\mu}_x$, equal to (4·1) with $(\hat{\beta}_0, \hat{\beta}_1)$ substituted for (β_0, β_1). The deviance residuals are those of McCullagh & Nelder (1983, p. 30). The data have been collapsed into ten categories of x values as in Table 10, except

Table 10. *Deviance residuals for the maximum likelihood fit of* (4·1) *to the data of Table 4; zero category disregarded*

	Category of x										Sum sq.
Poem	1	2	3–4	5–9	10–19	20–29	30–39	40–59	60–79	80–99	devs
JON	−0·28	−0·85	1·10	0·34	−1·05	0·40	0·79	−0·75	−0·71	0·97	6·02
MAR	−1·28	−0·16	1·09	0·62	−0·20	−0·10	−0·28	0·19	0·15	−0·45	3·63
DON	−0·44	0·66	−0·76	−0·06	0·52	1·03	−0·82	−0·05	−2·15	1·47	10·00
CYM	−0·52	−0·50	−0·49	0·48	0·62	−0·51	1·12	1·09	−1·91	−1·12	8·98
PUC	0·65	−2·10	−0·26	0·95	−0·87	1·06	−0·47	−1·02	1·40	−0·53	11·19
PHO	−0·40	0·24	0·78	−1·29	0·73	0·69	−0·50	0·15	−0·22	−0·52	4·09
SON	1·03	−1·89	−0·70	0·90	−0·67	−0·11	0·88	−0·91	0·57	−0·06	8·33
TAY	0·31	0·11	0·20	−0·74	−2·00	1·99	1·56	0·50	−1·03	−1·24	13·94
Sum sq. devs	3·92	9·49	4·42	4·46	7·54	7·06	6·31	3·93	12·19	6·65	

Bottom row, sum of squared residuals by column.

that the category $x=0$ is not considered. The sum of squared residuals shows no poem with an unusual value compared to a chi-squared with 8 degrees of freedom. Likewise summing the squared residuals down the columns reveals no discrepant categories.

6. Technical comments on the regression model and the slope test

The Poisson form for the counts m_x follows easily, as an excellent approximation, from the trapping assumptions made at the beginning of § 2. The regression form (4·1) is motivated by a simple alternative hypothesis for what constitutes non-Shakespearean authorship; we assume that a word with Shakespearean intensity parameter λ_s as in § 2, has non-Shakespearean intensity $\tilde{\lambda}_s$, where

$$\log \tilde{\lambda}_s = a_1 + (a_2+1)\log \lambda_s, \tag{6·1}$$

the constants (a_1, a_2) depending on the author but not on the word s. This is a simple log linear model in the natural parameter of a Poisson distribution.

Efron & Thisted (1976) showed that for Shakespeare the distribution $G(\lambda)$ appearing in (2·1) is well-approximated by an improper gamma distribution, with density function $dG(\lambda)$ proportional to

$$g(\lambda) = \lambda^{\alpha-1} e^{-\lambda\gamma/(1-\gamma)}. \tag{6·2}$$

The parameters in (6·2) were estimated to be $(\hat{\alpha}, \hat{\gamma}) = (-0{\cdot}3954, 0{\cdot}9905)$.

Lemma. *Under assumptions* (6·1), (6·2) *the expectations* μ_x *of* m_x *satisfy*

$$\log \mu_x \doteq \log \nu_x + \beta_0 + \beta_1 \log (x+\tfrac{1}{2}+\alpha), \tag{6·3}$$

where $\beta_0 = a_1 + a_2 \log \gamma$ *and* $\beta_1 = a_2$.

Proof. Together (2·2), (6·2), and the approximation $1-e^{-\lambda t} \doteq \lambda t$ give

$$\nu_x = St\gamma^{x+\alpha+1}(x+\alpha)!/x!.$$

Addition of the alternative hypothesis (6·1) gives

$$\mu_x \doteq S\int_0^\infty e^{-\lambda}\frac{\lambda^x}{x!}(1-e^{-\tilde{\lambda}t})\,dG(\lambda) \doteq St\,e^{a_1}\gamma^{x+\alpha+1+a_2}\frac{(x+\alpha+a_2)!}{x!}.$$

Therefore

$$\frac{\mu_x}{\nu_x} \doteq e^{a_1}\gamma^{a_2}\frac{(x+\alpha+a_2)!}{(x+\alpha)!},$$

and (6·3) follows from the approximation $d\log x!/dx \doteq \log(x+\frac{1}{2})$. □

The Lemma shows that (4·1) follows from the simple alternative hypothesis (6·1). It would seem better to use $\log(x+\frac{1}{2}+\alpha) = \log(x+0{\cdot}105)$ rather than $\log(x+1)$ in (4·1), but a small amount of numerical experimentation showed no difference in the results. The choice $\log(x+1)$ was used only because $x+1$ is the position of x in the sequence $0, 1, \ldots, 99$.

For estimating the slope β_1 it does not matter whether we condition on (m_+, m_0), as we now show. Conditioned on (m_+, m_0), the vector $m \equiv (m_1, \ldots, m_{99})$ has the multinomial distribution

$$m|(m_+, m_0) \sim \text{Mult}\{m_{(0)}, \pi(\beta_1)\}, \quad m_{(0)} \equiv m_+ - m_0 = m_1 + \ldots + m_{99},$$

where $\pi(\beta_1)$ is the vector with xth component

$$\pi_x(\beta_1) \equiv \frac{\mu_x(\beta_0, \beta_1)}{\sum \mu_y(\beta_0, \beta_1)} = \frac{\hat{\nu}_x(x+1)^{\beta_1}}{\sum \hat{\nu}_y(y+1)^{\beta_1}} \quad (x = 1, \ldots, 99).$$

This is a one-parameter exponential family with natural parameter β_1, so that the maximum likelihood estimate $\hat{\beta}_1$ satisfies

$$c'\{p - \pi(\hat{\beta}_1)\} = 0, \tag{6·4}$$

where c is the vector with xth component $\log(x+1)$, and $p \equiv m/m_{(0)}$, the vector of relative proportions of counts in categories $1, \ldots, 99$. The estimated standard error of $\hat{\beta}_1$, according to usual large-sample theory, is

$$\left[m_{(0)} \left\{ \sum \hat{\pi}_x c_x^2 - (\sum \hat{\pi}_x c_x)^2 \right\} \right]^{-\frac{1}{2}}, \tag{6·5}$$

where $\hat{\pi}_x \equiv \pi_x(\hat{\beta}_1)$.

Alternatively, we can consider the distribution of $m = (m_1, \ldots, m_{99})$ unconditionally, in which case its distribution is as described in (4·1) except only for $x = 1, \ldots, 99$. This is a two-parameter exponential family of distributions. The maximum likelihood value for $\hat{\beta}_1$ and its estimated standard error in this model are identical to those given by (6·4) and (6·5).

ACKNOWLEDGEMENTS

Ronald Thisted was supported by the National Science Foundation. Bradley Efron was supported by the National Science Foundation and the U.S. Public Health Service.

APPENDIX

Sources of poems analysed

[JON] 'An Elegy', number 40 in 'The Underwood' (Donaldson, 1975, pp. 188–9).
[MAR] Four short poems, three extracted from longer works (Williams, 1952, pp. 45–7).
[DON] 'The Ecstacy' (Williams, 1952, pp. 87–9).
[CUM] Apparitions' song from 'Cymbeline', Act 5, Scene 4: lines 30–92.
[PUC] Fairies' songs from 'A Midsummer Night's Dream', Act 3, Scene 2: lines 102–21, 396–9, 437–41, 448–64.
[PHO] 'The Phoenix and Turtle'
[SON] Sonnets 12–15.

The Shakespearean passages were selected to represent different styles and periods of Shakespeare's poetic writing. With the exception of the sonnets, these passages have more highly restrictive meter and rhyme schemes than other Shakespearean poetry, reflecting that found in the Taylor poem. The non-Shakespearean passages were chosen for their superficial similarity to the Taylor poem in terms of content and style.

REFERENCES

DONALDSON, I. (Ed.). (1975). *Ben Jonson: Poems.* Oxford University Press.
EFRON, B. & THISTED, R. (1976). Estimating the number of unseen species; How many words did Shakespeare know? *Biometrika* **63**, 435–7.
GOOD, I. J. & TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.

LELYVELD, J. (1985). A scholar's find: Shakespearean lyric. *New York Times* (November 24, 1985), 1, 12. With corrections of 'Editor's Note', (November 25, 1985), 2.

MCCULLAGH, P. & NELDER, J. A. (1983). *Generalized Linear Models.* London: Chapman and Hall.

MOSTELLER, F. & WALLACE, D. L. (1984). *Applied Bayesian and Classical Inference, The Case of the Federalist Papers.* New York: Springer-Verlag.

TAYLOR, G. (1985). Shakespeare's new poem: A scholar's clues and conclusions. *New York Times Book Review* (December 15), 11–4.

WILLIAMS, O. (Ed.). (1952). *Immortal Poems of the English Language.* New York: Washington Square Press.

[*Received May* 1986. *Revised October* 1986]