*Biometrika* (1976), **63**, 3, *pp.* 435–47
*With* 3 *text-figures*
*Printed in Great Britain*



# Estimating the number of unseen species: How many words did Shakespeare know?

BY BRADLEY EFRON AND RONALD THISTED
*Department of Statistics, Stanford University, California*

SUMMARY

Shakespeare wrote 31 534 different words, of which 14 376 appear only once, 4343 twice, etc. The question considered is how many words he knew but did not use. A parametric empirical Bayes model due to Fisher and a nonparametric model due to Good & Toulmin are examined. The latter theory is augmented using linear programming methods. We conclude that the models are equivalent to supposing that Shakespeare knew at least 35 000 more words.

*Some key words:* Empirical Bayes; Euler transformation; Linear programming; Negative binomial; Vocabulary.

## 1. INTRODUCTION

Estimating the number of unseen species is a familiar problem in ecological studies. In this paper the unseen species are words Shakespeare knew but did not use. Shakespeare's known works comprise 884 647 total words, of which 14 376 are types appearing just one time, 4343 are types appearing twice, etc. These counts are based on Spevack's (1968) concordance and on the summary appearing in an unpublished report by J. Gani & I. Saunders. Table 1 summarizes Shakespeare's word type counts, where $n_x$ is the number of word types appearing exactly $x$ times ($x = 1, \ldots, 100$). Including the 846 word types which appear more than 100 times, a total of

$$\sum_{x=1}^{\infty} n_x = 31\,534$$

different word types appear. Note that 'type' or 'word type' will be used to indicate a distinct item in Shakespeare's vocabulary. 'Total words' will indicate a total word count including repetitions. The definition of type is any distinguishable arrangement of letters. Thus, 'girl' is a different type from 'girls' and 'throneroom' is a different type from both 'throne' and 'room'.

How many word types did Shakespeare actually know? To put the question more operationally, suppose another large quantity of work by Shakespeare were discovered, say $884\,647t$ total words. How many new word types in addition to the original 31 534 would we expect to find? For the case $t = 1$, corresponding to a volume of new Shakespeare equal to the old, there is a surprisingly explicit answer. We will show that a parametric model due to Fisher, Corbet & Williams (1943) and a nonparametric model due to Good & Toulmin (1956) both estimate about 11 460 expected new word types, with an expected error of less than 150.

The case $t = \infty$ corresponds to the question as originally posed: how many word types did Shakespeare know? The mathematical model at the beginning of §2 makes explicit the sense of the question. No upper bound is possible, but we will demonstrate a lower bound

of approximately 35000 more word types in addition to the 31534 already observed. Our bound involves the theory of empirical Bayes estimation (Robbins, 1956; Good, 1953). It also involves linear programming in both a computational and a theoretical sense. This approach is similar to that taken by Harris (1959). More details are given in an unpublished report of the same title, available from the authors on request.

Table 1. *Shakespeare's word type frequencies*

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Row total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0+ | 14376 | 4343 | 2292 | 1463 | 1043 | 837 | 638 | 519 | 430 | 364 | 26305 |
| 10+ | 305 | 259 | 242 | 223 | 187 | 181 | 179 | 130 | 127 | 128 | 1961 |
| 20+ | 104 | 105 | 99 | 112 | 93 | 74 | 83 | 76 | 72 | 63 | 881 |
| 30+ | 73 | 47 | 56 | 59 | 53 | 45 | 34 | 49 | 45 | 52 | 513 |
| 40+ | 49 | 41 | 30 | 35 | 37 | 21 | 41 | 30 | 28 | 19 | 331 |
| 50+ | 25 | 19 | 28 | 27 | 31 | 19 | 19 | 22 | 23 | 14 | 227 |
| 60+ | 30 | 19 | 21 | 18 | 15 | 10 | 15 | 14 | 11 | 16 | 169 |
| 70+ | 13 | 12 | 10 | 16 | 18 | 11 | 8 | 15 | 12 | 7 | 122 |
| 80+ | 13 | 12 | 11 | 8 | 10 | 11 | 7 | 12 | 9 | 8 | 101 |
| 90+ | 4 | 7 | 6 | 7 | 10 | 10 | 15 | 7 | 7 | 5 | 78 |

Entry $x$ is $n_x$, the number of word types used exactly $x$ times. There are 846 word types which appear more than 100 times, for a total of 31534 word types.

## 2. THE BASIC MODEL

We use the species trapping terminology of Fisher's paper. Suppose that there exist $S$ species and that after trapping for one unit of time we have captured $x_s$ members of species $s$. Of course we only observe those values $x_s$ which are greater than zero. The basic distributional assumption is that members of each species $s$ enter the trap according to a Poisson process, the process for species $s$ having expectation $\lambda_s$ per unit time, so that $x_s$ has a Poisson distribution of mean $\lambda_s$ ($s = 1, \ldots, S$). Most of the calculations in this paper do not require the $S$ individual Poisson processes to be independent of one another. Whenever independence is required it will be specifically mentioned, and referred to as the 'independence assumption'.

Figure 1 gives a schematic representation of the situation. It is convenient to imagine the observation, or trapping, period as running from time $-1$ to time 0. We wish to extrapolate from the counts in $[-1, 0]$ to a time $t$ in the future. Let $x_s(t)$ be the number of times species $s$ appears in the whole period $[-1, t]$. The Poisson process assumption implies (i) that $x_s(t)$ has a Poisson distribution of mean $\lambda_s(1+t)$ and (ii) that, given $x_s(t)$, $x$ conditionally is binomial $\{x_s(t), 1/(1+t)\}$. For the situation described in the introduction, $t$ equals the total word count of newly discovered Shakespearean literature divided by 884647.

Assumption (i) is dispensable, but assumption (ii) is crucial. It says essentially that the time period $[-1, 0]$ is typical of the whole period $[-1, t]$. If the hypothetical newly discovered works of §1 were to consist entirely of business letters, we would not expect our results to be valid.

Let $G(\lambda)$ be the empirical cumulative distribution function of the numbers $\lambda_1, \ldots, \lambda_S$. Also, if $n_x$ is the number of species observed exactly $x$ times in $[-1, 0]$, let

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

and let $\Delta(t)$ be the expected number of species observed in $(0, t]$ but not in $[-1, 0]$, so that

$$\Delta(t) = S\int_0^\infty e^{-\lambda}(1-e^{-\lambda t})\,dG(\lambda). \tag{2·2}$$

We wish to estimate $\Delta(t)$, the expected number of new species to be found in the next $t$ time units. By substituting the expansion

$$1-e^{-\lambda t} = \lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} - \dots$$

into (2·2), and comparing the result with (2·1), we obtain the formal equality

$$\Delta(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 - \dots\ . \tag{2·3}$$

This intriguing result, which appears as formula (24) of Good & Toulmin (1956), is empirical Bayes in the sense Robbins (1956, 1968) originally attached to this term. It is related to an earlier result of Goodman (1949).

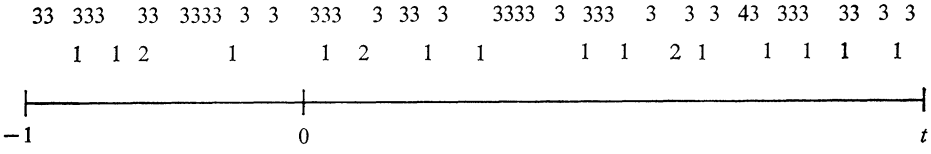


Fig. 1. The Poisson process model; $x_1 = 3$, $x_2 = 1$, $x_3 = 13$, $x_4 = 0$.

The right-hand side of (2·3) need not converge, but, if we assume that it does, expression (2·3) suggests the unbiased estimator for $\Delta(t)$

$$\hat{\Delta}(t) = n_1 t - n_2 t^2 + n_3 t^3 - \dots\ . \tag{2·4}$$

For the Shakespeare data with $t = 1$ this estimate is

$$\hat{\Delta}(1) = 11\,430. \tag{2·5}$$

Under the independence assumption a reasonable approximation, erring on the conservative side, is to take the $n_x$ themselves to be independent Poisson variates, with means $\eta_x$, in which case

$$\operatorname{var}\{\hat{\Delta}(1)\} = \sum_{x=1}^{\infty} \eta_x \doteq \sum_{x=1}^{\infty} n_x = 31\,534.$$

This gives $\hat{\Delta}(1)$ a standard deviation of 178.

The estimator $\Delta(t)$ is a function of the data only through the statistics $n_1, n_2, \dots$; the quantity $n_0$ is unobservable, being in fact almost the same as $\Delta(\infty)$. We are disregarding the labels connected with the observations $x_s$. All the estimates considered in this paper are of this form, but other authors have attempted more refined models; see McNeil (1973) and an unpublished paper by J. Gani and I. Saunders.

Unfortunately (2·4) is useless for values of $t$ larger than one. The geometrically increasing magnitude of $t^x$ produces wild oscillations as the number of terms increases. Good & Toulmin suggest the use of Euler's transformation to force convergence of the series. This idea is discussed in detail in §4. First though, we will examine Fisher's parametric empirical Bayes model in §3.

### 3. FISHER'S NEGATIVE BINOMIAL MODEL

Fisher *et al.* (1943) added the following assumptions to those at the beginning of §2.

*Fisher's assumption* 1. The cumulative distribution function $G(\lambda)$ is approximated by a gamma distribution with density function, for $c_{\alpha\beta} = \{\beta^{\alpha}\Gamma(\alpha)\}^{-1}$,

$$g_{\alpha\beta}(\lambda) = c_{\alpha\beta}\lambda^{\alpha-1}e^{-\lambda/\beta}. \tag{3·1}$$

*Fisher's assumption* 2. The parameters $\lambda_1, \ldots, \lambda_S$ are independent and identically distributed with density $g_{\alpha\beta}(\lambda)$.

Fisher's assumption 1 by itself gives most of the useful conclusions from this model. We shall note explicitly whenever Fisher's assumption 2 is invoked. Assumptions 1 and 2 together constitute a parametric empirical Bayes model in the sense of Efron & Morris (1973).

From (2·1) we obtain, for $\gamma = \beta/(1+\beta)$,

$$\eta_x = \eta_1 \frac{\Gamma(x+\alpha)}{x!\,\Gamma(1+\alpha)}\gamma^{x-1}. \tag{3·2}$$

Expression (3·2) is proportional to the negative binomial distribution with parameters $\alpha$ and $\gamma$, written to take advantage of the fact that for the unseen species problem the case $x = 0$ need not be considered. This allows the parameter $\alpha$ to take values less than zero, any value greater than $-1$ giving finite values to $\eta_1, \eta_2, \ldots$. The density $g_{\alpha\beta}(\lambda)$ is improper at the origin for $\alpha < 0$, and the expression (3·1) for $c_{\alpha\beta}$ is meaningless. Fisher particularly liked the choice $\alpha = 0$, which gives (3·2) the form known as the logarithmic distribution; see also Engen (1974) and Holgate (1969) for extended discussion.

We can write (2·2) in the form

$$\Delta(t) = \eta_1 \frac{\int_0^{\infty} e^{-\lambda}(1-e^{-\lambda t})\,dG(\lambda)}{\int_0^{\infty} \lambda e^{-\lambda}\,dG(\lambda)} \tag{3·3}$$

to avoid ambiguities in the case where $G$ is improper. By substituting (3·1) for $dG(\lambda)$ we obtain, in the obvious notation, $\Delta_{\alpha\gamma}(t) = -\eta_1\{(1+\gamma t)^{-\alpha}-1\}/(\gamma\alpha)$ unless $\alpha = 0$, in which case $\Delta_{0\gamma}(t) = (\eta_1/\gamma)\log(1+\gamma t)$.

If $\alpha > 0$, $\Delta_{\alpha\gamma}(t)$ approaches its limiting value $\eta_1/\alpha$ as $t$ goes to infinity. The improper cases $\alpha \leqslant 0$ have $\Delta_{\alpha\gamma}(t)$ increasing without bound as $t$ increases. The infinite spike of $g_{\alpha\beta}(\lambda)$ near $\lambda = 0$ produces an unbounded number of new species as longer and longer time periods are examined.

There is no reason to suppose that Fisher's parametric model will fit the Shakespeare data. It has only mathematical convenience and a limited amount of previous empirical successes to recommend it. In fact, the fit is extremely good. Substituting the values $\hat{\eta}_1 = 14376$, $\hat{\alpha} = -0{\cdot}3954$, $\hat{\gamma} = 0{\cdot}9905$, which are explained below, into (3·2) gives estimates $\hat{\eta}_x$ remarkably close to the observed $n_x$. To assess the accuracy of a fit such as Table 2 exhibits we need a theory of errors, and for that we need both the independence assumption mentioned at the beginning of §2 and also Fisher's assumption 2. Consider only the first $x_0$ values of $n_x$; $n_1, \ldots, n_{x_0}$. Denote their sum by $N_0$. Given $N_0$, the vector $(n_1, \ldots, n_{x_0})$ will have a multinomial distribution with $N_0$ trials and with vector of probabilities proportional to (3·2).

Table 3 shows the maximum likelihood fits, obtained by iterative search for various choices of $x_0$. The last column is Wilks's likelihood ratio statistic (Wilks, 1962, Chapter 13) for testing the adequacy of the two-parameter model based on (3·2). The sample sizes are enormous, the smallest being 23517, so that under the null hypothesis this statistic should

be distributed as a $\chi^2$ variate with $x_0 - 3$ degrees of freedom. We see that the fit is very good, even too good for $x_0 \leqslant 15$. With sample sizes of this magnitude, deviations of just a few percent from (3·2) would cause rejection.

All further calculations involving Fisher's model use the fitted parameter values for $x_0 = 40$,

$$\hat{\alpha} = -0{\cdot}3954, \quad \hat{\gamma} = 0{\cdot}9905. \tag{3·4}$$

We will use $\hat{\eta}_1 = n_1 = 14376$ rather than the fitted value $\hat{\eta}_1 = 14399$, which makes $\hat{\eta}_1 + \ldots + \hat{\eta}_{40}$ equal the observed sum 29660. However, in most of the calculations $\eta_1$ enters as a multiplicative constant, so that multiplication by $14399/14376 = 1{\cdot}0016$ converts the result. Note that the notation $\hat{\eta}_x$ will continue to mean any reasonable estimate of $\eta_x$. It will be mentioned when these are taken to be the maximum likelihood values from (3·2) and (3·4).

Table 2. *Maximum likelihood estimates for $\eta_x$ from Fisher's negative binomial model, and observed frequencies*

| | $x=1$ | $x=2$ | $x=3$ | $x=4$ | $x=5$ | $x=6$ | $x=7$ | $x=8$ | $x=9$ | $x=10$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\eta}_x$ | 14376 | 4305 | 2281 | 1471 | 1050 | 798 | 633 | 518 | 433 | 369 |
| $n_x$ | 14376 | 4343 | 2292 | 1463 | 1043 | 837 | 638 | 519 | 430 | 364 |

Table 3. *Maximum likelihood fits of the negative binomial model to the first $x_0$ values of $n_x$; $\hat{\beta} = \hat{\gamma}/(1-\hat{\gamma})$*

| $x_0$ | $\sum_{x=1}^{x_0} n_x$ | $\hat{\alpha}$ | $\hat{\gamma}$ | $\hat{\beta}$ | $\chi^2_{x_0-3}$ |
|---|---|---|---|---|---|
| 5 | 23517 | −0·3834 | 0·9795 | 47·82 | 0·027 |
| 10 | 26305 | −0·3906 | 0·9884 | 85·44 | 2·024 |
| 15 | 27521 | −0·3889 | 0·9861 | 70·78 | 3·815 |
| 20 | 28266 | −0·3901 | 0·9875 | 78·77 | 8·832 |
| 30 | 29147 | −0·3944 | 0·9899 | 97·92 | 16·874 |
| 40 | 29660 | −0·3954 | 0·9905 | 104·26 | 30·437 |

Unfortunately $\hat{\alpha} = -0{\cdot}3954$ puts us in the case where $\Delta_{\alpha\gamma}(t)$ goes to infinity as $t$ gets large. The data agree very well with a model we know must ultimately fail! However, we can still use (3·3) to estimate $\Delta(t)$ for finite values of $t$. For $t = 1$ we get

$$\hat{\Delta}(1) = \Delta_{-0{\cdot}3954,\,0{\cdot}9905}(1) = 11483.$$

This agrees with (2·5) to within 0·5 %.

For $t = 10$, $\Delta_{-0{\cdot}3954,\,0{\cdot}9905}(10) = 57704$, which is almost twice as large as Shakespeare's observed vocabulary. How accurate is this estimate? The hypothetical standard error from the negative binomial maximum likelihood estimation model, which we have not computed, is uninformative, since we know that that model must fail for large $t$. Sections 4 to 7 are devoted to finding nonparametric estimates of $\Delta(t)$ for large $t$, and assessing their accuracy.

## 4. Euler's transformation

Euler's transformation (Bromwich, 1955, p. 62) is a method of forcing oscillating series like (2·3) to converge rapidly. The substitution $t = u/(2-u)$ gives the formal relationship

$$\sum_{x=1}^{\infty} (-1)^{x+1} \eta_x t^x = \sum_{y=1}^{\infty} \xi_y u^y,$$

where

$$\xi_y = \sum_{x=1}^{y} \binom{y-1}{x-1} \frac{(-1)^{x+1}}{2^y} \eta_x = \frac{1}{2^y} \delta^y(\eta_1). \tag{4·1}$$

Here the backward difference operator is defined by

$$\delta^0(\eta_1) = \eta_1, \quad \delta^1(\eta_1) = \eta_1 - \eta_2, \quad \delta^2(\eta_1) = \eta_1 - 2\eta_2 + \eta_3, \quad \ldots .$$

Let

$$\Delta^{x_0}(t) = \sum_{x=1}^{x_0} (-1)^{x+1} \eta_x t^x, \quad \Delta^{x_0}(u) = \sum_{y=1}^{x_0} \xi_y u^y, \tag{4·2}$$

$$\Delta(t) = \lim_{x_0 \to \infty} \Delta^{x_0}(t), \quad \Delta(u) = \lim_{x_0 \to \infty} \Delta^{x_0}(u).$$

By definition $\Delta(t) = \Delta(u)$ if both limits exist. For $\eta_x$ positive, as here, the partial sums $\Delta^{x_0}(u)$ will usually converge more quickly to the common limit than the sums $\Delta^{x_0}(t)$. For $\Delta_{\alpha\gamma}(t)$ as given in (3·3), $\Delta^{x_0}(t)$ does not even converge, while the series $\Sigma_y \xi_y u^y$ converges in the nicest possible way, having in fact all nonnegative terms if $\alpha \leqslant 1$.

LEMMA. *For* $-1 < \alpha \leqslant 1, \Delta_{\alpha\gamma}(u) = \Sigma_y \xi_y u^y$ *has* $\xi_y \geqslant 0$ *for all* $y$.

The proof will be omitted.

Good & Toulmin (1956) suggest estimating $\xi_y$ by substituting $\hat{\eta}_x$ for $\eta_x$ in (4·1), and then using the Euler transformed series to estimate $\Delta(t)$,

$$\hat{\Delta}^{x_0}(u) = \sum_{y=1}^{x_0} \hat{\xi}_y u^y, \quad u = \frac{2t}{1+t}. \tag{4·3}$$

We have computed the first 20 values of $\hat{\xi}_y$ from (4·1) using $\hat{\eta}_x = n_x$ and also by using the maximum likelihood values (3·4). The latter are all positive, in accordance with the Lemma. The former are positive for $y = 1, \ldots, 9$, and negative for $y = 10, \ldots, 20$. However, all the negative values are within one-half a standard deviation of zero.

This suggests not taking $x_0$ greater than 9 if we intend to compute $\hat{\Delta}^{x_0}(u)$ from $\hat{\eta}_x = n_x$. The estimates $\hat{\xi}_y$ for $y > 9$ are within noise distance of zero, and we have, admittedly weak, theoretical reasons for believing the $\xi_y$ to be positive. The calculations of §5 will show $x_0 = 9$ to be a reasonable choice. The corresponding estimate of $\Delta(1)$ is $\hat{\Delta}^9(1) = 11\,441 \pm 147$, the standard error 147 being computed from (5·2). Calculation of $\hat{\Delta}^9(1)$ from the maximum likelihood estimates of $\hat{\eta}_x$, (3·4), gives $\hat{\Delta}^9(1) = 11\,460$ as the estimate. The question of assigning a standard deviation to the second of these estimates is a difficult one, but it is reasonable to say that the estimate is at least as accurate as the first one, and perhaps considerably more so.

In the present notation, (2·5) can be written as $\hat{\Delta}^\infty(1) = 11\,430 \pm 178$. Comparison of this with $\hat{\Delta}^9(1)$ above shows that we have reduced the standard deviation considerably by reducing $x_0$ from $\infty$ to 9. The price we pay, as Good & Toulmin noted, is in terms of bias. Thus $\hat{\Delta}^{x_0}(t)$ is not an unbiased estimate of $\Delta(t)$ for $x_0 < \infty$ because of the truncated terms in the series. The calculations of §5 will show that $\hat{\Delta}^9(1)$ can have a bias as large as $+8$ and as small as $-62$. This is with no assumptions on the form of $G(\lambda)$. Under the negative binomial model, the Lemma shows that $\hat{\Delta}^9(1)$ must have a negative bias, since all the terms we are ignoring are positive.

Taking both variance and bias into account, $\hat{\Delta}^9(1)$ is not noticeably superior to $\hat{\Delta}^\infty(1)$, except in computational effort. The choice of $x_0$ becomes far more crucial for values of $t > 1$, as §5 will show.

## 5. General linear estimators

There is another expression of the Euler transformation which makes obvious its effect on oscillating series. Substitution of (4·1) into the right-hand side of (4·2) shows, after some rearrangement, that $\Delta^{x_0}(u)$ is just the average of the oscillating series $\Delta^x(t)$ over values of $x$ distributed binomially $\{x_0, 1/(1+t)\}$. This averaging process is what smooths out the oscillations.

The estimator (4·3), with $\hat{\eta}_x = n_x$, is now seen to be of the form

$$\hat{\Delta} = \sum_{x=1}^{\infty} h_x n_x, \tag{5·1}$$

where, if $Z$ denotes a binomially distributed random variable with index $x_0$ and parameter $1/(1+t)$,

$$h_x = \begin{cases} (-1)^{x+1} t^x \operatorname{pr}(Z \geqslant x) & (x = 1, \ldots, x_0), \\ 0 & (x > x_0). \end{cases}$$

Notice that the naive estimator

$$\hat{\Delta}^{x_0}(t) = \sum_{x=1}^{9} (-1)^{x+1} n_x t^x$$

has $h_x = (-1)^{x+1} 10^x$ in this case, so that $h_9 = 10^9$, compared with $h_9 = 0{\cdot}424$ in Table 4! The Euler transformation drastically reduces $h_x$ for large $x$.

Table 4. *Euler coefficients in the general linear estimator* (5·1) *for* $x_0 = 9$ *and* $t = 10$

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $h_x$ | 5·759 | −19·421 | 41·539 | −59·152 | 57·155 | −37·190 | 15·653 | −3·859 | 0·424 |

We call estimators of the form (5·1) general linear estimators. We will calculate the variance of such an estimator from the independent Poisson assumption as

$$\operatorname{var}(\hat{\Delta}) = \sum_{x=1}^{\infty} h_x^2 \eta_x, \tag{5·2}$$

and note that this value may be somewhat large, as the calculations in an unpublished report by the authors demonstrate.

For each estimator (5·1) define the function

$$H(\lambda) = \sum_{x=1}^{\infty} h_x \lambda^x / x! \quad (0 < \lambda < \infty). \tag{5·3}$$

By (2·1) we have

$$E(\hat{\Delta}) = \sum_{x=1}^{\infty} h_x \eta_x = S \sum_{x=1}^{\infty} \int_0^{\infty} (h_x e^{-\lambda} \lambda^x / x!)\, dG(\lambda) = S \int_0^{\infty} e^{-\lambda} H(\lambda)\, dG(\lambda),$$

assuming, as will always be the case for the $h_x$ used below, that summation and integration can be interchanged. The bias of $\hat{\Delta}$ for estimating $\Delta(t)$ is, by (2·2),

$$E\{\hat{\Delta} - \Delta(t)\} = S \int_0^{\infty} e^{-\lambda} \{H(\lambda) - (1 - e^{-\lambda t})\}\, dG(\lambda). \tag{5·4}$$

It is convenient to rewrite (5·4) in a form which depends on $\eta_+ = \eta_1 + \eta_2 + \ldots$, rather than $S$, since we always have an easy estimate for $\eta_+$ available, namely $n_+ = \Sigma n_x$. Define

$$P = \int_0^\infty (1-e^{-\lambda})\,dG(\lambda), \quad d\tilde{G}(\lambda) = P^{-1}(1-e^{-\lambda})\,dG(\lambda).$$

Notice that $\eta_+ = SP$, by summation of $\eta_x$ in (2·1). That is, $P$ is just the expected proportion of the $\lambda_s$ having $x_s > 0$. Also $\tilde{G}$ can be thought of as the empirical cumulative distribution function of those $\lambda_s$ having $x_s > 0$, although strictly speaking this interpretation is only justified in the limiting case $S \to \infty$.

By multiplying and dividing (5·4) by $(1-e^{-\lambda})/P$ we obtain

$$E\{\hat{\Delta} - \Delta(t)\} = \eta_+ \int_0^\infty \frac{e^{-\lambda}}{1-e^{-\lambda}} \{H(\lambda) - (1-e^{-\lambda t})\}\, d\tilde{G}(\lambda). \tag{5·5}$$

The integrand

$$B_t(\lambda) = \frac{e^{-\lambda}}{1-e^{-\lambda}} \{H(\lambda) - (1-e^{-\lambda t})\} \tag{5·6}$$

determines the bias of $\hat{\Delta}$ for any $G(\lambda)$ or $\tilde{G}(\lambda)$.

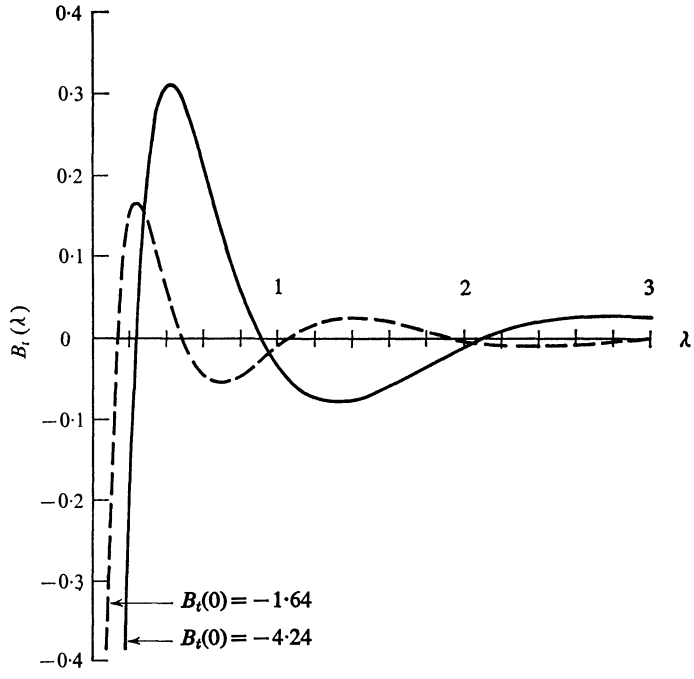


Fig. 2. The bias function $B_t(\lambda)$, equation (5·6), for $\hat{\Delta}^{x_0}(t)$, at $t = 10$; solid line, $x_0 = 9$, and dashed line, $x_0 = 19$.

For example, with $t = 1$, $x_0 = 9$ we compute $B_t(\lambda)$ to oscillate about zero, having its smallest value at $\lambda = 0$ and largest value at $\lambda = 0{\cdot}6$, $B_t(0) = -0{\cdot}00196$, $B_t(0{\cdot}6) = 0{\cdot}00024$. From (5·5) we see that the greatest negative bias occurs if $\tilde{G}$ puts all its mass at $\lambda = 0$, in which case the bias equals $-0{\cdot}00196\eta_+ \doteq -0{\cdot}00196 \times 31\,534 = -62$. The greatest positive bias occurs if $\tilde{G}$ puts all mass at $\lambda = 0{\cdot}6$, in which case it equals $0{\cdot}00024 \times 31\,534 = 8$. Of course, the data in Table 1 tell us that $\tilde{G}$ follows neither extreme for the Shakespeare word counts. In §§6 and 7 we employ such information to get better bounds in a systematic way.

Figure 2 shows $B_t(\lambda)$ at $t = 10$, for $x_0 = 9$ and for $x_0 = 19$. The bias situation is now much more serious. For $x_0 = 9$ the possible bias ranges from $-4{\cdot}24\eta_+$ to $0{\cdot}31\eta_+$. For $x_0 = 19$ the range is from $-1{\cdot}64\eta_+$ to $0{\cdot}15\eta_+$. This does not mean that $x_0 = 19$ is better than $x_0 = 9$. The respective estimators from equation (4·3) and their standard deviations from (5·2) are $\hat{\Delta}^9(10) = 45\,188 \pm 3994$ and $\hat{\Delta}^{19}(10) = 53\,867 \pm 702\,566$. Its huge variance makes $\hat{\Delta}^{19}(10)$

useless. The choice of $x_0$ must take into account both bias and variance. For this case $x_0 = 9$ seemed to be as good or better than any other choice, though admittedly the criterion of goodness is vague.

We need not restrict attention to linear estimators of the form (4·3). An attempt to choose a best linear estimator $\hat{\Delta}(10) = h_1 n_1 + \ldots + h_{x_0} n_{x_0}$ is described in unpublished work by the authors. This search yielded no estimator noticeably superior to $\hat{\Delta}^9(10)$.

## 6. Lower bounds for $\Delta(t)$

As $t$ gets large it becomes more and more difficult to estimate a reasonable upper bound for $\Delta(t)$. Suppose that Shakespeare actually had $10^6$ word types with $\lambda_s = 10^{-6}$. These types would have almost no effect on our data set. The expected number of them occurring in our sample is only 1. However, for $t = 10^6$ an expected fraction $1 - e^{-1} = 0{\cdot}632$ of them would be observed. This type of counterexample can be pushed arbitrarily far. Unfortunately, the trouble begins for $t$ values much smaller than $10^6$. We see this in Fig. 2, where the possible negative bias is already very large for $t = 10$.

Table 5. *Lower bound estimates for* $\Delta(t)$ *based on linear transformations of* $\hat{\Delta}^{x_0}(u)$, $x_0 = 9$

| $t$ | $a$ | $b$ | Lower bound estimate (6·2) | St. dev. (5·2) | Estimate − st. dev. |
|---|---|---|---|---|---|
| 1 | 0·999 | 0·0001 | 11454 | 147 | 11307 |
| 3 | 0·979 | 0·002 | 25143 | 986 | 24157 |
| 5 | 0·939 | 0·007 | 31974 | 1966 | 30008 |
| 8 | 0·879 | 0·010 | 36554 | 2965 | 33588 |
| 10 | 0·850 | 0·012 | 38015 | 3397 | 34618 |
| 12 | 0·827 | 0·014 | 38927 | 3713 | 35214 |
| 15 | 0·801 | 0·016 | 39784 | 4048 | 35736 |
| 20 | 0·772 | 0·017 | 40580 | 4408 | 36172 |
| 30 | 0·742 | 0·019 | 41331 | 4793 | 36538 |
| 60 | 0·710 | 0·022 | 42061 | 5212 | 36848 |
| 120 | 0·694 | 0·023 | 42411 | 5433 | 36977 |

The situation is better for lower bounds. Equation (5·3) shows that $\hat{\Delta} = \Sigma h_x n_x$ satisfies $E(\hat{\Delta}) \leqslant \Delta(t)$ if, for all $\lambda \geqslant 0$,

$$H(\lambda) \leqslant 1 - e^{-\lambda t}. \tag{6·1}$$

In other words, the linear estimator $\hat{\Delta}$ will be a lower bound for $\Delta(t)$ in expectation, no matter what $G$ happens to be, if $H(\lambda)$ is everywhere less than $1 - e^{-\lambda t}$.

As we saw in §5, the estimators based on Euler's transformation do not satisfy (6·1). However, given $\hat{\Delta} = \Sigma h_x n_x$ we can always make a linear transformation $h_x^0 = a h_x - b$ $(x = 1, 2, \ldots)$ which gives, through (5·3), $H^0(\lambda) = aH(\lambda) - b(e^{\lambda} - 1)$. The corresponding new estimator is

$$\hat{\Delta}^0 = \sum_{x=1}^{\infty} h_x^0 n_x = a\hat{\Delta} - b n_+, \tag{6·2}$$

where $n_+ = \Sigma n_x$ as before.

Table 5 shows the lower bounds obtained in this way from the Euler estimators (5·1), with $x_0 = 9$, for various choices of $t$. The constants $a$ and $b$ were chosen so that $H^0$ satisfied (6·1). Subject to this constraint, $a$ and $b$ were selected to maximize (6·2) with $\hat{\eta}_x$ in place of $n_x$, $\hat{\eta}_x$ the maximum likelihood estimates obtained from (3·2) and (3·4). The resulting value

of $\hat{\Delta}^0$ is tabulated as the 'lower bound estimate'. The standard deviation from (5·2) appears in the next column, followed by the estimate minus one standard deviation.

Table 5 shows that this reasonably conservative lower bound for $\Delta(t)$ fails to get much larger as $t$ grows from 10 to 120. As we shall see in §7, it is impossible to get a substantially larger lower bound for $t$ approaching infinity, even using more general linear estimators. This seems to say that the Shakespeare data, unaided by parametric assumptions like Fisher's assumption 1, runs out of predictive power for $t$ greater than 10.

A potential flaw in Table 5 is that the estimates and standard deviations are derived ignoring the fact that $a$ and $b$ depend on the data, since they are chosen so that (6·2) is maximized for the data set at hand. This point is considered more carefully in §7, and is shown not to make much difference.

## 7. LINEAR PROGRAMMING BOUNDS

The method employed in §6 to find $\hat{\Delta}$ satisfying $E(\hat{\Delta}) \leqslant \Delta(t)$ can be approached more generally as a linear programming problem.

*Program* 1. Choose $h_1, \ldots, h_{x_0}, h_{x_0+1}$ to maximize

$$\hat{\Delta} = \sum_{x=1}^{x_0} h_x \hat{\eta}_x + h_{x_0+1} \sum_{x=x_0+1}^{\infty} \hat{\eta}_x \tag{7·1}$$

subject to the constraints, for $\lambda > 0$,

$$H(\lambda) = \sum_{x=1}^{x_0} h_x \lambda^x/x! + h_{x_0+1} \sum_{x=x_0+1}^{\infty} \lambda^x/x! \leqslant 1 - e^{-\lambda t}. \tag{7·2}$$

Condition (7·2) guarantees, by (6·1), that $E(\hat{\Delta}) \leqslant \Delta(t)$ for any $G$. Subject to this constraint, (7·1) requires maximization of the estimated value at a likely value of the true parameters $\eta_1, \eta_2, \ldots$ . In this section we take $\hat{\eta}_x$ to be the maximum likelihood values from (3·2) and (3·4) for $x = 1, \ldots, x_0$ and set

$$\sum_{x=x_0+1}^{\infty} \hat{\eta}_x = \sum_{x=x_0+1}^{\infty} n_x.$$

Other reasonable choices of $\hat{\eta}_x$ give almost identical answers.

Program 1 was solved on the IBM 360/67 computer at Stanford using the IBM program MPS/360. The infinite number of constraints in (7·2) was replaced by the discrete set

$$H(\lambda_l) \leqslant 1 - e^{-\lambda_l t}, \quad \lambda_l = 2^{\frac{1}{16}l - 10} \quad (l = 0, \ldots, 272) \tag{7·3}$$

($\lambda_0 = 2^{-10}, \lambda_{272} = 128$). As before, $x_0 = 9$ was used for most of the calculations. These choices were based on a small amount of numerical experimentation.

For the case $t = \infty$ the resulting optimum coefficients $h_x$ were substituted into (7·1) to obtain the lower bound estimate $\hat{\Delta}(\infty) = 59568$ for Shakespeare's total unobserved vocabulary. Unfortunately, the standard error for $\hat{\Delta}(\infty)$, calculated from (5·2) on the assumption that the $h_x$ are fixed constants, is the enormous value, 204784. This might seem to render $\hat{\Delta}(\infty)$ useless, but we shall see that this is not actually so.

The linear programming problem dual to program 1 (Hillier & Lieberman, 1974, p. 90) or, rather, the dual to the discretized version (7·1) and (7·3), is as follows.

*Program* 2. Choose $S > 0$ and a discrete distribution function $G(\lambda)$ with support on the set $\{\lambda_1, \lambda_2, \lambda_l, \ldots, \lambda_L\}$ to minimize

$$\Delta(t) = S \int_0^{\infty} e^{-\lambda}(1 - e^{-\lambda t})\, dG(\lambda) \tag{7·4}$$

subject to the constraints

$$S\int_0^\infty (e^{-\lambda}\lambda^x/x!)\,dG(\lambda) = \hat{\eta}_x \quad (x = 1, \ldots, x_0), \quad S\int_0^\infty e^{-\lambda} \sum_{x=x_0+1}^{\infty} (\lambda^x/x!)\,dG(\lambda) = \sum_{x=x_0+1}^{\infty} \hat{\eta}_x. \tag{7·5}$$

The dual Program 2 finds the 'least favourable situation' in that it selects $S$ and $G$ to minimize $\Delta(t)$ subject to the constraint that the expected word counts $\eta_1, \ldots, \eta_{x_0}$ and the sum of their successors equal certain specified values. Program 2 is nearly identical to the problem considered by Harris (1959). With the $\hat{\eta}_x$ chosen as before we see that, by the duality

Table 6. *Points of support of minimizing distribution G in Program 2 at* $t = \infty$

| | $l$ = 120–121 | $l$ = 167–168 | $l$ = 191–192 | $l$ = 208–209 | $l$ = 226–227 |
|---|---|---|---|---|---|
| $\lambda$ | 0·1806 | 1·384 | 3·914 | 8·175 | 17·830 |
| $dG$ | 0·7444 | 0·1220 | 0·0507 | 0·0294 | 0·0535 |

Table 7. *Lower and upper bounds on* $\Delta(t)$ *calculated by solving the linear programming problem* (7·4) *and* (7·6); $x_0 = 9$, $c = 1$

| $t$ | Lower bound on $\Delta(t)$ | Upper bound on $\Delta(t)$ |
|---|---|---|
| 1 | 11205 | 11732 |
| 3 | 23828 | 29411 |
| 5 | 29898 | 45865 |
| 10 | 34640 | 86600 |
| 20 | 35530 | 167454 |
| $\infty$ | 35554 | $\infty$ |

theorem, Program 2 has the same solution as Program 1. For $t = \infty$ solving Program 2 gives $\hat{\Delta}(\infty) = 59568$, as before. The minimizing distribution $G$ has its support at 10 of the $\lambda_i$ values, occurring in five adjacent pairs, given in Table 6. Of course we do not believe that $\eta_x = \hat{\eta}_x$ exactly, but we can loosen the constraints to take into account our uncertainty, say by taking

$$\hat{\eta}_x - c\sqrt{\hat{\eta}_x} \leqslant S\int_0^\infty (e^{-\lambda}\lambda^x/x!)\,dG(\lambda) \leqslant \hat{\eta}_x + c\sqrt{\hat{\eta}_x} \quad (x = 1, \ldots, x_0), \tag{7·6}$$

and similarly for the last constraint in (7·5). Here $c$ measures approximately how many standard deviations we allow the fitted values of $\eta_x$ to vary from $\hat{\eta}_x$.

Solution of (7·4) and (7·6) for $t = \infty$, $x_0 = 9$ and $c = 1$ gives a minimum value of $\hat{\Delta} = 35554$, which is quite consistent with the last column of Table 4.

We now have a believable lower bound on $\Delta(\infty)$. The choice $c = 1$ may seem optimistic, but we have reason to believe the true $\eta_x$ to be nearer $\hat{\eta}_x$ than (2·6) indicates. The issue of concern to us in §6, namely, that of choosing the estimator from the data and then ignoring that selection process in setting confidence intervals, has disappeared. The linear programming method yields a lower bound directly as a function of the unknown parameters $\eta_x$. Confidence bounds on $\eta_x$ of the type (7·6) then yield a bound on $\hat{\Delta}$ in the usual way. For those preferring a still more conservative bound, $c = 2$ gives $\hat{\Delta}(\infty) = 30845$ with $x_0 = 9$.

Table 7 gives lower and upper bounds on $\Delta(t)$ obtained from (7·4) and (7·6) with $x_0 = 9$, $c = 1$. For the upper bound the 'minimize' in (7·5) is simply changed to 'maximize'. The agreement of the lower bounds with the last column of Table 5 is remarkable. This is important since Table 5 is much easier to calculate than Table 7.

## 8. CONCLUSIONS

Figure 3 displays the different estimates of $\Delta(t)$. Our experience with the Shakespeare data can be summarized as follows.

(i) Estimate $\hat{\Delta}(\infty) = 35\,000$ is a reasonably conservative lower bound for the amount of vocabulary Shakespeare knew but did not use.

(ii) An estimate of $\Delta(t)$ can be made very accurately for $t \leqslant 1$, but the uncertainties magnify quickly as $t$ grows larger. Without a parametric model the data give very little additional information for $t$ larger than 10.

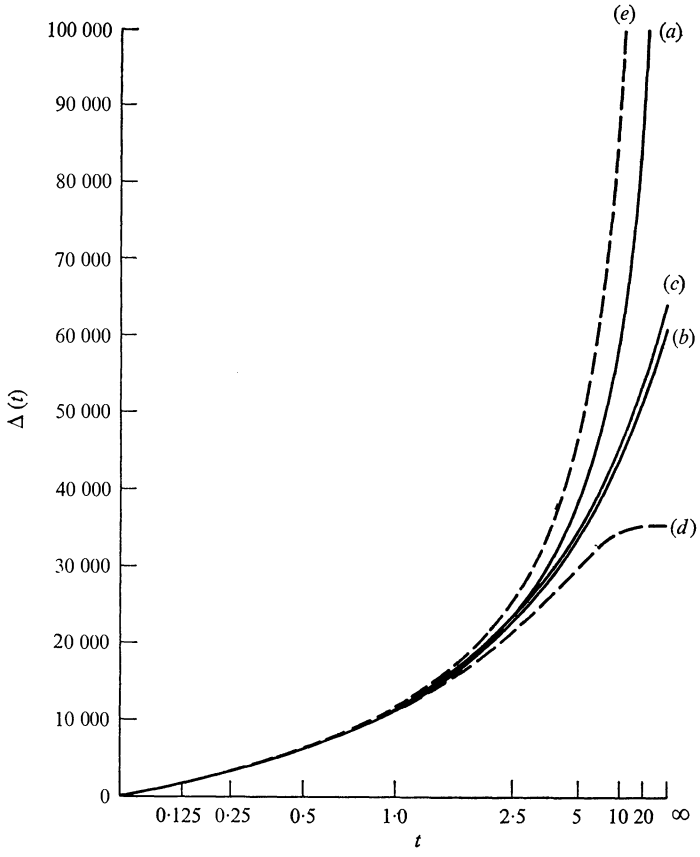


Fig. 3. Different estimates of $\Delta(t)$ for the Shakespeare data: (*a*) Fisher's negative binomial model with parameters (3·4); (*b*) Euler transformation (4·4), $x_0 = 9$, $\hat{\xi}_y$ from $\hat{\eta}_x = n_x$; (*c*) as (*b*), but with $\hat{\xi}_y$ from maximum likelihood values (3·2) and (3·4); (*d*) lower bound estimates from linear program (7·4) and (7·6), $c = 1$; (*e*) upper bound, as (*d*).

(iii) Fisher's negative binomial model fits the data extraordinarily well. However the linear programming approach produces other empirical Bayes solutions which also fit the observed data, and give smaller estimates of $\Delta(t)$ for $t > 1$.

(iv) All the methods give very similar answers for $t \leqslant 1$.

(v) Euler's transformation performs well compared to more elaborate techniques.

This paper was inspired by a lecture by J. Gani; P. Diaconis contributed many useful ideas and references.

## References


Bromwich, T. (1955). *An Introduction to the Theory of Infinite Series*, 2nd edition. London: Macmillan.

Efron, B. & Morris, C. (1973). Stein's estimation rule and its competitors – an empirical Bayes approach. *J. Am. Statist. Assoc.* **68**, 117–30.

Engen, S. (1974). On species frequency models. *Biometrika* **61**, 263–70.

Fisher, R. A., Corbet, A. S. & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12**, 42–58.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–64.

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.

Goodman, L. A. (1949). On the estimation of the number of classes in a population. *Ann. Math. Statist.* **20**, 572–9.

Harris, B. (1959). Determining bounds on integrals with applications to cataloging problems. *Ann. Math. Statist.* **30**, 521–48.

Hillier, F. & Lieberman, G. (1974). *Introduction to Operations Research*, 2nd edition. San Francisco: Holden-Day.

Holgate, P. (1969). Species frequency distributions. *Biometrika* **56**, 651–60.

McNeil, D. (1973). Estimating an author's vocabulary. *J. Am. Statist. Assoc.* **68**, 92–6.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp.* **1**, 137–63.

Robbins, H. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Statist.* **39**, 256–7.

Spevack, M. (1968). *A Complete and Systematic Concordance to the Works of Shakespeare*, Vols. 1–6. Hildesheim: George Olms.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley.


[*Received June* 1975. *Revised December* 1975]