ESTIMATING THE NUMBER OF UNSEEN SPECIES
(HOW MANY WORDS DID SHAKESPEARE KNOW?)

BY

BRADLEY EFRON and RONALD THISTED

TECHNICAL REPORT NO. 70
JULY 25, 1975

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ESTIMATING THE NUMBER OF UNSEEN SPECIES
(HOW MANY WORDS DID SHAKESPEARE KNOW?)

by

BRADLEY EFRON and RONALD THISTED

TECHNICAL REPORT NO. 70
July 25, 1975

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MPS74-21416

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Also partially supported by Public Health Service Grant 1 RO1 GM21215-01
Report No. 9
ESTIMATING THE NUMBER OF UNSEEN SPECIES
(HOW MANY WORDS DID SHAKESPEARE KNOW?)

by

Bradley Efron and Ronald Thisted

Abstract
Shakespeare wrote 31,534 different words, of which 14,376 appear only once, 4,343 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 and Toulmin are examined. The latter theory is augmented using linear programming methods. We conclude that Shakespeare knew at least 35,000 more words.

Acknowledgement
This paper was inspired by a lecture by J. Gani. P. Diaconis contributed many useful ideas and references.
ESTIMATING THE NUMBER OF UNSEEN SPECIES

(HOW MANY WORDS DID SHAKESPEARE KNOW?)

by

Bradley Efron and Ronald Thisted

1. Introduction

Shakespeare's known works comprise 884,647 total words, of which 14,376 are types* appearing just one time, 4,343 are types appearing twice, etc. These counts are based on Spevack's concordance [16] and on the summary appearing in [6]. Table 1 summarizes Shakespeare's word type counts,

\[ n_x \equiv \text{number of word types appearing exactly } x \text{ times}, \]

for \( x \) equals 1 to 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.

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 884,647 \( \cdot t \) total words. How many

---

*"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 "throne room" is a different type from both "throne" and "room".
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 [5] and a nonparametric model due to Good and Toulmin [8] both estimate about 11,460 expected new word types, with an expected error of less than 150. A much less accurate estimate is possible for \( t=10 \), approximately 45,000 expected new word types, but with the range of reasonable values as low as 35,000 and as high as 87,000.

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

Table 1. Shakespeare's word type frequencies. Entry \( x \) is \( n_x \), the number of word types used exactly \( x \) times. 846 word types appear more than 100 times, for a total of 31,534 word types.

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 Section 2 makes explicit the sense of the question.)
No upper bound is possible, but we will demonstrate a lower bound of approximately 35,000 more word types in addition to the 31,534 already observed.

Estimating the number of unseen species, in this case the number of unobserved word types, is a familiar problem in ecological studies. Corbet, in the first part of the paper involving Fisher [5], trapped butterflies in Malaya. At the end of his study he had observed 118 species once, 74 twice, 44 three times, etc., the first fifteen values of $n_x$ being

<table>
<thead>
<tr>
<th>$x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_x$</td>
<td>118</td>
<td>74</td>
<td>44</td>
<td>24</td>
<td>29</td>
<td>22</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>6</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

How many more species would Corbet expect to see if he trapped for another $t$ units of time? Fisher answers the question in terms of a parametric model. Good and Toulmin [8] rightly question the uncritical acceptance of this model, and propose a nonparametric approach. (See also [7], [15].)

We shall use the Shakespeare data to investigate the theoretical questions involved, and to see how the two methods compare in this case. This data set has some extraordinarily nice features which make it possible to examine the problem under the most favorable conditions. Finally, we shall add to the theory of [5] and [8] in order to provide a dependable lower bound on the number of unseen species. Our bound involves the theory of empirical Bayes estimation as put forth in Robbins [14] and Good [7]. It also involves linear programming in both a computational and a theoretical sense. This approach is similar to that taken in Harris [10].
Capture-recapture problems of the type discussed in [1] resemble the unseen species problem. For example, five lists of heroin addicts are available, from which one can count how many addicts appear once, twice, ..., five times. The question is how many addicts have not appeared on any of the lists? This can be considered a discretized version of the problem discussed here, in which the species, here individual addicts, are recorded as present or absent in each list rather than having their total number of appearances counted. We have not attempted to compare the methodology of this paper with traditional capture-recapture methods.

2. The Basic Model

We use the species trapping terminology of Fisher's paper [5]. Suppose that there exist \( S \) species: 1, 2, 3, ..., \( s \), ..., \( S \), 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

\[
(2.1) \quad x_s \sim \text{Poisson}(\lambda_s), \quad s = 1, 2, \ldots, S.
\]

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 that (i) $x_s(t) \sim$ Poisson ($\lambda_s[1+t]$) and that (ii) given $x_s(t)$, $x_s$ conditionally is Binomial ($x_s(t), 1/(1+t)$).

\[
\begin{array}{cccccccccccccc}
33 & 333 & 33 & 33 & 3 & 33 & 3 & 33 & 3 & 333 & 3 & 3 & 3 & 43 & 333 & 33 & 3 & 3
\end{array}
\]

\[
\begin{array}{cccccccccccccc}
1 & 1 & 2 & 1 & 1 & 2 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}
\]

\[-1 \quad \quad 0 \quad \quad t\]

Figure 1. The Poisson process model. In this example $x_1 = 3$, $x_2 = 1$, $x_3 = 13$, $x_4 = 0$.

Assumption (i) is dispensable. It is possible, though clumsy, to derive most of what follows without any distributional assumptions on the $x_s(t)$ values. (See the end of this Section.) Assumption (ii), however, is crucial. It says essentially that the time period $[-1,0]$ is typical of the whole period $[-1,t]$. If the hypothetical newly discovered Shakespeare of Section 1 consisted entirely of business letters, we would not expect our results to be valid. In asking how many word types Shakespeare knew we are really asking how large a vocabulary he was drawing on for his plays and sonnets.
Assumption (ii) is more likely to be valid in the large than in the small. Evidence in [6] suggests that the historical plays are richer in vocabulary than the comedies. If we found just one more play and it was a history, we might expect (ii) to overestimate $x_s$, once-used word types for example having greater than average probability of appearing in the new sample. (Notice that "t" in this case is defined to be the total word types in the newly discovered work divided by 884,647.) Since here we are trying to peer into Shakespeare's mind and estimate how many word types he might have used in writing what he actually did write, Assumption (ii) seems at least plausible. The results of Section 3 lend it indirect support.

A natural alternative to the Poisson process assumption is the compound Poisson process, in which bunches of the same species are trapped according to individual Poisson processes, but the bunches can be of different sizes. Without going into detail, this assumption would tend to increase our estimates of the unseen vocabulary. It is difficult, though not impossible, to reconcile the compound model with the results of Section 3. This model is interesting in its own right, but will not be pursued further here.

Good and Toulmin [8] assume that each unit observed is the result of a single independent multinomial draw, belonging to species $s$ with some fixed probability $\pi_s$, $\sum_{s=1}^{S} \pi_s = 1$. When the number of draws, in this case 884,647, becomes large this leads to approximately the same model as above, with the added result that the individual Poisson processes are asymptotically independent. Assumption (ii) is tacitly
satisfied in their model since draws actually observed are assumed typical of those contemplated in the future. The point of view in [8] is "empirical Bayes", in the nonparametric sense Robbins originally attached to that term. Nothing is assumed about the values \( \pi_1, \pi_2, \ldots, \pi_S \), but the data \( x_1, x_2, \ldots, x_S \) are used to draw inferences about them. We shall follow the same approach with the parameters \( \lambda_1, \lambda_2, \ldots, \lambda_S \).

Let \( G(\lambda) \) be the empirical c.d.f. of the numbers \( \lambda_1, \lambda_2, \ldots, \lambda_S \),

\[
G(\lambda) = \frac{\# \{ \lambda < \lambda \}}{S}.
\]

Also define

\[
(2.2)
\eta_x \equiv E \{ n_x \} = E \# \{ \text{species observed exactly } x \text{ times in } [-1,0] \} = S \int_0^\infty e^{-\lambda} \lambda^x / x! \, dG(\lambda)
\]

and

\[
(2.3)
\Delta(t) \equiv E \# \{ \text{species observed in } (0,t] \text{ but not in } [-1,0] \} = S \int_0^\infty e^{-\lambda} (1 - e^{-\lambda t}) \, dG(\lambda).
\]

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

\[
(2.5)
1 - e^{-\lambda t} = \lambda t - \frac{\lambda^2 t^2}{2!} + \frac{\lambda^3 t^3}{3!} + \ldots
\]

into (2.4), and comparing with (2.3), gives the formal equality.
\[ \Delta(t) = \eta_1 t - \eta_2 t^2 + \eta_3 t^3 \ldots \]

This intriguing result appears as formula (24) in Good and Toulmin [8]. It is related to an earlier result in Goodman [9].

The right-hand side needn't converge, but assuming it does it suggests the unbiased estimator for \( \Delta(t) \)

\[ \hat{\Delta}(t) = n_1 t - n_2 t^2 + n_3 t^3 \ldots \]

For the Shakespeare data with \( t=1 \) this gives

\[ \hat{\Delta}(1) = 11,430. \]

Under the independence assumption a reasonable approximation, erring on the conservative side, is to take the \( n_x \) themselves to be independent Poisson variates,

\[ \text{ind} \quad n_x \sim \text{Poisson} (\eta_x), \quad x = 1, 2, \ldots, \]

in which case

\[ \text{var} \hat{\Delta}(1) = \sum_{x=1}^{\infty} \eta_x \sim \sum_{x=1}^{\infty} n_x = 31,534. \]

(See the discussion at the end of Section 3.) This gives \( \hat{\Delta}(1) \) a standard deviation of 178.

The estimator \( \Delta(t) \) is a function of the data only through the statistics \( n_1, n_2, \ldots, n_x, \ldots \) (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 [6], [13].

Unfortunately (2.7) 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 and Toumlin [8] suggest the use of Euler's transformation to force convergence of the series. This idea is discussed in detail in Section 4. First though, we will examine Fisher's parametric empirical Bayes model in Section 3.

Returning to the discussion at the beginning of this section, let us now assume that \( m_x \) of the species were trapped exactly \( x \) times in the time interval \([-1,t]\). Define \( p = 1/(1+t), q = t/(1+t) \). Then solely under "Assumption (ii)", that given \( x_s(t), x_s \sim \text{Binomial} (x_s(t), p) \), we have

\[
\begin{align*}
\eta_0 &= m_0 + qm_1 + q^2m_2 + q^3m_3 + \ldots \\
\eta_1 &= pm_1 + 2pqm_2 + 3pq^2m_3 + \ldots \\
\eta_2 &= p^2m_2 + 3p^2qm_3 + \ldots \\
\eta_3 &= p^3m_3 + \ldots
\end{align*}
\]
(2.11)

etc. Remembering that \( q/p = t \) this gives

\[
(2.12) \quad \eta_0 - \eta_1 t + \eta_2 t^2 - \eta_3 t^3 + \ldots = m_0
\]

Since \( \eta_0 - m_0 \) equals the expected number of species seen in \((0,t]\) that were not seen in \([-1,0]\), we have derived (2.6) using only Assumption (ii).
3. **Fisher's Negative Binomial Model**

Fisher [5] added the following assumptions to those at the beginning of Section 2:

F1. The c.d.f. $G(\lambda)$ is approximated by a gamma distribution with density function

$$
(3.1) \quad g_{\alpha, \beta}(\lambda) = c_{\alpha, \beta} \lambda^{\alpha-1} e^{-\lambda/\beta} \quad (c_{\alpha, \beta} = [\beta^\alpha \Gamma(\alpha)]^{-1}).
$$

F2. The parameters $\lambda_1, \lambda_2, \ldots, \lambda_s, \ldots, \lambda_S$ are i.i.d. from $g_{\alpha, \beta}(\lambda)$. Actually F1 by itself gives most of the useful conclusions from this model. It will be explicitly noted whenever F2 is invoked. F1 and F2 together constitute a parametric empirical Bayes model in the sense of [3].

From (2.3) we have

$$
\eta_x = S c_{\alpha, \beta} \int_0^\infty \frac{\lambda^{x+\alpha-1} e^{-\lambda(1+1/\beta)}}{x!} d\lambda,
$$

$$
(3.2) \quad = S c_{\alpha, \beta} \frac{\Gamma(x+\alpha)}{x!} \gamma^{x+\alpha},
$$

where we have defined

$$
(3.3) \quad \gamma = \frac{\beta}{1+\beta}.
$$

Expression (3.2) is proportional to the negative binomial distribution with parameters $\alpha$ and $\gamma$. The last form of (3.2) is useful since for this problem we need only consider values of $x > 0$. 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 $\alpha = 0$, which gives (3.2) the form known as logarithmic distribution. See also Engen [4] and Holgate [12].

The parameter $\gamma$ goes from zero to one as $\beta$ goes from zero to infinity. The limiting case $\beta = \infty$, $\gamma = 1$ gives the improper density $g_{\alpha}(\lambda) \propto \lambda^{\alpha-1}$. Formula (3.2) is still valid in its last form.

The improper cases of (3.1) can be thought of as referring to an infinite population of possible $\lambda$ values. We can still imagine that the finite integral $\int_A^B \lambda^{\alpha-1} e^{-\lambda/\beta} d\lambda$ approximates the number of $\lambda$ values between any two positive numbers $A < B$. Here $Sc_{\alpha, \beta}$ is an arbitrary positive scaling constant. Formulas like the first line in (3.2) still have their correct intuitive meaning. Holgate [12] shows that there are no hidden pitfalls in dealing with improper forms for $g(\lambda)$.

We can write (2.4) in the form

$$\Delta(t) = \eta_1 \frac{\int_0^\infty e^{-\lambda} (1 - e^{-\lambda t}) dG(\lambda)}{\int_0^\infty \lambda e^{-\lambda} dG(\lambda)}.$$  

(3.4)

to avoid ambiguities in the case where $G$ is improper. Substituting (3.1) for $dG(\lambda)$ gives (in the obvious notation)
\[ \Delta_{\alpha, \gamma}(t) = \begin{cases} \frac{\eta_1}{\gamma \alpha} \left[ 1 - \frac{1}{(1+\gamma t)^{\alpha}} \right] & \alpha > 0 \\ \frac{\eta_1}{\gamma} \log(1+\gamma t) & \alpha = 0 \\ \frac{\eta_1}{-\gamma \alpha} [(1+\gamma t)^{-\alpha} - 1] & \alpha < 0 \end{cases} \]

(The cases \( \alpha \leq 0 \) are derived by first computing \( \frac{d\Delta_{\alpha, \gamma}(t)}{dt} = S \int_0^\infty \lambda e^{-\lambda (1+t)} dG(\lambda) \). Notice that \( \frac{\Delta_{\alpha, \gamma}(t/\gamma)}{\Delta_{\alpha, 1}(t)} = \text{constant} \). The parameter \( \alpha \) controls the shape of the function \( \Delta_{\alpha, \gamma}(t) \), \( \gamma \) acting only as a scale parameter.

Figure 2. \( \Delta_{\alpha, \gamma}(t) \) approaches the limit \( \frac{\eta_1}{\alpha} \) for \( \alpha > 0 \), but goes to infinity for \( \alpha \leq 0 \).
Figure 2 shows $\Delta_{\alpha, \gamma}(t)$ approaching its limiting value $\eta_1/\alpha$ as $t$ goes to infinity, for the case $\alpha > 0$. The improper cases $\alpha \leq 0$ have $\Delta_{\alpha, \gamma}(t)$ increasing without bound as $t$ increases. For $\alpha \leq 0$ 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 unbelievably good. Substituting the values $\hat{\eta}_1 = 14,381$, $\hat{\alpha} = -0.3901$, $\hat{\gamma} = 0.9875$ (which are explained below) into the last line of (3.2) gives estimates $\hat{n}_x$ remarkably close to the observed $n_x$.

\[
\begin{array}{cccccccccccc}
x & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\hat{n}_x & 14381 & 4331 & 2295 & 1479 & 1054 & 800 & 633 & 516 & 431 & 367 \\
n_x & 14376 & 4343 & 2292 & 1463 & 1043 & 837 & 638 & 519 & 430 & 364 \\
\end{array}
\]

\[
\begin{array}{cccccccccccc}
x & 11 & 12 & 13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 \\
\hat{n}_x & 316 & 276 & 244 & 217 & 194 & 175 & 159 & 145 & 132 & 122 \\
n_x & 305 & 259 & 242 & 223 & 187 & 181 & 179 & 130 & 127 & 128 \\
\end{array}
\]

To assess the accuracy of a fit like (3.6) we need a theory of errors, and for that we need both the independence assumption mentioned at the beginning of Section 2 and also F2. Consider only the first $x_0$ values of $n_x, n_1, n_2, \ldots, n_{x_0}$. Given $\sum_{x=1}^{x_0} n_x$, $(n_1, n_2, \ldots, n_{x_0})$ will
have a multinomial distribution with \( \sum_{x=1}^{x_0} n_x \) trials and vector of probabilities proportional to (3.2). That is, the probability of category \( x, x = 1, 2, \ldots, x_0 \), will equal

\[
\pi_{\alpha, \gamma}(x) \equiv c [\Gamma(x+\alpha)/(x! \Gamma(1+\alpha))] \gamma^{x-1},
\]

where \( c \) is chosen to make the probabilities sum to one.

Table 2 shows the maximum likelihood fits, obtained by iterative search for various choices of \( x_0 \). The last column is Wilks' maximum likelihood ratio statistic \([17, \text{Ch. 13}], (-2 \log \text{likelihood ratio}), \) for testing the adequacy of the two parameter model (3.7). The sample sizes are enormous, the smallest being 23,517, so under the null hypothesis (3.7) this statistic should be distributed as a \( \chi^2_{x_0-3} \) variate. We see that the fit is very good, even too good for \( x_0 \leq 15 \). With sample sizes of this magnitude, deviations of just a few percent from (3.7) would cause rejection.

<table>
<thead>
<tr>
<th>( x_0 )</th>
<th>( n_x )</th>
<th>( \hat{\alpha} )</th>
<th>( \hat{\gamma} )</th>
<th>( \hat{\beta} = \frac{\hat{\alpha}}{1-\hat{\gamma}} )</th>
<th>Goodness of Fit Statistic ( \sim \chi^2_{x_0-3} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>23517</td>
<td>-.3834</td>
<td>.9795</td>
<td>47.82</td>
<td>0.027</td>
</tr>
<tr>
<td>10</td>
<td>26305</td>
<td>-.3906</td>
<td>.9884</td>
<td>85.44</td>
<td>2.024</td>
</tr>
<tr>
<td>15</td>
<td>27521</td>
<td>-.3889</td>
<td>.9861</td>
<td>70.78</td>
<td>3.815</td>
</tr>
<tr>
<td>20</td>
<td>28266</td>
<td>-.3901</td>
<td>.9875</td>
<td>78.77</td>
<td>8.832</td>
</tr>
<tr>
<td>30</td>
<td>29147</td>
<td>-.3944</td>
<td>.9899</td>
<td>97.92</td>
<td>16.874</td>
</tr>
<tr>
<td>40</td>
<td>29660</td>
<td>-.3954</td>
<td>.9905</td>
<td>104.26</td>
<td>30.437</td>
</tr>
</tbody>
</table>
The values $\hat{a} = -0.3901$, $\hat{\gamma} = 0.9875$ used to generate (3.6) are the maximum likelihood estimates for $x_0 = 20$. The choice $\hat{\eta}_1 = 14,381$ combined with $\hat{a}$ and $\hat{\gamma}$ in the last line of (3.2) makes $\sum_{1}^{20} \hat{\eta}_x = 28,266$, the observed value, so in this sense it is also maximum likelihood. Actually the fitted numbers $\hat{\eta}_x$ for $x \leq 20$ are insensitive to the choice of $x_0$. Using $\hat{a}$, $\hat{\gamma}$, and $\hat{\eta}_1$ from $x_0 = 40$ changes the entries $\hat{\eta}_x$ in (3.6) by about one tenth of one percent.

The negative binomial model does not fit as well for large values of $x$. Using $\hat{a}$, $\hat{\gamma}$, and $\hat{\eta}_1$ fitted with $x_0 = 40$ gives small but consistent underestimates for higher values of $x$.

<table>
<thead>
<tr>
<th></th>
<th>50</th>
<th>60</th>
<th>70</th>
<th>80</th>
<th>90</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sum$</td>
<td>$\sum$</td>
<td>$\sum$</td>
<td>$\sum$</td>
<td>$\sum$</td>
<td>$\sum$</td>
</tr>
<tr>
<td>$\hat{\eta}_x$</td>
<td>314</td>
<td>216</td>
<td>155</td>
<td>115</td>
<td>88</td>
<td>69</td>
</tr>
<tr>
<td>$\eta_x$</td>
<td>331</td>
<td>227</td>
<td>169</td>
<td>122</td>
<td>101</td>
<td>78</td>
</tr>
</tbody>
</table>

The values of $\eta_x$ are declining slightly more slowly than predicted by (3.7). As we shall see, only the values of $\eta_x$ for $x \leq 20$ play a necessary role in the estimation of $\Delta(t)$, so the issue isn't crucial here. All further calculations involving Fisher's model use the fitted parameter values for $x_0 = 40$.

(3.9) \[ \hat{a} = -0.3954, \quad \hat{\gamma} = 0.9905. \]

We will use

(3.10) \[ \hat{\eta}_1 = \eta_1 = 14,376. \]
rather than the fitted value \( \hat{\eta}_1 = 14,399 \) for \( x_0 = 40 \), but in most of the calculations \( \hat{\eta}_1 \) enters as a multiplicative constant, so multiplying by \( 14,399/14,376 = 1.0016 \) converts the result. NOTE. The notation \( \hat{\eta}_x \) will continue to mean any reasonable estimate of \( \eta_x \). It will be specifically noted when these are taken to be the MLE values from (3.2), (3.9) - (3.10).

Unfortunately \( \hat{\alpha} = -0.3954 \) puts us on the dotted line going to infinity in Figure 2. The data agree very well with a model we know must ultimately fail! However, we can still use (3.5) to estimate \( \Delta(t) \) for finite values of \( t \). For \( t=1 \) we get

\[
(3.11) \quad \hat{\Delta}(1) = \Delta_{-0.3954,.9905}(1) = 11,483 .
\]

This agrees with (2.8) to within one-half of one percent.

For \( t=10, \Delta_{-0.3954,.9905}(10) = 57,704, \) 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 - 7 are devoted to finding nonparametric estimates of \( \Delta(t) \) for large \( t \), and assessing their accuracy.

It is easy to calculate the variances and covariances of \( \eta_1, \eta_2, \ldots \) under Fl and the independence assumption:

\[
(3.12) \quad \text{Var}_{\alpha,\gamma} \eta_x = \eta_x (1-c_{\alpha,x}) , \quad \text{Cov}_{\alpha,\gamma} (\eta_x, \eta_y) = -\sqrt{\eta_x \eta_y} \ c_{xy}
\]

where
\[
\begin{align*}
(3.13) \quad c_{xy} &= \frac{\Gamma(x+y+\alpha)}{\Gamma(x+\alpha) \Gamma(y+\alpha) x! y!} \left( \frac{\gamma}{1+\gamma} \right)^{\frac{x+y}{2}} \left( \frac{1}{1+\gamma} \right)^{\frac{x+y}{2} + \alpha}.
\end{align*}
\]

These results follow by writing \( n_x = \sum_{s=1}^{\infty} I_x(s) \), where \( I_x(s) \) equals 1 or 0 as \( x_s \) does or does not equal \( x \). The independence assumption says that the \( I_x(s) \) variates are independent of each other for different choices of \( s \). Since \( I_x(s) \) has variance \( \left[ e^{-\lambda_s} \lambda_s^x / x! \right] [1 - e^{-\lambda_s} \lambda_s^x / x!] \) we have

\[
(3.14) \quad \text{Var}_{\alpha, \gamma} n_x = \sum_{s=1}^{\infty} \text{Var}_{\alpha, \gamma} I_x(s)
\]

\[
= sc_{\alpha, \beta} \int_0^\infty \frac{e^{-\lambda x / x!}}{(1 - e^{-\lambda x / x!})^{\gamma-1} \lambda^{\gamma-1} e^{-\lambda / \beta}} d\lambda
\]

which reduces to (3.12) – (3.13). The covariance is calculated in a similar way. For \( \hat{\alpha}, \hat{\gamma} \) as in (3.9) some values of \( c_{xy} \) are

\[
\begin{array}{cccccccccc}
   & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\
\hline
x & & & & & & & & & & \\
\hline
1 & .80 & .83 & .85 & .87 & .88 & .89 & .90 & .90 & .91 & .91 \\
2 & .88 & .90 & .90 & .90 & .91 & .91 & .91 & 1 & & \\
3 & .93 & .93 & .93 & .93 & .93 & .93 & .93 & .93 & .93 & .93 \\
4 & .96 & .96 & .96 & .96 & .96 & .96 & .96 & .96 & .96 & .96 \\
5 & .97 & .97 & .97 & .97 & .97 & .97 & .97 & .97 & .97 & .97 \\
6 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 \\
7 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 \\
8 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 \\
9 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 \\
10 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 & .98 \\
\hline
\end{array}
\]

NOTE. If we add assumption F2 to F1 and the independence assumption, and if we assume the parameter values (3.9), then (2.9) is satisfied. (This is the limiting case of multinomial sampling with infinite sample size and one missing category, \( x=0 \), having probability approaching 1.) In this case \( \text{Var}_{\alpha, \gamma} n_x = \eta_x \), a value always greater
than that given in (3.12). The reason for the discrepancy is simple; the calculation for (3.12) assumes that the \( \lambda \)'s have fixed values empirically distributed as in (3.1). \( \Pi_2 \) chooses the \( \lambda \)'s randomly from (3.1), thus adding variance to the \( n_x \). This phenomenon may be related to the overfitting observed in Table 2.

4. Euler's Transformation

Euler's transformation [2] is a method of forcing oscillating series like (2.6) to converge rapidly. The substitution

\[
(4.1) \quad t = \frac{u}{2-u}, \quad 0 \leq u < 2 = \frac{u}{2}(1 + \frac{u}{2} + (\frac{u}{2})^2 + \ldots)
\]

gives the formal relationship

\[
(4.2) \quad \sum_{x=1}^{\infty} (-1)^{x+1} \eta_x t^x = \sum_{y=1}^{\infty} \xi_y u^y
\]

where

\[
(4.3) \quad \xi_y = \frac{\xi_y}{\xi_{y-1}} \frac{\eta_{y-1}}{2^y} \quad \eta_x = \frac{1}{2^y} \delta^y[\eta_1].
\]

The notation "\( \delta \)" indicates the backward difference operator, \( \delta^0[\eta_1] = \eta_1 \), \( \delta^1[\eta_1] = \eta_1 - \eta_2 \), \( \delta^2[\eta_1] = \eta_1 - 2\eta_2 + \eta_3 \), etc.

Let

\[
(4.4) \quad \Delta^0(t) = \sum_{x=1}^{x_0} (-1)^{x+1} \eta_x t^x, \quad \Delta^0(u) = \sum_{y=1}^{y_0} \xi_y u^y,
\]
and \( \Delta(t) \equiv \lim_{x_0 \to \infty} \Delta_0^x(t) \), \( \Delta(u) \equiv \lim_{x_0 \to \infty} \Delta_0^x(u) \). By definition \( \Delta(t) = \Delta(u) \) if both limits exist. For \( \eta_x \) positive, as here, the partial sums \( \Delta_0^x(u) \) will usually converge more quickly to the common limit than the sums \( \Delta_0^x(t) \). It may even happen that the right side of (4.2) exists while the left side doesn't.

Consider \( \Delta_{\alpha, \gamma}(t) \) as given in (3.5). From (2.6) and (3.2) we obtain

\[
(4.5) \quad \Delta_{\alpha, \gamma}(t) = \eta_0 t \sum_{x=1}^{\infty} (-1)^{x+1} \frac{\Gamma(x+\alpha)}{x! \Gamma(1+\alpha)} (\gamma t)^{x-1}
\]

which diverges for \( \gamma t > 1 \). For \( \alpha \leq 1 \) the series \( \sum_{y=1}^{\infty} \xi_y u^y \) converges in the nicest possible way, having in fact all non-negative terms:

**Lemma.** For \(-1 < \alpha \leq 1\), \( \Delta_{\alpha, \gamma}(u) = \sum_{y=1}^{\infty} \xi_y u^y \) has \( \xi_y \geq 0 \) for all \( y \).

**Proof.** From (3.2) we have

\[
(4.6) \quad \eta_x(\alpha, \gamma) = c \eta_x(\alpha, 1) \gamma^x, \quad c \equiv \text{constant not depending on } x,
\]

again using obvious notation to indicate the dependence of \( \eta_x \) on \( \alpha \) and \( \gamma \). (Formally \( c = \eta_1(\alpha, \gamma)/(\eta_1(\alpha, 1) \cdot \gamma) \), where \( \eta_1(\alpha, 1) \) is an arbitrary scaling factor.) An induction argument gives

\[
(4.7) \quad \delta^k[\eta_x(\alpha, 1)] = \eta_x(\alpha, 1)(-1)^{k} \frac{(\alpha-1)(\alpha-2) \ldots (\alpha-k)}{(x+1)(x+2) \ldots (x+k)}
\]

and

\[
(4.8) \quad \delta^k[\gamma^x] = (1-\gamma)^k \gamma^x.
\]
For $\alpha \leq 1$ both (4.7) and (4.8) are non-negative for all $x \geq 1$ and $k \geq 0$. The difference calculus then says that $\delta^k[\eta_x(\alpha, \gamma)] = \delta^k[c \eta_x(\alpha, 1) \cdot \gamma^x]$ is always non-negative, implying the Lemma by the last line of (4.3).

Good and Toulmin [8] suggest estimating $\Delta(t)$ by

$$
(4.9) \quad \hat{\Delta}^X_0(u) = \sum_{y=1}^{x_0} \hat{\xi}_y u^y, \quad u = \frac{2t}{1+t}
$$

where

$$
(4.10) \quad \hat{\xi}_y = \sum_{x=1}^{y} \frac{(y-1)}{2^y} \frac{(-1)^{x+1}}{x-1} \hat{\eta}_x.
$$

The $\hat{\eta}_x$ in (4.10) might be taken to be the nonparametric estimators $\eta_x$, the maximum likelihood estimates from a parametric model like (3.1), or some other smoothed version of the $\eta_x$. Smoothing will not be an important issue here since the Shakespeare data is already so close to a smooth parametric model.

Table 3 shows the first 20 values of $\hat{\xi}_y$ computed for the $\eta_x$ and also from the maximum likelihood values (3.9) - (3.10). The latter are all positive, in accordance with the Lemma. The former are positive for $y = 1, 2, \ldots, 9$, and negative for $y = 10, 11, \ldots, 20$. However, all of the negative values are within one-half standard deviation of zero (computed either from (3.12) - (3.13) or from (2.9)).

Table 3 suggests not taking $x_0$ greater than 9 if we intend to compute $\hat{\Delta}^X_0(u)$ from $\hat{\eta}_x = \eta_x$. 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 Section 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 (5.4a).

<table>
<thead>
<tr>
<th>( y )</th>
<th>( \hat{\xi}_y ) from ( \hat{n}_x = n_x ) (Standard error)</th>
<th>( \hat{\xi}_y ) from MLE values of ( \hat{n}_x ) (3.9)-(3.10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7188.0 (60)</td>
<td>7188.0</td>
</tr>
<tr>
<td>2</td>
<td>2508.3 (34)</td>
<td>2511.7</td>
</tr>
<tr>
<td>3</td>
<td>997.8 (23)</td>
<td>1001.5</td>
</tr>
<tr>
<td>4</td>
<td>422.5 (17)</td>
<td>424.5</td>
</tr>
<tr>
<td>5</td>
<td>185.8 (14)</td>
<td>186.4</td>
</tr>
<tr>
<td>6</td>
<td>83.3 (11)</td>
<td>83.7</td>
</tr>
<tr>
<td>7</td>
<td>36.7 (10)</td>
<td>38.23</td>
</tr>
<tr>
<td>8</td>
<td>14.8 (8)</td>
<td>17.66</td>
</tr>
<tr>
<td>9</td>
<td>4.2 (8)</td>
<td>8.23</td>
</tr>
<tr>
<td>10</td>
<td>-0.7 (7)</td>
<td>3.865</td>
</tr>
<tr>
<td>11</td>
<td>-2.8 (6)</td>
<td>1.825</td>
</tr>
<tr>
<td>12</td>
<td>-3.4 (6)</td>
<td>0.866</td>
</tr>
<tr>
<td>13</td>
<td>-3.2 (5)</td>
<td>0.413</td>
</tr>
<tr>
<td>14</td>
<td>-2.8 (5)</td>
<td>0.1974</td>
</tr>
<tr>
<td>15</td>
<td>-2.3 (5)</td>
<td>0.0947</td>
</tr>
<tr>
<td>16</td>
<td>-1.8 (4)</td>
<td>0.0455</td>
</tr>
<tr>
<td>17</td>
<td>-1.3 (4)</td>
<td>0.0220</td>
</tr>
<tr>
<td>18</td>
<td>-0.9 (4)</td>
<td>0.01061</td>
</tr>
<tr>
<td>19</td>
<td>-0.6 (4)</td>
<td>0.00513</td>
</tr>
<tr>
<td>20</td>
<td>-0.4 (3)</td>
<td>0.00249</td>
</tr>
</tbody>
</table>

Table 3. Estimated Euler coefficients (4.10) for the Shakespeare data.
Even if we don't believe the negative binomial model, and ultimately we can't, there is good reason to prefer the second column of Table 3 for use in (4.9) - (4.10). This is in the spirit of the James-Stein estimator of several parameters [3]. We are "shrinking" the vector of unbiased estimates \( \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{x_0} \) toward an a priori reasonable guess, the negative binomial values, and can expect to do better than the unbiased estimates in terms of squared error risk if the guess is a good one. In this case the guess appears to be very good, and we shrink all the way to the vector of negative binomial MLE values. For this data set the choice between the two columns of Table 3 makes little practical difference. Calculating \( \hat{\Delta}^9(1) \) from the second column of Table 3 gives

\[(4.12) \quad \hat{\Delta}^9(1) = 11,460\]

as the estimate. The question of assigning a standard deviation to (4.12) is a difficult one, but it is reasonable to say that it is at least as accurate as (4.11), and perhaps considerably more so.

In the present notation, (2.8) - (2.10) can be written as \( \hat{\Delta}^\infty(1) = 11,430 \pm 178 \). Comparing this with (4.11) shows that we have reduced the standard deviation considerably by reducing \( x_0 \) from \( \infty \) to 9. The price we pay, as Good and Toulmin noted, is in terms of bias. \( \hat{\Delta}^{x_0}(t) \) is not an unbiased estimate of \( \Delta(t) \) for \( x_0 < \infty \) because of the truncated terms in the series. The calculations of Section 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 \( \lambda \).
Under the negative binomial model, the Lemma tells us 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 Section 5 will show.

5. General Linear Estimators

There is another expression of the Euler transformation which makes obvious its effect on oscillating series. Substituting (4.3) into the right side of (4.4) gives, after some rearrangement,

\begin{equation}
\Delta^x(0) = \sum_{x=1}^{x_0} (-1)^{x+1} \eta_x t^x \cdot P\{\text{Binomial}(x_0, \frac{1}{1+t}) > x\}.
\end{equation}

That is

\begin{equation}
\Delta^x(0) = E\Delta^x(t), \quad x \sim \text{Binomial}(x_0, \frac{1}{1+t}).
\end{equation}

The Euler transformation $\Delta^x(0)$ is just the average of the oscillating series $\Delta^x(t)$ over values of $x$ binomially distributed with mean $x_0/(1+t)$ and variance $x_0 t/(1+t)^2$. This averaging process is what smooths out the oscillations.

The estimator (4.9), with $\hat{\eta}_x = n_x$, is now seen to be of the form

\begin{equation}
\hat{\Delta} = \sum_{x=1}^{\infty} h_x n_x.
\end{equation}
where \( h_x = (-1)^{x+1} t^x \) \( \Pr(\text{Binomial}(x_0, 1/(1+t)) \geq x) \) for \( x = 1, 2, \ldots, x_0 \) and \( h_x = 0 \) for \( x > x_0 \). For \( x_0 = 9 \), \( t = 10 \) the \( h_x \) are

\[
\begin{array}{cccccccccc}
\hline
x & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\
h_x & 5.759 & -19.421 & 41.539 & -59.152 & 57.155 & -37.190 & 15.653 & -3.859 & 0.424 \\
\hline
\end{array}
\]

Notice that the naive estimator \( \hat{\Delta}(t) = \sum_{x=1}^{9} (-1)^{x+1} h_x t^x \) has \( h_9 = (-1)^{x+1} 10^x \) in the case, so that \( h_9 = 10^9 \), compared with \( h_9 = 0.430 \) in (5.4)!

The Euler transformation drastically reduces \( h_x \) for large \( x \).

We call estimators of the form (5.3) \textit{general linear estimators}. We will calculate the variance of such an estimator from (2.9),

\[
(5.4a) \quad \text{Var} \hat{\Delta} = \sum_{x=1}^{\infty} h_x^2 \eta_x,
\]

realizing that this value may be somewhat large, as (3.12) shows.

For each estimator (5.3) define the function

\[
(5.5) \quad H(\lambda) = \sum_{x=1}^{\infty} h_x \lambda^x / x!, \quad 0 < \lambda < \infty.
\]

By (2.3) we have

\[
(5.6) \quad \text{E} \hat{\Delta} = \sum_{x=1}^{\infty} h_x \eta_x = S \sum_{x=1}^{\infty} \int_{0}^{\infty} h_x e^{-\lambda} \lambda^x / x! \, dG(\lambda)
\]

\[
= S \int_{0}^{\infty} e^{-\lambda} H(\lambda) \, dG(\lambda)
\]

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 (2.4),
(5.7) \[ E\{\Delta - \Delta(t)\} = S \int_0^\infty e^{-\lambda} [H(\lambda) - (1-e^{-\lambda t})] \, dG(\lambda) \]

which, for \( t=\infty \), becomes

(5.8) \[ E\{\Delta - \Delta(\infty)\} = S \int_0^\infty e^{-\lambda} [H(\lambda) - 1] \, dG(\lambda) \]

It is convenient to rewrite (5.7) and (5.8) in a form which depends on

(5.9) \[ \eta_+ \equiv \sum_{x=1}^\infty \eta_x \]

rather than \( S \), since we always have an easy estimate for \( \eta_+ \) available, namely \( \eta_+ \equiv \sum_{x=1}^\infty \eta_x \). Define

(5.10) \[ P \equiv \int_0^\infty (1-e^{-\lambda}) \, dG(\lambda) \]

and

(5.11) \[ d\tilde{G}(\lambda) \equiv \frac{1-e^{-\lambda}}{P} \, dG(\lambda) \]

Notice that \( \eta_+ = SP \) by summation of \( \eta_x \) in (2.3), or by direct interpretation of the integral (5.10). That is, \( P \) is just the expected proportion of the \( \lambda_s \) having \( x_s > 0 \). \( \tilde{G} \) can be thought of as the empirical c.d.f. of those \( \lambda_s \) having \( x_s > 0 \), though strictly speaking this interpretation is only justified in the limiting case \( S \to \infty \).

Multiplying and dividing (5.7) - (5.8) by \((1-e^{-\lambda})/P \) gives
\[ E(\hat{\Delta}(\lambda)) = \eta_+ \int_0^\infty \frac{e^{-\lambda}}{1-e^{-\lambda}} [H(\lambda) - (1-e^{-\lambda t})] \, d\tilde{G}(\lambda) \]

(5.12)

\[ E(\hat{\Delta}(\infty)) = \eta_+ \int_0^\infty \frac{e^{-\lambda}}{1-e^{-\lambda}} [H(\lambda) - 1] \, d\tilde{G}(\lambda) . \]

The integrand

(5.13) \[ B_t(\lambda) \equiv \frac{e^{-\lambda}}{1-e^{-\lambda}} [H(\lambda) - (1-e^{-\lambda t})] \]

determines the bias of \( \hat{\Delta} \) for any \( G(\lambda) \) (or \( \tilde{G}(\lambda) \)). For example, with \( t=1 \), \( x_0 = 9 \) we compute \( B_t(\lambda) \) to be

(5.14)

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>0</th>
<th>.1</th>
<th>.2</th>
<th>.3</th>
<th>.4</th>
<th>.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_t(\lambda) )</td>
<td>-.00196</td>
<td>-.00108</td>
<td>-.00047</td>
<td>-.00010</td>
<td>-.00012</td>
<td>-.00021</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>.6</th>
<th>.7</th>
<th>.8</th>
<th>.9</th>
<th>1.0</th>
<th>1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>( B_t(\lambda) )</td>
<td>.00024</td>
<td>.00023</td>
<td>.00018</td>
<td>.00013</td>
<td>.00007</td>
<td>.00003</td>
</tr>
</tbody>
</table>

\( B_t(\lambda) \) oscillates more slowly and more closely about zero as \( \lambda \) increases toward infinity.

From (5.12) and (5.14) we see that the greatest negative bias occurs if \( \tilde{G} \) puts all its mass at \( \lambda = 0 \), in which case the bias equals \( -.00196 \cdot \eta_+ \approx -.00196 \cdot 31,534 = -62 \). The greatest positive bias occurs if \( \tilde{G} \) puts all mass at \( \lambda = 0.6 \), in which case it equals \( .00024 \cdot 31,534 = 8 \). Of course, the data in Table 1 tell us that \( \tilde{G} \) follows neither extreme for the Shakespeare word counts. In Sections 6 and 7 we employ such information to get better bounds in a systematic way.
Figure 3 graphs $B_t(\lambda)$ 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 $-4.24 \cdot \eta_4$ to $0.31 \cdot \eta_4$. For $x_0 = 19$ the range is from $-1.64 \cdot \eta_4$ to $0.15 \cdot \eta_4$. This doesn't mean that $x_0 = 19$ is better than $x_0 = 9$. The respective estimators (4.9) and their standard deviations from (5.4a) are $\hat{\Delta}^9(10) = 45,188 \pm 3,994$, $\hat{\Delta}^{19}(10) = 53,867 \pm 702,566$. 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 needn't restrict attention to linear estimators of the form (4.9). An attempt to choose a "best" linear estimator $\hat{\Delta}(10)$

$$\hat{\Delta}(10) = \sum_{x=1}^{x_0} h_x n_x$$

was made along the following lines: various choices of $x_0$ and $G^0$, a preferred guess of the true $G$, were selected. ($G^0$ was chosen from the gamma family.) Then $h_1, h_2, \ldots, h_{x_0}$ were selected to minimize a weighted sum of variance and bias squared, say $c_0 \int_0^\infty B_{10}(\lambda) dG^0(\lambda) + c_1 \sum_{x=1}^{x_0} h_x^2 n_x^0$. Standard matrix methods give the optimum choice. This search yielded no estimator noticeably superior to $\hat{\Delta}^9(10)$. Again it must be admitted that the criteria were vague and the search non-exhaustive. Nonetheless, it was encouraging how well $\hat{\Delta}^9(10)$ stood up to competition. We return to this point in Sections 6 and 7, where methods based on the Euler transformation will be seen to give answers as good as those derived from more general (and more difficult) approaches.
Figure 3. The bias function $B_t(\lambda)$, equation (5.13), for $\Delta^x_0(t)$, at $t=10$, $x_0 = 9$ and $x_0 = 19$.

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 Shakespeare actually had one million word types with $\lambda_s = 1/1,000,000$. These one million types would have almost no effect on our data set. The expected number of
them occurring in our sample is only $1$. However, for $t = 1,000,000$

an expected fraction $1 - e^{-1} = .632$ of them would be observed. This

type of counterexample can be pushed arbitrarily far. Unfortunately,
the trouble begins for $t$ values much smaller than $1,000,000$. We
see this in Figure 3, where the possible negative bias is already
very large for $t = 10$. The situation is better for lower bounds.

This section and the one which follows are mainly concerned with
obtaining dependable underestimate of $\Delta(t)$, thus providing a con-

servative guess of how many new species will appear in the next $t$
time intervals. (Some upper bounds are given in Table 5.)

Equations (5.7) - (5.8) show that $\hat{\Delta} = \sum_{x=1}^{\infty} h_x n_x$ satisfies

\[ E\hat{\Delta} \leq \Delta(t) \text{ if } H(\lambda) \leq 1 - e^{-\lambda t} \text{ for all } \lambda \geq 0 \]

(6.1)

\[ E\hat{\Delta} \leq \Delta(\infty) \text{ if } H(\lambda) \leq 1 \text{ for all } \lambda \geq 0 \]

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 Section 5, the estimators based on Euler's transforma-
tion don't satisfy (6.1). However, given $\hat{\Delta} = \sum_{x=1}^{\infty} h_x n_x$ we can always

make a linear transformation

\[ h_x^0 = ah_x - b, \quad x = 1, 2, \ldots \]

(6.2)

which gives, through (5.4)

\[ H^0(\lambda) = ah(\lambda) - b(e^\lambda - 1) \]

(6.3)
The corresponding new estimator is

\[(6.4) \quad \hat{\Delta}^0 = \sum_{x=1}^{\infty} h^0_x n_x = a\Delta - b n_+ \]

where \(n_+ = \sum_{x=1}^{\infty} n_x\) as before.

Table 4 shows the lower bounds obtained in this way from the Euler estimators (5.3), with \(x_0 = 9\), for various choices of \(t\). The constants \(a\) and \(b\) in (6.3) were chosen so that \(H^0\) satisfied the top line of (6.1). Subject to this constraint, \(a\) and \(b\) were selected to maximize (6.4) with \(\hat{\eta}_x\) in place of \(n_x\), \(\hat{\eta}_x\) the maximum likelihood estimates obtained from (3.2) and (3.9) – (3.10).

The resulting value of \(\hat{\Delta}^0\) is tabulated as the "Lower Bound Estimate". The standard deviation from (5.4a), \(\left(\sum_{x=1}^{\infty} (h^0_x)^2 \hat{\eta}_x^2\right)^{1/2}\) appears in the next column, followed by the estimate minus one standard deviation. Of course, we could subtract two or more standard deviations to get an even more conservative lower bound, but as explained in the paragraph following (4.11) there is reason to believe that \(\Sigma h^0_x \hat{\eta}_x\) is a more accurate estimator of \(\Sigma h^0_x n_x\) than (2.9) indicates.

Looking at the last column of Table 4 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 Section 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, unaidered by parametric assumptions like PI, runs out of predicative power for \(t\) greater than 10.
<table>
<thead>
<tr>
<th>t</th>
<th>a</th>
<th>b</th>
<th>Lower Bound Estimate (6.4)</th>
<th>Standard Deviation (5.4a)</th>
<th>Estimate -S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.999</td>
<td>.0001</td>
<td>11454</td>
<td>147</td>
<td>11307</td>
</tr>
<tr>
<td>3</td>
<td>.979</td>
<td>.002</td>
<td>25143</td>
<td>986</td>
<td>24157</td>
</tr>
<tr>
<td>5</td>
<td>.939</td>
<td>.007</td>
<td>31974</td>
<td>1966</td>
<td>30008</td>
</tr>
<tr>
<td>8</td>
<td>.879</td>
<td>.010</td>
<td>36554</td>
<td>2965</td>
<td>33588</td>
</tr>
<tr>
<td>10</td>
<td>.850</td>
<td>.012</td>
<td>38015</td>
<td>3397</td>
<td>34618</td>
</tr>
<tr>
<td>12</td>
<td>.827</td>
<td>.014</td>
<td>38927</td>
<td>3713</td>
<td>35214</td>
</tr>
<tr>
<td>15</td>
<td>.801</td>
<td>.016</td>
<td>39784</td>
<td>4048</td>
<td>35736</td>
</tr>
<tr>
<td>20</td>
<td>.772</td>
<td>.017</td>
<td>40580</td>
<td>4408</td>
<td>36172</td>
</tr>
<tr>
<td>30</td>
<td>.742</td>
<td>.019</td>
<td>41331</td>
<td>4793</td>
<td>36538</td>
</tr>
<tr>
<td>60</td>
<td>.710</td>
<td>.022</td>
<td>42061</td>
<td>5212</td>
<td>36848</td>
</tr>
<tr>
<td>120</td>
<td>.694</td>
<td>.023</td>
<td>42411</td>
<td>5433</td>
<td>36977</td>
</tr>
</tbody>
</table>

Table 4. Lower bound estimates for $\Delta(t)$ based on linear transformations of $\hat{\Delta}^X_0(u)$, $x_0 = 9$.

A small point deserves mention here: in substituting $\hat{n}_x$ for $n_x$ in (6.4), the question arises of what to do for the larger values of $x$, remembering that the negative binomial model falls off too quickly in the upper tail. As a matter of fact, one calculates from (3.9) - (3.10) and (3.2) that $\sum_{x=10}^{\infty} \hat{n}_x = 4654$ while the data give

(6.5)  $\sum_{x=10}^{\infty} n_x = 5593$.

The estimates in Table 4 were actually calculated from (6.4) using $\hat{n}_x$ in place of $n_x$ for $x = 1, 2, \ldots, 9$, but with $n_x$ equal to their observed values for $x \geq 10$ (so (6.5) was satisfied.) Fortunately, this makes almost no difference in the estimates in this
case (nor does using (6.4) as given forgetting about the $\hat{\eta}_x$ altogether). For smaller data sets though, the Stein-type estimation involving $\hat{\eta}_x$ may be quite important.

A potentially disturbing flaw in Table 4 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 (6.4) is maximized for the data set at hand.) This point is considered more carefully in Section 7, and is shown not to make much difference.

7. Linear Programming Bounds

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

1. Choose $h_1, h_2, \ldots, h_{x_0}, h_{x_0+1}$ to maximize

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

subject to the constraints

$$H(\lambda) \equiv \sum_{x=1}^{x_0} h_x \lambda^x / x! + h_{x_0+1} \sum_{x=x_0+1}^{\infty} \lambda^x / x!$$

$$\leq 1 - e^{-\lambda t} \text{ for } \lambda > 0$$

Condition (7.2) guarantees, by (6.1), that $E \left\{ \sum_{x=1}^{x_0} h_x \eta_x + h_{x_0+1} \sum_{x=x_0+1}^{\infty} \eta_x \right\} \leq \Delta(t)$ no matter what $G$ is. Subject to this constraint, (7.1) says to maximize 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 MLE.
values (3.2), (3.9) – (3.10) for \( x = 1, 2, \ldots, x_0 \), and set \( \hat{\eta}_x \)
equal to \( \sum_{x=x_0+1}^{\infty} \eta_x \) as at (6.5). (Once again other reasonable choices of \( \hat{\eta}_x \) give almost identical answers.)

Pl was solved on Stanford's 360/67 computer using the IBM program MPS/360. The infinite number of constraints in (7.2) was replaced by the discrete set

\[
H(\lambda, t) \leq 1 - e^{-\lambda t}, \quad \lambda = 2^\ell / 16 - 10, \quad \ell = 0, 1, \ldots, 272
\]

\( (\lambda_0 = 2^{-10}, \lambda_{272} = 128) \). 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 found to be

\[
\begin{array}{cccccc}
x & 1 & 2 & 3 & 4 & 5 \\
h_x & 13.342 & -115.45 & 552.78 & -1673.78 & 3330.99 \\
\end{array}
\]

\[
(7.4)
\]

\[
\begin{array}{cccccc}
x & 6 & 7 & 8 & 9 & 10 \\
h_x & -4372.37 & 3650.72 & -1757.76 & 369.91 & -0.06637 \\
\end{array}
\]

Substituting into (7.1) gives the lower bound estimate

\[
(7.5) \quad \hat{\Delta}(\infty) = 59,568
\]

for Shakespeare's total unobserved vocabulary. Unfortunately, the standard error for (7.5), calculated from (5.4a) pretending that the \( h_x \) are fixed constants, is a whopping 204,784. This would seem to render (7.5) useless, but we shall see that this is not actually the case.
The dual linear programming problem [11] to P1 (more exactly, the
dual to the discretized version (7.1), (7.3)) is

P2. Choose \( S > 0 \) and a discrete c.d.f. \( G(\lambda) \) with support on
the set \( \{ \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_L \} \) to minimize

\[
\Delta(t) = S \int_0^\infty e^{-\lambda} (1 - e^{-\lambda_t}) \, dG(\lambda)
\]

subject to the constraints

\[
S \int_0^\infty e^{-\lambda} \frac{\lambda^x}{x!} \, dG(\lambda) = \hat{n}_x, \quad x = 1, 2, \ldots, x_0 \quad \text{and}
\]

\[
S \int_0^\infty e^{-\lambda} \sum_{x=x_0+1}^\infty \frac{\lambda^x}{x!} \, dG(\lambda) = \sum_{x=x_0+1}^\infty \hat{n}_x.
\]

The dual program P2 finds the "least favorable situation" in that
it selects \( S \) and \( G \) to minimize \( \Delta(t) \) subject to the expected word
bounds \( \hat{n}_1, \hat{n}_2, \ldots, \hat{n}_{x_0} \) and \( \sum_{x=x_0+1}^\infty \hat{n}_x \) equaling certain specified
values. (P2 is nearly identical to the problem considered in Harris
[10].) With the \( \hat{n}_x \) chosen as before we get, by the duality theorem,
that P2 has the same solution as P1. For \( t=\infty \) solving P2 gives
\( \hat{\Delta}(\infty) = 59,568 \) as at (7.5). The minimizing distribution \( G \) has its
support at 10 of the \( \lambda_j \) values, occurring in five adjacent pairs.

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>120-121</th>
<th>167-168</th>
<th>191-192</th>
<th>208-209</th>
<th>226-227</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda )</td>
<td>1.1806</td>
<td>1.384</td>
<td>3.914</td>
<td>8.175</td>
<td>17.830</td>
</tr>
<tr>
<td>( dG )</td>
<td>.7444</td>
<td>.1220</td>
<td>.0507</td>
<td>.0294</td>
<td>.0535</td>
</tr>
</tbody>
</table>
The minimizing value of \( S \) equals \( \hat{\Delta}^{(\infty)} + \sum_{x=1}^{\infty} \hat{\eta}_x = 59,568 + 31,523 = 91,091 \). (For \( t=\infty \), adding (7.6) and (7.7) gives \( S = \hat{\Delta}^{(\infty)} + \sum_{x=1}^{\infty} \hat{\eta}_x \), so that the estimated total vocabulary equals the observed plus the estimated unobserved vocabulary.)

If we believed that the true \( \eta_x \) were exactly as given in the right side of (7.7) then we would have to accept 59,568 as a lower bound on \( \Delta^{(\infty)} \), the amount of unobserved vocabulary. This is the smallest value consistent with (7.7). Of course we don't believe \( \eta_x = \hat{\eta}_x \) exactly, but we can loosen the constraints to take into account our uncertainty, say

\[
\hat{\eta}_x - c \sqrt{\hat{\eta}_x} \leq S \int_0^\infty e^{-\lambda} \lambda^x/x! \, dG(\lambda) \leq \hat{\eta}_x + c \sqrt{\hat{\eta}_x}, \quad x = 1, 2, \ldots, x_0
\]

(7.9) and

\[
\sum_{x=x_0+1}^{\infty} \hat{\eta}_x - c \sqrt{\sum_{x=x_0+1}^{\infty} \hat{\eta}_x} \leq S \int_0^\infty e^{-\lambda} \sum_{x=x_0+1}^{\infty} \lambda^x/x! \, dG(\lambda)
\]

\[
\leq \sum_{x=x_0+1}^{\infty} \hat{\eta}_x + c \sqrt{\sum_{x=x_0+1}^{\infty} \hat{\eta}_x}.
\]

Here \( c \) measures how many (approximate) standard deviations we allow the fitted values of \( \eta_x \) to vary from \( \hat{\eta}_x \).

Solving (7.6), (7.9) for \( t=\infty, x_0 = 9, c=1 \) gives a minimum value of

(7.10) \( \hat{\Delta} = 35,554 \),

which is quite consistent with the last column of Table 4. The minimizing \( G \) has its support at seven \( \lambda_L \) values occurring in three pairs and at \( \lambda_L = 128 \).
\[ \hat{\lambda} \quad \begin{array}{cccc}
136-137 & 183-184 & 207-208 & 272 \\
\lambda & 0.3613 & 2.768 & 7.829 & 128 \\
dG & 0.7517 & 0.1246 & 0.0531 & 0.0707 \\
\end{array}
\] (7.11)

\[ S = 67,088 \text{ is the minimizing value of the estimated total vocabulary.} \]

The solution has the constraints (7.9) at their lower limits for \( x = 1, 3, \) and 6, and at their upper limits for \( x = 2, 4, 8, \) and 10. (Choosing \( x_0 = 7 \) gives \( \hat{\lambda} = 35,306 \), while \( x_0 = 11 \) gives \( \hat{\lambda} = 35,586 \). In the latter case \( G \) is almost identical to (7.11) except that the largest support value occurs at \( \lambda = 228-229, \lambda = 19.444 \).)

We now have a believable lower bound on \( \Delta(\infty) \). The choice \( c=1 \) may seem optimistic, but as explained before we have reason to believe the true \( \eta_x \) to be nearer \( \hat{\eta}_x \) than (2.9) indicates. The question that bothered us in Section 6, 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 (7.9) then yield a bound on \( \hat{\Delta} \) in the usual way. (For those preferring a still more conservative bound, \( c=2 \) gives \( \hat{\Delta}(\infty) = 30,845 \) with \( x_0 = 9 \).)

Table 5 gives lower and upper bounds on \( \Delta(t) \) obtained from (7.6), (7.9) with \( x_0 = 9, c=1 \). For the upper bound the "minimize" in (7.6) is simply changed to "maximize". The agreement of the lower bounds with the last column of Table 4 is remarkable. This is important since Table 4 is much easier to calculate than Table 5.
Table 5. Lower and upper bounds on $\Delta(t)$ calculated by solving the linear programming problem (7.6), (7.9); $x_0 = 9$, $c=1$.

<table>
<thead>
<tr>
<th>$t$</th>
<th>lower bound on $\Delta(t)$</th>
<th>upper bound on $\Delta(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11,205</td>
<td>11,732</td>
</tr>
<tr>
<td>3</td>
<td>23,828</td>
<td>29,411</td>
</tr>
<tr>
<td>5</td>
<td>29,898</td>
<td>45,865</td>
</tr>
<tr>
<td>10</td>
<td>34,640</td>
<td>86,600</td>
</tr>
<tr>
<td>20</td>
<td>35,530</td>
<td>167,454</td>
</tr>
<tr>
<td>$\infty$</td>
<td>35,554</td>
<td>$\infty$</td>
</tr>
</tbody>
</table>

Since Robbins' first papers empirical Bayes work has systematically avoided dealing directly with the empirical c.d.f. $G$. Clever methods which go directly from data to answer, such as (2.7), have been used to finesse the difficult problem of estimating $G$. The methods of this section are atypical in this respect: we estimate $G$ directly.

8. Summary

Figure 4 displays the different estimates of $\Delta(t)$ we have developed. To summarize our experience with the Shakespeare data,

1) $\hat{\Delta}(\infty) = 35,000$ is a reasonably conservative lower bound estimate for the amount of vocabulary Shakespeare knew but didn't use.

2) $\Delta(t)$ can be estimated very accurately for $t < 1$, 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.

3) 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 \( t > 1 \).

4) All the methods give very similar answers for \( t < 1 \).

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

---

**Figure 4.** Different estimates of \( \Delta(t) \) for the Shakespeare data. A: Fisher's negative binomial model with parameters (3.9), (3.10). B: Euler transformation (4.9), \( x_0 = 9, \hat{\beta}_y \) from first column of Table 3. C: Same, but with \( \hat{\beta}_y \) from second column of Table 3. D: Lower bound estimates from linear program (7.6), (7.9), \( c=1 \). E: Upper bound, same method.
REFERENCES


[8] Good, I. J. and Toulmin, G. H. "The number of new species, and the increase in population coverage, when a sample is increased", Biometrika, 43 (January 1956), 45-63.


