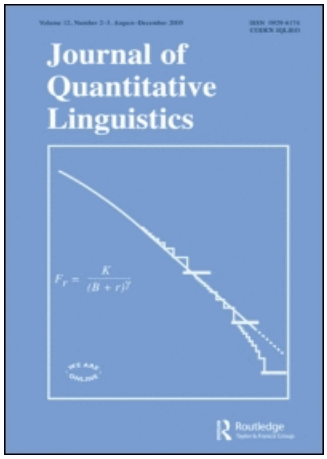

Journal of Quantitative Linguistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100702

Turing's formula revisited*

Online Publication Date: 01 August 2007
To cite this Article: Zhang, Zhiyi and Huang, Hongwei (2007) 'Turing's formula revisited*', Journal of Quantitative Linguistics, 14:2, 222 - 241
To link to this article: DOI: 10.1080/09296170701514189
URL: http://dx.doi.org/10.1080/09296170701514189

Journal of Quantitative Linguistics
2007, Volume 14, Numbers 2–3, pp. 222 – 241
DOI: 10.1080/09296170701514189

Turing's Formula Revisited*

Zhiyi Zhang and Hongwei Huang
University of North Carolina at Charlotte

ABSTRACT

A simple frequentist's justification of Turing's formula, an improvement to Turing's formula by means of reduced bias, a clarification of the relationships among various objects related to Turing's formula, a conservative confidence interval to Turing's target, and a conservative testing procedure using observed rank-frequencies under a hypothesized known infinite-dimensional multinomial distribution are given in this paper. As an example, the authorship of the nine-stanza poem "Shall I Die?" is tested against Shakespeare's canon and statistically significant evidence is found for a difference in word type usage.

INTRODUCTION

Turing's formula is an estimator proposed by Good (1953), but largely credited to Turing, to estimate "the population (relative) frequency of an arbitrary species that is represented r times in a sample". Of special interest is the case of $r = 0$, in which the estimation target becomes "the total population proportion of species that are not represented in a sample". The problem of estimating probability of a previously unobserved event may be encountered analogously in various fields such as population biology, speech recognition, risk management, etc., where quantifying the likelihoods of rare, but probable, events is essential. It is not difficult to imagine that any reasonable answer to the problem (Turing's formula being one) would be highly valuable.

There are four basic objects related to Turing's formula: the rank distribution (a random target denoted by μ_r), the expected rank distribution

*Address correspondence to: Zhiyi Zhang, Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA.
Tel: (704) 687-4549. E-mail: zzhang@uncc.edu

0929-6174/07/1402–30222$16.00

(denoted by π_r), the rank-frequency distribution (denoted by n_r), and the expected rank-frequency distribution (denoted by η_r). All these objects are defined in Section 2. The perspective implied by Turing's formula is really an intriguing one. The formula's power has long been recognized in practice. There are a large number of articles in the literature across many applied fields citing Turing's formula. However, there also seems to be a persistent lack of satisfaction among practitioners regarding Turing's formula, not in its power, but in its anti-intuitiveness. It began with the formula's introduction in a hard-to-read paper by Good (1953) before empirical Bayes was coined by Robbins (1955). It was not until 32 years later that Nádas (1985) explained that the rationale given by Good (1953) was empirical Bayes in nature. This connection helped a lot in re-reading Good's paper, but shed no light on how the formula would look like in the world of the frequentists where intuition is more easily found by many.

Ironically it was Robbins (1968), who coined empirical Bayes and gave a frequentist's interpretation to Turing's formula. Yet Robbins (1968) gave the interpretation in a twisted light of an added iid observation to the sample at hand which tilted the original platform slightly.

Our initial motivation for the study was to find a simple and direct frequentist's justification for Turing's formula. In the process, it became clear that Turing's formula utilizes only a part of the relevant information available, and that an improvement was possible.

Furthermore we realized that the observed rank-frequency distribution is useful beyond Turing's formula. In the case of a small multinomial sample (relative to the number of categories in the population) it may be utilized to test a general change in population distribution.

To summarize, the following are some of the things provided in this paper:

1. A simple frequentist's (as opposed to Bayesian) justification to Turing's formula.
2. An improvement of Turing's formula by means of reduced bias.
3. A clarification of relationships among μ_r, π_r, n_r and η_r.
4. A conservative confidence interval for the Turing target.
5. A conservative test based on n_r under the null hypothesis of a known infinite-dimensional multinomial distribution.
6. An application of the proposed test to the nine-stanza poem "Shall I Die?" (possibly Shakespearean) against Shakespeare's canon, in which significant statistical evidence is found to support a difference in word type usage.

The methodological discussion is given in the next section. The example of "Shall I Die?" appears in the following section. Several remarks are given in the final section.

TURING'S FORMULA REVISITED

Let us assume that a population of infinite elements (animals, words, etc.) can be characterized into an infinite number of categories (species, word types, etc.), C_1, $C_2, \ldots,$ each of which has an unknown population relative frequency, p_s, $s = 1, 2, \ldots, \infty, \ldots$, subjecting to $0 \leq p_s \leq 1$ for all

$$s \text{ and } \sum_{s=1}^{\infty} p_s = 1.$$

A sample of size N is taken from the population and the observed sample counts corresponding to the population categories are $y_1, y_2, \ldots$. We note that a population with finite categories, S, can be considered as a special case, which seems to be the study model in most of the publications regarding Turing's formula.

The estimation target of Turing's formula is often understood as "the total population (relative) frequency of all species not represented in the sample". The word "not" is underlined in the previous sentence because Good (1953) actually named the complementary target without the word "not". He did say that "the chance that the next animal sampled will belong to a new species" might be estimated by the probability rule of a complement. Directly translating this estimation target will give it the following expression:

$$\mu_0 = \sum_{s=1}^{\infty} p_s I[y_s = 0] \tag{1}$$

which, as Robbins (1967) pointed out, is a random variable and not a parameter in the usual statistical sense. Robbins (1967) also suggested that the estimation of Formula (1) could be profitably approached by that of its expectation

$$\pi_0 = E(\mu_0) = \sum_{s=1}^{\infty} p_s (1 - p_s)^N. \tag{2}$$

For further development, let us introduce several notations. First consider a data-based partition of the index set $D = \{s;\ s = 1, 2, \ldots\}$ into $D_0, \ldots, D_r, \ldots, D_N$ such that s is in D_r if and only if $y_s = r$, and let n_r be the number of indices in set D_r. Given the sample, n_r is the number of categories observed exactly r times. For each r, $r = 1, 2, \ldots, N$, the expected total probability of categories represented exactly r times in the sample may be expressed as

$$\pi_r = \binom{N}{r} \sum_{s=1}^{\infty} p_s^{r+1} (1 - p_s)^{N-r}. \tag{3}$$

The expected value of n_r may be expressed as

$$E(n_r) = E\left(\sum_{s=1}^{\infty} I[y_s = r] \right) = \binom{N}{r} \sum_{s=1}^{\infty} p_s^r (1 - p_s)^{N-r}. \tag{4}$$

Now let us observe a relationship between π_r and $E(n_r)$ by re-writing Formula (3):

$$\pi_r + \binom{N}{r} \binom{N}{r+1}^{-1} \pi_{r+1} = \binom{N}{r} \binom{N}{r+1}^{-1} E(n_{r+1}) \tag{5}$$

or

$$\pi_r + \frac{r+1}{N-r} \pi_{r+1} = \frac{r+1}{N-r} E(n_{r+1}). \tag{6}$$

When $r = 0$, Formula (6) becomes

$$\pi_0 + \frac{\pi_1}{N} = \frac{E(n_1)}{N}. \tag{7}$$

Formula (2.7) suggests that n_1/N is an unbiased estimate of $\pi_0 + \pi_1/N$ Since $\pi_1 \in [0,1]$, for large N,

$$\hat{\pi}_0 = \frac{n_1}{N} \tag{8}$$

provides a reasonable estimate for π_0 Formula (8) is the legendary and widely-used Turing's formula for the total probability of unseen categories. This formula was given an interpretation by Good (1953) under the finite dimensional multinomial distribution with an argument

based on a certain prior distribution (which requires S to be finite). Good's interpretation was later clarified by Nádas (1985) as a form of empirical Bayes approach. The authors of this paper would like to think that the simple expression of Formula (7) provides a direct frequentist's justification to Turing's formula. Turing's formula has a powerful implication: the probability of unobserved categories can be inferred, subject to a bias, by probabilistic characteristics of observed categories in a purely nonparametric setting.

Turing's Formula Improved

From Formula (7) it is clearly evident that Turing's formula is biased. In fact, the bias is exactly

$$\frac{\pi_1}{N} = \sum_{s=1}^{\infty} p_s^2 (1 - p_s)^{N-1}.$$

This bias can be reduced. To see this, let us re-write Formula (2) by repeatedly applying Formula (5).

$$\begin{aligned}
\pi_0 &= \binom{N}{1}^{-1} E(n_1) - \binom{N}{1}^{-1} \pi_1 \\
&= \binom{N}{1}^{-1} E(n_1) - \binom{N}{2}^{-1} E(n_2) + \binom{N}{2}^{-1} \pi_2 = \cdots \\
&= \sum_{k=1}^{r} (-1)^{k+1} \binom{N}{k}^{-1} E(n_k) + (-1)^{r+2} \binom{N}{r}^{-1} \pi_r = \cdots \\
&= \sum_{k=1}^{N} (-1)^{k+1} \binom{N}{k}^{-1} E(n_k) + (-1)^{N+2} \sum_{s=1}^{\infty} p_s^{N+1}.
\end{aligned}$$

These expressions suggest that, for any r, $r = 1, \ldots, N$,

$$\hat{\pi}_{0,r} = \sum_{k=1}^{r} (-1)^{k+1} \binom{N}{k}^{-1} n_k \tag{9}$$

may be used as an estimate for π_0 with bias

$$\binom{N}{r}^{-1} \pi_r,$$

which is easily seen to be progressively (as r increases) smaller, including the bias in $\hat{\pi}_{0,N}$ (namely $\sum_{s=1}^{\infty} p_s^{N+1}$), provided that $\max_s\{p_s\} < 1/2$. This sufficient condition for shrinking bias is really a superficial one since, in practice, if a category probability is greater than 1/2 it can be easily reduced by a refinement of the category. As a consequence, we propose

$$\hat{\pi}_{0,N} = \sum_{r=1}^{N} (-1)^{r+1} \binom{N}{r}^{-1} n_r \tag{10}$$

as an alternative to be used in place of Turing's formula.

To illustrate the amount of bias reduced from the Turing's formula in Formula (8) to $\hat{\pi}_{0,N}$ in Formula (10), their respective biases are calculated for several particular distributions and are compared in Tables 1, 2 and 3. The distributions used are

1. A uniform distribution with $S = 100$ and $p_s = 0.01$ for all $s = 1, \ldots, S$.
2. A discretized Pareto: $p_1 = p_2 = 1/3$, and $p_s = 2/[4(s-1)^2 - 1]$ for $s \geq 3$.
3. A discretized exponential: $f(x) = \exp(-x/10)/10$ and $p_s = \exp(-(s-1)/10) - \exp(-s/10)$ for $s \geq 1$.

Table 1. Bias comparison under uniform.

N	1	2	3	4	5	100	500
π_0	0.9900	0.9801	0.9703	0.9606	0.9510	0.3660	0.0066
$\hat{\pi}_0$ bias	0.0100	0.0099	0.0098	0.0097	0.0096	0.0037	6.6e-5
$\hat{\pi}_{0,N}$ bias	0.0100	0.0001	1.0e-6	1.0e-8	1.0e-10	1.0e-200	0
$\hat{\pi}_0$ bias ratio	0.0101	0.0101	0.0101	0.0101	0.0101	0.0101	0.0101
$\hat{\pi}_{0,N}$ bias ratio	0.0101	0.0001	1.0e-6	1.0e-8	1.1e-10	2.7e-200	0

Table 2. Bias comparison under discretized Pareto.

N	1	2	3	4	5	10	50
π_0	0.7548	0.5863	0.4695	0.3875	0.3293	0.2008	0.0877
$\hat{\pi}_0$ bias	0.2452	0.1685	0.1169	0.0819	0.0582	0.0142	0.0009
$\hat{\pi}_{0,N}$ bias	0.2452	0.0767	0.0250	0.0083	0.0027	1.1e-5	9.3e-25
$\hat{\pi}_0$ bias ratio	0.3248	0.2874	0.2489	0.2114	0.1769	0.0705	0.0099
$\hat{\pi}_{0,N}$ bias ratio	0.3248	0.1308	0.0533	0.0213	0.0083	5.6e-5	1.1e-23

Table 3. Bias comparison under discretized exponential.

N	1	2	3	4	5	10	50
π_0	0.9500	0.9034	0.8599	0.8191	0.7811	0.6240	0.1952
$\hat{\pi}_0$ bias	0.0450	0.0466	0.0436	0.0407	0.0381	0.0275	0.0038
$\hat{\pi}_{0,N}$ bias	0.0450	0.0033	0.0002	2.0e-5	1.6e-6	8.7e-12	8.0e-5
$\hat{\pi}_0$ bias ratio	0.0526	0.0516	0.0507	0.0497	0.0487	0.0441	0.0195
$\hat{\pi}_{0,N}$ bias ratio	0.0526	0.0037	0.0003	2.4e-5	2.1e-6	1.4e-11	4.1e-52

Other Reasonable Estimates

In passing, we state two other reasonable estimates for the mean of the total probability of the unobserved categories in a sample. Since

$$\pi_0 = E(\mu_0) = \sum_{s=1}^{\infty} p_s(1-p_s)^N$$
$$= \sum_{k=1}^{N} (-1)^{k+1} \binom{N}{k}^{-1} E(n_k) + (-1)^{n+2} \sum_{S=1}^{\infty} p_s^{N+1},$$

$$\tilde{\pi}_0 = \sum_{s=1}^{\infty} \hat{p}_s(1-\hat{p}_s)^N \tag{11}$$

and

$$\tilde{\tilde{\pi}}_0 = \sum_{k=1}^{N} (-1)^{k+1} \binom{N}{k}^{-1} n_k + (-1)^{N+2} \sum_{s=1}^{\infty} \hat{p}_s^{N+1} \tag{12}$$

where $\hat{p}_s = y_s/N$, are all intuitively reasonable.

More generally, we note that, for $r = 1, \ldots, N-1$, writing $\mu_r = \sum_{s=1}^{\infty} p_s I[y_s = r]$, since

$$\pi_r = E(\mu_r) = \binom{N}{k} \sum_{s=1}^{\infty} p_s^{r+1}(1-p_s)^{N-r}$$
$$= \sum_{k=r+1}^{N} (-1)^{k+r+1} \binom{N}{r} \binom{N}{k}^{-1} E(n_k)$$
$$+ (-1)^{N+r} \binom{N}{r} \sum_{s=1}^{\infty} p_s^{N+1}, \tag{13}$$

$$\tilde{\pi}_r = \binom{N}{r} \sum_{s=1}^{\infty} \hat{p}_s^{r+1}(1-\hat{p}_s)^{N-r}, \tag{14}$$

and

$$\tilde{\tilde{\pi}}_r = \sum_{k=r+1}^{N} (-1)^{k+r+1} \binom{N}{r} \binom{N}{k}^{-1} n_k + (-1)^{N+r} \binom{N}{r} \sum_{s=1}^{\infty} \hat{p}_s^{N+1} \qquad (15)$$

are all intuitively reasonable.

In fact, Formula (14) and Formula (15) all have their respective appeals. For example, Formula (14) always produces positive estimates; and Formula (15) attempts to further adjust for the bias of the type of estimates in Formula (10).

Relationship among Various Objects

For the clarity of presentation, we also note that there are four distinct, but related, objects involved in this paper:

1. Turing's object: $\mu = (\mu_0, \ldots, \mu_r, \ldots, \mu_N)'$, where $\mu_r = \sum_{s=1}^{\infty} p_s I[y_s = r]$. The components of Turing's object are functions of random variables and parameters, and hence are not directly observable. Given a sample, Turing's object may be reasonably referred to as the rank distribution. By the constraint, $\sum_{r=0}^{N} \mu_r = 1$, Turing's object is N-dimensional.
2. The expected rank distribution: $\pi = E(\mu) = (\pi_0, \ldots, \pi_N)'$ also N -dimensional.
3. The rank-frequency distribution: $n = (n_1, \ldots, n_N)'$ an $(N-1)$-dimensional object by the constraint, $\sum_{r=1}^{N} r n_r = N$.
4. The expected rank-frequency distribution: $\eta = (\eta_1, \ldots, \eta_N)' = (E(n_1), \ldots, E(n_N))'$, an $(N-1)$-dimensional object. For any $r = 1, \ldots, N$, n_r is an unbiased estimate of $E(n_r)$.

In the existing literature $n = (n_1, \ldots, n_N)'$ is sometimes referred to as the rank-frequency distribution, although there exists a certain amount of misunderstanding of the object (see Adamic & Huberman, 2002). n is also the subject of Zipf's Law (see Zipf, 1932, 1949). Turing's formula clearly is an attempt to link n to μ, and therefore to π Many subsequent works in quantitative linguistics, e.g. Efron and Thisted (1976), Thisted and Efron (1987), and Gale and Sampson (1995), modelled their study problems based on n_r's.

The relationship among these four objects may be summarized as follows.

1. $\pi_r = E(\mu_r)$ for $r = 0, \ldots, \mathrm{N}$; and

2. by (2.6),

$$\eta_- = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_{r+1} \\ \vdots \\ \eta_{N-1} \end{pmatrix} = \begin{pmatrix} N & 1 & & & & & \\ & \frac{N-1}{2} & 1 & & & & \\ & & \ddots & \ddots & & & \\ & & & \frac{N-r}{r+1} & 1 & & \\ & & & & \ddots & \ddots & \\ & & & & & \frac{2}{N-1} & 1 \end{pmatrix} \times \begin{pmatrix} \pi_0 \\ \pi_1 \\ \vdots \\ \pi_{N-1} \end{pmatrix} = A\pi_- \tag{16}$$

where the $(N-1) \times N$ matrix A has non-zero elements on the main diagonals as shown above, and zero elsewhere.

The above expressions clarifies how η and, π and hence μ, are related.

Specifically, an unbiased linear estimate in n_r of π_0 does not exist. However, Formula (16) does reveal that η_r reflexes the local characteristics of π at the corresponding r since η_r is a linear combination of two neighbouring π_r's.

A Confidence Interval for π_0

The covariance matrix of the vector, $n_- = (n_1, \ldots, n_r, \ldots, n_{N-1})'$, denoted by $\Sigma = (\sigma_{i,j})_{(N-1)\times(N-1)}$, can be expressed in terms of p, $s = 1, \ldots,$ i.e. for any r, r_1, $r_2 = 1, \ldots, N-1$, and $r_1, \neq r_2$,

$$\begin{aligned} \sigma_{r,r} &= Var(n_r) = E(n_r^2) - [E(n_r)]^2 \\ &= \sum_{s=1}^{\infty} \binom{N}{r} p_s^r (1-p_s)^{N-r} + I[r \le N/2] \\ &\quad \times \sum_{s \neq t} \binom{N}{r} \binom{N-r}{r} p_s^r p_t^r (1-p_s-p_t)^{N-2r} \\ &\quad - \left[\sum_{s=1}^{\infty} \binom{N}{r} p_s^r (1-p_s)^{N-r} \right]^2, \end{aligned} \tag{17}$$

and

$$\begin{aligned}\sigma_{r_1,r_2} &= Cov(n_{r_1}, n_{r_2}) = E[n_{r_1} n_{r_2}] - [E(n_{r_1})][E(n_{r_2})] \\ &= I[r_1 + r_2 \le N]\binom{N}{r_1}\binom{N-r_1}{r_2} \\ &\quad \times \sum_{s\neq t} p_s^{r_1} p_t^{r_2}(1-p_s-p_t)^{N-r_1-r_2} \\ &\quad - \left[\sum_{s=1}^{\infty}\binom{N}{r_1} p_s^{r_1}(1-p_s)^{N-r_1}\right]\left[\sum_{s=1}^{\infty}\binom{N}{r_2} p_s^{r_2}(1-p_s)^{N-r_2}\right]. \quad (18)\end{aligned}$$

Letting

$$c_0 = \left(\binom{N}{1}^{-1}, \cdots, (-1)^{r+1}\binom{N}{r}^{-1}, \cdots, (-1)^{N+1}\binom{N}{N}^{-1}\right)',$$

$\hat{\pi}_{0,N} = c_0' n$ and $Var(\hat{\pi}_{0,N}) = c_0' \Sigma c_0$.

Since $\hat{\pi}_{0,N}$ is an unbiased estimate of $\theta = \pi_0 + (-1)^N \sum_{s=1}^{\infty} p_s^{N+1}$, by Chebychev's Inequality, we have

$$P\left(\hat{\pi}_{0,N} - \sqrt{\frac{c_0'\Sigma c_0}{\alpha}} + (-1)^N \sum_s^{\infty} p_s^{N+1} < \pi_0 < \hat{\pi}_{0,N} + \sqrt{\frac{c_0'\Sigma c_0}{\alpha}} + (-1)^N \sum_s^{\infty} p_s^{N+1}\right) \ge 1-\alpha.$$

Approximating Σ and $\sum_{s=1}^{\infty} p_s^{N+1}$ by replacing the p_s s with $\hat{p}_s$ s, we have the following approximate and conservative confidence interval for π_0, which covers π_0 with probability at least $(1-\alpha)\times 100\%$,

$$\left[\hat{\pi}_{0,N} + (-1)^N \sum_{s=1}^{\infty} \hat{p}_s^{N+1}\right] \pm \sqrt{\frac{c_0'\hat{\Sigma} c_0}{\alpha}} \quad (19)$$

where $\hat{\Sigma}$ is the estimated Σ.

Testing Hypothesis

Under a hypothesized distribution, H_0, i.e. all $\{p_s; s=1, \ldots\}$ are known, $E_0(n\_)$ and Σ_0 are known, and can be exactly calculated based on Formulas (4), (17) and (18). For any $(N-1)$-dimensional constant

vector c and a significance level α, by Chebychev's Inequality, we have a conservative simple test which rejects H_0 if

$$c'[n_- - E_0(n_-)][n_- - E_0(n_-)]'c \geq \frac{c'\Sigma_0 c}{\alpha}. \tag{20}$$

In practice, there are options in choosing c to serve specific purposes.

Letting $c_1 = (1, \ldots, 1)'$, the test statistic, $c_1'n_- = \sum_{r=1}^{N-1} n_r$, is simply the total number of different categories represented in the sample. This test can then be thought of as a test for category coverage. In fact, the coverage statistic is widely used in practice to estimate the total number of categories in the population; that, of course, implies that there are S finite but unknown categories in the population to begin with – a special case of the model under consideration. Interested readers may refer to Lewontin and Prout (1956), Darroch (1958), Harris (1968), Johnson and Kotz (1977), Marchand and Schroeck (1982), Holst (1981), Darroch and Ratcliff (1980), Esty (1982, 1983, 1985, 1986a, 1986b), and Chao and Lee (1992).

If c_2 is taken to be a vector of zeros and ones with the ones distributed over a particular range of r, then the test can be considered as one of category coverage for that specific range. Similarly, the category coverage of two separate ranges of r may be compared by letting c_3 to be a vector of 0, 1 and -1, with the 1 and -1 distributed over the two contrasting ranges respectively.

Another way of choosing c may be profitably approached by the structure of Σ_0. Let $\lambda_1, \ldots, \lambda_N$ and $v_1, \ldots, v_N$ be the eigenvalues (in descending order) and their corresponding (normalized) eigenvectors respectively. Any of the eigenvectors, or any linear combination of a subset of them, may be chosen to be c. One particularly attractive property of the eigenvectors is that $v_r'n_-$, $r = 1, \ldots, N$, are uncorrelated and therefore convey "orthogonal" information in the sample. In fact, a component-wise examination on

$$z_r = v_r'[n_- - E_0(n_-)]/\sqrt{\lambda_r}, \tag{21}$$

$r = 1, \ldots, N-1$, could reveal detailed departure from H_0 if existed.

WAS "SHALL I DIE?" PENNED BY SHAKESPEARE?

On 14 November 1985, Shakespearean scholar Gary Taylor discovered a nine-stanza poem in a bound folio volume that had been in the collection

of the Bodleian Library since 1755 (Lelyweld, 1985; Taylor, 1985). The first stanza starts with "Shall I die? Shall I fly?", and hence the poem is often referred to in the literature as "Shall I Die?". The poem has 428 words, excluding the stanza heading – the Roman numerals "I" through "IX".

As an illustrative example of the results found in the preceding section, we will test the hypothesis (H_0) that "Shall I Die?" is by Shakespeare by gauging the observed rank-frequency distribution from the poem in question against the word-frequency distribution derived from Shakespeare's "complete" works, known as the Shakespearean canon, and refer to it as "the canon" hereafter. There exist several versions of the canon word frequency list. The differences are largely due to slightly different definitions of words and punctuations. The version we have was received on 12 July 2005 from D. Jurafsky of Stanford University. This list has a total of 883,140 words, classified into 34,782 word types.

Shakespearean Canon Processed

Suppose we classify the canon words into 34,783 categories: 34,782 word types in the canon and one category including words known to Shakespeare but not observed in the canon, and let the category probabilities be $\{p_s;\ s=1,\ldots,34783\}$. We then could estimate these category probabilities, using the canon word list. There are several options here. The first option is to assign zero probability to the category of unobserved words in the canon – although it is clear that its true value must be positive, and the observed relative frequency to each of the 34,782 word types. This approach would allow us to work neatly with 34,783 well-defined categories. However, this approach has a somewhat discomforting consequence: with a zero probability for the unobserved category in the canon, we cannot exactly pretend that the estimated p_s s form the true word type distribution in Shakespeare's vocabulary. However, the zero category probability makes very little difference in subsequent development since the statistical properties of the observed n_r s in the "Shall I Die?" sample are largely dictated by their mean vector and their covariance matrix, to which the value of one small individual p_s has extremely little impact. Nevertheless, for those who do feel uncomfortable with assigning zero probability to the unobserved category in the canon,

there is a second option. By the second option, the p_s s may be estimated in two steps:

1. Estimating the total probability of categories in D_r, namely π_r, $r = 1, \ldots, 883140$, by means of, e.g. Formula (15).
2. Estimating each individual category probability, in D_r, by $\approx\!\!\!\!\!\!\sim \pi_r / n_r$. This estimate is based on a further "symmetry assumption" (see Nádas, 1985). The symmetry assumption basically says that the probabilities of the categories that are represented an equal number of times in the sample are identical.

We note that the second option is primarily serving the purpose of having a non-zero probability for the unobserved category in the canon. When this is achieved, all the other category probabilities need to be re-adjusted. Since by the second step of the second option, a non-zero value is assigned to the total probability of each and every group, D_r, including those with no observations ($n_r = 0$). This "lack of resolution" again causes a technical problem: the estimated (non-zero) probability mass cannot be distributed to any categories. This difficulty may be avoided by a re-definition of the canon categories. For example, in the canon list, $n_r \neq 0$, for $r \leq 178$. One could therefore re-define categories by combining together all categories belonging to $D_{179}, \ldots, D_{883140}$ as one new category, and call it D_∞. Of course, the new category need not start from 179 to 883,140. In fact, it need not include a subset of D_rs in any particular order, so long as the remaining groups receive reasonable probability estimates. For the example at hand, letting $D_{179}, \ldots, D_{883140}$, in a new category makes intuitive sense since it represents more common word types in the canon. The total probability of all categories in D_∞ is estimated to be $1 - \sum_{r=0}^{178} \approx\!\!\!\!\!\!\sim \pi_r$.

For illustration purpose, we proceed with the second option. From the canon, the n_rs, are summarized in Table 4.

There are 34,178 (the canon has 34,782) word types covered in Table 4, excluding 604 word types in D_∞ (the last entry in Table 4 with*). Now we can re-group the categories by defining 34,180 $(1 + 34{,}178 + 1)$ categories in the canon: the category of unobserved word types (1), plus the categories covered in Table 4 (34,178), and plus the new category D_∞ (1).

Group probabilities by Formula (15) are given in Table 5. By the symmetry assumption, each category probability is given by the group probability divided by the number of categories in that group.

Table 4. n_r of the Canon.

r	0	1	2	3	4	5	6	7	8	9
00+		15876	4646	2512	1622	1181	910	664	577	454
10+	380	336	283	266	242	214	218	189	139	145
20+	141	122	120	118	101	103	116	88	93	77
30+	69	62	66	45	63	60	55	56	49	47
40+	42	49	33	46	34	35	38	33	41	25
50+	34	28	18	38	30	35	21	35	28	22
60+	22	22	23	18	19	16	19	22	18	20
70+	14	21	20	25	12	11	20	11	14	14
80+	17	8	7	14	12	15	18	12	18	11
90+	7	11	6	14	9	5	10	11	9	7
100+	7	8	4	11	4	7	13	5	7	10
110+	7	9	12	10	5	4	7	8	8	8
120+	4	5	4	9	3	8	7	4	4	4
130+	10	5	5	5	2	4	5	4	8	6
140+	8	6	5	5	6	3	4	2	3	4
150+	3	6	2	1	7	6	4	5	2	3
160+	9	2	3	6	3	1	3	5	5	2
170+	5	1	5	7	6	6	1	5	6	604*

Table 5. Group probabilities (%) for D_r of the canon.

r	0	1	2	3	4	5	6	7	8	9
00+	1.80	1.05	0.85	0.73	0.67	0.62	0.53	0.52	0.46	0.43
10+	0.42	0.38	0.39	0.38	0.36	0.39	0.36	0.28	0.31	0.32
20+	0.29	0.30	0.31	0.27	0.29	0.34	0.27	0.29	0.25	0.23
30+	0.22	0.24	0.17	0.24	0.24	0.22	0.23	0.21	0.21	0.19
40+	0.23	0.16	0.22	0.17	0.18	0.20	0.18	0.22	0.14	0.19
50+	0.16	0.16	0.23	0.18	0.28	0.13	0.23	0.18	0.15	0.15
60+	0.15	0.16	0.13	0.14	0.12	0.14	0.17	0.14	0.16	0.11
70+	0.17	0.16	0.21	0.10	0.09	0.17	0.10	0.12	0.13	0.15
80+	0.07	0.06	0.13	0.11	0.14	0.18	0.12	0.18	0.11	0.07
90+	0.11	0.06	0.15	0.10	0.05	0.11	0.12	0.10	0.08	0.08
100+	0.09	0.05	0.13	0.05	0.08	0.16	0.06	0.09	0.12	0.09
110+	0.11	0.15	0.13	0.06	0.05	0.09	0.11	0.11	0.11	0.05
120+	0.07	0.06	0.13	0.04	0.11	0.10	0.06	0.06	0.06	0.15
130+	0.07	0.07	0.08	0.03	0.06	0.08	0.06	0.13	0.09	0.13
140+	0.10	0.08	0.09	0.10	0.05	0.07	0.03	0.05	0.07	0.05
150+	0.10	0.03	0.02	0.12	0.11	0.07	0.09	0.03	0.05	0.16
160+	0.04	0.06	0.11	0.06	0.02	0.06	0.09	0.10	0.04	0.10
170+	0.02	0.10	0.14	0.12	0.12	0.02	0.10	0.12	0.10	68.64*

For example, the probability for each of the 15,876 categories in D_1 is 0.0105/15,876. The last entry in Table 5 is the probability of D_∞. Tables 4 and 5 together give estimates of probabilities to all 34,180 categories of the cannon.

"Shall I Die?" Processed

There are $N = 428$ words in "Shall I Die?". Classifying these words by the 34,180 categories of the canon, the sample rank-frequency distribution, n_r, $r = 1, \ldots, 428$, are calculated. All n_rs are zero except the following four:

r	1	2	3	428
n_r	161	8	1	1

Testing Hypothesis

Let us pretend that the category probabilities calculated from Table 5 were true (H_0). We may then calculate the expected rank-frequency distribution under H_0 by Formula (4). The first three expected values are

r	1	2	3
$E(n_r)$	122.12	2.09	0.11

The other $E(n_r)$, $r = 4, \ldots, 427$, are graphically represented in Figure 1.

Now let us make a visual observation regarding the sample characteristics and those of the canon. Firstly, $n_1 = 161$ and $E(n_1) = 122$, a considerable difference. Secondly, $n_2 = 8$ and $E(n_2) = 2$, also a difference. Thirdly, Figure 1 suggests that if any n_r is to be non-zero, it is likely, under H_0, to be for a r in the ranges of $0 \le r \le 20$ or $260 \le r \le 320$. The fact that we have $n_{248} = 1$ in the sample demonstrates another disagreement with H_0. In summary, the comparison seems to suggest that the "Shall I Die?" poem includes more "uncommon" word types (and therefore fewer "common" word types) than these suggested by the canon adjusted for sample size. Our attempt here is to assess the statistical significance of the difference observed taking into consideration the intrinsic random variation.

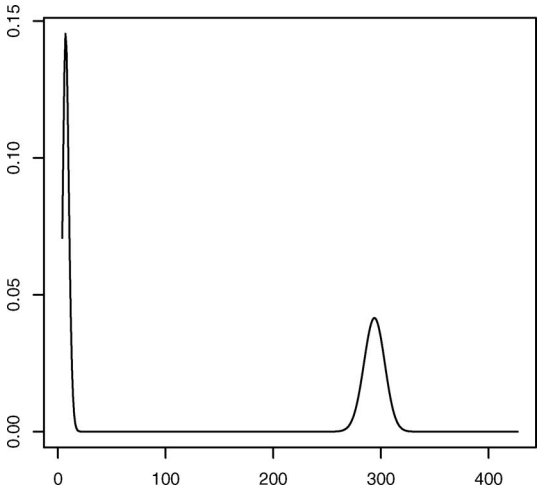

Fig. 1. Expected rank-frequency distribution under H_0.

Tests Based on Word Type Coverage

Let c_1 be a 427-dimensional vector of ones. $c_1'n_- = \sum_{r=1}^{427} n_r$ is simply the total number of different word types (not counting repeats) appeared in the sample. The data on hand yield $c_1'n_- = 171$, $c_1'E(n_-) = 126.27$, and by Formula (17) and Formula (18), $c_1'\Sigma_0 c_1 = 85.272$. The left-hand-side of Formula (20) is 2000.77; at $\alpha = 0.05$, the right-hand-side of Formula (20) is 1715.44; and therefore Formula (20) is true. This leads to a rejection of H_0. Recalling that Formula (20) is based on Chebyshev's Inequality and therefore a conservative test, the rejection to H_0 indicates that the p-value of the test at hand is at least 0.05. We therefore conclude that there is sufficient statistical evidence in "Shall I Die?" suggesting a difference in pattern of word type usage in comparison to that of the canon.

This test, curiously, does not take into consideration the frequencies of the word types in the sample. This property is quite refreshing and somewhat anti-intuitive. On one hand, it is perfectly natural to compare sample relative frequencies of the categories to their hypothesized population counterparts. On the other hand, when a sample is small and the population categories are very numerous, the comparison of category relative frequencies becomes difficult. One must (artificially) re-aggregate the population categories to have sufficient number of observations in each category for large sample theories to be approximately valid.

The Shakespearean canon has 34782 word types and 15876 of them were each used one single time. The extremely "thin" probability distribution of the rare word types is what exactly defines Shakespearean "style" (for lack of better words). Any aggregation of these categories, albeit necessary at times, would put a crimp on Shakespearean style. We propose the test in Formula (20) for its small-sample nature. We imagine it typical in testing authorship to have a relatively small writing sample compared to a hypothesized word type distribution.

Suppose one wishes to focus only on the uncommon word types in the sample to test H_0 by letting c_2 be a vector with the first 10 components being ones and the remaining components being zeros. Then

$$\sum_{r=1}^{10} n_r = 170, \qquad \sum_{r=1}^{10} E_0(n_r) = 125.12,$$

and ${c_2}'\Sigma_0 c_2 = 85.802$. The left-hand-side of Formula (20) is 2015.11, and at $\alpha = 0.05$, the right-hand-side of Formula (20) is 1716.04. Formula (20) is true.

Let c_3 be such that the first 10 components are ones and the remaining 417 are negative ones. Then

$$\sum_{r=1}^{10} n_r - \sum_{r=11}^{427} n_r = 169, \qquad \sum_{r=1}^{10} E(n_r) - \sum_{r=11}^{427} E(n_r) = 123.96,$$

and ${c_3}'\Sigma_0 c_3 = 85.592$. The left-hand-side of Formula (20) is 2028.60, and at $\alpha = 0.05$, the right-hand-side of Formula (2.20) is 1731.84. Formula (20) is true.

Let v_1, v_2, and v_3 be the first three eigenvectors of Σ_0 corresponding to the first, second and third largest eigenvalues respectively. Let $c_4 = v_1$, $c_5 = v_2$, $c_6 = v_3$, and $c_7 = v_1 + v_2 + v_3$. The following is a summary of the associated statistics.

c	v_1	v_2	v_3	$v_1 + v_2 + v_3$
$z = \frac{c'[n_- - E_0(n_-)]}{\sqrt{c'\Sigma_0 c}}$	3.9758	3.1198	4.7306	4.8109
(20) at $\alpha = 0.05$	false	false	true	true
(20) at $\alpha = 0.10$	true	true	true	true

Did Shakespeare write "Shall I Die?"? Unfortunately we still cannot answer this question, even with the statistical significance demonstrated by the above test. What the test has shown is a difference in style regarding the word type usage. Then again, the canon is mostly based on Shakespeare's plays and not his poems. (We are speculating here!) The mere difference in genre would naturally lead, and might indeed have led, to a discrepancy in style.

The "Shall I Die?" poem merely serves as an illustrative example. It is not our intention to prove or to disprove Shakespeare's authorship of the poem. Our survey on the information of "Shall I Die?" seems to suggest that the debate on the authorship is current and active. A large volume of rich and many-faceted discussion may be found in the literature and on the Internet regarding this issue. We particularly appreciate the discussions beyond the dimension of statistics, and fully realize that any statistical procedure, at its best, can only offer a narrow perspective on such an issue. Nevertheless, we suggest that, at least from that one narrow perspective, "Shall I Die?" does not seem to fit the canon well.

REMARKS

Under H_0 of the canon, the p-value of the test of the type in Formula (20) may be approximately assessed by computer simulation, provided that the pseudo-random number generator used is sound. Let an iid sample of size $N = 428$ be repeatedly taken from the canon distribution, and the left-hand-side of Formula (20) be computed and compared to the observed statistic. In the case of c_1, the observed statistic is 2000.77. The proportion of times the left-hand-side of Formula (20) exceeding the observed statistic is approximately equal to the p-value of the test. Our simulation studies show that, in 10,000 trials, not even one single time the simulated statistic exceeded 2000.77. This suggests that the departure of the "Shall I Die?" poem from the canon is quite significant.

One of the more active areas where Turing's formula is frequently used is related to estimation of the total number of population categories (S) under the assumption that S is finite but unknown. For the equal-probable case (i.e. $p_s = 1/S$, for $s = 1, \ldots, S$) the estimate by Darroch (1958) seems quite satisfactory. For the unequal-probable case, Darroch and Ratcliff (1980) proposed one estimate and Chao and Lee (1992) offered several refined versions. These estimates strive to take the

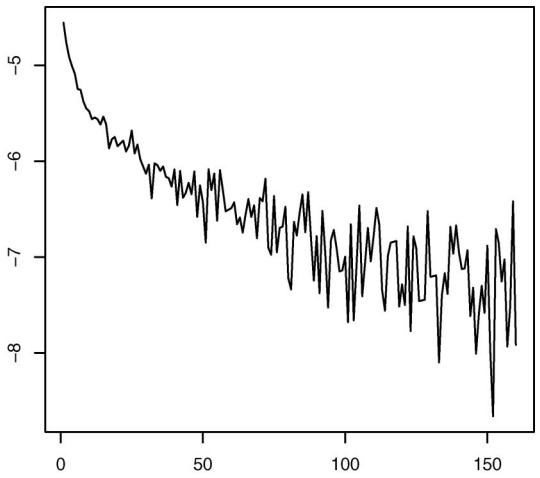

Fig. 2. $\log(n_r/N)$ vs. r of Shakespearean canon.

"unevenness" of the probability distribution into consideration, and all use Turing's formula in the denominators. When applied to the equal-probable case, it is clear that these estimates over-estimate S, some times severely, for relatively small N. We believe the over-estimation is partially caused by the over-estimation of Turing's formula ($\hat{\pi}_0$) for π_0, and hence by the under-estimation of $1 - \hat{\pi}_0$ for $1 - \pi_0$, which is the denominator of these estimates. Therefore the proposed estimate of this paper Formula (10) could immediately improve the performance of these proposed by Darroch and Ratcliff (1980) and Chao and Lee (1992).

Incidentally the Shakespearean canon provides a counter-example to Zipf's law. Figure 2 shows the plot of $\log(n_r/N)$ versus r for $r = 1, \ldots, 160$. A nonlinear relationship is quite clear for lower values of r.

REFERENCES

Adamic, L. A., & Huberman, B. A. (2002). Zipf's law and the internet. *Glottometrics*, *3*, 143 – 150.

Darroch, J. N. (1958). The multiple recapture census I: Estimation of a closed population. *Biometrika*, *45*, 343 – 359.

Darroch, J. N., & Ratcliff, D. (1980). A note on capture-recapture estimation. *Biometrics*, *36*, 149 – 153.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, *63*(3), 435 – 447.

Esty, W. W. (1982). Confidence intervals for the coverage of low coverage samples. *The Annals of Statistics*, *10*, 190–196.

Esty, W. W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *The Annals of Statistics*, *11*, 905–912.

Esty, W. W. (1985). Estimation of the number of classes in a population and the coverage of a sample. *Mathematical Scientist*, *10*, 41–50.

Esty, W. W. (1986a). The size of a coinage. *Numismatic Chronicle*, *146*, 185–215.

Esty, W. W. (1986b). The efficiency of Good's nonparametric coverage estimator. *The Annals of Statistics*, *14*, 1257–1260.

Gale, W. A., & Sampson, G. (1995). Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, *2*(3), 217–237.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237–264.

Harris, B. (1968). Statistical inference in the classical occupancy problem unbiased estimation of number of classes. *JASA*, *63*, 837–847.

Holst, L. (1981). Some assymptotic results for incomplete multinomial or Poison samples. *Scandinavian Journal of Statistics*, *8*, 243–246.

Johnson, N. L., & Kotz, S. (1977). *Urn Models and their Applications: an Approach to Modern Discrete Probability Theory*. New York: John Wiley.

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

Lewontin, R. C., & Prout, T. (1956). Estimation of the number of different classes in a population. *Biometrika*, *12*, 211–223.

Marchand, J. P., & Schroeck, P. E. (1982). On the estimation of the number of equally likely classes in a population. *Communication in Statistics, Part A – Theory and Methods*, *11*, 1139–1146.

Nádas, A. (1985). On Turing's formula for word probability. IEEE Transactions on Acoustics, Speech, and Signal Processing. *ASSP-33*, 1414–1416.

Robbins, H. E. (1955). An empirical Bayes approach to statistics. *Proceedings of 3rd Berkeley Symposium on Statistical Probability*, *1*, 157–164.

Robbins, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Statistics*, *39*, 1, 256–257.

Taylor, G. (1985). Shakespeare's new poem: A scholar's clue and conclusion. *New York Times Book Review* (15 December), 11–14.

Thisted, R., & Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika*, *74*(3), 445–455.

Zipf, G. K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addision-Wesley.