ON ESTIMATING THE NUMBER OF UNSEEN SPECIES: HOW MANY EXECUTIONS WERE THERE?

BY

P. J. BICKEL

AND

J. A. YAHAV

TECHNICAL REPORT NO. 43
JUNE 1985

*RESEARCH PARTIALLY SUPPORTED BY
OFFICE OF NAVAL RESEARCH CONTRACT N00014-80-C-0163

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA

ON ESTIMATING THE NUMBER OF UNSEEN SPECIES: HOW MANY EXECUTIONS WERE THERE?*

By PETER J. BICKEL

Department of Statistics, University of California, Berkeley

and JOSEPH A. YAHAV

Department of Statistics, The Hebrew University, Jerusalem

SUMMARY

We consider the problem of estimating the number of distinct executions witnessed by the total refugee communities in three locations. We give lower bounds to these numbers and prove asymptotic normality for the estimator considered.

Some key words: Lower bounds; Empirical Bayes; Asymptotic normality; Dependent variables

*Research partially supported by Office of Naval Research contract N00014-80-C-0163.

ON ESTIMATING THE NUMBER OF UNSEEN SPECIES: HOW MANY EXECUTIONS WERE THERE?*

P. J. Bickel and J. A. Yahav

1. INTRODUCTION

Our interest in this problem was aroused by the question of how to estimate the number of distinct executions that took place in South Vietnam after the takeover by the Hanoi government in 1975.

J. Desbarats of the Institute of East Asian Studies and Karl D. Jackson of the Department of Political Science at the University of California at Berkeley provided us with data taken from a sample of refugees living in Chicago, Orange County and San Francisco:

Table 1

	Chicago	Orange County	San Francisco
Sample size	98	840	1160
# distinct executions	30	67	36
# distinct executions reported once	24	61	28
# distinct executions reported twice	5	3	7
# distinct executions reported 3 times	0	1	0
# distinct executions reported 4 times	1	0	0
# distinct executions reported 5 times	0	1	0
# distinct executions reported 6 times	0	1	1
# reported executions	38	81	48

On the basis of this data we were interested in estimating the number of distinct executions witnessed by the total refugee communities in

*Research partially supported by Office of Naval Research contract N00014-80-C-0163.

each of the those locations. Extrapolation to Vietnam is of course questionable, since the refugee populations cannot be considered as random samples from the Vietnam population.

The problem of estimating the number of unseen species was discussed by Fisher, Corbet and Williams in (1943), where a parametric empirical Bayes model was devised. Good (1953) developed this model without assuming a parametric family of priors on the parameters. Good and Toulmin (1956) and Efron and Thisted (1976) augmented the Good approach by constructing better estimators. Goodman (1949) studied a version of the problem in which the the total population size is known and constructed the unique unbiased estimator for the number of species.

Our data differs from that modelled by previous authors in two respects:

(i) The bulk of the sample consists of individuals who have seen no executions. If we unrealistically consider the known text of Shakespeare, the Efron-Thisted example, as a sample from all his writings, then every word written corresponds to one (and only one) type. The obvious lack of independence of successive words of the text make this assumption unrealistic and renders understandable Efron and Thisted's unwillingness to pursue distribution theory.

(ii) There are individuals reporting more than one execution, unlike the species problem where each individual belongs to only one species. We accommodate these special features in a general sampling with replacement model in section 2. We derive lower bounds $\underline{k}$ of different types to the unknown number k of "executions". These bounds are readily estimable by simple estimators $\hat{\underline{k}}$. We maintain this

grisly terminology for convenience since our results of course apply to the estimation of the number of unseen species. The reason for looking at estimable lower bounds rather than trying to estimate k directly are discussed here and in section 2. In section 3 we proceed further and derive a normal approximation to the distribution of $\hat{\underline{k}}$ when the sample size and k are large, for the simple model in which each individual sees at most one execution -- the type of model considered by previous writers but with "nonresponse" permitted. The asymptotic variance $\sigma^2_{\underline{k}}$ of $\hat{\underline{k}}$ is hard to estimate but we can find an upper bound $\bar{\sigma}^2_{\underline{k}}$ to $\sigma^2_{\underline{k}}$ such that $\bar{\sigma}^2_{\underline{k}}$ is easily estimated. Finally we propose lower confidence bounds of the form $\hat{\underline{k}} - z_{1-\alpha}\hat{\bar{\sigma}}_{\underline{k}}$. In section 4 we discuss the application of our techniques to the Vietnam and Efron-Thisted data. Proofs for the asymptotic approximations claimed are given in an appendix. Our results are central limit theorems for sums of functions of the components of multinomial vectors. They are based on work of Stack (1957) and are related to later work of Morris (1975) although Morris' conditions seem to fail for the indicator functions that we need to apply them to.

2. SIMPLE LOWER BOUNDS AND THEIR ESTIMATORS

Given the refugee population sizes, we ignore that the samples were obtained without replacement. The situation can be described as follows: Let $S = \{\text{All vectors } (\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_k) \text{ of } k \text{ 0's and 1's}\}$, F = a set of probabiilty distributions on S. Let $U_1,U_2,\ldots,U_n$ be independent and identically distributed according to $P \in F$, $U_i = (\varepsilon_{i1},\varepsilon_{i2},\ldots,\varepsilon_{ik})$. Let α be a random permutation of $\{1,2,\ldots,k\}$ (fixed for all n). Let $X_i = \{\alpha(j): \varepsilon_{ij} = 1\}$. $X_1,X_2,\ldots,X_n$ are our observations. The correspondence with the "execution" situation is the following:

(i) The executions are labelled $1,2,\ldots,k$, the sampled individuals $1,2,\ldots,n$.

(ii) $\varepsilon_{ij} = 1$ if and only if the i^{th} individual in the sample saw the j^{th} execution so U_i is the record of which executions were seen by individual i, and which were not. U_i is of course not observable.

(iii) X_i is the list of executions witnessed by individual i, with the "names", the $\alpha(j)$, containing no information about the magnitude of k.

Let $N_s = \sum_{i=1}^{n} I\{U_i = s\}$, the number of sampled individuals whose execution record is s. If $\pi_s = P[\{s\}]$ is the proportion in the population of such individuals, then $N = \{N_s: s \in S\}$ has a multinomial $(n,\pi_s;s \in S)$ distribution. All statistics we consider are functions of N which are also functions of $(X_1,X_2,\ldots,X_n)$. In the Fisher et al., Good, Good and Toulmin and Efron and Thisted situations (with dependence between words ignored) this model is further restricted by F making only s with exactly k-1 zeros possible. An observation is exactly one species or word, and we can replace S by $\{1,2,\ldots,k\}$, N by $(N_1,N_2,\ldots,N_k)$ where N_j is the number of times j occurs. $(N_1,N_2,\ldots,N_k)$ has a multinomial $(n,\pi_1,\pi_2,\ldots,\pi_k)$ distribution. The observable sufficient statistic is given by $\{N_{\alpha(j)}: N_{\alpha(j)} > 0\}$. In the "simplified" execution problem, the 0 vector is also admitted and we can replace S by $\{0,1,\ldots,k\}$ where the outcome 0 means that the sampled individual witnessed no executions, etc.
We shall refer to this as the <u>simple</u> model.

The total number of distinct executions listed T_n is a natural underestimate of k. If $N_j = \sum_{i=1}^{n} \varepsilon_{ij}$, the number of individuals in the sample that witnessed execution j, $T_n = \sum_{j=1}^{k} I\{N_j > 0\}$,

$$E[T_n] = \sum_{j=1}^{k} (1 - (1-\pi_j)^n) = \sum_{j=1}^{k} \sum_{r=0}^{n-1} \pi_j (1-\pi_j)^r . \qquad (2.1)$$

Denote $E[T_n]$ by θ_n. Then, the bias

$$k - E[\theta_n] = \sum_{r=n}^{\infty} \sum_{j=1}^{k} \pi_j (1-\pi_j)^r . \tag{2.2}$$

More generally, for $m \geq 1$, denote the expected number of distinct executions listed in a sample of size m, by

$$\theta_m = \sum_{j=1}^{k} \sum_{r=0}^{m-1} \pi_j (1-\pi_j)^r ,$$

so $\theta_n = E[T_n]$. Then, as a limit, $\theta_\infty = k$ and we define $\theta_0 = 0$. For $m \leq n$, θ_m is estimable unbiasedly on the basis of our sample of size n by the U-statistic,

$$\hat{\theta}_m = \binom{n}{m} \sum_{j=1}^{k} I\{ \sum_{r=1}^{m} \varepsilon_{i_r j} > 0 \} ,$$

the average number of distinct executions listed in subsamples of size m drawn from the sample of n. So, $T_n = \hat{\theta}_n$. Finally, let

$$\Delta_m = \theta_m - \theta_{m-1} = \sum_{j=1}^{k} \pi_j (1-\pi_j)^{m-1} ,$$

the expected number of executions seen by the first person in a sample of size m and no others, which is unbiasedly estimable by $\hat{\Delta}_m = \hat{\theta}_m - \hat{\theta}_{m-1}$. In the simple model, Efron and Thisted and previous writers pointed out that if $n\pi_j = \lambda_j$, $j = 1,2,\ldots,k$ are moderate and n is large, we can approximate and simplify the model by taking N_j independent Poisson (λ_j). This approach is discussed further in Section 4.

PROPOSITION 1. *Even in the simple model no unbiased estimator of* k *exists.*

Proof. We can write an estimator $\delta(N_0,N_1,\ldots,N_k)$ with the understanding that

$$\delta(n_0,n_1,\ldots,n_k) = \delta(n_0',n_1',\ldots,n_k') \tag{2.3}$$

if $t = \sum_{j=1}^{k} I\{n_j > 0\} = \sum_{j=1}^{k} I\{n'_j > 0\}$ and the nonzeros of $n_{j_1},\ldots,n_{j_k}$ are a permutation of the nonzeros of $n'_{j_1},\ldots,n'_{j_k}$. Unbiasedness for $k = 1$ then requires $\sum_{i=0}^{n} \binom{n}{i}\pi^i(1-\pi)^{n-i}\delta(i,n-i) = 1$ for all $0 < \pi < 1$. Hence $\delta(i,n-i) \equiv 1$. However $k = 2$ forces

$$\sum_{\substack{i,j \\ 0\le i+j\le n}} \binom{n}{i,j}\pi_1^i\cdot\pi_2^j(1-\pi_1-\pi_2)^{n-i-j}\delta(i,j,n-i-j) = 2$$

for all $0 < \pi_1, \pi_2$ and $\pi_1 + \pi_2 < 1$. Hence $2 = \delta(i,j,n-i-j)$. But (2.3) requires $\delta(i,0,n-i) = \delta(i,n-i)$ for $i > 0$, and we have a contradiction. Q.E.D.

Here is our principal lower bound $\underline{k}_1$. Let

$$M_r = \Big(\sum_{j=1}^{k}\pi_j\Big)^{-1}\sum_{j=1}^{k}\pi_j(1-\pi_j)^r = \frac{\Delta_{r+1}}{\Delta_1}\,.$$

The bias is then, from (2.2), $k - \theta_n = \sum_{r=n}^{\infty}\Delta_1 M_r \ge \Delta_1\cdot\sum_{r=n}^{\infty} M_{n-1}^{\frac{r}{n-1}} = \Delta_1\cdot M_{n-1}^{\frac{n}{n-1}}(1 - M_{n-1}^{\frac{1}{n-1}})^{-1} = \Delta_n[(\frac{\Delta_1}{\Delta_n})^{\frac{1}{n-1}} - 1]^{-1}$. Our lower bound is

$$\underline{k}_1 = \theta_n + \Delta_n[(\frac{\Delta_1}{\Delta_n})^{\frac{1}{n-1}} - 1]^{-1}\,.$$

Evidently $\underline{k}_1 = k$ if and only if $\pi_j \equiv \pi$, $1 \le j \le k$. $\underline{k}_1$ can be thought of as the first in a hierarchy of bounds. Thus,

$$\underline{k}_1 = \theta_n + \min\{\sum_{j=1}^{k}(1-\pi_j)^n:\ \sum_{j=1}^{k}\pi_j = \Delta_1,\ \sum_{j=1}^{k}\pi_j(1-\pi_j)^{n-1} = \Delta_n,\ 0 \le \pi_j \le 1,\ 1 \le j \le k\}\,. \qquad (2.4)$$

We can similarly define

$$\underline{k}(i_1,i_2,\ldots,i_r) = \theta_n + \min\{\sum_{j=1}^{k}(1-\pi_j)^n:\ \sum_{j=1}^{k}\pi_j(1-\pi_j)^{i_t-1} = \Delta_{i_t},\ t = 1,2,\ldots,r\}\,.$$

Of course $\underline{k}(1,2,\ldots,n)$ is the best of these bounds.

ASSUMPTION A. *Suppose that as* $k \to \infty$, $n(k) \to \infty$ *and* π_j *vary in such a way that the empirical distributions* $G_n(\cdot)$ *of* $n\pi_1, n\pi_2, \ldots, n\pi_k$ *converge weakly to* $G_0(\cdot)$, *a probability distribution on* $(0,\infty)$. *Moreover, suppose that*

$$\frac{n}{k}\Delta_1 \to \int_0^\infty \lambda dG_0(\lambda) \ . \tag{2.5}$$

Then,

$$\frac{\theta_n}{k} \to \int_0^\infty (1-e^{-\lambda}) dG_0(\lambda) \tag{2.6}$$

and

$$\frac{n}{k}\Delta_{[tn]+1} \to \int_0^\infty \lambda e^{-t\lambda} dG_0(\lambda) \ , \quad 0 \le t < 1 \ . \tag{2.7}$$

In this case if $\bigcup_{r=1}^{\infty} \{0 = t_1 < t_2 < t_3 < \cdots < t_r\}$ is dense in $[0,1]$

$$\lim_{\substack{k\to\infty \\ r\to\infty}} \frac{1}{k}\, \underline{k}(t_1,t_2,\ldots,t_r) \ge 1 - \int_0^\infty e^{-\lambda} dG_0(\lambda) + \min_G \{\int_0^\infty e^{-\lambda} dG(\lambda) : \int_0^\infty \lambda e^{-\lambda t} dG(\lambda) = \int_0^\infty \lambda e^{-t\lambda} dG_0(\lambda),\ 0 \le t \le 1\} = 1 \ , \tag{2.8}$$

since the measure $\lambda dG_0(\lambda)$ is uniquely determined by its Laplace transform on an interval. This suggests that consistent estimators of $\int_0^\infty e^{-\lambda} dG_0(\lambda)$ can be constructed by choosing $r(k) \to \infty$ slowly and estimating $\underline{k}([t_1n]+1,[t_2n]+1,\ldots,[t_rn]+1)$ by $\hat{\underline{k}}$ in which $\Delta_{[t_1n]+1}, \Delta_{[t_2n]+1}, \ldots, \Delta_{[t_rn]+1}$ are replaced by the corresponding points from a completely monotone approximation to $\hat{\Delta}_{[nt]+1}$ (Feller [2], p. 439).

Now,

$$\hat{\underline{k}}_1 = T_n + \frac{1}{n}\sum_{j=1}^{k} I_{\{N_j=1\}}\left[\left(\sum_{j=1}^{k} N_j / \sum I_{\{N_j=1\}}\right)^{1/n-1} - 1\right]^{-1} \quad (2.9)$$

Calculation of $\hat{\underline{k}}$ is eased by noting that for this linear programming problem only G concentrating on at most r points need be considered. This approach will be considered elsewhere. It is closely related to Efron and Thisted's maximization of negatively biased linear estimators of k.

We analyze only $\underline{k}_1$ and $\hat{\underline{k}}_1$ further in what follows, but note several other easily estimable nonlinear lower bounds. Thus from (2.1),

$$\frac{\theta_n}{k} \leq 1 - (1 - k^{-1}\sum_{j=1}^{k} \pi_j)^n = 1 - (1 - \frac{\Delta_1}{k})^n \ . \quad (2.10)$$

Let $\underline{k}_2$ be the integer part of the unique finite root $\geq \Delta_1$ of:

$$\frac{\theta_n}{k} = 1 - (1 - \frac{\Delta_1}{k})^n \ . \quad (2.11)$$

The existence and unicity of such a root and $\underline{k}_2 \leq k$ follow from (2.6) and the fact that $g(x) = \theta_n x - 1 + (1-\Delta_1 x)^n$ is convex on $[0,\Delta_1^{-1}]$, $g(0) = 0$ and $g(\frac{1}{k}) \leq 0$. The estimate $\hat{\underline{k}}_2$ obtained by substituting $\hat{\theta}_n$, $\hat{\Delta}_1$ in (2.11) is closely related to the Petersen-Chapman-Darroch estimate (see Seber [7]) which solves

$$\frac{\hat{\theta}_n}{k} = 1 - \prod_{i=1}^{n} (1 - \frac{M_i}{k})^+ \ , \quad (2.12)$$

where $M_i = \sum_{j=1}^{k} \varepsilon_{ij}$, the length of i's list. This estimate is appropriate if $P[\varepsilon_{ij} = 1]$ does not depend on j though it may depend on i. Substituting $\theta_n = E[\hat{\theta}_n]$ and $E[M_i] = \Delta_1$ in, (2.12) leads

to (2.11). Again $\underline{k}_2 = k$ if and only if $\pi_j = \pi$, $1 \le j \le k$.

Another bound is

$$\underline{k}_3 = \theta_n^2/n^2 \sum_{j=1}^{k} \pi_j^2 = \theta_n^2/n^2(\Delta_1 - \Delta_2) .$$

This bound is not comparable to $\underline{k}_1$, $\underline{k}_2$ and becomes sharp when $k \to \infty$ and all $n\pi_j$ are small. It gives some insight to see what happens to $\underline{k}_i$, $i = 1,2,3$ under assumption A. From (2.5)-(2.7) we have

$$\frac{\underline{k}_1}{k} \to \rho_1 = \int (1-e^{-\lambda}) dG_0(\lambda) + \left[\int \lambda e^{-\lambda} dG_0(\lambda)\right]\left[\log \int \lambda dG_0(\lambda) - \log \int \lambda e^{-\lambda} dG_0(\lambda)\right]^{-1}$$

since $n\left[\left(\frac{\Delta_1}{\Delta_n}\right)^{\frac{1}{n-1}} - 1\right] \sim n\left\{\left[\frac{\int \lambda dG_0(\lambda)}{\int \lambda e^{-\lambda} dG_0(\lambda)}\right]^{1/(n-1)} - 1\right\}$,

$$\frac{\underline{k}_2}{k} \to \rho_2$$

where $x = \frac{1}{\rho_2} > 1$ satisfies

$$h(x) = \left[1 - \int_0 e^{-\lambda} dG_0(\lambda)\right] \cdot x + \exp\left[-x \int_0^{\infty} \lambda dG_0(\lambda)\right] - 1 = 0 ,$$

$$\frac{\underline{k}_3}{k} \to \rho_3 = \frac{\left(1 - \int e^{-\lambda} dG_0(\lambda)\right)^2}{\int \lambda^2 dG_0(\lambda)} .$$

PROPOSITION 2. $\rho_j \le 1$, $j = 1,2,3$, *with equality for* $j = 1,2$ *if and only if* G_0 *is a point mass. Moreover,*

$$\rho_1 \ge \rho_2 . \tag{2.13}$$

Proof. The claim for ρ_1 is equivalent to

$$\frac{\int \lambda e^{-\lambda} dG_0(\lambda)}{\int e^{-\lambda} dG_0(\lambda)} \le \log \frac{\int \lambda dG_0(\lambda)}{\int \lambda e^{-\lambda} dG_0(\lambda)} . \tag{2.14}$$

Let $dQ(\lambda) = e^{-\lambda}dG_0(\lambda)[\int e^{-\lambda}dG_0(\lambda)]^{-1}$. So (2.14) is equivalent to $\int \lambda dQ(\lambda) \leq \log \frac{\int \lambda e^{\lambda} dQ(\lambda)}{\int \lambda dQ(\lambda)}$. We have

$$\int \lambda e^{\lambda} dQ(\lambda) \geq \int \lambda dQ(\lambda) \cdot e^{\int \lambda dQ(\lambda)} \tag{2.15}$$

which holds by the strict convexity of λe^{λ} with strict inequality valid unless Q is a point mass. (2.14) follows easily from (2.15). To prove the claim for ρ_2, note that $h(0) = 0$, $h''(x) > 0$, $h(\infty) = \infty$ implies that h has exactly one positive root. Since

$$h(1) = -\int e^{-\lambda}dG_0(\lambda) + \exp\{-\int \lambda dG_0(\lambda)\} < 0 ,$$

unless G_0 is a point mass, $\frac{1}{\rho_2} \geq 1$ and the result follows. Finally $[\int (1-e^{-\lambda})dG_0(\lambda)]^2 \leq \int (1-e^{-\lambda})^2 dG_0(\lambda) \leq \int \lambda^2 dG_0(\lambda)$ yields $\rho_3 \leq 1$.

To prove (2.13) note that

$$\underline{k}_2 \leq \theta_n + \sum_{j=1}^{k} (1-\pi_j)^n = k$$

if $\sum_{j=1}^{k} \pi_j = \Delta_1$. Therefore,

$$\begin{aligned} \underline{k}_2 &\leq \theta_n + \min\{\sum_{j=1}^{k} (1-\pi_j)^n : \sum_{j=1}^{k} \pi_j = \Delta_1\} \\ &\leq \theta_n + \min\{\sum_{j=1}^{k} (1-\pi_j)^n : \sum_{j=1}^{k} \pi_j = \Delta_1, \sum_{j=1}^{k} \pi_j (1-\pi_j)^{n-1} = \Delta_n\} \\ &= \underline{k}_1 \end{aligned}$$

by (2.4) Q.E.D.

We do not pursue $\underline{k}_2$, $\underline{k}_3$ at this time although the assumptions that lead to the approximate normality of $\hat{\underline{k}}_1$ also lead to normality of $\hat{k}_i$, $i = 2,3$.

3. APPROXIMATION TO VARIANCES AND DISTRIBUTIONS OF ESTIMATORS

As we suggested in section 2 we want to make large k approximations to the distribution of $(\hat{\underline{k}}_1 - \underline{k}_1)/\sqrt{k}$. Embed our situation in a sequence indexed by s. Suppose that $k(s) \to \infty$, $n(s) \to \infty$ and

ASSUMPTION $\tilde{A}$. (i) $G_s(x) = \frac{1}{k} \sum_{j=1}^{k} I\{n\pi_j \leq x\} \to G(x)$ as $s \to \infty$ *where*

(ii) $\lim_{x\to\infty} G(x) = 1$, $\lim_{x\to 0} G(x) = 0$,

and

(iii) $\sup_s \int_0^\infty t^4 dG_s(t) < \infty$.

LEMMA 1. *Under Assumption $\tilde{A}$, $\hat{\underline{k}}$ may be linearized asymptotically:*

$$\hat{\underline{k}}_1 - \underline{k}_1 = \sum_{j=1}^{k} (I\{N_j=0\} - (1-\pi_j)^n) - A_n \sum_{j=1}^{k} (I\{N_j=1\} - n\pi_j(1-\pi_j)^{n-1}) + B_n \sum_{j=1}^{k} (N_j - n\pi_j) + o_p(k^{-1/2}) \tag{3.1}$$

where

$$\begin{aligned} A_n &= c_n^{-1} + d_n c_n^{-2} \cdot \frac{n}{n-1} \\ B_n &= d_n^{n-2} \cdot c_n^{-2} \cdot \frac{n}{n-1} \\ c_n &= n(d_n - 1) \\ d_n &= \left(\frac{\Delta_1}{\Delta_n}\right)^{1/n-1} \end{aligned} \tag{3.2}$$

Proof. Write $\hat{\underline{k}}_1 = g_n(\hat{\theta}_n, n\hat{\Delta}_n, n\hat{\Delta}_1)$

$\underline{k} = g_n(\theta_n, n\Delta_n, n\Delta_1)$.

Taylor expand around $(\theta_n, n\Delta_n, n\Delta_1)$ and note that the sums above are just $(\hat{\theta}_n - \theta_n)$, $n(\hat{\Delta}_n - \Delta_n)$, $n(\hat{\Delta}_1 - \Delta_1)$. The remainder is $o_p(k^{-1/2})$

since under $\tilde{A}$, the second derivatives of g_n are uniformly bounded in n, k in a neighborhood of $(\theta_n,\Delta_n,\Delta_1)$ and the sums are $O_p(k^{-1/2})$ by Theorems 1-3 of the appendix. (The centering constants are easily shown to differ from those of the theorems by $o(k^{-1/2})$. Q.E.D.

The asymptotic variance of the right hand side of (3.1) is

$$\sigma_1^2 \sim k\Big\{\int_0^\infty [e^{-x}(1-e^{-x}) + A_n^2 xe^{-x}(1-xe^{-x}) + B_n^2 x]dG(x)$$
$$+ \int_0^\infty xe^{-x}[2A_n x - 2B_n - 2A_nB_n(1-x)]dG(x)$$
$$- \frac{1}{\int_0^\infty x\, dG(x)}\Big[\int_0^\infty (A_n xe^{-x}(x-1) + B_n x - xe^{-x})dG(x)\Big]^2\Big\}$$

An upper bound on σ_1^2 is:

$$\bar{\sigma}_1^2 = k\Big\{\int_0^\infty [(1-e^{-x}) + A_n^2 xe^{-x} + xB_n^2]dG(x) \qquad (3.3)$$
$$+ \int_0^\infty xe^{-x}[2A_n x - 2B_n - 2A_nB_n(1-x)]dG(x)$$
$$- \frac{1}{\int_0^\infty x\, dG(x)}\Big[\int_0^\infty (A_n xe^{-x}(x-1) + B_n x - xe^{-x})dG(x)\Big]^2\Big\} .$$

The upper bound of the asymptotic variance can be estimated consistently using $\tilde{A}$ as follows:

$$k\int_0^\infty (1-e^{-x})dG(x) \approx \sum_j I\{N_j \geq 1\} = \hat{\theta}_n$$
$$k\int_0^\infty x\, dG(x) \approx \sum_j N_j = n\hat{\Delta}_1$$
$$k\int_0^\infty xe^{-x}dG(x) \approx \sum_j I\{N_j=1\} = n\hat{\Delta}_n .$$
$$k\int_0^\infty x^2e^{-x}dG(x) \approx 2\sum_j I\{N_j=2\} .$$

A_n, B_n can be estimated consistently by using $\hat{\Delta}_1$ and $\hat{\Delta}_n$ in the definition of A_n and B_n. All of these claims are consequences of Theorem 1 of the appendix.

THEOREM 1. *In the simple model, if Assumption $\tilde{A}$ holds,* $\dfrac{\hat{k} - k_1}{\hat{\sigma}_1}$ *has as* $s \to \infty$ *a limiting normal distribution with mean* 0 *and variance* ≤ 1.

Assumption $\tilde{A}$ is uncheckable as it stands. However, a heuristic argument suggests an estimable parameter to guide us. For k large $\tilde{A}$(ii) suggests that the N_j's are approximately independent Poisson($n\pi_j$) where $n\pi_j \leq M$, $j = 1,2,\ldots,k$. Then $\max_j N_j$ should be stochastically smaller than $\max_j N_j'$ where the N_j' are independent Poisson(M). It is well known that

$$\frac{(\max_j N_j') \log\log k}{\log k} = O_p(1) \tag{3.4}$$

and the same should hold for $\max_j N_j$. Similarly $\tilde{A}$ suggests that $O_p(1)$ is not $o_p(1)$. Since $\sum_{j=1}^{k} N_j$ is of the order of k, we finally conclude that the approximations of the theorem are reasonable if $\alpha = (\max_{j\geq 1} N_j)\cdot(\log\log \sum_{j=1}^{k} N_j)[\log \sum_{j=1}^{k} N_j]^{-1}$ is moderate -- not too large and not too small. Further theoretical and Monte Carlo work is needed to make this statement precise. It is however notable that for our data α ranges from 4.7 to 20.5. For the Efron-Thisted data $\alpha = 2747$.

Although the Steck theory is no longer applicable for the full model, it seems plausible here too that under suitable conditions, as $n, k \to \infty$ the N_j behave like (dependent) Poisson $n\pi_j$ variables and that $(\hat{\theta}_n, n\hat{\Delta}_1, n\hat{\Delta}_n)$ still exhibit limit normal behaviour. However the

possibility that two or more executions can appear on the same list of an individual witness leads to a more elaborate upper bound on the variance. One has to introduce $P(j'|j)$ as the probability that execution j' is witnessed by an individual chosen at random given that this individual witnessed execution j. If $P(j'|j)$ is of the order of $\pi_{j'}$ or smaller, then formula (3.3) for an upper bound of the variance of our estimator $\hat{\underline{k}}$ does apply. If $P(j'|j)$ for most (j',j) is of order one, then the calculation of the variances of $\Sigma I\{N_j=1\}$, $\sum N_j$ and $\Sigma I\{N_j>0\}$ become somewhat more involved taking account of the covariances that can be estimated.

4. APPLICATION OF LOWER BOUND ESTIMATES TO VIETNAM DATA AND COMPARISON WITH OTHER TECHNIQUES

Table 2. *Lower bounds for the Vietnam data*

	Chicago	Orange County	San Francisco
$T_n = \sum_{j=1}^{k} I\{N_j>0\}$	30	67	36
$\hat{\underline{k}}_1$	82	282	88
$\hat{\underline{k}}_2$	80	210	79
$\hat{\underline{k}}_3$	53	76	36
$\hat{\sigma}_1(\hat{\underline{k}}_1)$	29.9	69.9	27.4

The authors cited in the introduction considered the simple model. They argued that $D = \sum_{j=1}^{k} (1-\pi_j)^n$ may be approximated by $\sum_{j=1}^{k} e^{-\alpha_j}$ where $\alpha_j = n\pi_j$, for n large. Good (1953) observed that $\sum_{j=1}^{k} e^{-\alpha_j}$ may be approximated by $\Delta(t) = \sum_{j=1}^{k} e^{-\alpha_j}(1-e^{-\alpha_j t})$, for t large. $\Delta(t)$ can be interpreted as the expected number of species that would be seen in an additional sample of size $n\cdot t$, and were not seen in the initial sample. Now $\Delta(t)$ can be expanded as:

$$\Delta(t) = \sum_{j=1}^{k} e^{-\alpha_j} \left(\sum_{i=1}^{\infty} (-1)^{i+1} \frac{(\alpha_j t)^i}{i!} \right) \tag{4.1}$$

and $\sum_{j=1}^{k} \frac{(\alpha_j t)^i}{i!} \cdot e^{-\alpha_j}$ can be estimated unbiasedly by $t^i \sum_{i=1}^{k} I\{N_j=i\}$. The resulting $\hat{\Delta}(t)$ can always be defined since $\sum_j I\{N_j=i\} = 0$ for $i > \max(N_1,\ldots,N_k)$, and it is not hard to see that it is always an unbiased estimate of $\Delta(t)$. However $\hat{\Delta}(t)$ becomes unstable as t grows out of $[0,1)$, the region of absolute convergence of the series in $\hat{\Delta}$. It does not converge as $t \to \infty$ even though $\Delta(t) \to \Delta(\infty)$, the quantity we want to estimate.

Good and Toulmin (1956) and Efron and Thisted (1976) considered (biased) estimators based on approximating $\Delta(t)$ by a partial sum in i of a series obtained by applying a summability method to the original series (4.1). Efron and Thisted's methods gave what seem to be reasonable estimates for the number of words Shakespeare knew but did not use. In estimating $\Delta(\infty)$ their method requires setting two tuning constants, (i) t, such that $\Delta(t) \approx \Delta(\infty)$, and (ii) the number of terms taken in the partial sum mentioned above. The constants used for the Shakespeare data applied to ours give substantially lower answers than ours. On the other hand $\hat{\underline{k}}_1$ gives lower results than those of Efron and Thisted, for the Shakespeare data. This is not surprising since we expect the π_j are far from equal in that case. Here is a combination of the two methods which evades the need for the tuning constants and yields high values for both data sets. Use $\hat{\Delta}(1)$ to estimate $\Delta(1)$ and then estimate the remainder by our method., Formally:

$$\Delta(t) = \sum_{j=1}^{k} \sum_{r=n}^{n(t+1)-1} \pi_j (1-\pi_j)^r \tag{4.2}$$

Arguing as in section 2,

$$D \geq \Delta(t) + \frac{\Delta_1 \cdot (\frac{\Delta_n}{\Delta_1})^{\frac{n(t+1)}{n-1}}}{1 - (\frac{\Delta_n}{\Delta_1})^{1/n-1}} \tag{4.3}$$

and for large n

$$D \geq \Delta(t) + \frac{\sum_{j=1}^{k} \alpha_j e^{-\alpha_j}}{\log \frac{\sum \alpha_j}{\sum \alpha_j e^{-\alpha_j}}} \cdot \left(\frac{\sum \alpha_j e^{-\alpha_j}}{\sum \alpha_j} \right)^t .$$

Using our previous estimator and Good's estimator for $\Delta(1)$, we get

$$\hat{\underline{D}} = \hat{\Delta}(1) + \frac{(\sum_{j=1}^{k} I\{N_j=1\})^2}{(\sum_{j=1}^{k} N_j) \log \frac{\sum N_j}{\sum I\{N_j=1\}}} . \tag{4.4}$$

The results are consistent with Table 2.

	$\hat{\underline{k}}_1$	$\hat{\underline{D}} + T_n$
Chicago	82	82
Orange County	283	287
San Francisco	88	86

Using (4.4) to estimate the number of words Shakespeare knew but did not use increases our lower bound estimate by almost 300% but still is only about one-third of Efron and Thisted's estimate.

We finally compare our lower bounds with estimates obtained by a parametric empirical Bayes method proposed by Fisher et al (1943).

We assume π_j, $j = 1,2,\ldots,k$ i.i.d. with density $f(x) = (b+1)(1-x)^b$, $0 \le x \le 1$ and $b > 0$. As $k \to \infty$

$$\frac{1}{k}\sum_{j=1}^{k} \pi_j \xrightarrow{P} \frac{1}{b+2} ,$$

$$\frac{1}{k}\sum_{j}^{k} \pi_j(1-\pi_j)^{n-1} \xrightarrow{P} \frac{(b+1)}{(b+n+1)(b+n)} ,$$

and

$$\frac{1}{k}\sum_{j=1}^{k} (1-\pi_j)^{n} \xrightarrow{P} \frac{(b+1)}{(b+n+1)} .$$

Note that

$$\frac{\sum_j \pi_j(1-\pi_j)^{n-1}}{\sum_j \pi_j} \xrightarrow{P} \frac{(b+2)(b+1)}{(b+n+1)(b+n)} \approx \frac{b^2}{(b+n)^2} .$$

Use the method of moments to estimate

$$n\sum_{j=1}^{k} \pi_j \quad \text{by} \quad \sum_{j=1}^{k} N_j ,$$

$$n\sum_{j}^{k} \pi_j(1-\pi_j)^{n-1} \quad \text{by} \quad \sum_{j=1}^{k} I\{N_j=1\} ,$$

b by

$$\hat{b} = n\sqrt{\sum_j I\{N_j=1\} / \sum_j N_j}\ \left(1 - \sqrt{\sum_j I\{N_j=1\} / \sum_j N_j}\right)^{-1} \qquad (4.5)$$

and D by

$$\hat{D} = \frac{1}{n}\sum_j N_j \cdot \frac{(\hat{b}+1)(\hat{b}+2)}{(\hat{b}+n+1)} \approx \sum_j N_j \frac{(\hat{b}/n)^2}{(\hat{b}/n)+1} . \qquad (4.6)$$

From our data, we obtain

Table 3. *Lower bound and parametric empirical Bayes estimates*

	Chicago	Orange County	San Francisco
$\hat{6}/n$	3.8265	6.5548	2.4691
$\hat{D}$	116	461	84
$\underline{\hat{D}}_1$	53	216	52

We see that the parametric empirical Bayes approach yields estimates that are about twice our lower bound estimates.

APPENDIX A

We let k, n and the π_j depend on a hidden index s, so that as $s \to \infty$, $n(s) \to \infty$, $k(s) \to \infty$ and $\pi_j(s) \to 0$. We suppress the dependence on s when it is clearly understood.

ASSUMPTION $\tilde{A}$. (i) $G_s(x) = \frac{1}{k} \sum_{j=1}^{k} I\{n\pi_j \leq x\} \to G(x)$.

(ii) $\lim_{x\to\infty} G(x) = 1$, $\lim_{x\to 0} G(x) = 0$.

(iii) $\sup_s \int_0^\infty t^4 dG_s(t) < \infty$.

$\tilde{A}$ is a blanket assumption for this section.

Let $h: I \to R$, where $I = \{0,1,2,\ldots\}$, be bounded. Let $N = (N_0,N_1,\ldots,N_k)$ be multinomial $(n,\underset{\sim}{\pi})$, $\underset{\sim}{\pi} = (\pi_0,\pi_1,\ldots,\pi_k)$. We will differentiate between three situations and state corresponding Theorems 1-3.

1. $\pi_0 \to 0$, $k\pi_0 \to \infty$
2. $\underline{\lim}\, \pi_0 > 0$, $k\pi_0 \to \infty$
3. $k\pi_0 = O(1)$

Before stating the theorems, we will prove two lemmas.

LEMMA 1. *If* $\pi_0 \to 1$, *then* $\frac{k}{n} \to 0$. *Otherwise* $0 < \underline{\lim}\, \frac{k}{n} \leq \overline{\lim}\, \frac{k}{n} < \infty$.

Proof. By $\tilde{A}$,

$$\frac{n}{k}(1-\pi_0) = \frac{1}{k} \sum_{j=1}^{k} n\pi_j \to \int_0^\infty t\, dG(t) \tag{1}$$

where $0 < \int_0^\infty t\, dG(t) < \infty$.

LEMMA 2. *Let* $M = n - N_0$. *Then*

$$\frac{M}{k} \xrightarrow{p} \int_0^\infty t \, dG(t) \ . \tag{2}$$

If $k\pi_0 \to \infty$, *then*

$$\frac{M - n\sum_{j=1}^{k} \pi_j}{\sqrt{n\pi_0(1-\pi_0)}} \rightsquigarrow N(0,1) \ .$$

Proof.
$$E[\frac{M}{k}] = \frac{1}{k} \sum_{j=1}^{k} n\pi_j \to \int_0^\infty t \, dG(t)$$

$$Var(\frac{M}{k}) = \frac{1}{k^2} n \cdot \pi_0(1-\pi_0) \sim \frac{1}{k}\pi_0 \int_0^\infty t \, dG(t) \to 0 \ .$$

Hence (2) is proved. To prove the asymptotic normality, we note that $k\pi_0 \to \infty$ implies that $n(1-\pi_0)\pi_0 \to \infty$ by (1) and hence $\frac{N_0 - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} \to N(0,1)$. But $M - n\sum_{j=1}^{k} \pi_j = -(N_0 - n\pi_0)$ which proves (3).

Given N_0 let $\tilde{N}_j$, $j = 1,2,\ldots,k$ be independent Poisson with parameter $M\mu_j$ and $\mu_j = \pi_j(1-\pi_0)^{-1}$. Let

$$w_a(t) = \sum_{r=0}^{\infty} h^a(r) e^{-t} t^r (r!)^{-1} \ , \quad a = 1,2$$

$$w_{11}(t) = \sum_{r=0}^{\infty} r \cdot h(r) e^{-t} t^r (r!)^{-1} \ .$$

and
$$w(t) = w_2(t) - (w_1(t))^2 \ .$$

THEOREM 1. *If* $k\pi_0 \to \infty$ *and* $\pi_0 \to 0$ *then*

$$\frac{\sum_{j=1}^{k} \{h(N_j) - E[h(\tilde{N}_j)|N_0]\}}{\sigma(N_0)} \rightsquigarrow N(0,1) \tag{4}$$

where

$$\sigma^2(N_0) = \sum_{j=1}^{k} \mathrm{Var}(h(\tilde{N}_j)|N_0) - \frac{(\sum_{j=1}^{k} \mathrm{Cov}(\tilde{N}_j, h(\tilde{N}_j)|N_0))^2}{M} \tag{5}$$

$$= k\left(\int_0^\infty w(t)\,dG(t) - \frac{(\int_0^\infty (w_{11}(t) - t w_1(t))\,dG(t))^2}{\int_0^\infty t\,dG(t)}\right) + o_p(k) .$$

$$\sum_{j=1}^{k} E[h(\tilde{N}_j)|N_0] = \sum_{j=1}^{k} w_1(n\pi_j) + o_p(k^{1/2}) . \tag{6}$$

Proof. We establish the correspondence with Steck's theorem 2.2. All computations are conditional on N_0. The conditioning vector V_s of Steck is one dimensional and lattice given N_0.

$$V_s = \sum_{j=1}^{k} N_j - M \sum_{j=1}^{k} \mu_j .$$

The vector U_s is one dimensional

$$U_s = \sum_{j=1}^{k} (h(\tilde{N}_j) - E[h(\tilde{N}_j)|N_0]) .$$

To prove (4) we need to check conditions (ii)-(v), of Steck's Theorem 2.2.

Condition (ii).

$$\sum_{j=1}^{k} \mathrm{Var}(\tilde{N}_j|N_0) \xrightarrow{p} \infty$$

$$\sum_{j=1}^{k} \mathrm{Var}(h(\tilde{N}_j)|N_0) \xrightarrow{p} \infty$$

We note that $\sum_{j=1}^{k} V(\tilde{N}_j|N_0) = M$ and by Lemma 2 converges to ∞. As for the second term,

$$\sum_{j=1}^{k} \mathrm{Var}(h(\tilde{N}_j)|N_0) = \sum_{j=1}^{k} w(\tfrac{M}{k}\mu_j) = k\int_0^\infty w(\tfrac{M}{k}\cdot\frac{t}{\int_0^\infty v dG_k(v)})dG_k(t) \ .$$

But $t \to w(at)$ is bounded and continuous in (a,t) since h is bounded. Using (2) of Lemma 2 we get

$$\sum_{j=1}^{k} w(\tfrac{M}{k}\mu_j) \sim k\int_0^\infty w(t)dG(t) \to \infty \ ,$$

since $w(t) > 0$ for $t > 0$.

Condition (iii). Argument of Steck's theorem 2.4 shows that (iii) will hold provided that:

$$P\lim \mathrm{Corr}^2\Big(\sum_{j=1}^{k}\tilde{N}_j, \sum_{j=1}^{k} h(\tilde{N}_j)|N_0\Big) < 1 \qquad (7)$$

$$P\lim \frac{1}{k}\mathrm{Var}\Big(\sum_{j=1}^{k} h(N_j)|N_0\Big) = \int_0^\infty w(t)dG(t) > 0 \qquad (8)$$

$$P\lim \frac{1}{k}\mathrm{Var}\Big(\sum_{j=1}^{k} N_j|N_0\Big) = \int_0^\infty t dG(t) > 0$$

(where P lim denotes limit in probability) and

$$\underline{\lim}\ \frac{1}{k}\sum_{j=1}^{k} I\{\varepsilon \le n\pi_j \le \tfrac{1}{\varepsilon}\} > 0 \quad \text{for some} \quad \varepsilon > 0 \ . \qquad (9)$$

But (7) equals $$\frac{\Big(\sum_{j=1}^{k} w_{11}(\tfrac{M}{k}\mu_j) - M\sum_{j=1}^{k} w_1(\tfrac{M}{k}\mu_j)\Big)^2}{M\sum_{j=1}^{k} w(\tfrac{M}{k}\mu_j)} \ .$$

Arguing as for (ii), (7) is equivalent to

$$\frac{\Big(\int_0^\infty w_{11}(t)dG(t) - \int_0^\infty w_1(t)dG(t)\cdot\int_0^\infty t dG(t)\Big)^2}{\int_0^\infty t dG(t)\cdot\int_0^\infty w(t)dG(t)} < 1. \qquad (10)$$

Let $T \sim G$ and N given T be Poisson(T). The expression in (10) equals $\frac{E^2[Cov(h(N),N|T)]}{E[Var(h(N)|T)]\cdot E[Var(N|T)]}$. Since h is bounded, h is not linear in N and hence for every T such that $T > 0$ we have

$$|Cov(h(N),N|T)| < Var^{1/2}(h(N)|T)Var^{1/2}(N|T) .$$

Hence

$$|E[Cov(h(N),N|T)]| < E[Var^{1/2}(h(N)|T)\cdot Var^{1/2}(N|T)]$$

but

$$E[Var^{1/2}(h(N)|T)Var^{1/2}(N|T)] \leq E^{1/2}[Var(h(N)|T)]\cdot E^{1/2}[Var(N|T)]$$

Since $P(T>0) > 0$, (7) follows. (8) was essentially proved under (ii). (9) is immediate by A(ii).

Condition (iv). Follows from (7).

Condition (v) (Part 1). This reduces to

$$\frac{1}{k^2}\sum_{j=1}^{k} E[(h(\tilde{N}_j) - E[h(\tilde{N}_j)|N_0])^4|N_0] \xrightarrow{p} 0 , \qquad (11)$$

$$\frac{1}{k^2}\sum_{j=1}^{k} E[(\tilde{N}_j - E[\tilde{N}_j|N_0])^4|N_0] \xrightarrow{p} 0 . \qquad (12)$$

(11) is immediate since h is bounded while (12) is equal to

$$\frac{1}{k^2}\sum_{j=1}^{k}\{M\mu_j + 3(M\mu_j)^2\} \sim \frac{1}{k}\left(\int_0^\infty t dG(t) + 3\int_0^\infty t^2 dG(t)\right) \to 0 .$$

Condition (v) (Part 2). This reduces to

$$\overline{\lim}\sup_j \frac{E[(\tilde{N}_j - E[\tilde{N}_j|N_0])^4|N_0]}{k} \leq C < \infty . \qquad (13)$$

But (13) equals

$$\overline{\lim}\ \sup_j \frac{M\mu_j + 3(M\mu_j)^2}{k} \le \overline{\lim}\ \frac{1}{k} \sum_{j=1}^{k} \{M\mu_j + 3(M\mu_j)^2\}$$
$$\sim \int_0^\infty t\, dG(t) + 3\int_0^\infty t^2 dG(t) < \infty .$$

Steck's conditions have been checked and (5) has been proved in checking (iii).

Proof of (6). Note first that if $|h| \le M$ so are $|w|$, $|w'|$ and $|w''|$. Since $E[h(\tilde{N}_j)|N_j] = w_1(M\mu_j)$, we can Taylor expand around $n\pi_j$ to get

$$\sum_{j=1}^{k} \{w_1(M\mu_j) - w_1(n\pi_j)\} = \sum_{j=1}^{k} w_1'(n\pi_j)(M\mu_j - n\pi_j) + O_p\Big(\sum_{j=1}^{k} (M\mu_j - n\pi_j)^2\Big) .$$

But

$$\sum_{j=1}^{k} w'(n\pi_j)(M\mu_j - n\pi_j) = (M - n(1-\pi_0)) \sum_{j=1}^{k} w'(n\pi_j)\mu_j \qquad (14)$$
$$\sim k\int_0^\infty t w'(t) dG(t) \left(\frac{M}{n(1-\pi_0)} - 1\right)$$
$$= O_p\left(k \cdot \left(\frac{\pi_0}{(1-\pi_0)n}\right)^{1/2}\right) = O_p((k\pi_0)^{1/2})$$
$$= o_p(k^{1/2})$$

since $\pi_0 \to 0$. The remainder is $O_p((M - n(1-\pi_0))^2 \sum_{j=1}^{k} \mu_j^2) =$ $O_p(n\pi_0(1-\pi_0)kn^{-2}) = o_p(1)$. Q.E.D.

THEOREM 2. *If* $\underline{\lim}\ \pi_0 > 0$,

$$\frac{\sum_{j=1}^{k} \{h(N_j) - w_1(n\pi_j)\}}{\tau(N_0)} \rightsquigarrow N(0,1) \qquad (15)$$

where

$$\tau^2(N_0) = \sum_{j=1}^{k} \mathrm{Var}(h(\tilde{N}_j)|N_0) - \frac{\left(\sum_{j=1}^{k} \mathrm{Cov}(\tilde{N}_j, h(\tilde{N}_j)|N_0)\right)^2}{M}$$

$$+ E\left(\sum_{j=1}^{k} w_1(M\mu_j) - w_1(n\pi_j)\right)^2$$

$$= K\left\{\int_0^\infty w(t)\,dG(t) - \frac{\left(\int_0^\infty (w_{11}(t) - tw_1(t)\,dG(t)\right)^2}{\int_0^\infty t\,dG(t)} + \pi_0 \frac{\int_0^\infty tw'(t)\,dG(t)}{\int_0^\infty t\,dG(t)}\right\}$$

$$+ o(k) .$$

Proof. We only used $k\pi_0 \to \infty$ in the proof of Theorem 1 save for the expression in (6). That term now yields

$$\sum_{j=1}^{k} \{w_1(M\mu_j) - w_1(n\pi_j)\} = \sum_{j=1}^{k} w'(n\pi_j)(M\mu_j - n\pi_j)$$

$$+ \sum_{j=1}^{k} w''(n\pi_j)\tfrac{1}{2}(M\mu_j - n\pi_j)^2 + O_p \sum_{j=1}^{k} |M\mu_j - n\pi_j|^3 .$$

By (14)

$$E\left(\sum_{j=1}^{k} w'(n\pi_j)(M\mu_j - n\pi_j)\right)^2 \sim \frac{k^2}{n}\,\pi_0(1-\pi_0)^{-1}\left(\int_0^\infty tw'(t)\,dG(t)\right)^2$$

$$\sim \frac{k\pi_0\left(\int_0^\infty tw'(t)\,dG(t)\right)^2}{\int_0^\infty t\,dG(t)} .$$

Also

$$\sum_{j=1}^{k} w''(n\pi_j)(M\mu_j - n\pi_j)^2 = (M - n(1-\pi_0))^2 \sum_{j=1}^{k} w''(n\pi_j)\mu_j^2 \qquad (16)$$

$$\sim (M - n(1-\pi_0))^2 \int_0^\infty w''(t)t^2\,dG(t)\,\frac{1}{k\left(\int_0^\infty t\,dG(t)\right)^2}$$

$$= O_p(1) .$$

Finally,

$$E\Big(\sum_{j=1}^{k}|M\mu_j-n\pi_j|^3\Big) \sim E|M-n(1-\pi_0)|^3\,\frac{\int_0^\infty t^3\,dG(t)}{k^2\big(\int_0^\infty t\,dG(t)\big)^3}$$

$$= O(E^{3/4}(M-n(1-\pi_0))^4\cdot k^{-2}) = O(k^{-1/2})\ .$$

The result follows. Q.E.D.

THEOREM 3. *If* $\pi_0 \to 0$ *but* $k\pi_0 = O(1)$ *then*

$$\frac{\sum_{j=1}^{k}\{h(N_j)-w_1(n\pi_j)\}}{\sigma_n^2} \to N(0,1)$$

where

$$\sigma_n^2 = \sum_{j=1}^{k}\mathrm{Var}(h(N_j^*)) - \frac{1}{n}\Big(\sum_{j=1}^{k}\mathrm{Cov}(N_j^*,h(N_j^*))\Big)^2 \qquad (17)$$

$$\sim k\left(\int_0^\infty w(t)\,dG(t) - \frac{\big(\int_0^\infty (w_{11}(t)-tw_1(t))\,dG(t)\big)^2}{\int_0^\infty t\,dG(t)}\right)$$

and the N_j^* *are independent* Poisson$(n\pi_j)$

Proof. By Lemma 1, $\pi_0 \to 0$ implies $n\pi_0 = O(1)$ and $N_0 = o_p(k^{1/2})$. Proceed as in the proof of Theorem 1 by applying Steck's theorem with $U_s = \sum_{j=0}^{\infty}(h(N_j^*)-Eh(N_j^*))$ and $V_s = \sum_{i=0}^{\infty}(N_j^*-n\pi_j)$. Note that

$$\sum_{j=0}^{k}\mathrm{Var}(h(N_j^*)) - \frac{1}{n}\Big(\sum_{j=0}^{k}\mathrm{Cov}(N_j^*,h(N_j^*))\Big)^2$$

$$\sim k\left(\int_0^t w(t)\,dG(t) - \frac{\big(\int_0^\infty (w_{11}(t)-tw_1(t))\,dG(t)\big)^2}{\int_0^\infty t\,dG(t)}\right)$$

and

$$h(N_0^*) - E[h(N_0^*)] = o_p(k^{1/2})$$

and Theorem 3 follows. This covers the important case $\pi_0 = 0$. Q.E.D.

Finally we shall use

COROLLARY. *If* $\tilde{A}$ *holds and* h *is bounded,*

$$\frac{\sum_{j=1}^{k} h(N_j)}{k} \to \int_0^\infty w_1(t)dG(t) .$$

Proof. Since $\frac{1}{k}\sum_{j=1}^{k} w_1(n\pi_j) \to \int_0^\infty w_1(t)dG(t)$ the result follows immediately from Theorems 1-3. Q.E.D.

References

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species. Biometrika 63, 435-447.

Feller, W. (1966). An Introduction to Probability Theory and its Applications, 2nd Ed., New York, Wiley.

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

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

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

Goodman, L. A. (1949). On the estimation of the number of classes in a population. Ann. Math. Statist. 20, 572-579.

Morris, C. (1975). Central limit theorems for multinomial sums Ann. Statist. 3, 165-188.

Seber, G.A.F. (1973). The Estimation of Animal Abundance. New York, Hafner Press.

Steck, G. P. (1957). Limit theorems for conditional distributions. University of California Publications in Statistics 2:12, 235-284.