The Annals of Statistics
1987, Vol. 15, No. 4, 1466–1481

DISTRIBUTIONS ON PARTITIONS

BY ROBERT KEENER,[1] EDWARD ROTHMAN[2] AND NORMAN STARR

University of Michigan

A two parameter family of distributions on partitions is obtained by mixing a multinomial with a symmetric Dirichlet distribution. Estimates for the parameters are proposed and studied in various asymptotic limits.

1. Introduction. This paper considers estimation of the parameters for a family of distributions on partitions. To obtain these distributions, imagine an infinite population classified into an unknown number $m \geq 1$ of species. Further suppose that this population evolves with time and that the species proportions $p_1, \ldots, p_m$ have an equilibrium distribution which is a symmetric Dirichlet distribution $D(A, \ldots, A)$. At some point in time, after equilibrium has been achieved, a researcher takes a random sample from the population. The sample size n is partitioned by the species represented in the sample. Thus the distribution of this random partition is that induced by mixing the multinomial distribution with $D(A, \ldots, A)$. In situations where the "names" of the species have no import, the partition is the only relevant data from the experiment.

This model or family of distributions arises naturally in genetics as the sampling distribution for neutral alleles. In that context, the species are types of alleles at a genetic locus. See Rothman and Templeton (1980) for further discussion, or Wright (1969) for a derivation based on diffusion approximations.

In most applications, it is perhaps more natural to view a random partition as induced from a multinomial sample with *fixed* but unknown proportions $p_1, \ldots, p_m$. An empirical Bayes approach in these problems leads naturally to estimation for the model considered in this paper. This approach is one way to allow an unknown number of species and avoid overparameterization.

Instead of denoting partitions as the ordered species counts, it is more convenient to represent a partition as a sequence $G(1), G(2), \ldots$, where $G(i)$ is the number of species which occur i times. $D = \sum_{i=1}^{\infty} G(i)$ is the number of species observed and the sample size is $n = \sum_{i=1}^{\infty} iG(i)$. The distributions under consideration are given by

$$(1.1)\quad P(G(i) = g(i),\ i = 1,2,\ldots) = \binom{m}{d} \frac{d!}{\prod_{i=1}^{\infty} g(i)!} \frac{\prod_{i=1}^{\infty} \binom{i + A - 1}{i}^{g(i)}}{\binom{n + mA - 1}{n}},$$

Received March 1985; revised March 1987.
[1]Research supported in part by NSF Grant MCS-8102080.
[2]Research supported in part by Department of Energy Contract E(11-1)(2828).
AMS 1980 *subject classifications*. Primary 62E20; secondary 62F10.
Key words and phrases. Partitions, convergence in distribution, maximum likelihood estimation, unbiased estimation, Ewens' sampling formula, Maxwell–Boltzmann, Bose–Einstein.

where $d = \Sigma g(i)$ and $x! = \Gamma(x + 1)$ is used to define combinations when the arguments are not integers. The unknown parameters in this model are $A > 0$ and $m = 1, 2, \ldots$. To obtain (1.1), let Y_i be the number of times species i is observed. Then

$$P(Y_i = y_i,\ i = 1, \ldots, m) = E\binom{n}{y_1 \cdots y_m} \prod_{i=1}^{m} p_i^{y_i}$$

$$= \prod_{i=1}^{m} \binom{A + y_i - 1}{y_i} \Big/ \binom{mA + n - 1}{n}. \tag{1.2}$$

Equation (1.1) follows by a simple counting argument.

As limiting and special cases, the model (1.1) incorporates three of the most natural distributions for partitions. If $A = 1$, the distribution is that induced by Bose–Einstein sampling as the probabilities in (1.2) all equal $\binom{m+n-1}{n}^{-1}$. A Bayesian approach to inference in this case has been studied by Hill (1979). As $A \to \infty$ the distributions converge weakly to the distribution induced by Maxwell–Boltzmann sampling (see Theorem 3.7). Estimation in this case is considered by Lewontin and Prout (1956); see Johnson and Kotz (1969). Finally, if $m \to \infty$ and $A \to 0$ so that $mA \to \omega$, the distributions converge weakly to Ewens' sampling formula (see Theorem 3.8).

Other authors have studied model (1.1), especially Chen (1980, 1981a, b). He gives an alternative derivation of (1.1) that may be more natural or convenient in some settings (cf. the proof of Theorem 3.3). In this derivation, $Y_1, \ldots, Y_m$ begin as i.i.d. with a negative binomial distribution. Equation (1.1) then gives the distribution of $G(1), G(2), \ldots$ conditional on $\Sigma_1^m Y_i = n$. His main concern in the (1980) paper is degenerate convergence of $G(i)/m$ to Zipf's law. All three papers have results about joint normality for the $G(i)$ as $n \to \infty$ with $m/n \to \lambda$: weak convergence in the (1980) paper, local convergence in (1981b) and weak convergence for a more general model in (1981a). These results complement our Theorems 3.3 and 3.4 which give asymptotic normality for D and for the maximum likelihood estimate of m in the same limit.

Different stochastic models for partitions that have been studied include an empirical Bayes model due to Fisher, Corbet and Williams (1943) and a nonparametric model due to Good and Toulmin (1956). Fisher's model is obtained as follows. Let $\lambda_1, \ldots, \lambda_m$ be i.i.d. from a gamma distribution with parameters A and β. Conditional on $\lambda_1, \ldots, \lambda_m$, $Y_1, \ldots, Y_m$ are independent with $Y_i \sim$ Poisson (λ_i). Call the positive Y's which determine the observed partition $\tilde{Y}_1, \ldots, \tilde{Y}_D$. Conditional on D, these are i.i.d. with

$$P(\tilde{Y} = x) \propto \frac{\Gamma(x + A)}{x!\Gamma(1 + A)} \left(\frac{\beta}{1 + \beta}\right)^{x-1}, \qquad x = 1, 2, \ldots. \tag{1.3}$$

This distribution is the negative binomial distribution conditioned to be positive and Fisher's model for partitions is that induced by this distribution for $\tilde{Y}_1, \ldots, \tilde{Y}_D$. The main difference between our model and Fisher's is that we have D a random variable and n constant, while Fisher has D constant and n

random. Although the derivation leading to Fisher's model only makes sense if $A > 0$, (1.3) gives probability distributions for partitions for any $A > -1$. In Section 4, we analyze word frequency data for Shakespeare to compare our methods with an analysis by Efron and Thisted (1976) using Fisher's model and Good and Toulmin's model. Fisher's model fits better than ours primarily because negative values for A are allowed. The m.l.e. for A in Fisher's model is -0.3954.

Section 2 considers estimation when A is known. In this case D is complete and sufficient. Estimates are given for m and $EG(j)$.

Asymptotic distribution theory is used in Section 3 to study the performance of these estimators in various limiting situations.

In Section 4, data sets are analyzed and Section 5 gives our conclusions.

2. Estimation with A known.

LEMMA 2.1. *The statistic D is complete and sufficient for the family* (1.1) *as m varies with A fixed. The distribution of D is given by*

$$P(D = d) = \binom{m}{d} H(d, n) \Big/ \binom{n + mA - 1}{n}, \tag{2.1}$$

where

$$H(d, n) = \sum_{j=0}^{d-1} (-1)^j \binom{d}{j} \binom{n + Ad - Aj - 1}{n}. \tag{2.2}$$

PROOF. By symmetry

$$P(D = d) = \binom{m}{d} P(Y_i \geq 1 \text{ for } i \leq d,\ Y_i = 0 \text{ for } i > d).$$

Using Boole's formula,

$$\begin{aligned} &P(Y_i \geq 1 \text{ for } i \leq d,\ Y_i = 0 \text{ for } i > d) \\ &\quad = \sum_{j=0}^{d-1} \binom{d}{j} (-1)^j P(Y_i = 0 \text{ for } i > d - j). \end{aligned} \tag{2.3}$$

Now, since $\sum_{i=1}^{d-j} p_i \sim \beta((d-j)A, (m-d+j)A)$, we have

$$\begin{aligned} P(Y_i = 0 \text{ for } i > d - j) &= E\left(\sum_{i=1}^{d-j} p_i \right)^n \\ &= \frac{\Gamma(mA)\Gamma(n + (d-j)A)}{\Gamma(n + mA)\Gamma((d-j)A)} \end{aligned}$$

and (2.1) follows. It is worth noting that when $d > n$, the left-hand side of (2.3) equals 0. Consequently $H(d, n) = 0$ for $d > n$, which is hard to verify directly from (2.2). D is sufficient because (1.1) divided by (2.1) is functionally independent of m. To show that D is complete, suppose $Ef(D) = 0$ for all $m \geq 1$. When $m = 1$, $D = 1$ a.s., so $f(1) = 0$. When $m = 2$, since $P(D = 2) > 0$ and

$P(D > 2) = 0$, we get $f(2) = 0$, etcetera. This argument shows that D is complete even for the restricted parameter space $m \in \{1, \ldots, n\}$. □

One estimate for m is the maximum likelihood estimator $\tilde{m}$.

LEMMA 2.2. *If $d = n$, the likelihood approaches 1 as $m \to \infty$ and there is no maximum likelihood estimate for m. If $d < n$, the likelihood attains its maximum at least once and at most twice, in which case the two values are adjacent integers.*

PROOF. Using Stirling's formula, as $m \to \infty$,

$$L(m) = \frac{m!}{(m-d)!}\frac{(mA-1)!}{(n+mA-1)!}\frac{n!}{d!}H(d,n) \sim \frac{m^d}{(mA)^n}\frac{n!}{d!}H(d,n).$$

When $d < n$, this approaches 0 as $m \to \infty$, so the maximum is attained. To finish the proof we will show that if m is allowed to vary over $[0, \infty)$, L will attain its maximum at a unique point m^* and decrease as m moves away from m^*. Suppose

$$\frac{\partial}{\partial m}\log L(m) = \sum_{i=0}^{d-1}\frac{1}{m-i} - \sum_{i=0}^{n-1}\frac{1}{m+i/A} = 0.$$

Then

$$\frac{\partial^2}{\partial m^2}\log L(m) = -\sum_{i=0}^{d-1}\frac{1}{(m-i)^2} + \sum_{i=0}^{n-1}\frac{1}{(m+i/A)^2} < 0.$$

Consequently L has no local minima and this completes the proof. □

An alternative estimator for m derived to be unbiased when $m \le n$ is given by

$$\hat{m}(D) = D + D\frac{H(D-1,n)}{H(D,n)}.$$

LEMMA 2.3. *$\hat{m}$ is unbiased for $m \le n$. Consequently $\hat{m}$ is UMVUE for $m \in \{1, \ldots, n\}$.*

PROOF. Using (2.1),

$$\begin{aligned} E\hat{m}(D) &= ED + \sum_{d=2}^{m} d\binom{m}{d}H(d-1,n)\Big/\binom{n+mA-1}{n} \\ &= ED + m\sum_{d=1}^{m-1}\binom{m-1}{d}H(d,n)\Big/\binom{n+mA-1}{n} \\ &= ED + m\binom{n+mA-A-1}{n}\Big/\binom{n+mA-1}{n}. \end{aligned}$$

The lemma now follows from the formula for ED given in Corollary 3.2. □

To use the estimator $\hat{m}$, a practitioner must calculate H. Using (2.2) for this task is not advisable because the alternating character of the sum may lead to numerical instability. A better method is to use the recurrence relation in the next lemma.

LEMMA 2.4. *$H(1, n) = \binom{n + A - 1}{n}$ and $H(n, n) = A^n$. For $2 \le d \le n$,*

$$H(d, n + 1) = \{(n + Ad)H(d, n) + AdH(d - 1, n)\}/(n + 1).$$

PROOF. The expression for $H(1, n)$ comes directly from (2.2). To show that $H(n, n) = A^n$, if $m = n$,

$$\begin{aligned} H(n, n)\Big/\binom{n + nA - 1}{n} &= P(D = n) \\ &= P(Y_i = 1, 1 \le i \le n) \\ &= n!E\prod_{i=1}^{n} p_i \\ &= \frac{n!A^n\Gamma(nA)}{\Gamma(n + nA)}. \end{aligned}$$

Finally, using (2.2),

$$\begin{aligned} H(d, n + 1) &= \sum_{j=0}^{d} (-1)^j \binom{d}{j}\binom{n + Ad - Aj}{n + 1} \\ &= \sum_{j=0}^{d} (-1)^j \left\{ \frac{n + Ad}{n + 1}\binom{n + Ad - Aj - 1}{n}\binom{d}{j} \right. \\ &\qquad \left. - \frac{Ad}{n + 1}\binom{d}{j - 1}\binom{n + A(d - 1) - A(j - 1) - 1}{n} \right\} \end{aligned}$$

and the lemma follows. □

The recursion can also be derived probabilistically by calculating the chance that $D = d$ in a sample of size $n + 1$ by conditioning on the value of D after sampling n. After observing d in a sample of size n, the chance the next observation is a new species is $[A(m - d)]/(n + mA)$.

Assessing how well the model fits a given data set is an important task because the estimate of m is very model dependent. One possible way to investigate the fit is to compare the observed $G(1), G(2), \ldots$ with estimates of their expected values $q_i = EG(j)$. Good estimates of $EG(j)$ are described in the next theorem. Using these estimates, a natural test statistic for a formal goodness of fit test is $\chi^2 = \Sigma(G(j) - \hat{q}_j)^2/\hat{q}_j$. An interesting open problem is to find approximations for the null distribution of χ^2 (perhaps in the limit used in Theorem 3.3). Standard results are not directly applicable due to the lack of independence in this model.

THEOREM 2.5. *The UMVUE for q_j is*

$$E(G(j)|D) = D\binom{j+A-1}{A-1}\frac{H(D-1,n-j)}{H(D,n)}.$$

PROOF. Let

$$S(d,n) = \left\{g \in \{1,2,\dots\}^{\infty} : \sum_{i=1}^{\infty} g(i) = d,\ \sum_{i=1}^{\infty} ig(i) = n\right\}.$$

The conditional distribution of G, given D, is given by

$$P(G = g|D = d) = \frac{d!}{\prod_{i=1}^{\infty} g(i)!} = \frac{\prod_{i=1}^{\infty}\binom{i+A-1}{A-1}^{g(i)}}{H(d,n)}$$

for $g \in S(d,n)$. Let $\tilde{g}(i) = g(i)$ for $i \neq j$ and $\tilde{g}(j) = g(j) - 1$. Then

$$\begin{aligned} E(G(j)|D = d) &= \sum_{g \in S(d,n)} \frac{d!}{\prod_{i=1}^{\infty}\tilde{g}(i)!}\frac{\prod_{i=1}^{\infty}\binom{i+A-1}{A-1}^{\tilde{g}(i)}}{H(d,n)}\binom{j+A-1}{A-1} \\ &= \sum_{\tilde{g} \in S(d-1,n-j)} \frac{(d-1)!}{\prod_{i=1}^{\infty}\tilde{g}(i)}\left\{\prod_{i=1}^{\infty}\binom{i+A-1}{A-1}^{\tilde{g}(i)}\right\}\frac{d\binom{j+A-1}{A-1}}{H(d,n)} \\ &= d\binom{j+A-1}{A-1}\frac{H(d-1,n-j)}{H(d,n)} \end{aligned}$$

and the theorem follows from the Lehmann–Scheffé theorem. □

The next corollary gives a formula for $EG(j)$ and can be used substituting $\tilde{m}$ for m to find the maximum likelihood estimates of these quantities.

COROLLARY 2.6.

$$EG(j) = m\binom{j+A-1}{A-1}\binom{n-j+mA-A-1}{n-j}\Big/\binom{n+mA-1}{n}.$$

PROOF. Follows immediately from (2.1) and Theorem 2.5. □

3. Asymptotic distribution theory. A key tool is the generating function for D given in the next lemma. Let

$$\theta^{(j)} = \frac{(n+mA-1-Aj)!(mA-1)!}{(n+mA-1)!(mA-1-Aj)!},$$

so that

$$P(D = d) = \sum_{j=1}^{d}(-1)^{d-j}\frac{m!}{(m-d)!j!(d-j)!}\theta^{(m-j)}. \tag{3.1}$$

Since $H(n, d) = 0$ for $d > n$, this formula holds for $1 \le d \le m$ even when $n < m$.

LEMMA 3.1. *For any complex* x,

$$Ex^D = \sum_{j=0}^{m-1} \binom{m}{j} x^{m-j}(1-x)^j \theta^{(j)}.$$

PROOF. Using (3.1),

$$\begin{aligned} Ex^D &= \sum_{d=1}^{m} \sum_{j=1}^{d} (-1)^{d-j} \frac{m! x^d \theta^{(m-j)}}{(m-d)!\, j!\, (d-j)!} \\ &= \sum_{j=1}^{m} \sum_{d=j}^{m} (-x)^{d-j} \binom{m-j}{d-j} \binom{m}{j} \theta^{(m-j)} x^j \\ &= \sum_{j=1}^{m} (1-x)^{m-j} x^j \binom{m}{j} \theta^{(m-j)} \end{aligned}$$

and the lemma follows. □

Differentiation in Lemma 3.1 gives the next corollary.

COROLLARY 3.2. *The mean and variance of* D *are*

$$ED = m(1 - \theta^{(1)})$$

and

$$\operatorname{Var} D = m\left(\theta^{(1)} + (m-1)\theta^{(2)} - m\theta^{(1)^2}\right).$$

Define

$$\theta = \left(\frac{mA - 1}{n + mA - 1}\right)^A.$$

Using Stirling's formula, $E(m - D) = m\theta + O(1)$ as $m \to \infty$ uniformly in n, and keeping an extra term in Stirling's formula, $\operatorname{Var} D = \sigma^2 m + O(1)$ as $m \to \infty$ uniformly in n, where $\sigma^2 = \theta(1-\theta) - A\theta^2(1 - \theta^{1/A})$.

THEOREM 3.3. *If* $m, n \to \infty$ *so that* $m/n \to K > 0$, *then*

$$P(D = d) = \frac{1}{\sigma\sqrt{m}} \exp\left\{-\frac{(d - m + m\theta)^2}{2m\sigma^2}\right\} + o\left(\frac{1}{\sqrt{m}}\right),$$

uniformly in d. *Also if* $d, n \to \infty$ *so that* $\lambda = d/n \to \lambda_0 \in (0,1)$, *then*

$$H(d, n) \sim (1 + A\lambda^*)^n \left(\frac{\lambda}{\lambda^* - \lambda}\right)^d \sqrt{\frac{A\lambda(\lambda^* - \lambda)}{n(1 + A\lambda^*)\sigma^{*2}}},$$

where $\lambda^* = \lambda^*(d, n)$ *is the unique positive solution of*

$$\lambda = \lambda^*\left\{1 - \left(\frac{A\lambda^*}{1 + A\lambda^*}\right)^A\right\} \tag{3.2}$$

and $\sigma^{*2} = (\lambda^* - \lambda)[\lambda - A(\lambda^* - \lambda)/(1 + A\lambda^*)]/\lambda^{*2}$. *In this same limit,*

$$E[G(j)|D = d] \sim d\binom{j + A - 1}{A - 1}\frac{\lambda^* - \lambda}{\lambda}\Big/(1 + A\lambda^*)^j$$

and $\hat{m}(d) \sim n\lambda^*$.

PROOF. It will be convenient to use an alternative stochastic derivation of the model (1.1) for this proof. Let $X(p)$ have the negative binomial distribution

$$P(X(p) = k) = p^A(1 - p)^k\frac{\Gamma(A + k)}{k!\Gamma(A)}$$

for $k = 0, 1, \ldots$, and let $X_1(p), \ldots, X_m(p)$ be i.i.d. as $X(p)$. Let

$$N(p) = \sum_1^m X_i(p),$$

$$D(p) = \#\{i \le m: X_i(p) > 0\}$$

and

$$G^*(j, p) = \#\{i \le m: X_i(p) = j\},$$

for $j = 1, 2, \ldots$. Then a straightforward calculation shows that

$$P(G^*(i, p) = g(i), i = 1, 2, \ldots | N(p) = n) = P(G(i) = g(i), i = 1, 2, \ldots).$$

Letting $Z_i(p) = I\{X_i(p) > 0\}$, $D(p) = \sum_{i=1}^m Z_i(p)$ and by a local central limit theorem for lattice distributions [such as Theorem 22.1 of Bhattacharya and Rao (1976)],

$$\begin{aligned} &P(D(p) = d, N(p) = n) \\ &\quad = m^{-1}\{\operatorname{Var} Z(p)\operatorname{Var} X(p)\}^{-1/2} \\ &\qquad \times\phi\left(\frac{d - mEZ(p)}{\sqrt{m\operatorname{Var} Z(p)}}, \frac{n - mEX(p)}{\sqrt{m\operatorname{Var} X(p)}}, \rho(p)\right) + o(1/m), \end{aligned}$$

uniformly in d and n as $m \to \infty$, where ϕ is the standard bivariate normal density

$$\phi(x, y, \rho) = \frac{1}{\sqrt{2\pi(1 - \rho^2)}}\exp -\frac{1}{2}\left(\frac{x^2 - 2\rho xy + y^2}{1 - \rho^2}\right)$$

and

$$\begin{aligned} \rho(p) &= \operatorname{Cor}(X(p), Z(p)) \\ &= \left\{\frac{Ap(1 - p)p^A}{1 - p^A}\right\}^{1/2}. \end{aligned}$$

Inspection of the proof of this local central limit theorem shows that the result stated holds uniformly for p in a compact subset of $(0, 1)$. By a one-dimensional version of the same theorem,

$$P(N(p) = n) = \frac{1}{\sqrt{2\pi m \operatorname{Var} X(p)}} \exp\left\{ -\frac{(n - mEX(p))^2}{2m \operatorname{Var} X(p)} \right\} + o\left(\frac{1}{\sqrt{m}} \right),$$

uniformly in n and p in a compact subset of $(0, 1)$ as $m \to \infty$. By division, taking $p = mA/(n + mA)$ so that $EN(p) = n$,

$$P\left(D\left(\frac{mA}{n + mA} \right) = d \middle| N\left(\frac{mA}{n + mA} \right) = n \right)$$

$$= \frac{1}{\sqrt{m}\,\sigma} \exp\left\{ -\frac{(d - m + m\theta)^2}{2m\sigma^2} \right\} + o\left(\frac{1}{\sqrt{m}} \right),$$

uniformly in d as $m, n \to \infty$ with $m/n \to K > 0$, proving the first assertion of the theorem. The second assertion follows from this and (2.1) using Stirling's formula. Finally,

$$E[G(j)|D = d] = d\binom{j + A - 1}{A - 1} \frac{m - d + 1}{d} \frac{n!}{(n - j)!} \frac{(n + mA - 1 - j)!}{(n + mA - 1)!} R,$$

where R is the ratio of $P(D = d - 1)$ with a sample of $n - j$ to $P(D = d)$ with a sample of n. Taking $m = [n\lambda^*]$, $R \to 1$ and the approximation for $E(G(j)|D = d)$ follows easily from Stirling's formula (a derivation from the approximation for H is also possible but more tedious). The approximation for $\hat{m}$ can be obtained in a similar fashion. □

Asymptotic normality of D is a special case of Corollary 3.6 of Holst (1981). The local version of this result, given in Theorem 3.3, is necessary to approximate H.

Our next result gives the limiting distribution of the maximum likelihood estimator.

THEOREM 3.4. *If $m, n \to \infty$ so that $m/n \to \lambda$, the distribution of $(\tilde{m} - m)/\sqrt{m}$ converges to $N(0, \theta_0^2/\sigma_0^2)$, where $\theta \to \theta_0$.*

PROOF. We will verify the stronger assertion that

$$\frac{\tilde{m} - m}{\sqrt{m}} - \frac{D - m(1 - \theta)}{\sqrt{m}} \frac{\theta}{\sigma^2} \to 0, \tag{3.3}$$

in probability. Let $d = d_n$ be a sequence such that

$$\frac{d + m\theta - m}{\sqrt{m}} = a_n = O(1)$$

and (abusing notation) let $\tilde{m}_n$ be the m.l.e. corresponding to d_n. Then

$L(\tilde{m})/L(\tilde{m}-1) \geq 1$ and $L(\tilde{m}+1)/L(\tilde{m}) \leq 1$. By Stirling's formula (suppressing the dependence of d and $\tilde{m}$ on n),

$$\frac{L(\tilde{m})}{L(\tilde{m}-1)} = \frac{\tilde{m}}{\tilde{m}-d}\left(\frac{\tilde{m}A-1}{n+\tilde{m}A-1}\right)^A\left(1+O\left(\frac{1}{\tilde{m}}\right)\right)$$

and

$$\frac{L(\tilde{m}+1)}{L(\tilde{m})} = \frac{\tilde{m}+1}{\tilde{m}+1-d}\left(\frac{\tilde{m}A+A-1}{n+\tilde{m}A+A-1}\right)^A\left(1+O\left(\frac{1}{\tilde{m}}\right)\right).$$

We next claim that $\tilde{m}/m \to 1$. If not, we can assume (taking a subsequence) that $\tilde{m}/m \to K \neq 1$. If $K < \infty$, then

$$\lim\frac{L(\tilde{m})}{L(\tilde{m}-1)} = \lim\frac{L(\tilde{m}+1)}{L(\tilde{m})} = \frac{K}{K-1+\theta}\left(\frac{AK}{AK-1/\lambda}\right)^A \neq 1,$$

a contradiction. If $K = \infty$, then

$$\frac{L(\tilde{m})}{L(\tilde{m}-1)} = 1 - \frac{n-d}{\tilde{m}} + o\left(\frac{n}{\tilde{m}}\right),$$

also a contradiction. Let $\varepsilon = (\tilde{m}-m)/m$. Taylor expanding,

$$\frac{L(\tilde{m})}{L(\tilde{m}-1)} = \frac{L(\tilde{m}+1)}{L(\tilde{m})} + O\left(\varepsilon^2+\frac{1}{m}\right) = 1 - \frac{\sigma^2\varepsilon}{\theta^2} + \frac{a}{\theta\sqrt{m}} + O\left(\varepsilon^2+\frac{1}{m}\right).$$

Using this, it is easy to derive a contradiction unless (3.3) holds and the theorem follows. □

We suspect that $(\hat{m}-m)/\sqrt{m}$ is asymptotically normal as $n \to \infty$ with $m/n \to K$ and believe a proof could be constructed along the following lines. Keep extra terms in the expansions used to prove Theorem 3.3 and show that

$$P(D=d) = \frac{\exp(-d^{*2}/2)}{\sqrt{2\pi m\sigma^2}}\left(1+\frac{1}{\sqrt{n}}\gamma(d^*, m/n)\right) + O\left(\frac{1}{n}\right), \tag{3.4}$$

where $d^* = (d+\theta m-m)/\sqrt{m\sigma^2}$ and γ is a smooth function of both its arguments. This result then implies that

$$\frac{\hat{m}-m}{\sqrt{m}} - \frac{\theta}{\sigma^2\sqrt{m}}(d+\theta m-m) \to 0,$$

in probability. This gives a limiting distribution for $(\hat{m}-m)/\sqrt{m}$ which is $N(0, \theta_0^2/\sigma_0^2)$, so $\tilde{m}$ and $\hat{m}$ are asymptotically equivalent in this limit assuming (3.4) holds.

Our next limit gives a Poisson approximation for the distributions of D.

THEOREM 3.5. *If $n \to \infty$ and $m \sim (\lambda^A n/A)^{1/(A+1)}$, the distribution of $m-D$ converges to $P(\lambda)$.*

PROOF. Using Lemma 2.1,

$$P(m - D = k) = \sum_{j=0}^{m-k-1} (-1)^j \frac{m!\theta^{(j+k)}}{k!j!(m-k-j)!}.$$

In this limit, $m\theta \to \lambda$ and $(m!/(m-k-j)!)\theta^{(j+k)} \to \lambda^{j+k}$. The theorem follows from dominated convergence. To dominate the summands, use

$$\frac{m!}{(m-k-j)!} \le m^{k+j} \quad \text{and} \quad \theta^{(j+k)} \le \theta^{j+k}. \qquad \square$$

In this limit $\hat{m}$ and $\tilde{m}$ behave as indicated in the next corollary.

COROLLARY 3.6. *If $n \to \infty$ and $m \sim (\lambda^A n/A)^{1/(A+1)}$, then $\hat{m} - (D + \lambda) \to 0$ in probability and $\tilde{m} - D \to [\lambda]$ (the greatest integer $\le \lambda$) provided λ is not an integer.*

PROOF. By the last theorem, if $m - d = O(1)$, then $\hat{m}(d) - d = \lambda + o(1)$. The result about $\tilde{m}$ follows by approximating the likelihood by $e^{-\lambda}\lambda^{m-d}/(m-d)!$. Details are similar to those for Theorem 3.4 and are omitted. $\square$

Our last two results deal with limits where the joint distributions of the G's converge.

THEOREM 3.7. *As $A \to \infty$,*

$$P(G(i) = g(i),\ i = 1,2,\ldots) \to \frac{m!n!}{(m-d)!m^n} \Big/ \prod_i \{g(i)!(i!)^{g(i)}\}$$

and

$$P(D = d) \to \frac{\mathbb{S}_n^{(d)} m!}{(m-d)!m^n},$$

where $\mathbb{S}_n^{(d)}$ is a Stirling number of the second kind [Abramowitz and Stegun (1970)].

PROOF. The first result follows easily from (1.1) and the second from (2.1) and (2.2) using the identity

$$\mathbb{S}_n^{(d)} = \frac{1}{d!} \sum_{j=0}^{d} (-1)^{d-j} \binom{d}{j} j^n.$$

The limiting distribution for D is called Arfwedson's (1951) distribution and is discussed by Johnson and Kotz (1969). $\square$

THEOREM 3.8. *If $m \to \infty$ and $A \to 0$ so that $mA \to \omega$, then*

$$P(G(i) = g(i),\ i \ge 1) \to \omega^d \Big/ \left\{ \binom{n + \omega - 1}{n} \prod_{i \ge 1} [g(i)!i^{g(i)}] \right\}$$

and

$$P(D = d) \to \frac{(\omega - 1)!}{(n + \omega - 1)!}\omega^d \left|S_n^{(d)}\right|,$$

where $S_n^{(d)}$ *is a Stirling number of the first kind.*

PROOF. The first assertion follows easily from (1.1) using the fact that $\Gamma(\cdot)$ has a simple pole at -1. Another proof is given by Watterson (1976). The limiting distribution is Ewens' (1972) sampling formula which arises frequently in genetic models without selection. He derives the distribution for D. For an independent derivation, use Lemma 3.1 to show that

$$Ex^D \to \frac{(n - 1 + \omega x)!(\omega - 1)!}{(n + \omega - 1)!(\omega x - 1)!}.$$

The result then follows from the identity

$$Z(Z + 1) \cdots (Z + n - 1) = \sum_{m=0}^{n} \left|S_n^{(m)}\right| Z^m. \qquad \square$$

In this limiting situation, D is complete and sufficient for ω and the limiting distribution for D forms an exponential family: Ewens gives tables for finding confidence intervals for ω.

In practice, A is usually unknown and it is natural to try to estimate m and A jointly using maximum likelihood. If either of these last two distributions fits the data, it may be the case that the likelihood is maximized as $A \to \infty$ or as $m \to \infty$, $mA \to \omega$. This happens in some of the data sets we have considered and can wreak havoc on computer programs which ignore this possibility. There are several interesting open questions concerning joint estimation using maximum likelihood. In particular, are the estimates unique and are they consistent and asymptotically normal in the limit $n \to \infty$ with $m/n \to \lambda$? These questions seem rather challenging, especially uniqueness with the discrete character of m to worry about.

4. Examples. The first two data sets considered are from Mosteller and Wallace (1964) and concern word usage by Hamilton and Jefferson. In these data sets, $G(s)$ is the number of manuscripts in which a specific word ("can" for Hamilton and "may" for Jefferson) occurs exactly s times. In these data sets the number m of manuscripts surveyed is known, so the estimators $\tilde{m}$ and $\hat{m}$ can be compared with the true value. Also if the authors use "can" and "may" a fixed proportion of the time and the manuscripts are all of comparable length, then neglecting "pigeonholing," the distribution of the partition should be that induced by Bose–Einstein sampling, so the true value of A should be roughly 1.

In the data for Madison, $m = 262$, $n = 172$ and $d = 106$. Assuming $A = 1$, the estimates for m are $\hat{m} = 272.1$ and $\tilde{m} = 274$, both close to the correct value. The log of the likelihood at $\tilde{m}$ is $\log L = -8.273$. With $A = \infty$, the estimates are $\hat{m} = 160.7$ and $\tilde{m} = 161$ with $\log L = -9.464$, and under Ewens' sampling

TABLE 1
Madison: Uses of "may"

		Model					
		$A = 1$		$A = \infty$		ESF	A unknown
s	$G(s)$	$\widehat{EG}$	$\widetilde{EG}$	$\widehat{EG}$	$\widetilde{EG}$	$\widetilde{EG}$	$\widetilde{EG}$
1	63	65.09	65.12	59.35	59.27	69.76	63.36
2	29	25.27	25.14	31.80	31.67	20.81	27.00
3	8	9.72	9.67	11.21	11.22	8.25	10.18
4	4	3.70	3.70	2.92	2.96	3.67	3.59
5	1	1.40	1.41	0.60	0.62	1.74	1.21
6	1	0.52	0.54	0.10	0.11	0.86	0.39

formula (ESF), $\tilde{\omega} = 116.7$ and $\log L = -10.203$. Simultaneous maximum likelihood estimation of m and A produces a mildly surprising result, $\tilde{m} = 217$ and $\tilde{A} = 1.998$ with $\log L = -8.079$. The likelihood is only 18% smaller than the likelihood when $A = 1$, and the estimates are far from their correct values. This behavior occurs in other data sets and will be discussed in the concluding remarks section. Table 1 gives m.l.e. estimates $\widetilde{EG}$ and UMVUE estimates $\widehat{EG}$ for $EG(s)$ under the four models fitted. Note that joint estimation of m and A does not degrade the estimates for $EG(s)$ and that the m.l.e. and UMVUE estimates are generally quite close.

In the Hamilton data set, $m = 247$, $n = 139$ and $d = 90$. When $A = 1$, $\hat{m} = 250.2$, $\tilde{m} = 253$ and $\log L = -8.541$. When $A = \infty$, $\hat{m} = 145.6$, $\tilde{m} = 146$ and $\log L = -12.247$. Under Ewens' sampling formula, $\tilde{\omega} = 109.4$ and $\log L = -8.338564$. Finally, estimating both m and A, $\tilde{m} = 10{,}000{,}001$, $\tilde{A} = 1.094 \times 10^{-5}$ and $\log L = -8.338546$. Again these estimates are far from the correct values. Table 2 gives the estimates for $EG(s)$.

Our next example deals with species of birds on Malaysia. The data were communicated to us by Dr. Marina Wong. For these data, $n = 702$ and $d = 83$. When $A = 1$, $\hat{m} = 93.97$, $\tilde{m} = 93$ and $\log L = -58.2117777$. When $A = \infty$,

TABLE 2
Hamilton: Uses of "can"

		Model					
		$A = 1$		$A = \infty$		ESF	A unknown
s	$G(s)$	$\widehat{EG}$	$\widetilde{EG}$	$\widehat{EG}$	$\widetilde{EG}$	$\widetilde{EG}$	$\widetilde{EG}$
1	60	58.04	58.12	53.97	53.84	61.48	61.48
2	20	20.76	20.62	25.75	25.62	17.21	17.21
3	5	7.33	7.28	8.04	8.07	6.41	6.41
4	2	2.55	2.56	1.85	1.89	2.67	2.67
5	2	0.88	0.89	0.33	0.35	1.19	1.19
6	1	0.30	0.31	0.05	0.05	0.55	0.55

TABLE 3
Species of birds

		Model					
		$A = 1$		$A = \infty$		ESF	A unknown
s	$G(s)$	$\widehat{EG}$	$\widetilde{EG}$	$\widehat{EG}$	$\widetilde{EG}$	$\widetilde{EG}$	$\widetilde{EG}$
1	24	9.7	9.5	0.1	0.1	23.5	19.2
2–3	12	16.2	15.9	2.4	2.4	18.7	18.4
4–5	12	12.6	12.5	10.0	10.0	9.4	10.3
6–9	8	17.6	17.4	42.1	42.1	10.4	12.1
10–13	8	10.7	10.6	24.2	24.3	5.9	7.0
14–20	7	9.6	9.6	4.0	4.0	5.8	6.9
21–27	6	4.0	4.0	0.0146	0.013	3.2	3.6
28–37	6	2.0	2.0	—	—	2.6	2.7
> 37	0	0.7	0.8	—	—	3.4	2.6

$\hat{m} = 93.0167$, $\tilde{m} = 83$ and $\log L = -318.2087725$. Under Ewens' sampling formula, $\tilde{\omega} = 24.28$ and $\log L = -51.0344975$. Finally, estimating both m and A, $\tilde{m} = 189$, $\tilde{A} = 0.189678$ and $\log L = -50.1369247$. Estimates for $EG(s)$ (summed over various ranges of s) are given in Table 3. From these estimates and from $\log L$, Ewens' sampling formula fits these data nearly as well as the model where $m = \tilde{m}$ and $A = \tilde{A}$.

Our final example concerns word frequency data for Shakespeare compiled by Spevack (1968) and analyzed by Efron and Thisted (1976). For these data, $n = 884{,}647$ and $d = 31{,}534$. With such a large data set, we found it impossible (with our programs at least) to find unbiased estimates using the recursion formula for Lemma 2.4. When $A = 1$ (the Bose–Einstein model), from (2.2),

$$H(d, n) = \binom{n-1}{d-1}$$

and using this identity, most of our results simplify considerably, allowing easy estimation in this case. Due to these difficulties with the size of the data set, we fit our model using maximum likelihood estimation for the Maxwell–Boltzmann ($A = \infty$) and Ewens' sampling formula cases and using the unbiased estimates for the Bose–Einstein case ($A = 1$). The results are given in Table 4 along with the fitted values under Fisher's model obtained by Efron and Thisted [these values are close but not exactly the maximum likelihood estimates under Fisher's model; see Efron and Thisted (1976) for more details] and fitted values for a conditional version of our model. When $A = 1$, $\hat{m} = 32{,}699$, and when $A = \infty$, $\hat{m} = 31{,}534$ and under Ewens' sampling formula, $\tilde{\omega} = 6385.4$.

From Table 4 we see that as A decreases, the fitted number of singletons increases, but never reaches a value acceptably close to 14,376. Fisher's model, by allowing negative values for A, can be fitted adequately. When the m.l.e. of A is negative, there is no estimate for m in Fisher's model.

TABLE 4

		Model				
		$A = 1$	$A = \infty$	ESF	Fisher[a]	Conditional
s	$G(s)$	$\widehat{EG}$	$\widehat{EG}$	$\widehat{EG}$	$\widehat{EG}$	$\widehat{EG}$
1	14376	1124.0	0.0	6339.7	14376	14249
2	4343	1084.0	0.0	3147.1	4305	4165
3	2292	1045.3	0.0	2083.1	2281	2199
4	1463	1008.1	0.0	1551.1	1471	1420
5	1043	972.1	0.0	1232.0	1050	1017
6	837	937.5	0.0	1019.3	798	777
7	638	904.1	0.1	867.4	633	619
8	519	871.8	0.2	753.6	518	509
9	430	840.8	0.6	665.0	433	429
10	364	810.8	1.7	594.2	369	368

[a] These estimates are from Efron and Thisted (1976).

In our model, if we condition on the value of D, we get

$$P(G(i) = g(i),\ i = 1,2,\ldots | D = d) = \frac{d! \prod_{i=1}^{\infty} \binom{i + A - 1}{i}^{g(i)}}{H(d,n) \prod_{i=1}^{\infty} g(i)!}.$$

This gives a family of distributions and as in Fisher's model, values of A in $(-1, 0)$ make sense. Unfortunately, by the sufficiency, once we condition on D, all information concerning m is lost.

Expected values for the $G(j)$ under this conditional model are given in Theorem 2.5 and the approximation of $E(G(j)|D = d)$ given in Theorem 3.3 is still accurate although the proof needs some modification. [Conditional on $D = d$, the nonzero $X(p)$'s are conditionally i.i.d. with mass function $p^A(1 - p)^k \Gamma(A + k)/\{k!\Gamma(A)(1 - p^A)\}$ for $k = 1, 2, \ldots$. Then $G(j)$ and N are asymptotically jointly normal and the proof proceeds as the proof of Theorem 3.3.] λ^* when $A < 0$ is the unique *negative* solution of $\lambda = \lambda^*\{1 - (A\lambda^*/(1 + A\lambda^*))^A\}$. This approximation was used to obtain the fitted values in the last column of Table 4. A was obtained by minimum χ^2 on these 10 cells. The fitted value for A was -0.4149, close to the value of -0.3954 obtained by Efron and Thisted. Although our conditional model has only one free parameter (A), compared with two parameters (A, p) for Fisher's model, the fit obtained is nearly as good.

5. Conclusions. The Dirichlet multinomial model (1.1) for partitions fits the data sets in Section 4 quite well. The maximum likelihood and unbiased estimates were always close and there is little reason to advocate one over the other. For the first three data sets considered, Ewens' sampling formula fits reasonably well. Whenever this happens, it will be difficult to estimate m and A separately with much accuracy. This occurs because pairs m, A with constant

product mA give nearly identical distributions for the partition by Theorem 3.8. In the final data set, negative values for A seem necessary to obtain a good fit. These can be obtained by conditioning on D in our model or by fitting Fisher's model. Using either approach, there is no reasonable estimate for m.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1970). *Handbook of Mathematical Functions*. Dover, New York.

ARFWEDSON, G. (1951). A probability distribution connected with Stirling's second class numbers. *Skand. Aktuarietidskr.* **34** 121–132.

BHATTACHARYA, R. N. and RAO, R. R. (1976). *Normal Approximation and Asymptotic Expansions*. Wiley, New York.

CHEN, W. (1980). On the weak form of Zipf's law. *J. Appl. Probab.* **17** 611–622.

CHEN, W. (1981a). Limit theorems for general size distributions. *J. Appl. Probab.* **18** 139–147.

CHEN, W. (1981b). Some local limit theorems in the symmetric Dirichlet-multinomial urn models. *Ann. Inst. Statist. Math. A* **33** 405–415.

EFRON, B. and THISTED, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoret. Population Biol.* **3** 87–112.

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. Animal Ecol.* **12** 42–58.

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.

HILL, B. M. (1979). Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *J. Amer. Statist. Assoc.* **74** 668–679.

HOLST, L. (1981). Some conditional limit theorems in exponential families. *Ann. Probab.* **9** 818–830.

JOHNSON, N. L. and KOTZ, S. (1969). *Distributions in Statistics: Discrete Distributions*. Houghton Mifflin, Boston.

LEWONTIN, R. C. and PROUT, T. (1956). Estimation of the number of different classes in a population. *Biometrics* **12** 211–223.

MOSTELLER, F. and WALLACE, D. L. (1964). Inference and disputed authorship. *The Federalist*. Addison-Wesley, Reading, Mass.

ROTHMAN, E. D. and TEMPLETON, A. R. (1980). A class of models of selectively neutral alleles. *Theoret. Population Biol.* **18** 135–150.

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

WATTERSON, G. A. (1976). The stationary distribution of the infinitely-many neutral alleles diffusion model. *J. Appl. Probab.* **13** 639–651.

WRIGHT, S. (1969). *Evolution and the Genetics of Population* **2**. Univ. Chicago Press, Chicago.

DEPARTMENT OF STATISTICS
1444 MASON HALL
UNIVERSITY OF MICHIGAN
419 SOUTH STATE STREET
ANN ARBOR, MICHIGAN 48109