THEORETICAL POPULATION BIOLOGY 18, 135–150 (1980)

A Class of Models of Selectively Neutral Alleles*

EDWARD D. ROTHMAN

Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109

AND

ALAN R. TEMPLETON

Department of Biology, Washington University, St. Louis, Missouri 63130

Received August 8, 1979

Using exchangeability as a statistical analog of neutrality, we derive a generalized sampling distribution for neutral alleles. The distribution depends upon a parameter that determines the underlying marginal distribution of the number of copies of a neutral allele and that can range from zero to infinity. The sampling model of Ewens (1972) is a special case characterized by an extreme value (0) of this parameter. Two other special cases are considered, one of which seems to be applicable to populations with a structure like that of the Yanomama Indians of South America. We then investigate the expected frequency spectra under these three special cases and discover that all three models yield a broad range of possible spectra with overlap between the special cases. We finally show that Ewens' sampling model cannot be used to construct tests of neutrality versus selection tending to maintain polymorphisms, but it can be used to construct tests of directional selection versus neutrality plus selection tending to yield polymorphic states.

1. INTRODUCTION

One impact of the use of electrophoresis to detect genetic variability in natural populations was to focus attention on the neutralist versus the selectionist hypotheses as explanations of genetic variability. This debate has caused several workers to propose a wide assortment of tests that could potentially distinguish between these hypotheses (Ewens, 1977). One such test is based on Ewens' (1972) model of the sampling theory of selectively neutral alleles. This theory rests upon four basic assumptions:

* This work was supported in part by the Department of Energy, ERDA Contract E(11-1)-2828.

0040-5809/80/050135-16$02.00/0

(i) neutrality of all alleles at a locus;

(ii) a fixed population size that is large compared to the sample size;

(iii) a stationary stochastic process of mutation and drift; and

(iv) a potentially infinite number of alleles (only unique alleles result from mutation).

Although additional assumptions are included in the development of the model, replacement of population size, N, with an effective population size, N_e, generalizes the model sufficiently to serve a broadened class of applications. Ewens (1972) pointed out, however, that assumption (iv) is probably violated for electrophoretic data because electrophoretic techniques lack a total ability to differentiate alleles. Thus, the data consist of electromorph allele categories rather than unique alleles. However, in the past few years, the ability to resolve alleles has increased significantly (Coyne, 1976; Singh *et al.*, 1976; Johnson, 1977a). Consequently, Watterson (1978) applied Ewens' sampling theory and some of his own extensions of this theory to these new data sets characterized by enhanced allelic resolution. Moreover, Kingman (1978a) has shown that Ewens' formula is valid even when some allelic categories are pooled, due to lack of resolution, as long as the pooling is not too extensive. Consequently, the sampling theory for the infinite allele model is becoming more and more appropriate. It is therefore important to examine this sampling theory in more detail and, if possible, generalize it so as to be more robust when applied to "real" populations.

Much work investigating the robustness and underlying assumptions of Ewens' sampling theory has already been done. Watterson (1976) has shown that Ewens' sampling formula can be obtained as the limit of a mixture of a multinomial with a Dirichlet. As discussed in the work of Rothman *et al.* (1974), the multinomial–Dirichlet distribution has sufficient flexibility in its parameters to describe a wide range of underlying population structures. However, Ewens' sampling formula is a specific limit of the multinomial–Dirichlet; in particular, it is the "Poisson–Dirichlet" limit (Kingman, 1977). Although the family of multinomial–Dirichlet distributions has great robustness, it is still not clear how robust the subfamily of Poisson–Dirichlet limits is and how applicable these limits are to "real" populations.

In this regard, some simulations of an American Indian population are of interest (Li *et al.*, 1978; Neel and Rothman, 1978). Although only certain aspects of Ewens' model may be examined and although the simulations are themselves only "models" (though more complex), a lack of agreement is found. Specifically, Li *et al.* (1978) and Neel and Rothman (1978) are better able to estimate the expected time to extinction of a mutant allele and mutation rate by assuming that the number of copies produced by a mutant allele is distributed according to a geometric distribution. A consequence of this assump-

tion is that the joint distribution of the allele numbers in a population is of a form (Bose–Einstein) substantially different from Ewens' model, as we will show in this paper. However, two restrictive comments are worthwhile: first, the joint distribution of all alleles was not obtained empirically for the simulation; and second, the simulation may, in fact, not be in better agreement with reality than Ewens' model. A clearer distinction will hopefully be possible when more empirical data become available. Nonetheless, this result does indicate that certain population structures cannot be adequately described by Ewens' sampling formula.

Another difficulty in interpreting Ewens' sampling formula was revealed by Gillespie (1977). He developed a model of selection in a random environment that yields Ewens' sampling formula. Hence, the null hypothesis for Ewens' sampling formula is apparently not neutrality, but neutrality plus some types of selection. Examining Gillespie's selection model in more detail, we discovered that his model and Ewens' model do share one assumption: the random variables (numbers in allele categories) are finitely exchangeable; i.e., the joint mass function of the allele numbers is invariant to permutations of the allele labels. As Kingman (1978b) has pointed out, Ewens' sampling formula (as well as a wide range of other possible sampling formulas we will develop in this paper) assumes exchangeability. Moreover, Watterson (1977) developed a model of symmetric overdominance that yields a sampling distribution different from Ewens' formula, but nevertheless with the property of finite exchangeability.

Whether or not this confoundment of neutrality with exchangeability is a critical problem when applied to real populations remains to be seen. It is certainly natural to assume that all neutral alleles must have the statistical property of exchangeability; in fact, exchangeability can be regarded as the very essence of neutrality. But as Gillespie's and Watterson's work demonstrates, the set of neutral alleles is only a subset of the exchangeable alleles. However, it is doubtful that selection will often result in the degree of symmetry Gillespie (1977) and Watterson (1977) assumed, and consequently the number of selected systems that have allele exchangeability in real populations is probably quite small. Hence, we will develop in this paper a general sampling theory for neutral alleles under the assumption that exchangeability is an appropriate statistical analog to neutrality. In any event, this theory will be valid for neutral alleles since exchangeability is a fundamental property of neutrality.

2. General Model

We first provide a class of models for the selectively neutral hypothesis. This general class of models is obtained by noting that in a population of size $2N$ with K alleles, $K = 1, 2, \ldots, 2N$, the joint mass function of the allele

frequencies must be finitely exchangeable under neutrality. That is, for each K,

$$P(f(A_1) = n_1, \ldots, f(A_K) = n_K \mid K, 2N) = P(f(A_{L1}) = n_1, \ldots, f(A_{LK}) = n_K \mid K, 2N),$$

where the alleles are labeled $A_1, \ldots, A_k$, $f(A_i)$ is the number of copies of A_i in the population, and $L_1, \ldots, L_k$ is any permutation of the indices $(1, 2, \ldots, K)$. The class of joint probability mass functions for finitely exchangeable random variables has been characterized by deFinetti (1931) to be mixtures of hypergeometrics. This may be written

$$P(f(A_1) = n_1, \ldots, f(A_K) = n_K \mid K, 2N) = \frac{\sum \prod_{i=1}^{K} \binom{T_i}{n_i} g(T_1, \ldots, T_K)}{\binom{S}{2N}}, \quad (2.1)$$

where the sum is over all vectors $(T_1, \ldots, T_K) \geqslant (n_1, \ldots, n_K)$ and $\sum T_i = S$. Here $g(T_1, \ldots, T_K)$ is a joint probability mass function, and (2.1) holds for all $n_i \geqslant 1, \sum_{i=1}^{K} n_k = 2N$.

Although Eq. (2.1) is the most general form of a model for the selectively neutral hypothesis, we here restrict attention to a subset of models. The particular subset that we choose is sufficiently general so as to provide adequate approximations to most models contained in (2.1). Furthermore we obtain Ewens' model as a limiting case within this subset.

The particular subset of interest is found first by choosing g as a mixture of a multinomial $(S; p_1, \ldots, p_K)$ and an exchangeable Dirichlet. This is a special case of a model proposed for population structure in Rothman *et al.* (1974). It is,

$$P(n_1, \ldots, n_K \mid K, 2N) = \frac{\prod_{i=1}^{K} \binom{n_i + A - 1}{A - 1}}{\sum_{j=1}^{K} (-1)^{K-j} \binom{K}{j} \binom{Aj + 2N - 1}{2N}}, \quad (2.2)$$

where $n_i > 0$ and $\sum_{i=1}^{K} n_i = 2N$, and $A > 0$ is a parameter. (In this paper $\binom{\alpha}{\beta}$ means $\Gamma(\alpha + 1)/[\Gamma(\beta + 1)\,\Gamma(\alpha - \beta + 1)]$ where $\Gamma(\cdot)$ is the gamma function.) As already mentioned, Kingman (1977, 1978b) and Watterson (1976) have shown that Ewens' sampling formula is a particular limit of (2.2). In the next two sections we study the properties of (2.2), and then turn to applications and discussion in the remaining section.

3. Genesis and Some Special Cases

In the previous section, we argued that the property of exchangeability alone may require that the sampling formula of neutral alleles be of the form

(2.2). In this section we present another derivation of (2.2) that will provide more insight into the underlying population genetic assumptions. Moreover, we will examine three special cases of (2.2), one of which is the Ewens sampling formula.

Many of the classical single-deme population models assume that the progeny distribution is Poisson (Karlin and McGregor, 1968). So, the marginal distribution (i.e., not conditioned upon population size or other allele numbers) of the number of copies of a neutral allele at any point in time given the numbers in the previous generation is also Poisson distributed with a parameter depending upon the number of copies in the previous generation. Since the time of origin of a neutral allele is unknown, the progeny distribution itself may not be Poisson (Kojima and Kelleher, 1962), the population may be genetically subdivided (Rothman *et al.*, 1974), and the validity of the assumption of a constant Poisson mean may not be reasonable through time (Templeton, 1977) or space (Karlin, 1969), it is very difficult to decide exactly what marginal distribution would be most appropriate for "real" populations. Therefore, we will generate a robust family of possible marginal distributions by regarding the Poisson parameter as a random variable. For convenience, suppose the Poisson mean, λ, has a gamma distribution such that $E(\lambda) = \bar{\lambda}$ and $\mathrm{Var}(\lambda) = \bar{\lambda}^2/A$. Mixing this gamma distribution with the Poisson implies the n_i's are marginally independent negative binomials;

$$P(n_i = k) = \binom{A + k - 1}{A - 1}\left(\frac{\bar{\lambda}}{\bar{\lambda} + A}\right)^k\left(\frac{A}{\bar{\lambda} + A}\right)^A, \qquad i = 1, 2,..., \quad (3.1)$$

where A (as will soon be evident) is the same A parameter appearing in (2.2). The parameter $\bar{\lambda}$ will not appear in any of our final results since the number of alleles, K, in our finite sample of size $2N$ is a sufficient statistic. The formulation given by (3.1) now allows us to interpret A as a mixing parameter that defines the degree of variability present in the underlying Poissons. As A gets smaller, the amount of heterogeneity in the Poisson parameters increases.

From the assumption that the joint distribution of n_i, $i = 1, 2,...,$ are independently distributed according to (3.1), we can derive the distribution conditional on the sample size of $2N$ genes yielding K alleles such that $n_i > 0$ for all i to be

$$P(n_1,..., n_K \mid K, 2N) = \frac{\prod_{i=1}^{K}\binom{n_i + A - 1}{A - 1}}{\sum_{j=1}^{K}(-1)^{K-j}\binom{K}{j}\binom{Aj + 2N - 1}{2N}}. \qquad (3.2)$$

As is evident, (3.2) is identical to (2.2), but with this derivation it is clear that the A parameter defines the underlying marginal distribution of the number

of copies of a neutral allele. We now examine three special cases of (3.2), two of which represent extreme values of the A parameter ($A \to 0$ and $A \to \infty$) and one an intermediate A value ($A = 1$).

Consider first the case $A \to 0$. One can take the limit of (3.2) directly, but instead we return to (3.1). Taking the limit of the negative binomial as $A \to 0$ with the condition that $n_i > 0$, we get that the n_i's have log series distributions (Fisher *et al.*, 1943; Watterson, 1974):

$$P(n_i = j) = \frac{\theta^j}{j[-\ln(1-\theta)]}, \qquad j = 1, 2, \ldots . \tag{3.3}$$

Then the distribution of $\sum_{i=1}^{K} n_i$ is

$$P\left(\sum_{i=1}^{K} n_i = 2N\right) = \frac{\theta^{2N} K! \mid S_{2N}^{(K)} \mid}{(2N)!}, \tag{3.4}$$

where $S_{2N}^{(K)}$ is a Stirling number of the first kind (cf. Abramowitz and Stegun, 1965). From (3.3) and (3.4), we obtain

$$P(n_1, \ldots, n_K \mid K, 2N) = \frac{(2N)!}{K! \mid S_{2N}^{(K)} \mid} \cdot \left(\prod_{i=1}^{K} n_i\right)^{-1}. \tag{3.5}$$

This is Ewens' (1972) model which arose as a result of both a diffusion approximation to a Markov chain and a combinatorial argument made precise in the accompanying paper by Karlin and McGregor (1972). This allelic distribution implies that the most likely assemblage involves many alleles at low frequency and a few at high frequency.

The next case we consider is $A = 1$. Substituting $A = 1$ directly into (3.1) or (3.2) and noting that

$$\sum_{j=1}^{K} (-1)^{K-j} \binom{K}{j} \binom{j + 2N - 1}{2N} = \binom{2N-1}{K-1} \tag{3.6}$$

we obtain the Bose–Einstein allocation:

$$P(n_1, \ldots, n_K \mid K, 2N) = \frac{1}{\binom{2N-1}{K-1}}, \qquad n_i > 0, \qquad \sum_{i=1}^{K} n_i = 2N. \tag{3.7}$$

An alternative derivation to (3.7) using a Galton–Watson branching process reveals more clearly our interest in this special case. First, assume the number of copies produced by an allele in a single generation has a geometric distribution, a special case of the negative binomial. (Recall that there is evidence for this

being the distribution for the simulated Indian population of Li *et al.*, 1978.) Then,

$$\begin{aligned} P(j \text{ copies}) &= bc^{j-1}, \qquad j = 1, 2,..., \\ P(0 \text{ copies}) &= 1 - \frac{b}{1-c}, \end{aligned} \tag{3.8}$$

where $0 < b < 1 - c < 1$.

The number of copies in a Galton–Watson branching process at the tth generation is also geometric;

$$P(j \text{ copies at } t) = B_t \cdot C_t^{j-1}, \qquad j = 1, 2,...$$

$$P(0 \text{ copies at } t) = 1 - \frac{B_t}{1 - C_t}, \qquad t = 1, 2,...,$$

where

$$B_t = \begin{cases} m^t \left[\dfrac{1 - S_0}{m^t - S_0}\right]^2, & m \neq 1 \\ \left[\dfrac{(1-c)}{1 + (t-1)c}\right]^2, & m = 1, \end{cases}$$

$$C_t = \begin{cases} \dfrac{m^t - 1}{m^t - S_0} & m \neq 1 \\ \dfrac{ct}{1 + (t-1)c}, & m = 1, \end{cases}$$

$$S_0 = \frac{1 - b - c}{c(1-c)} \quad \text{and} \quad m = \frac{b}{(1-c)^2}.$$

Conditioning on K, $2N$, and $n_i > 0$ for all i yields (3.7), the Bose–Einstein distribution. Hence, the Bose–Einstein result may be more appropriate than Ewens' formula when applied to populations with a structure like the simulated Indian population. Note that in constrast to the log series (Ewens') model, the Bose–Einstein allocation places equal weight on all permutations.

The last special case we consider is the limit $A \to \infty$; that is, as (3.1) converges to a Poisson with no heterogeneity in its parameter. We may either take the limit of 3.2 as $A \to \infty$ or simply start with the n_i's being unconditionally independent Poissons. Conditioning first on $\sum_{i=1}^{K} n_i = 2N$ results in a multinomial distribution. Then conditioning on $n_i > 0$ for all i, we have

$$P(n_1 \ \ldots \ n_K \mid K \ 2N) = \frac{(2N)!}{k!\, \mathcal{S}_{2N}^{(K)}} \cdot \frac{1}{n_i!}, \tag{3.9}$$

where $\mathcal{S}_{2N}^{(K)}$ is a Stirling number of the second kind (Abramowitz and Stegun, 1965). The distribution described by (3.9) is called the Maxwell–Boltzman

allocation. Under this allocation, the most likely configurations are such that all n_i are approximately equal.

The relations between the three sampling models (3.5), (3.7), and (3.9) are now evident. For Ewens' formula, the marginal probability density function of n_i is assumed to be log series, for Bose–Einstein it is geometric, and for Maxwell–Boltzman it is Poisson. All of these can be viewed as special cases or limits of the assumption that the n_i are independent negative binomials. Moreover, the ordering (3.5), (3.7), and (3.9) corresponds to decreasing heterogeneity in the parameters of the Poissons underlying the negative binomials.

4. The Frequency Spectrum

Let $G(i)$, $i = 1, 2,...$, denote the number of alleles with i copies in the population. In this section we investigate the properties of these $G(i)$ with a view toward furthering our understanding of the models developed in Section 3.

First, in the general case where n_i are independent negative binomials we obtain

$$P(G(1), G(2),... \mid K, 2N) = \frac{K!}{\prod G(i)!} \frac{\prod \binom{i + A - 1}{A - 1}^{G(i)}}{\sum_{j=1}^{K} (-1)^{n-j} \binom{K}{j} \binom{A_j + 2N - 1}{2N}}, \quad (4.1)$$

where $\sum G(i) = K$ and $\sum iG(i) = 2N$.

For the three special cases considered in Section 3, we have from (4.1)

Maxwell–Boltzman:

$$P(G(1), G(2),... \mid K, 2N) = \frac{K!(2N - K)!}{\prod G(i)!} \cdot \prod \left[\frac{1}{(i-1)!}\right]^{G(i)}; \quad (4.2)$$

Bose–Einstein:

$$P(G(1), G(2),... \mid K, 2N) = \frac{K!}{\prod G(i)!} \cdot \frac{1}{\binom{2N - 1}{K - 1}}; \quad (4.3)$$

Ewens:

$$P(G(1), G(2),... \mid K, 2N) = \frac{(2N)!}{|S_{2N}^{(K)}|} \cdot \frac{1}{\prod G(i)!} \cdot \prod \left(\frac{1}{i}\right)^{G(i)}. \quad (4.4)$$

All mass functions are defined for $G(i) \geqslant 0$, $\sum G(i) = K$, and $\sum iG(i) = 2N$. Suppose now that $n_1,..., n_K$ are finitely exchangeable. Then, if a random sample is made,

$$P(n_1 = i \mid \tilde{G}) = G(i)/K.$$

Thus, the marginal distribution of n_1 is

$$P(n_1 = i) = E(G(i)/K)$$

with the expectation operator taken over the distribution of the G's. Hence, these expectations characterize the frequency spectrum of neutral alleles.

Consider now the expected proportions of alleles with i copies given N and K. These expectations are

Maxwell–Boltzman:

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) = \binom{2N-K}{i-1}; \tag{4.5}$$

Bose–Einstein:

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) = \frac{\binom{2N-i-1}{K-2}}{\binom{2N-1}{K-1}}; \tag{4.6}$$

Ewens:

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) = \frac{(2N)!}{K(2N-i)!} \cdot \frac{1}{i} \cdot \frac{|S_{2N-i}^{(K-1)}|}{|S_{2N}^{(K)}|}. \tag{4.7}$$

Taking the limit of (4.5), (4.6), and (4.7) as $N \to \infty$ but with $K/2N$ converging to a constant γ_0, we have

Maxwell–Boltzman:

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) \to \frac{\lambda^{i-1}e^{-\lambda}}{(i-1)!}, \qquad \lambda = \frac{1-\gamma_0}{\gamma_0}; \tag{4.8}$$

Bose–Einstein:

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) \to \gamma_0(1-\gamma_0)^{i-1}; \tag{4.9}$$

Ewens:

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) \to \frac{{\gamma_0}^i}{i[-\ln(i-\gamma_0)]}, \qquad i = 1, 2, \ldots. \tag{4.10}$$

Hence, the three models yield the truncated Poisson, geometric, and log series frequency spectra respectively, as expected from Section 3.

The above frequency spectra all assume $K/2N \to \gamma_0$, a constant, as $N \to \infty$. However, Kimura and Crow (1964) have shown that, under neutrality, the number of alleles in a finite population is a random variable, given $2N$. Con-

sequently, under neutral theory, $K/2N$ may well converge to a random variable, given $2N$, rather than a constant. Assume that $K/2N$ converges in law to a random variable γ with distribution $F(\gamma)$. With this assumption, Hill (1970) has shown that as $N \to \infty$, the Bose–Einstein result yields

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) \to \int_0^1 \gamma(1-\gamma)^{i-1}\, dF(\gamma). \tag{4.11}$$

When $F(\gamma) = \gamma$ (actually, only the behavior of $F(\gamma)$ near $\gamma = 0$ must be of this form for large i, cf. Hill and Woodroofe, (1975)), we obtain Yules law (Yule, 1924):

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) \to \frac{1}{i(i+1)}, \qquad i = 1, 2, \ldots . \tag{4.12}$$

More generally, if γ has a Beta(α, β) distribution, then

$$E\left(\frac{G(i)}{K} \,\middle|\, K, 2N\right) \to \frac{\alpha\Gamma(\alpha+\beta)}{\Gamma(\beta)} \frac{\Gamma(\beta+i-1)}{\Gamma(\alpha+\beta+i)}. \tag{4.13}$$

Corresponding results can be achieved for the Maxwell–Boltzman and Ewens cases. It is also interesting to note that the log series frequency spectrum associated with Ewens' model as $N \to \infty$ and $K/2N \to \gamma_0$ can also result from the Maxwell–Boltzman case (Hill, 1970) by choosing

$$F'(\gamma) = \begin{cases} 0, & \gamma < \epsilon \\ -[\gamma \ln \epsilon]^{-1}, & \gamma > \epsilon \end{cases} \qquad 0 < \epsilon < 1. \tag{4.14}$$

In addition to the case in which $K/2N$ converges in law to a random variable, we also consider the case in which $K \propto \ln N$ for Ewens' model. This case is of interest because under neutrality (Ewens, 1969)

$$E(K) = 2N\mu E(t_0), \tag{4.15}$$

where μ is the mutation rate and $E(t_0)$ is the mean time to extinction for a neutral mutant. As N gets large (such that $4N\mu = \theta$), most neutral mutants are lost and $E(t_0) \sim 2 \ln N$ (Ewens, 1969). Hence, as $N \to \infty$, $\mu \to 0$, and $4N\mu \to \theta$, we expect (Kimura and Ohta, 1969; Nei, 1977)

$$E(K) \simeq \theta \ln 2N.$$

In this case, using the asymptotic expansion for Stirling numbers of the first kind given by Moser and Wyman (1958), we can show that, for Ewens' model,

$$E(G(i) \mid K, 2N) \to \theta/i$$

with θ the solution of $K = \theta/\theta + \theta/(\theta + 1) + \cdots + \theta/(\theta + 2N - 1)$. Furthermore, the rth factorial moment converges to

$$\theta^r/i^r.$$

Our final result in this regard is that the covariance between $G(i)$ and $G(j)$, $i \neq j$, converges to zero. Thus, the $G(i)$ will behave approximately like independent Poisson random variables with mean (and hence variance) θ/i.

One implication of these results is that for those cases with $K/2N$ converging in probability to γ_0, a constant, the variance of $G(i)/K$ converges to zero; hence, a good fit to data of one of the marginal frequency spectra ((4.8), (4.9), or (4.10)) may be anticipated. On the other hand, where $K \propto \ln N$ as $N \to \infty$, or when $K/2N$ converges in law to a random variable γ with positive variance (the most likely outcomes under current neutral theory), the behavior of $G(i)$, i small, will seem quite erratic and a good fit of one of the marginal frequency spectra should not be anticipated, even if the "correct" underlying model is chosen.

Thus, depending upon the underlying model, the type of convergence of $K/2N$ and, in the log series case, on the relative magnitudes of K and N, a wide variety of allelic abundances curves may be expected. These curves may be proportional to θ/i; or θ^i/i or $1/i$, etc.

5. Applications and Discussion

An application of the results of Section 4 is found in Fig. 1 by using the *Colias meadii* data of Johnson (1977b) on the frequency spectrum of 103 electrophoretic variants (using the gel sciving technique). The expected frequencies under the truncated Poisson, geometric, and log series models were obtained by using the maximum likelihood estimate of γ_0 in Eqs. (4.8), (4.9), and (4.10). In addition, Yule's law (4.12) was used, but in that case no parameter needs to be estimated. As can be seen from Fig. 1, Ewens' log series model and Yule's law fit the data about equally well as both gave roughly equal, nonsignificant (at the 5% level) goodness-of-fit chi-squares. However, the Poisson and geometric distributions fit the data very poorly. From this observation, it might be tempting to conclude that Ewens' model is appropriate, while the Maxwell–Boltzman and Bose–Einstein models are not. However, recall that Yule's law arose under very general conditions from the Bose–Einstein allocations. Moreover, the log-series spectrum also results from the Maxwell–Boltzman allocation when (4.14) holds. This illustrates the difficulty of trying to infer which underlying model is "correct" from the frequency spectrum; the wide variety of possible spectra makes such an inference virtually impossible to make.

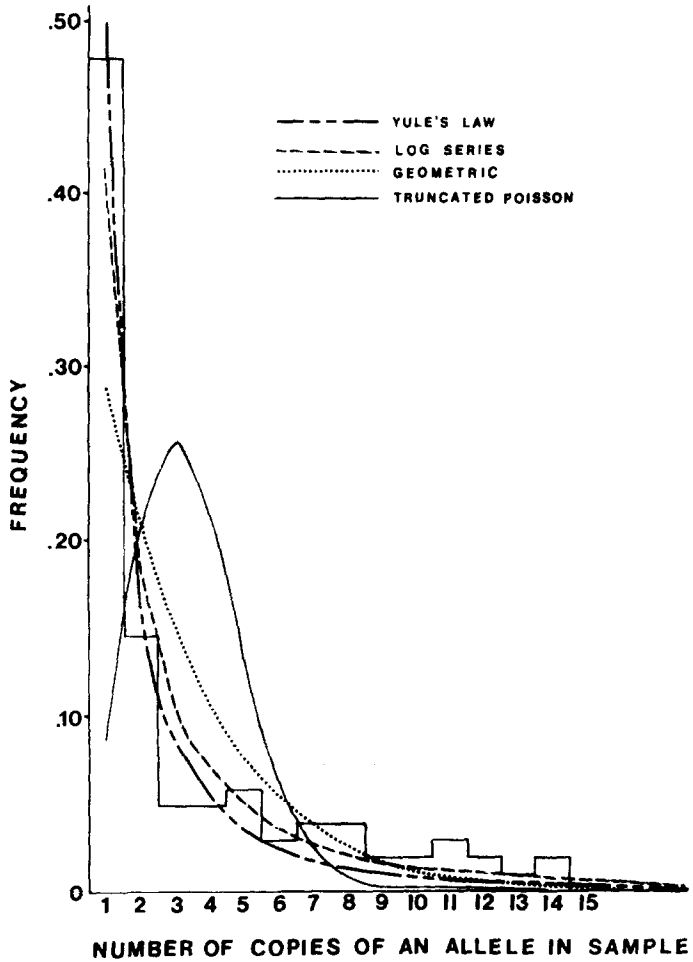

FIG. 1. Frequency distribution of variants detected by gel sieving as reported in Johnson (1977). The histogram tabulates, for a sample of 14 loci examined in 20 individuals, the number of observations in each frequency class. A truncated Poisson, geometric, and log series distributions were then fitted to the data using maximum likelihood estimates of the parameters. A Yule's law distribution is also graphed, but in that case no parameter need be estimated.

As mentioned in the Introduction, a possible application of Bose–Einstein statistics may be found in the simulation results of Li *et al.* (1978) of an American Indian population. The motivation for this application came not from a frequency spectrum, but rather from the observation that the number of copies produced by certain mutant alleles had a geometric distribution, a condition which we showed in Section 3 will lead to the Bose–Einstein allocation. Assuming the

Bose–Einstein model is appropriate for that simulated population, we can obtain an estimate of the expected time to extinction for neutral alleles. First, assume the singleton class (the class of alleles at a locus with only one copy in the population) is at equilibrium. Then, the expected number of copies entering this class represents a balance between new mutants and older ones that by chance enter the singleton class. Thus, given the number of alleles K and the mutation rate μ, then

$$\frac{2N\mu}{K} + \sum_{i \neq 1} P_{i1} E \frac{G(i)}{K} = E \frac{G(1)}{K}(1 - P_{11}),$$

where P_{i1} = the probability that an allele with i copies in a given generation will yield one copy in the next generation. Now, assuming rare variants produce a geometric number of copies (i.e., with probability mass function bc^{i-1}, $i = 1, 2,...$) and independence, then

$$P_{i1} = b^i.$$

Finally, using the results of Section 4 that Bose–Einstein allocation under very general conditions yields Yule's law,

$$E\left(\frac{G(i)}{K}\right) = \frac{1}{i(i+1)}.$$

Using an estimate of $b = 0.1768$ obtained from the simulation of Li *et al.* (1978) and using the relationship between $E(t_0)$ and $K/2N$ implied by (4.15), we obtain the estimate

$$E(t_0) \simeq 3.$$

This may be compared with the result of 2.82 using a more direct approach in Li *et al.* (1978). Thus, the assumption of Bose–Einstein allocation for this simulated population seems to work quite well.

The motivation for Ewens' model was to test the null hypothesis of neutrality. In this context, Ewens (1972) suggested that the information function

$$B = -\sum_{i=1}^{K} x_i \ln x_i$$

where $x_i = n_i/2N$ could serve both as an index of neutrality and also as a test statistic for the hypothesis of neutrality. Ewens (1977) noted that this test (and others such as the "heterozygosity measure" $H = \sum_{i=1}^{K} x_i(1 - x_i)$) would be too large under the null hypothesis of neutrality if heterosis were present (and we may add any other type of selection causing the allele distribution to be too even such as frequency-dependent selection favoring minority genotypes) and would be too small if there was selection for one allele and

against others. However, as mentioned in Section 3, the eveness of the allele frequency distribution increases as the parameter A increases. Hence, as A increases, the expected values of both B and H (and related measures depending upon the eveness of the distribution) also would increase. Consequently, a large value of B or H does not necessarily imply selection, but could be interpreted as meaning that a value of A greater than zero is more appropriate for describing the population. This means that Ewens' test statistic and similar measures cannot be used to test for the presence of heterosis, frequency-dependent selection, certain types of stochastic selection (e.g., Gillespie (1977)), and other types of selection tending to yield polymorphic states. Ewens (1979) has recently derived a test of "generalized neutrality" for distinguishing neutrality plus directional selection from selection tending to maintain a polymorphic state. However, in view of the discussion above, this test is also confounded with the A parameter, making it difficult to assign a particular biological significance to a large test value. However, because Ewens' model does represent an extreme value of the A parameter ($A \to 0$), very small values of his original test statistic and similar statistics would be unlikely under any of the neutral models considered in this paper. Hence, Ewens' statistic and similar measures can be used as a one-sided test of directional selection versus neutrality plus selection tending to maintain polymorphisms. Unfortunately, this contrast is not interesting with respect to the neutralist/selectionist controversy.

A specific test of neutrality versus heterosis was suggested by Watterson (1977). Watterson obtained the sampling distribution under the assumption that all homozygotes have a relative fitness of 1 and all heterozygotes a fitness of $1 + s$, with all other assumptions being essentially like those of Ewens (1972). On the basis of the Neyman–Pearson lemma that the most powerful test is based on the likelihood ratio, Watterson concluded that the most powerful test will be based upon the sample homozygosity statistic $F = \sum_{i=1}^{K} x_i^2$ for testing H_0: $s = 0$ versus H_1: $s = s_1$. However, this conclusion depends upon the assumption that Ewens' model adequately describes H_0. If the null sampling distribution is of the form (2.2), F values will tend to be depressed relative to the expectation under Ewens' model if $A > 0$, thus mimicking heterosis ($s > 0$) in Watterson's model. For example, the ratio of (2.2) with $A = 1$ over $A \to 0$ (the ratio of Bose–Einstein to Ewens' model) is, for $K \propto \ln N$,

$$\frac{(\ln 2N)^{K-1} \prod_{i=1}^{K} n_i}{\binom{2N}{K}}.$$

If the n_i have a distribution that is fairly even (as measured by a low F), this ratio could be greater than one. Hence, a significantly low F could be interpreted as either being due to heterosis or $A > 0$. Indeed, the richness of the

family of models defined by (2.2) and indexed by the parameter A indicates that a member of this family, given the data, will produce a satisfactory likelihood ratio when compared with a model for heterosis ($s > 0$) or any other type of selection model resulting in an allele distribution too even for Ewens' model. Hence, it is best to restrict Watterson's F statistics to a test of H_0: $s \geqslant 0$ versus H_1: $s < 0$ rather than H_0: $s = 0$ versus H_1: $s \neq 0$.

As the above discussion illustrates, it is very difficult to construct a statistic based upon cross-sectional data that actually tests the null hypothesis of neutrality. Many types of selection, and in particular selection that tends to maintain allelic diversity, the primary focus of the neutralist–selectionist debate, are confounded with the A parameter; that is, with the ecological, geographical, and population structural constraints that also influence the allelic distribution. In view of the importance the A parameter plays in defining the sampling model, we end this paper with a conjecture which we hope to investigate in a subsequent paper. If the usual chi-square goodness-of-fit test for equal frequencies is less than the appropriate degrees of freedom, the maximum likelihood estimator of A is infinity. When the chi-square is large (how large must be determined), the maximum likelihood estimator of A will be 0. Finally, in the intermediate case, the likelihood function will have a single mode, but will be quite flat, indicating that any choice of A, in particular $A = 1$, is appropriate.

Acknowledgments

We thank Warren Ewens for his critical reading of this manuscript and his excellent suggestions for improving an earlier draft.

References

Abramowitz, M., and Stegun, I. A. 1965. "Handbook of Mathematical Functions," Dover, New York.

Aretha, K. B., and Ney, P. W. 1971. "Branching Processes," Band 196, Springer–Verlag, Berlin/Heidelberg/New York.

Coyne, D. A. 1976. Lack of genic similarity between two sibling species of *Drosophila* as revealed by varied techniques, *Genetics* **84**, 593–607.

deFinetti, B. 1975. "Theory of Probability," Vol. 1, Wiley, London/New York.

Ewens, W. J. 1969. "Population Genetics," Methuen, London.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles, *Theoret. Pop. Biol.* 3, 87–112.

Ewens, W. J. 1977. Population genetics theory in relation to the neutralist–selectionist controversy, *Advan. Hum. Genet.* 8, 67–134.

Ewens, W. J. 1979. Testing the generalized neutrality hypothesis, *Theor. Pop. Biol.* 15, 205–216.

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**, 42–57.

GILLEPSIE, J. H. 1977. Sampling theory for alleles in a random environment, *Nature (London)* **266**, 443–445.

HILL, B. M. 1970. Zipf's Law and prior distributions for the composition of a population, *J. Amer. Statist. Assoc.* **69**, 1017–1026.

HILL, B. M., AND WOODROOFE, M. 1975. Stronger forms of Zipf's Law, *J. Amer. Statist. Assoc.* **70**, 212–219.

JOHNSON, G. B. 1977a. Assessing electrophoretic similarity, *Annu. Rev. Ecol. Syst.* **8**, 309–328.

JOHNSON, G. B. 1977b. Characterization of electrophoretically cryptic variation in the Alpine butterfly *Colias meadii, Biochem. Genet.* **15**, 665–693.

KARLIN, S. 1969. "A First Course in Stochastic Processes," Academic Press, New York.

KARLIN, S., AND MCGREGOR, J. 1968. The role of the Poisson progeny distribution in population genetic models, *Math. Biosci.* **2**, 11–17.

KARLIN, S., AND MCGREGOR, J. 1972. Addendum to a paper of W. Ewens, *Theor. Pop. Biol.* **3**, 113–116.

KIMURA, M., AND CROW, J. F. 1964. The number of alleles that can be maintained in a finite population, *Genetics* **49**, 725–738.

KIMURA, M., AND OHTA, T. 1969. The average number of generations until fixation of a mutant gene in a finite population, *Genetics* **61**, 763–771.

KINGMAN, J. F. C. 1977. The population structure associated with the Ewens sampling formula, *Theor. Pop. Biol.* **11**, 274–283.

KINGMAN, J. F. C. 1978a. Random partitions in population genetics, *Proc. Roy. Soc. London, Sect. A* **361**, 1–20.

KINGMAN, J. F. C. 1978b. Uses of exhangeability, *Ann. Probability* **6**, 183–197.

KOJIMA, K., AND KELLEHER, T. M. 1962. Survival of mutant genes, *Amer. Natur.* **96**, 329–343.

LI, F. H. F., NEEL, J. V., AND ROTHMAN, E. D. 1978. A second study of the survival of a neutral mutant in a simulated Amerindian population, *Amer. Natur.* **112**, 83–96.

MOSER, L., AND WYMAN, M. 1958. Asymptotic development of the Stirling numbers of the first kind, *J. London Math. Soc.* **33**, 133–146.

NEEL, J. V., AND ROTHMAN, E. D. 1978. Indirect estimates of mutation rates in tribal American Indians, *Proc. Nat. Acad. Sci. Usa* **75**, 5585–5588.

ROTHMAN, E. D., SING, C. F., AND TEMPLETON, A. R. 1974. A model for analysis of population structure, *Genetics* **78**, 943–960.

SINGH, R. S., LEWONTIN, R. C., AND FELTON, A. A. 1976. Genetic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in *Drosophila pseudoobscura*, *Genetics* **84**, 609–629.

TEMPLETON, A. R. 1977. Survival probabilities of mutant alleles in fine-grained environments, *Amer. Natur.* **111**, 951–960.

WATTERSON, G. A. 1974. Models for the logarithmic species abundance distributions, *Theor. Pop. Biol.* **6**, 217–250.

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

WTTERSON, G. A. 1977. Heterosis or neutrality? *Genetics* **85**, 789–814.

WATTERSON, G. A. 1978. Ana analysis of multi-allelic data, *Genetics* **88**, 171–179.

YULE, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S., *Philos. Trans. Roy. London Ser. B* **213**, 21–87.