THEORETICAL POPULATION BIOLOGY 3, 87–112 (1972)

The Sampling Theory of Selectively Neutral Alleles*

W. J. EWENS[†]

Department of Zoology, University of Texas at Austin, Austin, Texas, 78712

Received August 17, 1971

DEDICATED TO THE MEMORY OF KEN KOJIMA

In this paper a beginning is made on the sampling theory of neutral alleles. That is, we consider deductive and subsequently inductive questions relating to a sample of genes from a selectively neutral locus. The inductions concern estimation, confidence intervals and hypothesis testing. In particular the test of the hypothesis that the alleles being sampled are indeed selectively neutral will be considered. In view of the large amount of data currently being obtained by electrophoretic methods on allele frequencies and numbers, and the current interest in the possibility of extensive "non-Darwinian" evolution, such a sampling theory seems necessary. However, a large number of unsolved problems in this area remain, a partial listing being given towards the end of this paper.

MATHEMATICAL THEORY

A number of quantities will be considered in this paper and it is useful to gather together the notations that will be used consistently throughout. We define.

N = number of individuals in the parent (diploid) population (normally unknown)

N_e = effective size of parent population (normally unknown)

u = mutation rate to entirely new alleles (normally unknown)

n = number of individuals in sample (taken in generation t)

K = number of alleles in the population in generation t (an unknown random variable)

k = number of different alleles observed in the sample (a realized value of a random variable)

* Supported by USPHS Grant GM-15769.

[†] Current address: Mathematical Research Center, University of Wisconsin, Madison, WI 53706. Permanent address: Department of Mathematics, La Trobe University, Bundoora 3083, Victoria, Australia.

$\theta = 4N_e u$

$E_s(\cdot)$ = expected value of some sample variable, given the population structure at generation t

$E_p(\cdot)$ = expected value of some population variable in generation t

$E(\cdot) = E_p E_r(\cdot)$

m_e = effective number of alleles in the population

π_i = probability that the number of alleles observed in the sample is i, $(i = 1, 2, \ldots, 2n)$

n_i = number of genes in the sample of the i-th allelic type, $i = 1, \ldots, k$. Note that $\sum n_i = 2n$.

Consider a locus A in the population in question and suppose that alleles from the infinite series $A_1, A_2, \ldots$ can occur at the locus. Any gene is assumed to mutate, with probability u, to form an allele not currently existing (nor previously existing) in the population. [Note: this assumption is made for mathematical reasons and in practice the expression "(nor previously existing)" can be omitted.]

Before considering sample properties we review some properties of the population itself. [For an expanded version of what follows, see Ewens (1969, pp. 67–71).] Given the population configuration in generation $t - 1$, the configuration in generation t behaves in a stochastic fashion and we shall suppose [adapting a standard model due to Wright (1931)] that if in generation $t - 1$ the number of genes of any existing allele is a, then the probability that in generation t there are b genes of this allele is

$$p_{a,b} = \binom{2N}{b} \left\{ \frac{a(1-u)}{2N} \right\}^{b} \left\{ \frac{2N - a(1-u)}{2N} \right\}^{2N-b}, \qquad \begin{array}{l} a = 1, 2, \ldots, 2N, \\ b = 0, 1, 2, \ldots, 2N. \end{array} \tag{1}$$

Equation (1) is actually insufficient to specify the complete stochastic process under discussion: a proper specification is given in Karlin and McGregor (1967) and involves multinomial transition probabilities. The derivations given here will not use this complete specification but will rely only on (1). To this extent, then, our derivation is partially intuitive: when this is so a proper derivation is given in the accompanying paper (Karlin and McGregor, 1972).

We assume that sufficient time has elapsed so that a stationary situation has been reached. Then from the transition matrix generated by (1) we can define the "effective number of alleles" m_e in the population, defined as the reciprocal of the probability F that two genes drawn at random from the population are of the same allelic type: at equilibrium the latter quantity is (Crow and Kimura, 1964)

$$F = \{2N(1-u)^{-2} - 2N + 1\}^{-1} \simeq (1+4Nu)^{-1},$$

so that

$$m_e \simeq 1 + 4Nu. \tag{2}$$

More generally we have

$$m_e \approx 1 + 4N_e u, \tag{3}$$

where N_e is the effective population size. The error in (3) approaches zero as $N_e \to \infty$ in such a way that $\theta = 4N_e u$ is kept constant.

So far as other quantities are concerned, it seems necessary to use the diffusion approximation to the model (1). Aspects of the relevant diffusion theory will be found in Ewens (1969, Chapters 5 and 6), and we will be content here simply to reproduce the relevant results. First we consider the mean value $E(K)$ of the *actual* number K of alleles in the population in generation t: this is

$$E(K) = \theta^* \left[\int_{(2N)^{-1}}^{1} x^{-1}(1-x)^{\theta^*-1}\,dx\right], \tag{4}$$

where $\theta^* = 4Nu$. In somewhat greater generality we can write

$$E(K) = \theta \left[\int_{(2N)^{-1}}^{1} x^{-1}(1-x)^{\theta-1}\,dx\right], \tag{5}$$

where $\theta = 4N_e u$ and N_e is the variance-effective population size. A considerable portion of our subsequent discussion will relate to the parameter θ.

A statement rather more informative than (5) is that

$$\begin{gathered} E(K; x_1, x_2) = \theta \int_{x_1}^{x_2} x^{-1}(1-x)^{\theta-1}\,dx, \\ (2N)^{-1} \leqslant x_1 \leqslant x_2 \leqslant 1, \end{gathered} \tag{6}$$

is the mean number of alleles occurring in the population with frequency between x_1 and x_2. Clearly the function

$$f(x) = \theta x^{-1}(1-x)^{\theta-1} \tag{7}$$

is such that $f(x)\,\delta x$ is the probability that an allele will occur in the population with frequency in $(x, x+\delta x)$ (since we shall assume that for sufficiently small δx the probability of two alleles having frequency in this range can be ignored). The alternative interpretation that $f(x)\,\delta x$ is the mean number of alleles with frequency in $(x, x + \delta x)$ will also be used on occasion. By a slight abuse of language we shall call $f(x)$ the frequency spectrum of the process, and it will turn out that the approach to sampling problems from a frequency-spectrum point of view considerably facilitates all the arguments.

The shape of the frequency spectrum is strongly dependent on θ. If $\theta \ll 1$, most of the mass in the spectrum is near $x = 1$, indicating that it is most likely that one allele occurs at high frequency in the population, together with a small number of very low-frequency alleles. If $\theta \gg 1$, it is unlikely that any allele will occur with appreciable frequency and one is most likely to observe a comparatively large number of low-frequency alleles. It is a characteristic of the spectrum that one seldom observes a situation with two or three alleles of moderate frequency.

From (7), the probability that a gene drawn at random from the population is of an allele whose frequency in the population is in $(x, x + \delta x)$ is

$$xf(x)\,\delta x = \theta(1 - x)^{\theta-1}\,\delta x, \tag{8}$$

where we may now assume $0 \leqslant x \leqslant 1$ to a sufficient approximation.

A further use of the frequency spectrum is as follows. Suppose that the (random) frequencies of the various alleles occurring in the population in generation t are $p_1, p_2, \ldots$, and consider any function of the form $\Sigma_i \phi(p_i)$, where $\phi(p_i)$ is $O(p_i)$ at most near $p_i = 0$. We have shown above that

$$\begin{aligned}&\text{Pr[an allele occurs in the population with frequency in } (x, x + \delta x)]\\ &\qquad = \theta x^{-1}(1 - x)^{\theta-1}\,\delta x.\end{aligned}$$

It follows from this that

$$E \sum_i \phi(p_i) = \theta \int_0^1 \phi(x)\, x^{-1}(1 - x)^{\theta-1}\, dx, \tag{9}$$

where we may take the lower terminal as zero [because of the assumed nature of $\phi(\cdot)$] with negligible loss of accuracy. Thus, for example,

$$E\left(\sum p_i\right) = \theta \int_0^1 (1 - x)^{\theta-1}\, dx = 1$$

as we expect, while

$$\begin{aligned}&\text{Pr (two genes drawn at random are of same allelic type)}\\ &\qquad = E\left(\sum p_i^2\right) = \theta \int_0^1 x(1 - x)^{\theta-1}\, dx = (1 + \theta)^{-1}.\end{aligned} \tag{10}$$

This rederives (3) and indicates an interesting connection between $E(K)$ and m_e, namely that to a sufficiently close approximation,

$$\theta(1-x)^{\theta-1},\ (2N)^{-1} \leqslant x \leqslant 1$$

is a density function and to this degree of approximation, if z is a random variable from this distribution,

$$m_e = [E(z)]^{-1} \qquad \text{while} \quad E(K) = E(z^{-1}).$$

We now turn to properties of a sample from this population.

Sampling Properties

Suppose a sample of n individuals ($2n$ genes) is drawn from the population. It will be assumed that $n \ll N$ so that although sampling is without replacement, binomial formulae can be used to a sufficient approximation. Note in particular that this means that our results cannot necessarily be taken over to describe population properties, although some such translation does seem possible (see Appendix 1). The first question we ask is: what is the mean number of different alleles in the sample?

If k is the actual number of alleles in the sample, we can write

$$k = a_1 + a_2 + \ldots,$$

where $a_i = 1$ if the i-th allele in the population in generation t is represented in the sample,

$$a_i = 0 \text{ otherwise.}$$

Then

$$\begin{aligned} E(k) &= \sum E(a_i) \\ &= \sum E_p E_s(a_i) \\ &= \sum E_p[1-(1-p_i)^{2n}] \end{aligned}$$

(if we suppose that in the population the various alleles occurring in generation t have frequencies $p_1, p_2, \ldots$). Using (9), this may be written

$$\begin{aligned} E(k) &\simeq \theta \int_0^1 x^{-1}(1-x)^{\theta-1}\{1-(1-x)^{2n}\}\,dx \\ &= \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \cdots + \frac{\theta}{\theta+2n-1}. \end{aligned} \tag{11}$$

Note that if θ is extremely small we have $E(k) \simeq 1$, whereas if θ is extremely (and from a biological point of view unrealistically) large, then $E(k) \simeq 2n$ and we expact all genes in the sample to be of different allelic type. Both these observations agree with the observations made as a result of Eq. (7). Note that despite this agreement, the present model does not strictly apply as $\theta \to 0$ or $\theta \to \infty$.

It is valuable to tabulate $E(k)$ [given by (11)] for selected values of n and θ: a representative set of values is given in Table I.

In view of Eq. (5), it is of some interest to see whether an expression for $E(k)$ can be obtained in terms of an integral. The parallel with (5) makes it natural for us to try an expression of the form

$$E(k) = \theta \left[\int_{(2n)^{-1}}^{1} x^{-1}(1 - x)^{\theta-1}\, dx \right]. \tag{12}$$

Although the mathematical forms of the alternative expressions (11) and (12) are quite different, it is not difficult to show that they give values for $E(k)$ which are almost identical. Further, we can use the parallel with (6) to state that the mean number of alleles $E(k; n_1, n_2)$ to occur in the sample with (sample) frequency in (x_1, x_2) is given by

$$E(k; n_1, n_2) = \theta \int_{x_1}^{x_2} x^{-1}(1 - x)^{\theta-1}\, dx. \tag{13}$$

Note that we may argue in the reverse direction and state, using (11), that an alternative expression for the mean number of alleles in the population in generation t is

$$E(K) = \frac{\theta}{\theta} + \frac{\theta}{\theta + 1} + \cdots + \frac{\theta}{\theta + 2N - 1}. \tag{14}$$

Thus the mean number of alleles in the population *not* observed in the sample is

$$\frac{\theta}{\theta + 2n} + \frac{\theta}{\theta + 2n + 1} + \cdots + \frac{\theta}{\theta + 2N - 1}, \tag{15}$$

or alternatively, using (5) and (12),

$$\theta \int_{(2N)^{-1}}^{(2n)^{-1}} x^{-1}(1 - x)^{\theta-1}\, dx \simeq \theta \log(N/n) - \theta(\theta - 1)[(2n)^{-1} - (2N)^{-1}]. \tag{16}$$

TABLE I

Mean Number of Different Alleles Observed in a Sample of n Individuals (i.e., $2n$ genes) from a Selectively Neutral Population and with Selected Values of $\theta = 4N_eu$

	Value of $\theta = 4N_eu$																	
Value of n	0.001	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.5	2.0	5.0	10.0
10	1.00	1.04	1.34	1.65	1.95	2.22	2.48	2.72	2.96	3.18	3.39	3.60	3.79	3.98	4.51	5.29	8.46	11.33
20	1.00	1.04	1.41	1.79	2.16	2.50	2.83	3.14	3.44	3.73	4.01	4.28	4.54	4.79	5.52	6.61	11.45	16.50
30	1.00	1.05	1.45	1.88	2.28	2.66	3.03	3.38	3.72	4.05	4.37	4.68	4.98	5.27	6.11	7.39	13.30	19.90
40	1.00	1.05	1.48	1.93	2.36	2.78	3.17	3.55	3.92	4.28	4.63	4.97	5.29	5.62	6.54	7.96	14.65	22.42
50	1.01	1.05	1.50	1.98	2.43	2.87	3.28	3.69	4.08	4.46	4.83	5.19	5.54	5.88	6.87	8.39	15.72	24.44
60	1.01	1.05	1.52	2.01	2.49	2.94	3.38	3.80	4.21	4.60	4.99	5.37	5.74	6.10	7.14	8.75	16.59	26.12
70	1.01	1.05	1.54	2.05	2.53	3.00	3.45	3.89	4.31	4.73	5.13	5.52	5.91	6.28	7.37	9.06	17.31	27.56
80	1.01	1.06	1.55	2.07	2.57	3.05	3.52	3.97	4.41	4.83	5.25	5.66	6.05	6.44	7.57	9.32	17.98	28.81
90	1.01	1.06	1.56	2.10	2.61	3.10	3.58	4.04	4.49	4.93	5.36	5.77	6.18	6.58	7.74	9.56	18.56	29.93
100	1.01	1.06	1.57	2.12	2.64	3.14	3.63	4.10	4.56	5.01	5.45	5.88	6.30	6.71	7.90	9.77	19.07	30.93
150	1.01	1.06	1.61	2.20	2.76	3.31	3.83	4.35	4.85	5.34	5.81	6.28	6.74	7.19	8.51	10.57	21.06	34.83
200	1.01	1.07	1.64	2.26	2.85	3.42	3.98	4.52	5.05	5.57	6.07	6.57	7.06	7.54	8.94	11.14	22.48	37.63
250	1.01	1.07	1.66	2.30	2.92	3.51	4.09	4.65	5.20	5.74	6.27	6.79	7.30	7.81	9.27	11.59	23.59	39.82

It is now necessary to derive further characteristics of the sampling distribution of k, and we do this by finding an explicit expression for $\pi_i = \Pr(k = i)$. It will turn out that the probability distribution $\{\pi_i\} = \{\pi_1, \pi_2, \ldots, \pi_{2n}\}$ will be central to the whole discussion of drawing inferences about the population from the sample. Despite the similarity in form between (5) and (12), it does not necessarily follow that an argument similar to that below yields the distribution of K.

We find π_i by a variant of the "coupon collector's problem" (or the "law of succession"). We suppose the sample of $2n$ genes is drawn one by one and consider the probability that the $(j + 1)$-th gene drawn is of an allelic type not observed on the first j draws. Now if we are given the current allele frequencies $p_1, p_2, \ldots$ in the population, this probability is $\sum(1 - p_i)^j p_i$. Use of (9) then shows that the required unconditional probability is

$$\theta \int_0^1 (1 - x)^j x[x^{-1}(1 - x)^{\theta-1}]\, dx = \theta/(\theta + j). \tag{17}$$

Thus the probability that the $(j + 1)$-th gene *is* of one or other allelic type previously drawn is

$$j/(\theta + j). \tag{18}$$

Note also that the probability that the first j draws all yield the same allelic type is

$$\begin{aligned} E\sum_i p_i^j &= \theta \int_0^1 x^j\{x^{-1}(1 - x)^{\theta-1}\}\, dx \qquad \text{[using (9)]} \\ &= \theta(j - 1)!\, \Gamma(\theta)/\Gamma(j + \theta). \end{aligned}$$

Hence, given that the first j draws yield only one allelic form, the probability that the first $j + 1$ draws yield only one allelic form, is

$$\frac{[\theta j!\, \Gamma(\theta)/\Gamma(\theta + j + 1)]}{[\theta(j - 1)!\, \Gamma(\theta)/\Gamma(\theta + j)]} = \frac{j}{(\theta + j)},$$

which is identical to (18). Thus the condition that the first draws all yield the same allelic type does not alter the probability that on the $(j + 1)$-th draw a new type is drawn. We argue more generally that the number of different allelic types drawn on the first j draws has no bearing on the probability that on the $(j + 1)$-th draw a new type is obtained. This arises essentially from the fact [from (1)] that the behaviour of the frequency of any allelic type is Markovian, that is, is independent of the numbers and frequencies of other allelic types. A formal proof of this statement has been obtained by Karlin and McGregor and is given in an accompanying paper (1972). Thus the probability that the first j draws yield i allelic types (which we denote $q_{j,i}$) is given, for $i = 1$, by

$$q_{j,1} = \theta(j - 1)1/\theta(\theta + 1) \ldots (\theta + j - 1), \tag{19}$$

and for $i = j$ by

$$q_{j,j} = \theta^j/\theta(\theta + 1) \ldots (\theta + j - 1). \tag{20}$$

Further we have the recurrence relation

$$q_{j+1,i} = q_{j,i}\{j/(\theta + j)\} + q_{j,i-1}\{\theta/(\theta + j)\}. \tag{21}$$

This recurrence relation, together with the boundary conditions (19) and (20), is sufficient for calculation of $q_{j,i}$ for all (i, j), and in particular for $j = 2n$. It turns out that the most economical and useful way of writing down the solution $\pi_i (= q_{2n,i})$ is by introducing the polynomial

$$\begin{aligned} L(\theta) &= \theta(\theta + 1) \cdots (\theta + 2n - 1) \\ &= l_1\theta + l_2\theta^2 + \cdots + l_{2n}\theta^{2n} \end{aligned} \tag{22}$$

(say): if this is done, then

$$\pi_i = l_i\theta^i/(l_1\theta + l_2\theta^2 + \cdots + l_{2n}\theta^{2n}). \tag{23}$$

The coefficients l_i (written more fully $l_{i,2n}$) are the "Stirling's numbers of the first kind." Properties of these are given in David and Barton (1962, pp. 291–296), where the notation used is $D_{i,2n}$. The theory which follows has interesting parallels with the theory of "record-breaking," discussed in David and Barton (1962, pp. 178–182), and which also makes substantial use of Stirling's numbers. The polynomial $L(\theta)$ is of value in exploring the properties of the distribution $\{\pi_i\}$. Thus we have, for example,

$$\begin{aligned} E(k) &= \sum i\pi_i \\ &= \left(\sum il_i\theta^i\right)\Big/L(\theta) \\ &= \theta L'(\theta)/L(\theta) \\ &= \theta(d/d\theta)\log L(\theta) \\ &= \frac{\theta}{\theta} + \frac{\theta}{\theta + 1} + \cdots + \frac{\theta}{\theta + 2n - 1}, \end{aligned}$$

[using (22)], and this expression agrees with (11). Next we have

$$\begin{aligned} \mathrm{Var}(k) &= \sum i(i - 1)\pi_i + [E(k)] - [E(k)]^2 \\ &= \frac{\theta^2 L''(\theta)}{L(\theta)} + \frac{\theta L'(\theta)}{L(\theta)} - \left[\frac{\theta L'(\theta)}{L(\theta)}\right]^2 \\ &= \frac{\theta L'(\theta)}{L(\theta)} - \theta^2\left[\frac{d^2}{d\theta^2}\log L(\theta)\right] \\ &= E(k) - \left[\frac{\theta^2}{\theta^2} + \frac{\theta^2}{(\theta + 1)^2} + \cdots + \frac{\theta^2}{(\theta + 2n - 1)^2}\right]. \end{aligned} \tag{24}$$

Note that $\operatorname{Var}(k) \to 0$ as $\theta \to 0$ (as we should expect) and also as $\theta \to \infty$ (as again we expect): the most variable situation occurs when θ assumes intermediate values [although we see in Table II that for cases of biological interest, Var (k) increases with θ).

We may also use (23) to compute higher moments of the distribution $\{\pi_i\}$: in particular the third and fourth cumulants of k are

$$\kappa_3 = \theta S_1 - 3\theta^2 S_2 + 2\theta^3 S_3, \tag{25}$$

$$\kappa_4 = \theta S_1 - 7\theta^2 S_2 + 12\theta^3 S_3 - 6\theta^4 S_4, \tag{26}$$

where $S_j = (1/\theta^j) + [1/(\theta+1)^j] + \cdots + [1/(\theta+2n-1)^j]$.

It follows that for n large, the skewness and kurtosis of the distribution of k approach zero. In general, all standardized cumulants above the second approach zero as $n \to \infty$, indicating asymptotic normality of the distribution. Note, however, that our theory is designed and applies for values of n considerably less than N and approximate normality is possibly not reached for values of n arising in practice. Note that the skewness of the distribution of k is always positive for $\theta < 1$ and also positive for sufficiently large n: These may often be the cases of practical interest.

It has been remarked above that our derivations rely on the assumption that $n \ll N$ and cannot be taken over to describe population distributions. This is unfortunate since in computer simulations it is usually necessary to consider properties of these distributions. If our arguments apply approximately for population properties we would have

$$\Pr(K = i) = p_i\theta^i/P(\theta), \tag{27}$$

where

$$\begin{aligned} P(\theta) &= \theta(\theta+1)\cdots(\theta+2N-1) \\ &= p_1\theta + p_2\theta^2 + \cdots + p_{2N}\theta^{2N}. \end{aligned} \tag{28}$$

Computer simulations suggest that (27) is approximately correct for small θ but, for larger θ, yields a distribution whose variance is larger than that suggested by simulations.

It is possible to obtain numerical expressions for $\pi_i (i = 1, 2, \ldots 2n)$ by using the recurrence relation (21). Appendix 3 lists a FORTRAN program in which this is done for values of n up to 1000: the user of this program is required simply to write in the appropriate values of θ and n in statements 1 and 2.

The form (23) is useful to discuss a property of the distribution of k, namely that this distribution is *complete*: that is, there is no function of k whose expected

TABLE II

Variance of Number of Different Alleles Observed in a Sample of n Individuals (i.e., $2n$ genes) from a Selectively Neutral Population and with Selected Values of $\theta = 4N_eu$

	Value of $\theta = 4N_eu$																	
Value of n	0.001	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.5	2.0	5.0	10.0
10	0.00	0.04	0.33	0.61	0.85	1.06	1.26	1.43	1.59	1.74	1.88	2.00	2.12	2.23	2.52	2.90	3.95	4.20
20	0.00	0.04	0.40	0.74	1.06	1.34	1.60	1.84	2.06	2.27	2.47	2.66	2.84	3.00	3.47	4.12	6.48	8.01
30	0.00	0.05	0.44	0.83	1.18	1.50	1.80	2.08	2.34	2.59	2.83	3.05	3.27	3.47	4.05	4.88	8.16	10.82
40	0.00	0.05	0.47	0.88	1.26	1.61	1.94	2.25	2.54	2.82	3.08	3.33	3.58	3.81	4.46	5.42	9.42	13.03
50	0.01	0.05	0.49	0.93	1.33	1.70	2.05	2.38	2.70	2.99	3.28	3.55	3.82	4.07	4.79	5.85	10.42	14.84
60	0.01	0.05	0.51	0.96	1.39	1.78	2.14	2.49	2.82	3.14	3.44	3.73	4.01	4.29	5.05	6.21	11.26	16.37
70	0.01	0.05	0.52	1.00	1.43	1.84	2.22	2.58	2.93	3.26	3.58	3.88	4.18	4.47	5.28	6.51	11.98	17.71
80	0.01	0.06	0.54	1.02	1.47	1.89	2.29	2.66	3.02	3.37	3.70	4.02	4.33	4.63	5.48	6.77	12.60	18.88
90	0.01	0.06	0.55	1.05	1.51	1.94	2.35	2.73	3.10	3.46	3.80	4.13	4.45	4.77	5.65	7.00	13.16	19.94
100	0.01	0.06	0.56	1.07	1.54	1.98	2.40	2.80	3.18	3.54	3.90	4.24	4.57	4.89	5.81	7.21	13.66	20.89
150	0.01	0.06	0.60	1.15	1.66	2.14	2.60	3.04	3.46	3.87	4.26	4.64	5.01	5.37	6.41	8.01	15.61	24.64
200	0.01	0.07	0.63	1.21	1.75	2.26	2.74	3.21	3.66	4.10	4.52	4.93	5.33	5.72	6.84	8.58	17.01	27.36
250	0.01	0.07	0.65	1.25	1.81	2.35	2.86	3.34	3.82	4.27	4.72	5.15	5.57	5.98	7.17	9.02	18.10	29.50

value is zero (or any required constant). For if such a function $g(k)$ existed we would have, from (23),

$$\sum_{i=1}^{2n} g(i)\, l_i \theta^i = 0.$$

This equation can hold (for a continuum of θ values) only when $g(i) \equiv 0$. We shall make use of this property of completeness on several occasions in our later discussion.

Inference Properties

We now turn to the question of what inductions about the population can be made from the sample.

We consider first the maximum likelihood estimator of θ, given only the observed value k of alleles in the sample; (we shall show later that ignoring the numbers $n_i \cdots n_k$ with which these occur in the sample involves no loss of information). The likelihood of observing k alleles is, from (23),

$$l_k \theta^k / L(\theta),$$

and this is maximized with respect to variation in θ when

$$\begin{aligned} k &= \theta L'(\theta)/L(\theta) \\ &= \frac{\theta}{\theta} + \frac{\theta}{\theta + 1} + \cdots + \frac{\theta}{\theta + 2n - 1}. \end{aligned}$$

In other words, the maximum likelihood estimator of θ is that value for which the expected number of alleles equals the observed number of alleles. Thus maximum likelihood estimates can be found by using Table I backwards and interpolating; however for convenience we exhibit in Table III the maximum likelihood estimators of θ for selected values of k and n.

A "symmetric" 95% confidence interval (θ_0, θ_1) for θ can be found from Neyman–Pearson theory by computing θ_0 and θ_1 from

$$\Pr\{\text{less than } k \text{ alleles} \mid \theta = \theta_1\} = 0.025,$$

$$\Pr\{\text{more than } k \text{ alleles} \mid \theta = \theta_0\} = 0.025.$$

The numercial program given in Appendix 3 may be used (with some trial and error) to evaluate θ_0 and θ_1 from these equations.

A confidence interval of some special interest arises when a sample of n indi-

TABLE III

Maximum Likelihood Estimates of θ for a Given Observed Number (k) of Different Alleles in a Sample of Size n ($2n$ genes)

	Value of k											
Value of n	1	2	3	4	5	6	7	8	9	10	11	12
10	0.000	0.319	0.718	1.209	1.805	2.524	3.391	4.440	5.718	7.290	9.252	
20	0.000	0.256	0.553	0.894	1.279	1.709	2.187	2.715	3.296	3.936	4.638	5.409
30	0.000	0.230	0.492	0.782	1.104	1.455	1.836	2.247	2.690	3.165	3.674	4.217
40	0.000	0.215	0.455	0.720	1.008	1.319	1.653	2.011	2.390	2.793	3.219	3.669
50	0.000	0.204	0.431	0.678	0.945	1.232	1.538	1.862	2.204	2.565	2.944	3.342
60	0.000	0.197	0.414	0.849	0.902	1.172	1.459	1.762	2.080	2.414	2.763	3.127
70	0.000	0.191	0.400	0.626	0.868	1.125	1.397	1.684	1.982	2.295	2.622	2.962
80	0.000	0.186	0.389	0.607	0.840	1.087	1.347	1.620	1.905	2.203	2.512	2.834
90	0.000	0.182	0.379	0.591	0.816	1.055	1.306	1.568	1.842	2.127	2.424	2.730
100	0.000	0.178	0.371	0.578	0.797	1.029	1.272	1.526	1.790	2.065	2.350	2.644
150	0.000	0.166	0.343	0.532	0.731	0.939	1.157	1.382	1.616	1.858	2.107	2.364
200	0.000	0.158	0.326	0.504	0.691	0.886	1.088	1.297	1.514	1.737	1.966	2.201
250	0.000	0.152	0.314	0.485	0.663	0.848	1.040	1.239	1.444	1.654	1.870	2.091

viduals gives rise to a single allele and a confidence interval of the form $(0, \theta_1)$ is required for θ. If a 95% confidence interval is wanted, the value θ_1 satisfies

$$\theta_1(2n-1)!/L(\theta_1) = 0.05,$$

Appendix 3 again being useful for determination of θ_1.

The above inferences relate to the parameter θ, which is the only population characteristic about which inferences can be made from the sample. [Thus note, for example, that it is impossible, from (15) and (16), to make exact inferences about the number of alleles *not* observed in the sample, since the nonestimable quantity N is involved in the distribution of this quantity. On the other hand the form of (16) indicates that if an estimate of N can be made by independent methods, then a reasonably good estimate of this quantity can be found.] We now note that our inferences about θ have used n and k only: that is, they have ignored the numbers $n_1, n_2, \ldots, n_k$ with which the alleles appear in the sample. We now show that this procedure is justified by proving that k is a *sufficient statistic* for θ; in other words, given k, the distribution of $n_1, \ldots, n_k$ is *independent* of θ, and so knowledge of these quantities provides no further information about θ. This is done by noting (Appendix 2) that

$$\Pr\{n_1, \ldots, n_k\} = g(n_1, \ldots, n_k)\, \theta^k / L(\theta). \tag{29}$$

The conditional distribution of $n_1, \ldots, n_k$, given k, is thus independent of θ. Hence not only may the numbers $n_1 \ldots n_k$ be ignored when carrying out inferences about θ: standard Rao–Blackwell theory indicates that it is *inefficient* to make use of these quantities in such inferences.

The sufficiency and completeness of the distribution of k indicates that there is a unique "best" (minimum variance unbiased) estimator of any estimable function of θ, and that such an estimator will be a function of k only. One particular function of θ which one sometimes wishes to estimate is the "effective number of alleles" $1 + \theta$ [cf. Eq. (3)]. This is normally estimated by $1/\sum x_i^2$, where $x_i = n_i/(2n)$. The above shows that such an estimator is very *inefficient* in that it uses precisely the information in the sample that should be ignored. We now show that in fact $1 + \theta$ is not an estimable function, for if there were an unbiased estimator of $(1 + \theta)$ there would be a (unique) function $g(k)$ unbiasedly estimating $1 + \theta$. We would then have

$$\sum_{i=1}^{2n} g(i)\, l_i \theta^i = (1 + \theta) L(\theta)$$

for all θ, and this is impossible since the left-hand side is a polynominal of degree $2n$ at most whereas the right-hand side is a polynomial of degree $2n + 1$. Thus it appears that probably the most satisfactory estimate of the effective number of alleles is $1 + \hat{\theta}$, where $\hat{\theta}$ is the maximum likelihood estimator of θ, even though this is a biased estimator. It is however possible to find a unique "best" estimator of $(1 + \theta)^{-1}$. If we denote this estimator by G(k), we must have

$$\sum_{i=1}^{2n} G(i)\, l_i \theta^i = \theta(\theta + 2)(\theta + 3) \cdots (\theta + 2n - 1),$$

so that $G(2n) = 0$ and

$$G(j) = \frac{\text{coeff } \theta^{j-1} \text{ in } (\theta + 2)(\theta + 3) \cdots (\theta + 2n - 1)}{\text{coeff } \theta^{j-1} \text{ in } (\theta + 1)(\theta + 2) \cdots (\theta + 2n - 1)},$$
$$(j = 1, 2,..., 2n - 1).$$

In particular we have

$G(1)$ = estimate of $(1 + \theta)^{-1}$ if 1 allele is observed in a sample of n individuals
$= 1,$

$G(2)$ = estimate of $(1 + \theta)^{-1}$ if 2 alleles are observed in a sample of n individuals

$$= \left[\frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{(2n-1)}\right] \Big/ \left[1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{(2n-1)}\right],$$

$G(3)$ = estimate of $(1 + \theta)^{-1}$ if 3 alleles are are observed in a sample of n individuals

$$= \frac{\left[\dfrac{1}{2 \cdot 3} + \dfrac{1}{2 \cdot 4} + \cdots + \dfrac{1}{(2n-2)(2n-1)}\right]}{\left[\dfrac{1}{1 \cdot 2} + \dfrac{1}{1 \cdot 3} + \cdots + \dfrac{1}{(2n-2)(2n-1)}\right]},$$

and so on. In general, the only estimable functions of θ are linear combinations of functions of the form $1/H(\theta)$, where $H(\theta)$ is a polynomial in θ of the form $(\theta + a)(\theta + b) \cdots (\theta + c)$, where a, b, ..., c are positive integers satisfaying $1 \leqslant a < b < ... < c \leqslant 2n - 1$.

Index Function

It is sometimes desired to find an index function $I(k, x_1, ..., x_k)$ whose expected value is independent of θ and known for selectively neutral populations. It is sometimes even further desired that the index function be a function of k only. Such an index function could hopefully be used as a measure of departure from selective neutrality. Our previous theory shows that it is impossible to construct a nontrivial index function that is a function of k only. For if $E\Phi(k) = A$ (A

constant), then the completeness of the distribution of k (proved above) yields $\Phi(k) \equiv A$, which is trivial. We shall exhibit later a useful index function $L = L(k, n_1, ..., n_k)$, which of course depends nontrivially on $n_1, ..., n_k$ as well as k.

Hypothesis Testing

A large number of tests of hypotheses may be made in areas related to the above discussion, and we shall consider here only two such tests.

Firstly, it is possible to test hypotheses about the value of the parameter θ, assuming selective neutrality does hold. Such tests must be carried out using the number k *only*, and will normally be closely associated with the procedure for finding confidence intervals for θ outlined above. The theory of such tests is straightforward and is not considered further here.

Second, it is possible to devise a test of the hypothesis that the alleles sampled are selectively neutral. Here the numbers $n_1 ... n_k$ become relevant and indeed our test of the hypothesis is carried out essentially by testing whether the observed values $n_1 ... n_k$ conform reasonably to what is expected under selective neutrality. The nuisance parameter θ is eliminated from such tests by making the tests conditional on the observed value of k. Thus we have, from Appendix 2,

$$\Pr\{n_1, ..., n_k \mid k, 2n, \text{neutrality}\} = (2n)!/(k!\, l_k n_1 \cdot n_2 \cdots n_k). \tag{30}$$

More precisely, we assume that, once k is given, the alleles are labelled $\mathrm{A}_1 \cdots \mathrm{A}_k$ in some conventional manner: Eq. (30) yields the probability that there are n_1 genes of the allele labelled A_1, ..., n_k genes of the allele labelled A_k.

Equation (30) can now be used in principle to find the distribution of any test statistic $t(n_1, ..., n_k)$. Standard statistical methods normally yield an "optimal" test statistic once an alternative hypothesis is given. Here the alternative hypothesis is "selection exists," and it is not clear that this is sufficiently precise to yield an unambiguous statistic. Here we shall simply note that if heterotic selection exists, the values $n_1, ..., n_k$ will tend to be "close," whereas if one allele is selected favoured, one of the n_i values will tend to be very large, the others being quite small. Under selective neutrality a situation intermediate between these two should obtain. Thus if we define $x_i = n_i/2n$ (= frequency of the i-th allele in the sample), it seems reasonable to introduce as an index of neutrality, and also as a test statistic for the hypothesis of neutrality, the "information" function

$$B = -\sum_{i=1}^{k} x_i \log x_i\,. \tag{31}$$

We shall not attempt any mathematical justification for using B other than noting the well-known properties enjoyed by the information statistic.

The complete distribution of B under the neutrality theory is complicated, and we shall be satisfied here with finding only the mean and variance of this distribution. Defining

$$i!\, l_{i,j}/j! = w_{i,j}\ ,$$

eq. (30) shows that if H is the domain $n_i \geqslant 1$, $\Sigma n_i = 2n$, then

$$\sum_H (n_1 \cdots n_k)^{-1} = w_{k,2n}\ . \tag{32}$$

Then the distribution $p(n_1)$ of n_1, found by appropriate summation in (30), is

$$p(n_1) = w_{k-1,2n-n_1}/(n_1 w_{k,2n}), \qquad n_1 = 1,\dots, 2n - k + 1. \tag{33}$$

From (33) we have

$$\begin{aligned} &E\left[-\sum x_i \log x_i\right] \\ &\quad = E\left[\log 2n - (2n)^{-1} \sum n_i \log n_i\right] \\ &\quad = \log 2n - (2n)^{-1}\, k \sum_{j=1}^{2n-k+1} (\log j)\, w_{k-1,2n-j}/w_{k,2n}\ . \end{aligned} \tag{34}$$

Computation of (34) is possible using standard recurrence relations for Stirling numbers and is embodied in the FORTRAN program in Appendix 4.

Similarly, the joint distribution $p(n_1\, , n_2)$ of n_1 and n_2 is found (for $k \geqslant 3$) to be

$$p(n_1\, , n_2) = w_{k-2,2n-n_1-n_2}/(n_1 n_2 w_{k,2n}). \tag{35}$$

Joint use of (33) and (35) yields the variance of $B = -\sum x_i \log x_i$ under the neutrality theory for $k \geqslant 3$; (a rather simpler formula obtains when $k = 2$). In both cases the necessary computations are embodied in the program in Appendix 4.

Since the exact distribution of B is not known, an exact test of selective neutrality cannot yet be given. An approximate procedure assumes that $B/\log k$ has a beta distribution, the parameters of which can be found from $E(B)$ and Var (B): a standard transformation then yields a random variable with an approximate F distribution. Part of the print-out in the program in Appendix 4 provides the value of F so calculated, together with its (non-integer) degrees of freedom. Values of F significantly in excess of standard significance points indicate evidence of heterotic selection, while large values of $1/F$ (with reversed degrees of freedom) indicate selection favouring one allele.

An even more approximate test would be to calculate the standardized variable

$$L = [B - E(\mathrm{B})]/\sigma(B) \tag{36}$$

which, under the neutrality theory, has mean zero and standard deviation 1. An approximate two-standard deviation rule could be used to test for significance of L.

We exemplify these points by considering three hypothetical samples, each with $N = 350$, $k = 4$, but with different allele frequencies in the three cases.

Sample	Allele frequencies				L	F(d.f.)	Comment
1	0.35	0.30	0.20	0.15	2.36	34.44(3.30,4.36)	F significantly large
2	0.83	0.11	0.04	0.02	0.02	1.02(3.30,4.36)	not significant
3	0.99	0.005	0.0025	0.0025	−1.70	0.07(3.30,4.36)	F significantly small

We shall not, however, emphasize the F test and prefer rather to concentrate attention on the statistic L, not used as a test function but rather simply as a reasonable index function of non-neutrality. We do this because electrophoretic data usually comprise results from several loci, several species and several geographical locations. Thus while the test outlined above can be applied to any single such observation, problems naturally arise regarding piecing together the results of a number of such tests (all related in some logical fashion) and in computing an overall significance level. Since in any event we do not expect our procedures to be any more than an adjunct to other methods in testing for non-Darwinian evolution, it may often be best to compute L for all sets of observations and to draw inferences not only from the various absolute values of L, but also from the patterns exhibited by the values of L for all the data at hand. Naturally, inferences so drawn involve a degree of subjective decision and thus should be used with caution. Nevertheless, in a forthcoming paper (Ewens and Langley, 1972) which investigates actual data using the above analysis, this approach will to a large extent be used.

Discussion

The aspect of the above discussion likely to be of most practical interest is the test of the hypothesis of selective neutrality based on the frequencies of the various alleles in a sample from some population. It is therefore important to add some words of caution relating to the use of this test. Firstly, the test does not appear particularly powerful (in the statistical sense): as a consequence, one may often maintain the hypothesis of neutrality when in fact (mild) selection does occur. Associated with this remark is the fact that any frequency configuration, and in particular that pertaining to a selectively neutral situation, can be explained by one or other selective system. Secondly, much data now exists where the same locus is considered in related populations or species, or in the same population at different time points, or where different loci in the same population are studied. In some cases of this nature a quite subjective decision may be superior to one based on the objective procedure described above. Thus if a certain allele occurs at high frequency in a large number of related species, with other (low-frequency) alleles occurring in approximately the same frequencies in the different species, one is tempted, without any formal statistical analysis, to conclude that the high-frequency allele is selected for in all the species. This is a reasonable procedure, but in order to develop an objective statistical test and in order also to obtain a testing procedure more powerful than that outlined, it would be valuable to extend the above statitstical test to cover such situations.

A third problem relates to nonidentification. The above analysis assumes total ability to differentiate alleles, whereas it is likely that even present methods often identify two different but very similar alleles.

Further, complications due to linkage, fluctuation in population size, etc., have not been considered. It will probably be necessary to extend the above method to a considerable degree, taking all these factors into account, before highly reliable methods are attained. Nevertheless it is believed that the theory advanced in this paper will be of some use in obtaining these methods.

Summary

The question of making inferences from a sample of genes from a locus is considered. If the locus is supposed neutral, the complete probability distribution of the number of different alleles seen is obtained. It is shown that this number is a sufficient statistic for the parameter θ characterizing the distribution of allele number and allele frequency. This indicates that previous estimators of $1/(1 + \theta)$ have been based on inefficient procedures. The derivation of minimum variance-unbiased estimators for estimable functions of θ is discussed.

An index function is constructed which can be used as a measure of "non-neutrality".

The question of finding confidence limits for θ is solved and a computer program provided for this purpose.

Finally the question of testing hypotheses, in particular that of selective neutrality, is considered. A test is provided for the latter hypothesis. The extensions of this test which will be necessary before powerful procedures are available are also discussed.

Appendix 1

The theory given in this paper relates to a sample whose size is supposed considerably less than the population from which it was taken. In particular in connection with computer simulations, which cannot normally use a value of N in excess of 1000, it is of interest to ask how much of the theory can be used to describe population properties.

The similarity in form between (5) and (12) (and between (11) and (14)) show that an immediate translation appears possible so far as mean numbers of alleles are concerned. However, as remarked elsewhere in this paper, formula (24) (with K replacing k and N replacing n) appears to overestimate the true variance of K, at least for moderate and large values of θ. This possibly arises because the distribution of k incorporates two random components: first, the stochastic nature of the parent population and second, randomness associated with the sampling process.

Note however that computer simulations indicate that one important sample property is maintained in the population, namely, that estimation of functions of θ is carried out best by using K only and ignoring the frequencies of the various alleles in the population.

Appendix 2

We wish to establish formula (29). To do so it is convenient to introduce the notation, for example, that

$$\{m_1\mathrm{A}_1\,,\, m_2\mathrm{A}_2\,,\, m_3\mathrm{A}_3\,,\, m_4\mathrm{A}_2\,,\, m_5\mathrm{A}_1\,,\ldots\}$$

is an ordered collection of genes in which the first m_1 genes sampled are of the same allelic type, the next m_2 are of the same allelic type but different from the first, the next m_3 are of the same (third) allelic type, the next m_4 are all of the second allelic type, the next m_5 are all of the first allelic type found, and so on. We wish to find an expression for the probability that in our sample we have n_1 genes of one specified allelic type, ..., n_k of a k-th specified allelic type. To do this we shall first find an expression for

$$\Pr\{n_1A_1\,,\, n_2A_2\,,\ldots,\, n_kA_k\},$$

that is, for the probability that we first draw a run of n_1 genes of one allelic type, followed by a run of n_2 genes of a second allelic type, and so on, concluding with a run of n_k genes of a k-th allelic type. (Note that the probability in question thus relates to an *ordered* sequence of outcomes.) We shall use repeatedly the facts that if j draws have been made, the probability of finding on the $(j+1)$-th draw a new allelic type not seen on the first j draws is $\theta/(\theta+j)$, and that the probability of any ordered sequence is the same as that of any other ordered sequence of the same length with the same numbers of the various alleles.

Consider first the case $k = 2$. We know that

$$\Pr\{n_1A_1\,,\,1A_2\} = \theta^2(n_1-1)!/\{\theta(\theta+1)\ldots(\theta+n_1)\}, \tag{A1}$$

and we seek to prove by induction the formula

$$\Pr\{n_1A_1\,,\,n_2A_2\} = \theta^2(n_1-1)!(n_2-1)!/\{\theta(\theta+1)\ldots(\theta+n_1+n_2-1)\}. \tag{A2}$$

Suppose (A2) is true for $n = m$. Then since

$$\begin{aligned}\Pr\{n_1A_1\,,\,(m+1)\,A_2\} &= \Pr\{n_1A_1\,,\,mA_2\} - \Pr\{n_1A_1\,,\,mA_2\,,\,1A_1\}\\ &\quad - \Pr\{n_1A_1\,,\,mA_2\,,1A_3\}\\ &= \Pr\{n_1A_1\,,\,mA_2\} - \Pr\{(n+1)\,A_1\,,\,mA_2\}\\ &\quad - \Pr\{n_1A_1\,,\,mA_2\,,\,1A_3\}\\ &= \theta^2(n_1-1)!(m-1)!/\{\theta(\theta+1)\cdots(\theta+n_1+m-1)\}\\ &\quad - \theta^2n_1!(m-1)!/\{\theta(\theta+1)\cdots(\theta+n_1+m)\}\\ &\quad - \theta^3(n_1-1)!(m-1)!/\{\theta(\theta+1)\cdots(\theta+n_1+m)\}\\ &= \theta^2(n_1-1)!m!/\{\theta(\theta+1)\cdots(\theta+n_1+m)\},\end{aligned}$$

we see by induction that (A2) is true for arbitrary n_2. We now aim to show by induction on k that

$$\begin{aligned}&\Pr\{n_1A_1\,,\ldots,\,n_kA_k\}\\ &\quad= \theta^k(n_1-1)!\cdots(n_k-1)!/\{\theta(\theta+1)\cdots(\theta+n_1+\cdots+n_k-1)\}.\end{aligned} \tag{A3}$$

If (A3) is true for some value of k, we have

$$\begin{aligned}&\Pr\{n_1A_1\,,\ldots,\,n_kA_k\,,\,1A_{k+1}\}\\ &\quad= \theta^{k+1}(n_1-1)!\cdots(n_k-1)!/\{\theta(\theta+1)\cdots(\theta+n_1+\cdots+n_k)\}.\end{aligned}$$

Proceeding as above we find easily that

$$\begin{aligned}&\Pr\{n_1A_1\,,\ldots,\,n_{k+1}A_{k+1}\}\\ &\quad= \theta^{k+1}(n_1-1)!\cdots(n_{k+1}-1)!/\{\theta(\theta+1)\cdots(\theta+n_1+\cdots+n_{k+1}-1)\}.\end{aligned}$$

Comparison of this formula with (A3), together with the truth of (A3) for $k = 1$, indicates that (A3) is true for all k.

The probability required in Eq. (29) will differ from that given in (A3) only be a combinatorial multiplicative factor independent of θ. Thus Eq. (29) is true, and exact evaluation of $g(n_1, ..., n_k)$ is unnecessary.

It remains to establish Eq. (30). We make the condition that the sample of $2n$ genes has yielded exactly k allelic types. The probability that these appeared as a run of n_1 of one type, followed by a run of n_2 of a second type, and so on, is found by dividing the right-hand side in (A3) by that in (23) (putting $i = k$), and is thus

$$(n_1 - 1)! \cdots (n_k - 1)!/l_k .$$

Now suppose that these k different types had been labelled in some conventional manner as $A_1 \cdots A_k$. The probability, given k and $2n$, that the genes had been sampled as a run of n_1 of A_1, followed by a run of n_2 of A_2, and so on, is thus

$$(n_1 - 1)! \cdots (n_k - 1)!/k!l_k . \tag{A4}$$

Finally, we wish to disregard the order in which the genes were drawn to compute the conditional probability, given k and $2n$, that there were n_1 genes of type A_1, n_2 of type A_2, etc., in any arbitrary order. This will be found by myltiplying (A4) by

$$(2n)!/(n_1!n_2! \cdots n_k!)$$

and is thus

$$(2n)!/\{k!l_k n_1 n_2 \cdots n_k\}. \tag{A5}$$

This is the required Eq. (30). It should be kept in mind that Eq. (A5) relates to *labelled* alleles. That is, once k is given, it is assumed that the alleles are labelled $A_1 \cdots A_k$ in some conventional manner: the probability (A5) is then the probability that there are n_1 genes of the ellele labelled $A_1, ..., n_k$ of the allele labelled A_k.

A further consequence of equation (A5) is that a higher probability attaches to a case where $n_1 \cdots n_k$ differ widely than to a case where $n_1 \cdots n_k$ are close: this agrees generally with the discussion following Eq. (7), although the parallel is not immediate as one must take account also of the number of possible permutations for any configuration of $n_1, ..., n_k$.

Appendix 3

The following FORTRAN program prints out values of π_i [see Eq. (23)] for any θ and any n ($n \leqslant 1000$). It is required merely to insert the desired values at statements 1 and 2. (Note that minor program changes may be needed for some computers.)

```
  DIMENSION P(2000), Q(2000)
1 THETA =
2 N =
  P(1) = 1.
  Q(1) = 1.
  M = 2 * N
  DO 3 K = 2, M
  P(K) = 0.
  Q(K) = 0.
3 CONTINUE
4 FORMAT (*  VALUE OF K   PROBABILITY
 X     CUM     PROBABILITY      *)
  PRINT 4
  DO 6 J = 2, M
  A = J - 1
  Q(1) = Q(1) * A/(THETA + A)
  DO 5 I = 2, J
  K = I - 1
  Q(I) = (A * P(I) + THETA * P(K)/(THETA + A)
5 CONTINUE
  DO 6 I = 1, J
  P(I) = Q(I)
6 CONTINUE
  CUM  = 0.
  K = 0
7 K = K + 1
  CUM = CUM + P(K)
  PRINT  8, K, P(K), CUM
8 FORMAT (I 10, 2F 21.6)
  IF (CUM. LT. 0.99999) GO TO 7
9 CONTINUE
  END
```

Appendix 4

The following program prints out values of B, L, F and degrees of freedom for F (see text for definitions) for any set of data required. Values of n and k should be inserted in the first statement and values for x_1, x_2,... in the fifth, sixth,..., statements. In the unlikely event $x_1 = x_2 = \cdots x_k$, the user should not proceed with this program, and may assume immediately that his data yield significant evidence of heterotic selection. The program is written for the Univac 1108 computer of the University of Wisconsin and may require minor modifications for other computers. Double precision should be used if possible.

```
      PARAMETER N =, K =
      PARAMETER NN = 2 * N, N1 = NN - 1, K1 = K - 1, K2 = K - 2
      DIMENSION W(K, NN)
      DIMENSION X(K)
      X(1) =
      X(2) =
      ...
      D = K
      D1 = K1
      SUM = 0.
      DO 1 I = 1, K
      A = X(I)
      B = ALOG(A)
      SUM = SUM - A * B
    1 CONTINUE
      DO 2 I = 1, K
      W(I, I) = 1.
    2 CONTINUE
      DO 3 I = 1, K1
      J = I + 1
      DO 3 L = J, K
      W(L, I) = 0.
    3 CONTINUE
      DO 4 I = 2, NN
      Z = I
      W(1, I) = 1./Z
    4 CONTINUE
      DO 5 I = 2, K
      J = I + 1
      Z = I
      Y = M
      W(I, M) = (Y - 1.) * W(I, M - 1)/Y + Z * W(I - 1, M - 1)/Y
    5 CONTINUE
      SIGA = 0.
      SIGB = 0.
      SIGC = 0.
      SIGD = 0.
      DD = NN
      DO 6 I = K1, N1
      J = NN - I
      A = J
      C = ALOG(A)
      PP = A/DD
      R = ALOG(PP)
      P1 = 1. - PP
      R1 = ALOG(P1)
      SIGA = SIGA + C * W(K1, I)
      SIGB = SIGB + A * C * C * C * W(K1, I)
      SIGD = SIGD + (PP * R + P1 * R1) * (PP * R + P1 * R1) * W(K1, I)/
     1    (A * W(K, NN))
    6 CONTINUE
      SIGA = SIGA * D/W(K, NN)
      SIGB = SIGB * D/W(K, NN)
      AVINF = ALOG(DD) - SIGA/DD
      IF (K.EQ.2) GO TO 10
      L1 = NN - K + 1
      DO 7 I = 1, L1
      L2 = L1 + 1 - I
      DO 7 J = 1, L2
      L3 = NN - I - J
      SIGC = SIGC + ALOG(I) * ALOG(J) * W(K2, L3)
    7 CONTINUE
      SIGC = SIGC * D * D1/W(K, NN)
      BV = SIGB + SIGC - SIGA * SIGA
      VARINF = BV/(DD * DD)
      GO TO 11
   10 VARINF = SIGD - AVINF * AVINF
   11 CONTINUE
      SDINF = SQRT(VARINF)
      V = (SUM - AVINF)/(SDINF)
      BB = ALOG(K)
      X = SUM/BB
      TX = AVINF/BB
      VAX = VARINF/(BB * BB)
      DF1 = 2. * TX * (TX * (1. - TX)/VAX - 1.)
      DF2 = DF1 * (1. - TX)/TX
      FR = (1. - TX) * X/(TX * (1. - X))
      PRINT 9
```

```
9  FORMAT (1H1, 20X, 15H VALUE OF B ,5X, 15H VALUE OF L ,5X,
 1    15H  F RATIO     ,5X, 30H   DEGREES OF FREEDOM OF F)
   PRINT 8, SUM, V, FR, DF1, DF2
8  FORMAT (1HO, 20X, 3(F15.4, 5X), 2F15.4)
   END
```

Acknowledgments

My principal debt is to James F. Crow, who has made many valuable suggestions and given considerable help during this research. The work was carried out under the hospitality of K. Kojima, who also contributed many suggestions. I have had invaluable aid from Charles Langley and Professor F. N. David. The problem of construction of an index function and of forming objective statistical tests of the neutrality hypothesis was suggested by R. K. Selander. The final draft of this paper was prepared after the author received many incisive comments by S. Karlin. Professors Karlin and McGregor have also provided in an accompanying paper proofs of several key results used in this paper.

References

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

David, F. N. and Barton, D. E. 1962. "Combinatorial Chance," Griffin, London.

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

Ewens, W. J. and Langley, C. 1972. To appear.

Karlin, S. and McGregor, J. L. 1967. The number of mutant forms maintained in a population. *In* "Proc. 5th Berk. Symp. Math. Stat. and Prob.," Vol. IV, pp. 415–438, University of California Press.

Karlin, S. and McGregor, J. L. 1972. Addendum to a paper of W. Ewens, *Theor. Pop. Biol.* **3**, X X–X X (following paper).

Wright, S. 1931. Evolution in Mendelian Populations, *Genetics* **16**, 97–159.