THEORETICAL POPULATION BIOLOGY 7, 256–276 (1975)

# On the Number of Segregating Sites in Genetical Models without Recombination

G. A. WATTERSON

*Monash University, Clayton, Australia, 3168*

Received June 24, 1974

The distribution is obtained for the number of segregating sites observed in a sample from a population which is subject to recurring, new, mutations but not subject to recombination. After allowance is made for the different effective population sizes, the results apply approximately to three population models, due to Wright, Burrows and Cockerham, and Moran. Included as extreme special cases are the distributions of the number of segregating sites in the whole population and of the number of heterozygous sites in a diploid individual. Some results of Fisher, Haldane, Kimura, and Ewens concerning the means of the distributions for different models are confirmed, but the variances, and the distributions themselves, are new.

## 1. INTRODUCTION

In this section, we outline the assumptions we use to obtain results concerning the probability distribution for the number of segregating sites found in a sample of gametes. We then present the major results of the paper, whose proofs are given in later sections of the paper. We also compare these results with those obtained previously, and find that some marked differences appear. This section concludes with some comments connecting the present work on segregating sites with related work on the number of alleles found in a sample.

### 1.1. *Assumptions*

A *cistron* has been defined as a portion of DNA specifying a single polypeptide chain of an enzyme. As such, the cistron is the functional *gene*, and is known to consist of a large number of nucleotide sites. Recombination of genetic information, due to crossing over and breakage, is rare within a cistron, but mutation may occur at any of the nucleotide sites.

In the present paper we discuss the behaviour of idealized citrons in which there is *no* recombination between sites. Of course, this assumption is meaningful only if the gametes making up our population go through a diploid phase, as they do in Burrows and Cockerham's model to be described below. For



Copyright © 1975 by Academic Press, Inc.
All rights of reproduction in any form reserved.

our other models of interest, the sites could be fully linked, or on entirely separate chromosomes, and it would make no difference. Our *essential* assumption is that an offspring gamete inherits *all* its citron sites from *one* parental gamete. That being so, the sites could just as well be those which constitute the whole genetic information, not just those pertaining to a particular cistron.

In our work, cistrons each contain infinitely many sites. Each site may be subject to mutation during meiosis, with the total number of mutant sites, per cistron per generation, being a random variable having a Poisson distribution with mean $\nu$. Further, we assume that the numbers of mutations which occur to different gametes are independent. Because there are infinitely many sites, we also assume that no two mutations ever occur at the same site (even in different gametes) so that at each site there are only two possible nucleotides, the original wild type and the mutant type.

The assumptions concerning how one generation succeeds another are classified into three distinct models. In the first, to be studied in detail in Section 2, the generations are kept distinct, and each offspring gamete is a (possibly mutant) copy of one of the parental gametes chosen at random with replacement. We shall refer to this as *Wright's model.*

In Burrows and Cockerhams' (1974) model, the generations are also kept distinct, and in each generation, the parental gametes are paired off at random into diploids, and each diploid produces exactly two offspring gametes. This model is studied in Section 3.

The third model, Moran's (1958) model, considers a population in which generations are overlapping. Indeed, at each unit time, *one* randomly chosen gamete (or haploid individual) is replaced by one new gamete, the latter being formed as in Wright's model above. The model is studied in detail in Section 4.

For all three models, we assume that the number of gametes is held fixed at $2N$ throughout all time. The three models have different effective diploid sizes, however, as follows.

$$N_e = \begin{cases} N, & \text{for Wright's model,} \\ 2N - 1, & \text{for Burrows and Cockerham's model,} \\ \tfrac{1}{2}N, & \text{for Moran's model,} \end{cases} \tag{1.1}$$

and therein lies their interest. It is, in fact, possible to incorporate a diploid phase in Wright's model, but it is less relevant for Moran's model.

### 1.2. *The distribution of the Number of Segregating Sites*

We consider the sites of a particular cistron possessed by each of $2N$ gametes in the population. If a sample of $i$ gametes is chosen at random, we denote by $K_i$ the number of sites, in the cistron of interest, which are segregating across members of the sample, that is, sites at which both the wild type and the mutant nucleotide type are present. Of course, $K_{2N}$ denotes the number of sites

segregating across the whole population, and $K_2$ denotes the number of segregating ("heterozygous") sites in two gametes, which latter can be interpreted as the two gametes randomly chosen to make up a diploid individual according to a "random union of gametes" scheme.

If the cistron of a particular gamete gained a Poisson-distributed number of extra mutant sites during meiosis, mean $\nu$, then the probability that at least one site was so effected is the cistron mutation rate $u = 1 - e^{-\nu}$, $\doteqdot \nu$ if $\nu$ is small. We introduce the parameter

$$\theta = 4N_e u \doteqdot 4N_e \nu = 2(N_e/N)\nu_m , \tag{1.2}$$

in which $\nu_m = 2N\nu$ (in Kimura's (1969) notation) is the mean number of new mutant sites for the population, per generation. Our approximate results are obtained by keeping $\theta$ fixed but considering large values for $N_e$ and correspondingly small values for $u$ and $\nu$. The results quoted below assume also that the populations have reached statistical equilibrium.

The main result of the paper is that the probability generating function for $K_i$, in any of the three models in (1.1), is given approximately by

$$E(s^{K_i}) \doteqdot \begin{cases} \prod_{j=1}^{i-1} [1 + \theta(1-s)/j]^{-1}, & \text{for } i = 2, 3,..., \text{ but small,} \quad (1.3a) \\ \exp\{\theta[\log i + \tfrac{1}{2}g(i/2N)](s-1)\}\, \Gamma(\theta + 1 - \theta s), & i \text{ large.} \quad (1.3b) \end{cases}$$

In (1.3b), $g(i/2N)$ is a number close to 0, but it depends on the model in question, and on the fraction, $i/2N$, that the sample constitutes of the population.

We may deduce that the mean of $K_i$ is approximately

$$E(K_i) \doteqdot \begin{cases} \theta \sum_{j=1}^{i-1} 1/j, & \text{for } i = 2, 3,..., \text{ but small,} \quad (1.4a) \\ \theta[\log i + \gamma + \tfrac{1}{2}g(i/2N)], & \text{for } i \text{ large.} \quad (1.4b) \end{cases}$$

Here, $\gamma$ is Euler's constant

$$\gamma = -\Gamma'(1) = \lim_{i\to\infty}\left[\sum_{j=1}^{i-1} 1/j - \log i\right] = 0.57721566... .$$

The variance is

$$\mathrm{Var}(K_i) \doteqdot \begin{cases} E(K_i) + \theta^2 \sum_{j=1}^{i-1} 1/j^2, & \text{for } i = 2, 3,..., \text{ but small,} \quad (1.5a) \\ E(K_i) + \theta^2\pi^2/6, & \text{for } i \text{ large.} \quad (1.5b) \end{cases}$$

For small values of $i$ (that is, for small samples), the number of segregating sites, $K_i$, is approximately the sum of $i-1$ independent, geometrically distributed (but not identically distributed) random variables. For large values of $i$, $K_i$ is approximately Poisson (or normally) distributed, but it is not exactly so distributed.

The most interesting single probability generated by (1.3) is that for the event that the sample contains *no* segregating sites and hence that all gametes in the sample have the same complete cistronic type (i.e., are "monomorphic").

$$\Pr(K_i = 0) \doteq \begin{cases} \dfrac{(i-1)!}{(\theta+1)(\theta+2)\cdots(\theta+i-1)}, & i = 2, 3,..., \text{ but small}, \quad (1.6a) \\ \left(\dfrac{\exp\{-\frac{1}{2}g(i/2N)\}}{i}\right)^{\theta} \Gamma(\theta+1), & i \text{ large}. \quad (1.6b) \end{cases}$$

In particular, the probability that the whole population is monomorphic is

$$P_{\text{mono}} \equiv \Pr(K_{2N} = 0) \doteq \left(\frac{\exp\{-\frac{1}{2}g(1)\}}{2N}\right)^{\theta} \Gamma(\theta+1), \quad (1.7)$$

where, for the models of Wright, Burrows and Cockerham, and Moran, we find that $\exp\{-\frac{1}{2}g(1)\} = 0.9045$, 1.1485, and 1.000, respectively. (Note that these latter quantities are not even monotonically related to their models' effective sizes $N_e$, although, of course, the values of the parameter $\theta$ are so related).

The number, $K_2$, of segregating sites in the corresponding cistrons of a randomly chosen diploid individual may be seen from (1.3a), to have approximately a geometric distribution

$$\Pr(K_2 = n) \doteq \frac{1}{\theta+1}\left(\frac{\theta}{\theta+1}\right)^{n}, \qquad n = 0, 1, 2,..., \quad (1.8)$$

and in particular, the probability that the diploid is homozygous with respect to the whole cistron is

$$\Pr(K_2 = 0) \doteq 1/(\theta+1) = 1/(4N_e u + 1). \quad (1.9)$$

In later sections, the exact distributions and moments of $K_2$ are found for each model. Some approximate moments for all models are

$$E(K_2) \doteq \theta, \qquad \operatorname{Var}(K_2) \doteq \theta + \theta^2. \quad (1.10)$$

For the population as a whole, approximate moments for the number of segregating sites are, from (1.4b) and (1.5b),

$$E(K_{2N}) \doteq \theta[\log 2N + \gamma + \tfrac{1}{2}g(1)], \qquad \operatorname{Var}(K_{2N}) \doteq E(K_{2N}) + \theta^2\pi^2/6. \quad (1.11)$$

In particular, for Wright's model,

$$E(K_{2N}) \doteqdot \tfrac{1}{2}\theta[2 \log 2N + 2\gamma + 0.200645] = \tfrac{1}{2}\theta[2 \log 2N + 1.355076]. \quad (1.12)$$

The accuracy of some of our approximations, particularly of (1.7)–(1.10) and (1.12), but also of all results for Moran's model, is discussed later.

### 1.3. *Comparison with Previous Work*

Fisher (1930; see also (1958, p. 98)), studied populations in which each gamete produced, independently of other gametes, a Poisson-distributed number of offspring gametes, mean 1. He forced exactly one new mutation to occur, per generation, throughout the entire population. By a somewhat heuristic argument, implicitly applying an ergodic theorem to interpret the results for one site to many sites, and carrying the results for a branching process model over to a model of fixed population size of the Wright type, Fisher found that the mean number of segregating sites would be as given by (1.12) above, with $\theta = 2$. Fisher also showed that the number of sites having exactly $j$ of the mutant alleles present throughout the population would be about $2/j$, $j = 1, 2, 3, \ldots$. More generally, Haldane (1939) showed that in Fisher's branching process model with one mutation per generation, if the number of offspring gametes per parental gamete had mean 1, but variance $\sigma^2$ (and was not necessarily Poisson distributed), then the number of sites of mutant allele frequency $j$ would be about $2/(\sigma^2 j)$, and there would be an expected number of segregating sites equal to $(2/\sigma^2) \log 2N + c$, where $c$ is a constant and $2N$ is the number of gametes being considered. This is consistent with our result given in (1.11), because with $\nu_m = 1$ and $N_e = N/\sigma^2$ in (1.2), we find $\theta = 2/\sigma^2$. It is typical that our *mean* results agree with those of other authors for other models and other assumptions about mutation incidence. The same is not time for other aspects, however; see below.

The result (1.9) is standard; it is usually called the "inbreeding coefficient" under mutation; see e.g. (7.2.3) in Crow and Kimura (1970). However (1.8) appears to be new, as are the corresponding exact results mentioned in later sections. For independent sites with Poisson mutations, Ewens (1974) found that $K_2$ would be exactly Poisson distributed in Wright's model, with mean $\theta$ and (in contrast to (1.10)) variance $\theta$. The difference is due to correlation between sites in our models. Earlier, Kimura (1969) had found that the mean of $K_2$ would be $\theta$ and the variance would be $\frac{2}{3}\theta$, for an independent sites model. But he assumed that there was a *fixed* number $\nu_m$ of mutations per generation, and he did not specify how these were distributed among the sites. Indeed, he implicitly used an ergodic argument to convert results for one site to many sites, and he also ignored the variability of nucleotide frequencies at that site, as studied later, for instance, by Maruyama (1973) and Watterson (1974).

The approximately Poisson nature of $K_i$ for $i$ relatively large (cf. (1.3b)), may be compared with an exactly Poisson distribution in Ewens' (1974) work on independent sites.

The probabilities (1.6a) agree with Ewens' (1972) Eq. (19) for the probabilities that samples drawn from Wright's model should be monomorphic. Ewens obtained his expressions using diffusion approximations. But for an infinite-sites model with independence between sites, Ewens (1974) obtained

$$\Pr(K_i = 0) = \exp\left(-\theta \sum_{j=1}^{i-1} 1/j\right)$$

$$\doteqdot (\exp(-\gamma)/i)^{\theta} \qquad \text{for} \quad i \text{ large},$$

which have similarities with (1.6a) and (1.6b) but, when $i$ is large, they miss the factor $\Gamma(\theta + 1)$ and have a different numerator term ($\exp(-\gamma) \doteqdot 0.5615$ rather than our 0.9045 for Wright's model) compared with our (1.6b).

Our probability (1.7) for the population to be monomorphic differs from the previously published value

$$P_{\text{mono}} \doteqdot (1/2N)^{\theta}, \tag{1.13}$$

which was got by Kimura (1971) (Eq. (6.25) with $q = 1/2N$) using diffusion approximations and ergodic arguments.

The independence of sites assumption employed by Fisher (1930), Kimura (1969), Maruyama (1973), Ewens (1974), and others, seems a particularly dangerous one considering the biology of the situation, except when only mean results are required. Sites can be correlated due to their belonging within gametes, even if they are on different chromosomes and so are unlinked. Each gamete should inherit its sites from at most two parental gametes, rather than each site being chosen independently across the whole population. The same criticism then applies to an across-sites ergodic argument based on this assumption.

There is, however, a perfectly valid ergodic argument that the stationary mean number of segregating sites, $E(K_{2N})$, should equal the product of the mutation rate, $\nu_m$, per generation of $2N$ gamete deaths, and the mean number of generations, $E(T)$ say, that a given site with initial mutant relative frequency $p = 1/2N$ takes to become homozygous again, due to the drift towards fixation or loss of the mutant. Watterson (1962) found $E(T)$ by diffusion theory, and it is given approximately by

$$E(T) \doteqdot \begin{cases} -4N_e[p \log p + (1-p)\log(1-p)], & \text{or} \quad (1.14a) \\ 2(N_e/N)[\log 2N + 1]. & (1.14b) \end{cases}$$

For Moran's model, Watterson (1961) found the exact result; it is

$$E(T) = \sum_{j=1}^{2N-1} 1/j. \tag{1.15}$$

Note that, in view of (1.2), there is good agreement between the approximation to $E(K_{2N})$ as in (1.11) and $\nu_m E(T)$, from (1.14b). In Section 4 below we find $E(K_{2N})$ exactly, for Moran's model, and verify that it is exactly $\nu_m E(T)$ as given by (1.15).

For Wright's model, $E(T)$ is known exactly only by numerical calculations for low values of $N$ (see Knox (1962), Ewens (1964a), Carr and Nassar (1970)). For $2N_e = 2N = 50$, and $p = 1/50$, the correct figure is

$$E(T) = 9.12677,$$

whereas the approximations (1.14a) (1.14b) and (1.12) with $\theta = 2$, yield 9.80391, 9.82405 and 9.17912, respectively. Ewens' (1964a) improved version of (1.14a),

$$E(T) \doteqdot -4N[p \log p + (1-p)\log(1-p)] - \pi^2/18 - \log(2N-1)/6N$$

yields 9.22965. This is quite good agreement, but it is remarkable that what is essentially Fisher's (1930) result, (1.12), yields the best approximation!

### 1.4. *The Number of Alleles in a Sample*

If there are $K_i$ segregating sites in a sample of $i$ gametes, these could be distributed among the gametes to yield anything between 2 and $2^{K_i}$ different cistron types, which we shall call alleles. The one exception is when $K_i = 0$, for then there must be exactly one allele present, as we have already exploited in (1.6) and (1.7). Unlike the independent sites models of other authors, our infinite-sites models correspond directly to infinite-alleles models because we allow no recombination between sites. But, unfortunately, our results for $K_i$'s distribution do not tell us much in general about the distribution of $K_i^*$ (say), the number of alleles in a sample of $i$ gametes. (Our notation $K$ and $K^*$ corresponds to Ewens' (1974) notation of $k^*$ and $k$, respectively!.)

However, there are various results known for $K_i^*$ in the literature. Thus for Moran's model, Karlin and McGregor (1967) found the exact moments of $K_{2N}^*$, the mean being (in our notation)

$$E(K_{2N}^*) = \frac{2Nu}{1-u}\left(\frac{\Gamma'(2N/(1-u))}{\Gamma(2N/(1-u))} - \frac{\Gamma'(2Nu/(1-u))}{\Gamma(2Nu/(1-u))}\right)$$

$$\sim \theta[\log 2N - \Gamma'(\theta)/\Gamma(\theta)] \quad \text{as} \quad N \to \infty \quad \text{with} \quad 2Nu \to \theta. \tag{1.16}$$

The diffusion approximation method, which presumably applies to all three of our models, yields (see Ewens (1966, Eq. (6.42)))

$$E(K_{2N}^*) \doteq \theta \left[1 + \int_{1/2N}^{1} x^{-1}(1-x)^{\theta-1}\, dx\right],$$

or, by a different method (see Wright (1969, Eq. (14.14))),

$$E(K_{2N}^*) \doteq \theta[\log 2N - \gamma - \Gamma'(\theta)/\Gamma(\theta)].$$

These means bear strong similarities to those of $K_{2N}$; see (1.11) and (1.12).

We should note that the mean number of alleles in the population, $E(K_{2N}^*)$, is usually considerably greater than the "effective number of alleles," $n = 1/\Pr(K_2^* = 0) \doteq \theta + 1$, as discussed by Kimura and Crow (1964).

Ewens (1972, 1974) and Watterson (1974) have found (by diffusion methods and discrete methods, respectively) approximate moments and distributions for $K_{2N}^*$, and $K_i^*$ more generally, for the Wright and Moran models. (Note, however, that Watterson (1974) considered sampling *with* replacement.) The upshot of all of this work suggests that $K_i^* - 1$ is approximately Poisson or normally distributed, with mean of about $\theta \log i$, for reasonably large samples. This shows, in view of (1.3b) and (1.4b), that $K_i$ and $K_i^*$ are similar in their large-sample distributions, although, of course, this does not necessarily show that $K_i \doteq K_i^*$ for a particular sample. It would be of considerable interest to study the joint behaviour of $K_i$ and $K_i^*$, say for Moran's model in which the analysis is usually simpler than in other models. We do not attempt it here, however.

The remaining sections contain the proofs of the results outlined above, and some further matters of detail.

## 2. Wright's Nonoverlapping Generation Model

When a sample of $i$ gametes is chosen at random from a population of $2N$ gametes at generation $t$, the number, $J$, of distinct parental gametes possessed by the sample is a random variable restricted by $1 \leqslant J \leqslant i$. When $J = 1$, the sample individuals all had a common parent, and when $J = i$, the sampled gametes each had distinct parents. We follow Felsenstein (1971); see also Karlin (1968, p. 512) and Burrows and Cockerham (1974), in denoting the distribution of $J$ by

$$\Pr(J = j \mid i) = G_{i,j}, \qquad j = 1, 2, 3, \ldots, i. \tag{2.1}$$

The $G_{i,j}$'s will be given explicitly below for Wright's model.

Let $K_i^{(t)}$ denote the number of segregating (heterozygous) sites observed among the $i$ sampled gametes. Assuming that the $J$ distinct parents were chosen randomly from generation $t-1$, they possessed $K_J^{(t-1)}$ segregating sites. Thus

$$K_i^{(t)} = K_J^{(t-1)} + X_i^{(t)}, \tag{2.2}$$

where $X_i^{(t)}$ is the number of newly segregating sites among the offspring, due to mutations. We assume that, for $i \geqslant 2$, $X_i^{(t)}$ has a Poisson distribution with mean $i\nu$.

In terms of probability generating functions (pgf's), (2.2) can be written as the convolution-mixture

$$E(s^{K_i^{(t)}}) = \sum_{j=1}^{i} G_{i,j} E(s^{K_j^{(t-1)}})\, e^{i\nu(s-1)}, \qquad i \geqslant 2. \tag{2.3}$$

Of course, a sample of one individual possesses no segregating sites so that

$$K_1^{(t)} = 0 \qquad \text{and} \qquad E(s^{K_1^{(t)}}) = 1.$$

When statistical stationarity has been reached, we drop the time superscripts and (2.3) becomes

$$E(s^{K_i}) = \sum_{j=1}^{i} G_{i,j} E(s^{K_j})\, e^{i\nu(s-1)}, \qquad i \geqslant 2, \tag{2.4}$$

subject to the boundary condition

$$E(s^{K_1}) = 1. \tag{2.5}$$

We can write down, exactly, the pgf for the number, $K_2$, of heterozygous sites in a diploid formed by the union of two random gametes. From (2.4) and (2.5) we obtain

$$E(s^{K_2}) = G_{2,1} e^{2\nu(s-1)} / [1 - G_{2,2} e^{2\nu(s-1)}]. \tag{2.6}$$

For Wright's model, $G_{2,1} = 1/2N$ and $G_{2,2} = 1 - 1/2N$. This shows that $K_2$ has a compound geometric-Poisson distribution. Let $X_1, X_2, \ldots$, be independent Poisson variables with mean $2\nu$, pgf $E(s^X) = e^{2\nu(s-1)}$. Let $M$ be a variate with the geometric distribution

$$\Pr(M = m) = G_{2,1} G_{2,2}^{m-1}, \qquad m = 1, 2, 3, \ldots, \tag{2.7}$$

and pgf $E(s^M) = G_{2,1} s / [1 - G_{22} s]$.

Then

$$K_2 = X_1 + X_2 + \cdots + X_M \tag{2.8}$$

has the pgf given by (2.6).

The interpretation of (2.8) is that $M$ is the number of generations which have elapsed since the two gametes can be traced back to a common ancestor. (Recall Malécot's result, that in the absense of mutation and in finite randomly mating populations, all gametes are "identical by descent" and can be traced back to a common ancestor.) The variates $X_1, X_2, \ldots, X_M$ are the numbers of new mutations picked up (and passed on) by those pairs of ancestors in their respective generations.

The solution of (2.4) can be achieved in theory in succession:

$$E(s^{K_i}) = \sum_{j=1}^{i-1} G_{i,j} E(s^{K_j})\, e^{i\nu(s-1)}/[1 - G_{i,i} e^{i\nu(s-1)}], \qquad i \geqslant 2. \tag{2.9}$$

For Wright's model in particular, each of $i$ offspring gametes was produced by a randomly chosen parental gamete (with replacement). Thus a particular gamete has a binomially distributed number of offspring altogether, parameters $2N$ and $1/2N$, and the number which happen to belong to the sample is binomial, parameters $i$ and $1/2N$. (For $2N$ large, the offspring distribution is approximately Poisson, mean 1, as considered by Fisher in deriving (1.12).)

The probability that $j$ distinct parents were used to produce $i$ offspring satisfies the recurrence

$$G_{i,j} = (j/2N)\, G_{i-1,j} + [(2N - j + 1)/2N]\, G_{i-1,j-1}, \qquad 2 \leqslant j \leqslant i, \tag{2.10}$$

while

$$G_{1,1} = 1, \qquad G_{1,i} = (2N)!/[(2N)^i (2N - i)!], \qquad G_{i,1} = 1/(2N)^{i-1}. \tag{2.11}$$

These equations were given by Burrows and Cockerham (1974), and the probability distributions involved have been studied also by Stevens (1937), David (1950), Feller (1950), Thomas (1951), Arfwedson (1951), Craig (1953), Nicholson (1961), Kempthorne (1967), and Felsenstein (1971). It seems that the later authors were unfamiliar with the work of the earlier authors, in particular that the equations have an explicit solution, found in the context of occupancy problems to be

$$\begin{aligned} G_{i,j} &= \mathscr{S}_i^{(j)} 2N(2N-1)(2N-2)\cdots(2N-j+1)(2N)^{-i} \\ &= (2N)^{-i} \binom{2N}{j} \Delta^j(0)^i, \qquad 1 \leqslant j \leqslant i, \end{aligned} \tag{2.12}$$

where $\mathscr{S}_i^{(j)}$ is a Stirling number of the second kind. For each fixed $i$ and $2N$, $G_{i,j}: j = 1, 2, 3, \ldots, i$ forms a probability distribution whose factorial moments are known (see, e.g. Johnson and Kotz (1969, pp. 251–252)).

From now on we shall be assuming that $\theta \doteq 4N\nu$ is $O(1)$ and that $2N$ is large, so that for moderate values of $i$ we have

$$e^{i\nu(s-1)} \doteq 1 + (i\theta/4N)(s-1) \doteq 1. \tag{2.13}$$

Hence from (2.6) and (2.11), the approximate pgf for $K_2$ in Wright's model is

$$E(s^{K_2}) \doteq 1/[1 + \theta - \theta s], \tag{2.14}$$

the pgf of the geometric distribution (1.8).

Some exact moments for $K_2$ in Wright's model may be obtained from (2.6) and (2.11):

$$E(K_2) = 4N\nu \doteq \theta,$$

and

$$\mathrm{Var}(K_2) = 4N\nu + 16N^2\nu^2 - 8N\nu^2 \doteq \theta + \theta^2,$$

the approximations being the moments of (1.10).

For small values of $i$ and large values of $2N$, it is easy to see that the bulk of the probability distribution $\{G_{i,j}\}$ is located at $j = i - 1$ and $j = i$. Indeed,

$$G_{i,i-1} \doteq i(i-1)/4N, \qquad G_{i,i} \doteq 1 - i(i-1)/4N.$$

Using (2.9) and (2.13) we find that

$$E(s^{K_i}) \doteq E(s^{K_{i-1}})/[1 - (\theta/j)(s-1)], \qquad i = 2, 3, \ldots,$$

with solution being (1.3a) after using the boundary condition (2.5).

As mentioned in the introduction, (1.3a) shows that $K_i$ may, for small $i$, be treated as approximately the sum of $i - 1$ independent variates:

$$K_i \doteq Y_1 + Y_2 + \cdots + Y_{i-1} \quad \text{say,} \tag{2.15}$$

with $Y_j$ having the geometric distribution

$$\Pr(Y_j = n) = [1/(1 + \theta/j)](\theta/(j + \theta))^n, \qquad n = 0, 1, 2, \ldots. \tag{2.16}$$

The interpretation of $Y_j$ in (2.15) is that it is the number of new mutations occurring to the ancestors of our sample during those generations when there were exactly $j + 1$ distinct ancestors present. (Compare the particular interpretation given for (2.8); $K_2$ and $Y_1$ have essentially the same interpretation and distribution.) With this interpretation, (2.16) is also not exactly, but only approximately, correct.

The reason why (2.15) is not exact is that, for Wright's model, there may be no generation in which our sample had exactly $j + 1$ ancestors, and there is a

dependency between the number of branches occurring successively in the family tree. Moran's model below leads to an exact interpretation, for the very reason that all intermediate ancestor sizes must be visited.

The moments of $K_i$ quoted in (1.4a) and (1.5a) follow either from (1.3a), or from (2.15) and (2.16), because $E(Y_j) = \theta/j$, and $\mathrm{Var}(Y_j) = \theta/j + \theta^2/j^2$. For large values of $i$, but still small relative to $2N$, we find the further approximations

$$E(K_i) \doteq \theta[\log(i-1) + \gamma],$$
$$\mathrm{Var}(K_i) \doteq \theta[\log(i-1) + \gamma + \theta\pi^2/6].$$

The probabilities (1.6a) follow from (1.3a), because $\Pr(K_i = 0) = E(0^{K_i})$.

For future use, it may be noted that (1.3a) may be rewritten as

$$E(s^{K_i}) \doteq \Gamma(i)\,\Gamma(1+\theta-\theta s)/\Gamma(i+\theta-\theta s)$$
$$\sim i^{\theta(s-1)}\,\Gamma(1+\theta-\theta s), \qquad \text{as} \quad i \to \infty. \tag{2.17}$$

The major term, $i^{\theta(s-1)} = \exp\{\theta(\log i)(s-1)\}$ is the pgf of a Poisson distribution with mean $\theta \log i$; more rigorously, the result can be reformulated to show that $(K_i - \theta \log i)/(\theta \log i)^{\frac{1}{2}}$ is approximately a standard normal variate when $\log i$ is large (but we still assume $i \ll 2N$ at present).

The extension of the above results to values of $i$ of the same order of magnitude as $2N$ requires further analysis. The number, $J$, of distinct parents possessed by $i$ offspring no longer has its probabilities concentrated on the values $J = i - 1$ and $J = i$. In fact, the mean is

$$E(J \mid i) = \sum_{j=1}^{i} jG_{i,j} = 2N - 2N\left(1 - \frac{1}{2N}\right)^i,$$

and the variance is

$$\mathrm{Var}(J \mid i) = 2N(1-1/2N)^i + 2N(2N-1)(1-2/2N)^i - (2N)^2(1-1/2N)^{2i};$$

see Johnson and Kotz (1969, p. 252). Writing $i/2N = x$, the normalized variable $J/2N$ has moments

$$E(J/2N \mid x) = 1 - (1-1/2N)^{2Nx} \sim 1 - e^{-x}, \qquad \text{as} \quad N \to \infty,$$

and

$$\mathrm{Var}(J/2N \mid x) \sim (1/2N)\,e^{-x}(1-(1+x)\,e^{-x}) \to 0, \qquad \text{as} \quad N \to \infty.$$

Hence $J/2N$ converges in mean square (and in probability) to its asymptotic mean $1 - e^{-x}$ as $N \to \infty$.

Now Eq. (2.4) may be rewritten

$$P(x, s) = E[P(J/2N, s) \mid x]\, e^{2Nvx(s-1)}, \qquad 2Nx \geqslant 2,$$

where

$$P(x, s) = E(s^{K_i}) = E(s^{K_{2Nx}}). \tag{2.18}$$

For large values of $N$ and fixed $x$ we find

$$P(x, s) \doteqdot P(1 - e^{-x}, s)\, e^{\frac{1}{2}\theta x(s-1)}, \tag{2.19}$$

or, by means of the transformation (suppressing $s$)

$$\phi(y) = -(2/\theta(s-1)) \log P(1-y, s), \qquad y = 1 - x,$$

we have

$$\phi(e^{y-1}) - \phi(y) = 1 - y. \tag{2.20}$$

The functional equation (2.20) was studied by Fisher (1930), who gave its solution as

$$\phi(y) \doteqdot -2 \log(1-y) + \text{constant}.$$

More precisely, we may deduce from Fisher's work that

$$\phi(y) = -2 \log(1-y) - g(1-y) + a, \tag{2.21}$$

where $a$ is an arbitrary constant and

$$g(x) = -\log\Big(1 - \frac{x}{6} - \frac{x^2}{72} - \frac{2x^3}{1080} + \frac{3x^4}{108 \times 144} + \frac{4 \times 71x^5}{168 \times 72^2} + \frac{5 \times 8759x^6}{630 \times 720^2} - \frac{6 \times 31x^7}{81 \times 720^2} - \frac{7 \times 1637x^8}{1008 \times 720^2} + \frac{8 \times 20879093x^9}{9504 \times 840 \times 720^2} \cdots\Big). \tag{2.22}$$

A series, alternative to (2.22), may be obtained more explicitly as follows. Substituting (2.21) into (2.20) yields

$$g(x) = g(1 - e^{-x}) + x + 2 \log[(1 - e^{-x})/x]. \tag{2.23}$$

Differentiating both sides $n$ times, at $x = 0$, leads to the recurrence relation

$$g^{(n-1)}(0) = \frac{2}{n(n-1)} \Big\{ \sum_{m=1}^{n-2} g^{(m)}(0)(-1)^{m+n}\, \mathscr{S}_n^{(m)} + 2B_n/n \Big\},$$

where $\mathcal{S}_n^{(m)}$ is a second-kind Stirling number and $B_n$ is a Bernoulli number. Starting with $g(0) = 0$ and $g'(0) = B_2 = 1/6$, we can find the derivatives in succession, and hence the Taylor series

$$g(x) = \frac{1}{6}x + \frac{1}{36}x^2 + \frac{37}{36 \times 180}x^3 + \frac{205}{24 \times 10800}x^4 - \frac{21625}{1134000 \times 120}x^5$$
$$- 1.54402 \times 10^{-4}x^6 - 2.462 \times 10^{-5}x^7 + 3.1981 \times 10^{-5}x^8$$
$$+ 1.9962 \times 10^{-5}x^9 \cdots . \tag{2.24}$$

In terms of $P(x, s)$, the solution (2.21) of (2.19) may be written

$$P(x, s) \doteqdot \exp\{-\tfrac{1}{2}\theta(s-1)\phi(1-x)\}$$
$$= \exp\{\tfrac{1}{2}\theta(s-1)[2\log(x) + g(x) - a]\}. \tag{2.25}$$

To make the solution fully explicit, Fisher chose the constant so that $\phi(0) = 0$ held. This is not appropriate for us; instead we require the approximation (2.17) to hold. Irrespective of the value of $i$ ($i \ll 2N$), (2.17) fixes the constant $a$, and we obtain our particular solution

$$P(x, s) \doteqdot \exp\{\tfrac{1}{2}\theta(s-1)[2\log(2Nx) + g(x)]\}\, \Gamma(1 + \theta - \theta s). \tag{2.26}$$

The result (1.3b) is equivalent to the above.

For large values of $\log i$, (1.3b) indicates that $K_i$ is approximately Poisson distributed, mean $\theta \log i$. More exactly, by differentiating (1.3b) and using $\Gamma'(1) = -\gamma$, $\Gamma''(1) = \gamma^2 + \pi^2/6$, we find the moments (1.4b) and (1.5b).

The probabilities (1.6b) and (1.7) follow from (1.3b) by putting $s = 0$. The quantity $g(i/2N)$ which arises in various formulas can be evaluated approximately by (2.22); for instance this yields $g(1) \doteqdot 0.20057553$. However, Fisher (1930) was not convinced of the accuracy of using only the quoted terms in (2.22), and calculated $g(1)$ by a somewhat different method to be

$$g(1) = 0.20064507\ldots .$$

Essentially the same result follows, using (2.23) and (2.24) in conjunction, to aid convergence.

Hence $e^{-\frac{1}{2}g(1)} \doteqdot .90454562$ is the required numerator quantity for calculating $P_{\text{mono}}$ for Wright's model, using (1.7).

We may check on the accuracy of (1.7) because in numerical cases the probabilities $\Pr(K_i = 0)$, and in particular $\Pr(K_{2N} = 0)$, can be computed from (2.9) successively:

$$\Pr(K_i = 0) = \sum_{j=1}^{i-1} G_{i,j} \Pr(K_j = 0)\, e^{-i\nu}/[1 - G_{i,i}e^{-i\nu}].$$

Writing $\theta = 4N(1 - e^{-\nu})$, and

$$P_{\text{mono}} = (a(\theta, 2N)/2N)^{\theta},$$

we exhibit in Table I the exact values $a(\theta, 2N)$ for comparison with the approximating numerator term derived from (1.7):

$$a(\theta) = 0.90454562[\Gamma(\theta + 1)]^{1/\theta},$$

and the numerator, 1, in the diffusion approximation (1.13).

TABLE I

Exact ($a(\theta, 2N)$) and Approximate ($a(\theta)$) Numerator Terms for $P_{\text{mono}}$ in Wright's Model

| | $a(\theta, 2N)$ | | | | |
|---|---|---|---|---|---|
| $2N$ \ $\theta$ | 0.1 | 0.5 | 1.0 | 5.0 | 10.0 |
| 10 | .59507 | .71011 | .83290 | 1.40695 | 1.62915 |
| 100 | .55515 | .70846 | .88909 | 2.12494 | 3.40509 |
| 500 | .55069 | .70975 | .90024 | 2.28885 | 3.88181 |
| 1000 | .55005 | .71003 | .90213 | 2.31808 | 3.97254 |
| 2000 | .54972 | .71020 | .90321 | 2.33501 | 4.02624 |
| $a(\theta)$ | .54933 | .71043 | .90455 | 2.35650 | 4.09644 |

The approximation (1.7) is evidently very good. To indicate the extent to which the diffusion approximation (1.13) can differ from the true value of $P_{\text{mono}}$, and to indicate the wide range of values the latter can take, we find that when $2N = 1000$ and $\theta = 1/10, 1/2, 1, 5, 10$, then

$$P_{\text{mono}} = 0.4721, 0.02665, 9.021 \times 10^{-4}, 6.693 \times 10^{-14}, 9.788 \times 10^{-25},$$

and

$$(1/2N)^{\theta} = 0.50119, 0.03162, 10 \times 10^{-4}, 0.1 \times 10^{-14}, 0.00001 \times 10^{-25},$$

respectively, while (1.7) yields

$$0.4720, 0.02665, 9.045 \times 10^{-4}, 7.266 \times 10^{-14}, 13.307 \times 10^{-25}.$$

Clearly, for this relatively low value of $2N$, the approximations (1.13) and (1.7) lose accuracy for $\theta > 1$, but (1.7) is much more accurate than is (1.13).

## 3. Burrows and Cockerham's Nonoverlapping Generation Model

Burrows and Cockerham (1974) considered a model in which the $2N$ gametes of a generation are paired at random to yield $N$ diploid individuals. Each such diploid then produces exactly two gametes for the next generation. The offspring gametes could both be copies of one parent gamete, or of the other parent gamete, or each parental gamete could be copied once, with respective probabilities $\frac{1}{4}, \frac{1}{4}, \frac{1}{2}$. As usual, we assume no recombination among the gametes.

It is clear that the number, $J$, of distinct gametes which produced a randomly chosen sample of $i$ offspring gametes, has its values restricted by $\frac{1}{2}i \leqslant J \leqslant i$. Burrows and Cockerham did not quote the explicit formula, where $j^* = \min(j, N)$,

$$G_{i,j} = \frac{2^i N!\, i!\, (2N-i)!}{(i-j)!\, (2N)!} \sum_{l=i-j^*}^{[i/2]} \frac{2^{-3l}}{(l-i+j)!\, (i-2l)!\, (N-i+l)!}, \qquad (3.1)$$

for the probability that $J = j$, but they did give a recurrence relation and the explicit cases when $i = 2N$. We find in particular that

$$G_{1,1} = 1,$$
$$G_{2,1} = 1/(2(2N-1)), \qquad G_{2,2} = 1 - 1/(2(2N-1)),$$
$$G_{3,2} = 3/(2(2N-1)), \qquad G_{3,3} = 1 - 3/(2(2N-1)),$$
$$G_{4,2} = 3/(4(2N-1)(2N-3)), \qquad G_{4,3} = 3(4N-7)/(2(2N-1)(2N-3)),$$
$$G_{4,4} = 1 - (24N-39)/(4(2N-1)(2N-3)).$$

For the study of the distribution of $K_i$, the formulas (2.1) to (2.9) still apply. For instance, from (2.6) we now have

$$E(s^{K_2}) = e^{2\nu(s-1)}/[2(2N-1) - (4N-3)\, e^{2\nu(s-1)}]$$

with its exact moments,

$$E(K_2) = 4(2N-1)\nu,$$

and

$$\mathrm{Var}(K_2) = E(K_2) + 4(4N-3)^2\nu^2.$$

From (1.2), we write $\theta \doteq 4N_e\nu = 4(2N-1)\nu$, so that the mean and variance of $K_2$ are approximately $\theta$ and $\theta + \theta^2$, and the distribution itself is approximately the geometric distribution (1.8).

Similarly we can show that the results (1.3a), (1.4a), (1.5a), and (1.6a) also hold. However, large samples need further consideration. The distribution (3.1) is such that $J/2N$ converges in mean square to its asymptotic mean $x - \frac{1}{4}x^2$, where $x = i/2N$, as $N \to \infty$. Hence for this model, (2.19) is replaced by

$$P(x, s) \doteq P(x - \tfrac{1}{4}x^2, s)\, e^{\frac{1}{2}\theta x(s-1)},$$

and making the (new) substitution

$$\phi(y) = -(4/\theta(s-1)) \log P(1-y, s), \qquad y = 1-x,$$

we get instead of (2.20),

$$\phi(\tfrac{1}{4}(1+y)^2) - \phi(y) = 1 - y. \tag{3.2}$$

The solution of this recurrence is of the form (see Haldane (1939), Moran (1962, p. 113)):

$$\phi(y) = -4 \log(1-y) - 2g(1-y) + 2a, \quad \text{say,} \tag{3.3}$$

so that

$$P(x, s) \doteq \exp\{-\tfrac{1}{4}\theta(s-1)\phi(1-x)\},$$

a result which leads again to the solutions (2.25), (2.26), and finally (1.3b), by the same reasoning as for Wright's model.

The function $g(x)$ has been left unspecified above, and is *not* given by (2.22) or (2.24), as in Wright's model. A power series expansion is possible, because the derivatives $g^{(n)}(0)$ can be obtained in succession by differentiating repeatedly the functional equation

$$g(x) = g(x - \tfrac{1}{4}x^2) + (x/2) + 2\log(1 - \tfrac{1}{4}x), \tag{3.4}$$

got from (3.2) and (3.3). In fact,

$$\begin{aligned} g(x) = &-\frac{1}{4}x - \frac{1}{48}x^2 - \frac{5}{48 \times 24}x^3 - \frac{111}{120 \times 768}x^4 - \frac{343}{5!\,7680}x^5 \\ &- \frac{902}{7!\,1536}x^6 - \frac{9843}{7!\,224 \times 256}x^7 - 8.202 \times 10^{-6}x^8 \\ &- 1.059 \times 10^{-6}x^9 \cdots . \end{aligned} \tag{3.5}$$

The series (3.5) is more rapidly converging for $|x| < 1$ than for $x = 1$, so that $g(1)$ may be fairly accurately calculated by a combination of (3.4) and (3.5). We get $g(1) \doteq -0.2769$ so that $e^{-\frac{1}{2}g(1)} \doteq 1.1485$, for use in finding $P_{\text{mono}}$ as in (1.7), and $E(K_{2N})$ by (1.11).

## 4. Moran's Overlapping Generation Model

In Moran's model, the population at time $t$ consists of $2N - 1$ of the haploids present at time $t - 1$, together with one new haploid. This individual is the offspring of a parent haploid, randomly chosen from those existing at time $t - 1$,

and the offspring will inherit the parent's type except for a Poisson-distributed number of new mutant sites, mean $\nu$. Between $t-1$ and $t$, a randomly chosen individual dies to make room for the offspring.

Let $p_{i,k}^{(t)}$ be the probability that a random sample of $i$ individuals at time $t$ exhibits $k$ segregating sites, that is, $p_{i,k}^{(t)} = \Pr(K_i^{(t)} = k)$. It is a simple matter to deduce the recurrence relation

$$p_{i,k}^{(t)} = \frac{i}{2N}\left(\frac{i-1}{2N}\right)\sum_{l=0}^{k} p_{i-1,l}^{(t-1)} e^{-\nu}\frac{\nu^{k-l}}{(k-l)!} + \frac{i}{2N}\left(1-\frac{i-1}{2N}\right)\sum_{l=0}^{k} p_{i,l}^{(t-1)} e^{-\nu}\frac{\nu^{k-l}}{(k-l)!} + \left(1-\frac{i}{2N}\right)p_{i,k}^{(t-1)}, \quad (4.1)$$

for $i = 2, 3, \ldots, 2n$. The terms on the right arise from the possibilities that the sample contains both parent and offspring, or offspring but not parent, or does not contain the offspring, respectively. The stationary probabilities $p_{i,k}$ satisfy the above recursion with time superscripts removed.

Writing $P_i(s) = \sum_{k=0}^{\infty} p_{i,k} s^k$, we find from (4.1) the recurrence

$$P_i(s) = \frac{i}{2N}\left(\frac{i-1}{2N}\right)P_{i-1}(s)\, e^{\nu(s-1)} + \frac{i}{2N}\left(1-\frac{i-1}{2N}\right)P_i(s)\, e^{\nu(s-1)} + \left(1-\frac{i}{2N}\right)P_i(s), \qquad i = 2, 3, \ldots, 2N,$$

that is,

$$P_i(s) = \frac{(i-1)\, e^{\nu(s-1)}}{2N-(2N-i+1)\, e^{\nu(s-1)}} P_{i-1}(s), \qquad i = 2, 3, \ldots, 2N.$$

The exact solution, using the boundary condition $P_1(s) = 1$, is trivially

$$P_i(s) = \prod_{j=1}^{i-1}\left\{\frac{e^{\nu(s-1)}}{(2N/j)-((2N/j)-1)\, e^{\nu(s-1)}}\right\}, \qquad i = 2, \ldots, 2N. \quad (4.2)$$

Thus $K_i$ can be interpreted as the sum of $i-1$ independent random variables,

$$K_i = Y_1 + Y_2 + \cdots + Y_{i-1},$$

where $Y_j$ is the number of new mutations to the ancestors of our sample during those times when there were exactly $j+1$ such ancestors. Moreover, $Y_j$ has a compound geometric-Poisson distribution with pgf:

$$E(s^{Y_j}) = \frac{e^{\nu(s-1)}}{(2N/j)-((2N/j)-1)\, e^{\nu(s-1)}},$$

because if can be shown that $Y_j$ is the sum of a geometrically distributed number $M_j$ of independent Poisson variables:

$$Y_j = X_{1,j} + X_{2,j} + \cdots + X_{M_j,j},$$

where

$$E(s^{M_j}) = s/[(2N/j) - ((2N/j) - 1)s], \qquad E(s^{X_{i,j}}) = e^{\nu(s-1)}.$$

The $X$'s represent the numbers of mutations occurring in single births of our sample's ancestors, there being $M_j$ such births while there are $j + 1$ ancestors.

From (4.2), we find that for Moran's model, (1.4a) and (1.5a) are exact for *all* $i = 2, 3,..., 2N$ if we define $\theta = 2N\nu$, while (1.6a) would be exact if we define $\theta = 2N(e^{\nu} - 1)$. The approximations (1.3a, b), (1.4b), (1.5b), and (1.6b) are valid, if we take $g(i/2N) = 0$ whenever it arises. It is also the case that

$$E(K_{2N}) = 2N\nu \sum_{j=1}^{2N-1} 1/j = \nu_m E(T),$$

where $E(T)$ is the mean time for extinction or fixation of a mutant whose initial relative frequency is $1/2N$; see (1.15).

The probability that the population is monomorphic is, from (4.2), exactly

$$P_{\text{mono}} = \Gamma(2N)\,\Gamma\left(\frac{2Nu}{1-u} + 1\right)\Big/\Gamma\left(\frac{2Nu}{1-u} + 2N\right),$$

where $1 - u = e^{-\nu}$. This is a special case of (2.11) in Watterson (1974), where it was conjectured (and proved by Trajstman (1974)) that the number of different *types* (say $K^*_{2N}$) in the present model has a distribution

$$\Pr(K^*_{2N} = n) = \eta^n \mid S^{(n)}_{2N} \mid \Gamma(\eta)/\Gamma(\eta + 2N), \qquad n = 1, 2,..., 2N,$$

where $\eta = 2Nu/(1 - u)$, and $S^{(n)}_{2N}$ is a Stirling number of the first kind. Karlin and McGregor (1967) (4.35) gave moments of $K^*_{2N}$ consistent with the above distribution. In any case, $P_{\text{mono}} = \Pr(K^*_{2N} = 1)$. $K^*_{2N}$ has approximately a Poisson or normal distribution, mean (1.16), which is very similar to $K_{2N}$'s behaviour.

## Acknowledgment

I thank Mrs. M. Wu for help with the numerical work, and in particular for computing Table I. The paper has also benefited greatly from my extensive discussions with Professor W. J. Ewens, who supplied its motivation and made many suggestions.

## REFERENCES


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

BURROWS, P. M. AND COCKERHAM, C. C. 1974. Distributions of time to fixation of neutral genes, *Theor. Pop. Biol.* **5**, 192–207.

CARR, R. N. AND NASSAR, R. F. 1970. Effects of selection and drift on the dynamics of finite populations, II, *Biometrics* **26**, 221–227.

CRAIG, C. C. 1953. On the utilization of marked specimens in estimating populations of flying insects, *Biometrika* **40**, 170–176.

CROW, J. F. AND KIMURA, M. 1970. "An Introduction to Population Genetics Theory," Harper & Row, New York.

DAVID, F. N. 1950. Two combinatorial tests of whether a sample has come from a given population, *Biometrika* **37**, 97–110.

EWENS, W. J. 1964a. The pseudo-transient distribution and its uses in genetics, *J. Appl. Prob.* **1**, 141–156.

EWENS, W. J. 1964b. The maintenance of alleles by mutation, *Genetics* **50**, 891–898.

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

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

EWENS, W. J. 1974. A note on the sampling theory for infinite alleles and infinite sites models, *Theor. Pop. Biol.* **6**, 143–148.

FELLER, W. 1950. "An Introduction to Probability Theory and Its Applications," Vol. 1, Wiley, New York.

FELSENSTEIN, J. 1971. The rate of loss of multiple alleles in finite haploid populations, *Theor. Pop. Biol.* **2**, 391–403.

FISHER, R. A. 1930. The distribution of gene ratios for rare mutations, *Proc. Roy. Soc. Edinburgh* **50**, 205–220.

FISHER, R. A. 1958. "The Genetical Theory of Natural Selection," Dover, New York.

HALDANE, J. B. S. 1939. The equilibrium between mutation and random extinction, *Ann. Eugen.* **9**, 400–405.

JOHNSON, N. L. AND KOTZ, S. 1969 "Distributions in Statistics: Discrete Distributions," Houghton Mifflin, Boston.

KARLIN, S. 1968. Equilibrium behaviour of population genetics models with non-random mating. II, *J. Appl. Prob.* **5**, 487–566.

KARLIN, S. AND MCGREGOR, J. 1967. The number of mutant forms maintained in a population, *in* "Proc. 5th Berkeley Symp. Math. Stat. Prob. IV," pp. 415–438, Univ. of California Press, Berkeley, CA.

KEMPTHORNE, O. 1967. The concept of identity of genes by descent, *in* "Proc. 5th Berkeley Symp. Math. Stat. Prob. IV," pp. 333–348, Univ. of California Press, Berkeley, CA.

KIMURA, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations, *Genetics* **61**, 893–903.

KIMURA, M. 1971. Theoretical foundations of population genetics at the molecular level, *Theor. Pop. Biol.* **2**, 174–208.

KNOX, S. 1962. A study of a random-mating population of fixed size, Ph.D. Thesis, Virginia Polytechnic Institute, Blacksburg, VA.

MARUYAMA, T. 1973. The variance of the number of loci having a given gene frequency, *Genetics* **73**, 361–366.

MORAN, P. A. P. 1958. Random processes in genetics, *Proc. Camb. Phil. Soc.* **54**, 60–71.

MORAN, P. A. P. 1962. "The Statistical Processes of Evolutionary Theory," Clarendon Press, Oxford.

NICHOLSON, W. L. 1961. Occupancy probability distribution critical points, *Biometrika* **48**, 175–180.

STEVENS, W. L. 1937. Significance of grouping, *Ann. Eugen.* **8**, 57–69.

THOMAS, M. 1951. Some tests of randomness in plant populations, *Biometrika* **38**, 102–111.

TRAJSTMAN, A. C. 1974. On a conjecture of G. A. Watterson, *Adv. Appl. Prob.* **6**, 489–493.

WATTERSON, G. A. 1961. Markov chains with absorbing states: a genetic example, *Ann. Math. Statist.* **32**, 716–729.

WATTERSON, G. A. 1962. Some theoretical aspects of diffusion theory in population genetics, *Ann. Math. Statist.* **33**, 939–957.

WATTERSON, G. A. 1974. The sampling theory of selectively neutral alleles, *Adv. Appl. Prob.* **6**, 463–488.

WRIGHT, S. 1969. "Evolution and the Genetics of Populations," Vol. II, Univ. of Chicago Press, Chicago, IL.

*Printed in Belgium*