THE HOMOZYGOSITY TEST OF NEUTRALITY

G. A. WATTERSON

Monash University, Victoria 3168, Australia

Manuscript received March 8, 1977
Revised copy received August 22, 1977

ABSTRACT

An earlier paper showed that the homozygosity (of a population or sample) was a good statistic for testing departures from selective neutrality in the direction of heterozygote advantage or disadvantage. It is here shown that homozygosity is also influenced by the presence of deleterious alleles and by other departures from neutrality, but at a lower order of magnitude of effect if the selection coefficients are of the same small order of magnitude. Tables are provided for the significance points and moments of the homozygosity, under the null hypothesis of neutrality.

Section 1.

IN a previous paper, WATTERSON (1977), I showed that the homozygosity of a population, or of a sample, is a good statistic for testing the neutral hypothesis for alleles at a locus against the alternative of heterozygote advantage or disadvantage. M. NEI (personal communication) suggested that the homozygosity would be considerably influenced by the presence of (additive) deleterious alleles, if such were present. It is a purpose of this paper to investigate the magnitude of the latter effect. We shall also consider the effect of other selection schemes. Finally, we shall tabulate some critical points in the distribution of homozygosity, for use in determining the significance of the departure of a sample's homozygosity from expectation under the neutral alleles hypothesis.

NEI and his colleagues have used the sample heterozygosity to estimate the mutation parameter, and then they have compared the observed variance of sample heterozygosities across different species with the theoretically predicted variance, assuming neutrality. See, for instance, FUERST, CHAKRABORTY and NEI (1977). The conclusions of the present paper are as follows: for detecting selective differences between heterozygotes and homozygotes, or detecting the presence of deleterious alleles, the sample homozygosity (or heterozygosity) is a preferable statistic to use directly, rather than its variance. If both the heterozygote advantage (or disadvantage) on the one hand, and the additive effect of deleterious alleles on the other, are small, the homozygosity is more influenced by the former departure from neutrality than by the latter. The same comment applies to other population and sample characteristics.

We note below that for testing departures from neutrality other than heterosis or additive deleterious alleles, it may be useful to consider sample statistics other than the homozygosity.

The present study does not give as much detail on various series expansions as do the papers by WATTERSON (1977), concerning heterosis, or LI (1977) concerning deleterious alleles.

Section 2. Population models with selection.

2.1 The K*-allele population*: We start by assuming that there are K alleles possible at a particular locus, and subsequently consider the limiting case with infinitely many alleles ($K \to \infty$). Suppose that an individual of genotype A_iA_j has fitness $1 + s_{ij}$. The mean fitness of the population with allele frequencies $x_1, x_2, \ldots, x_K$ is defined by $\overline{W} = 1 + \sum_{i,j=1}^{K} s_{ij}x_ix_j$, *assuming random mating.*

We suppose that each gene, of whatever allelic type, has a mutation rate u per generation, and the chance of a gene of type A_i mutating to one of type A_j in particular is $u/(K-1)$ for each $j \neq i$. The selection and mutation effects will be assumed to be small. In fact, we rescale the parameters according to

$$\sigma_{ij} = 2N_e s_{ij}, \quad \varepsilon = 4N_e u/(K-1) \quad \text{and} \quad \Theta = 4N_e u \ , \tag{2.1}$$

where N_e is the effective population size, and we hold σ, ε and Θ fixed as $N_e \to \infty$. The stationary distribution for a diffusion model having these parameters is a special case of one which has been postulated by WRIGHT (1949, p. 383) and verified by KIMURA (1956). It has the density

$$\phi(x_1, x_2, \ldots, x_{K-1}) = C_K^{-1} \exp\left(\sum_{i,j=1}^{K} \sigma_{ij}x_ix_j\right)\left(\prod_{i=1}^{K} x_i\right)^{\varepsilon-1} \tag{2.2}$$

over the domain

$$0 \leqslant x_1, x_2, \ldots, x_{K-1} \leqslant 1 \ , \quad \sum_{i=1}^{K-1} x_i \leqslant 1 \ , \tag{2.3}$$

and in which $x_K \equiv 1 - x_1 - x_2 - \ldots - x_{K-1}$. The normalizing constant is

$$C_K = \int \ldots \int \exp\left(\sum_{i,j=1}^{K} \sigma_{ij}x_ix_j\right)\left(\prod_{1}^{K} x_i\right)^{\varepsilon-1} dx_1 \ldots dx_{K-1} \ , \tag{2.4}$$

the integral being over the region (2.3).

Suppose for the moment that the mutation parameter ε is known, but we desire to test for the presence of some selective effects against the null hypothesis of complete neutrality, $\sigma_{ij} = 0$ for all i,j. Suppose also that it is not known which observed allele frequency corresponds to which selective effect (*e.g.*, if some alleles are potentially deleterious, we do not know which alleles they are in our sample). The observed allele frequencies may be arranged as descending order statistics:

$$x_{(1)} \geqslant x_{(2)} \geqslant \ldots \geqslant x_{(K)} \ , \quad \text{with} \quad \sum_{1}^{K} x_{(i)} = 1 \ ,$$

but it is not known how to assign selection coefficients to them. The likelihood of this data is

$$f_K(x_{(1)},x_{(2)},\ldots,x_{(K-1)}) = \sum_{\mathbf{i}} \phi(x_{(i_1)},x_{(i_2)},\ldots,x_{(i_{K-1})}) \ ,$$

the summation being over all possible permutations

$$\mathbf{i} = (i_1,i_2,\ldots,i_K)$$

of the sequence $(1,2,3,\ldots,K)$.

It is easy to check, by expanding out the exponential term in (2.2) in powers of $\Sigma\Sigma\sigma_{ij}x_ix_j$ and summing over the permutations $\mathbf{i}$, that

$$\begin{aligned} f_K(x_{(1)},x_{(2)},\ldots,x_{(K-1)}) = C_K^{-1}K!\,\{1 &+ S_1F + S_2(1-F) \\ &+ \tfrac{1}{2}\,S_3(1-6F+3F^2+8G-6H) \\ &+ \tfrac{1}{2}\,S_4(F-F^2-2G+2H) \\ &+ \tfrac{1}{2}\,S_5(F^2-H) + \tfrac{1}{2}\,S_6(G-H) + \tfrac{1}{2}\,S_7H \\ &+ O\,(\sigma^3)\}\,(\prod_{i=1}^{K} x_{(i)})^{\varepsilon-1} \ , \end{aligned} \tag{2.5}$$

where S_1 and S_2 are $O(\sigma)$, the order of any of the σ_{ij}:

$$S_1 = \sum_{i=1}^{K} \sigma_{ii}/K \ , \ S_2 = \sum_{i\neq j}\sum \sigma_{ij}/[K(K-1)] \ ,$$

S_3 through S_7 are $O(\sigma^2)$:

$$\begin{aligned} S_3 &= \sum_{i,j,k,l}\sum\sum\sum \sigma_{ij}\,\sigma_{kl}/[K(K-1)(K-2)(K-3)] \ , \\ &\qquad \text{all different} \\ S_4 &= 2\sum_i\sum_j\sum_k (\sigma_{ii}\,\sigma_{jk} + 2\sigma_{ij}\,\sigma_{ik})/[K(K-1)(K-2)] \ , \\ &\qquad \text{all different} \\ S_5 &= \sum_{i\neq j}\sum (\sigma_{ii}\sigma_{jj} + 2\sigma_{ij}^2)/[K(K-1)] \ , \\ S_6 &= 4\sum_{i\neq j}\sum \sigma_{ij}\sigma_{jj}/[K(K-1)] \ , \\ S_7 &= \sum_i \sigma_{ii}^2/K \ , \end{aligned}$$

and where

$$F = \sum_1^K x_{(i)}^2, \ G = \sum_1^K x_{(i)}^2, \ H = \sum_1^K x_{(i)}^4.$$

A similar expansion in (2.4) leads to the series for C_K:

$$\begin{aligned} C_K = \frac{(\Gamma(\varepsilon))^K}{\Gamma(K\varepsilon)}\{1 &+ S_1\frac{\varepsilon+1}{K\varepsilon+1} + S_2\frac{(K-1)\varepsilon}{K\varepsilon+1} \\ &+ \tfrac{1}{2}\,S_3\frac{(K-1)(K-2)(K-3)\varepsilon^3}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)} \quad + \tfrac{1}{2}\,S_4\frac{(K-1)(K-2)(\varepsilon+1)\varepsilon^2}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)} \\ &+ \tfrac{1}{2}\,S_5\frac{(K-1)(\varepsilon+1)^2\varepsilon}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)} \quad + \tfrac{1}{2}\,S_6\frac{(K-1)(\varepsilon+2)(\varepsilon+1)\varepsilon}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)} \\ &+ \tfrac{1}{2}\,S_7\frac{(\varepsilon+3)(\varepsilon+2)(\varepsilon+1)}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)} \quad + O\,(\sigma^3)\} \ . \end{aligned} \tag{2.6}$$

Putting (2.5) and (2.6) together, it may be deduced that

$$f_K(x_{(1)},x_{(2)},\ldots,x_{(K-1)})$$

$$= \frac{\Gamma(K\varepsilon)}{(\Gamma(\varepsilon))^K} K! \{1 + (S_1 - S_2)\left(F - \frac{\varepsilon+1}{K\varepsilon+1}\right)$$
$$+ (S_1(\varepsilon+1) + S_2(K-1)\varepsilon)^2/(K\varepsilon+1)^2$$
$$+ \tfrac{1}{2} S_3 \left(1 - 6F + 3F^2 + 8G - 6H - \frac{(K-1)(K-2)(K-3)\varepsilon^3}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)}\right)$$
$$+ \tfrac{1}{2} S_4 \left(F - F^2 - 2G + 2H - \frac{(K-1)(K-2)(\varepsilon+1)\varepsilon^2}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)}\right)$$
$$+ \tfrac{1}{2} S_5 \left(F^2 - H - \frac{(K-1)(\varepsilon+1)^2\varepsilon}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)}\right)$$
$$+ \tfrac{1}{2} S_6 \left(G - H - \frac{(K-1)(\varepsilon+2)(\varepsilon+1)\varepsilon}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)}\right)$$
$$+ \tfrac{1}{2} S_7 \left(H - \frac{(\varepsilon+3)(\varepsilon+2)(\varepsilon+1)}{(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)}\right)$$
$$+ \quad O(\sigma^3)\} \left(\prod_{i=1}^{K} x_{(i)}\right)^{\varepsilon-1} . \qquad (2.7)$$

It s clear that if the selective parameters σ_{ij} are all small, the first-order selection term in the likelihood (2.7) most influences the ratio of the likelihood under selection, (2.7), to the likelihood under no selection. The statistic F, the homozygosity, is powerful to detect departures of $S_1 - S_2$ from 0, using the usual likelihood-ratio method of test. In fact, multiplying both sides of (2.7) by F and integrating over all possible $x_{(1)}, x_{(2)}, \ldots, x_{(K-1)}$ values leads to the equation for the expected homozygosity:

$$E(F) = E(F|\text{neutrality}) + (S_1 - S_2)\,\text{Var}(F|\text{neutrality}) + O(\sigma^2) , \qquad (2.8)$$

where the moments on the right are known (see, *e.g.*, STEWART 1976),

$$E(F|\text{neutrality}) = (\varepsilon+1)/(K\varepsilon+1) ,$$

and

$$\text{Var}(F|\text{neutrality}) = 2(K-1)(\varepsilon+1)\varepsilon/[(K\varepsilon+3)(K\varepsilon+2)(K\varepsilon+1)^2] .$$

We therefore conclude that F is a statistic helpful in detecting differences between the average fitness of homozygotes as measured by S_1 and the average fitness of heterozygotes as measured by S_2. It was advocated for exactly this purpose in WATTERSON (1977), in the special case when all homozygotes had the same fitness, and all heterozygotes had the same fitness, different from that of homozygotes. Considerably more detail was obtained for that situation.

It can be seen from (2.7) that if the difference $S_1 - S_2$ is not the major selective effect, then the higher order statistics G and H may be useful, as well as F, to detect departures from neutrality.

In at least one other special case, the deleterious alleles case, F remains preeminent as test statistic. Suppose that L alleles $A_1, A_2, \ldots, A_L$ are fully fit, whereas the alleles $A_{L+1}, \ldots, A_K$ are deleterious. We assume additivity in the sense that two deleterious alleles are twice as disadvantageous as one:

$$\sigma_{ij} = 0 \text{ if } i,j \leqslant L \ ,$$
$$\sigma_{ij} = -\sigma \text{ if } i \leqslant L < j \ , \tag{2.9}$$

and

$$\sigma_{ij} = -2\sigma \text{ if } L < i,j \ ,$$

where σ is a measure of disadvantage carried by a single deleterious allele. We may substitute these values into the formulas for S_1 to S_7, or what is perhaps easier, rework the problem from scratch, to obtain the likelihood

$$f_K(x_{(1)},x_{(2)},\ldots,x_{(K-1)})$$
$$= \frac{\Gamma(K\varepsilon)}{(\Gamma(\varepsilon))^K} K! \{1 + 2\sigma^2 \frac{L(K-L)}{K(K-1)} \left(F - \frac{\varepsilon+1}{K\varepsilon+1}\right) + O(\sigma^3)\} \left(\prod_{i=1}^{K} x_{(i)}^{\varepsilon-1}\right) . \tag{2.10}$$

Notice that in this situation, because $S_1 = S_2$ due to there being no dominance, we find that the leading selection term is of order $O(\sigma^2)$ rather than $O(\sigma)$. This may be contrasted with (2.5) in general, and in particular with the special case discussed by WATTERSON (1977) when all heterozygotes had the same advantage over all homozygotes, so that

$$\sigma_{ii} = 0 \ , \ \sigma_{ij} = \sigma \text{ for all } i \neq j \ . \tag{2.11}$$

Then we find that $S_1 = 0$ and $S_2 = \sigma$, so that

$$f_K(x_{(1)},x_{(2)},\ldots,x_{(K-1)})$$
$$= \frac{\Gamma(K\varepsilon)}{(\Gamma(\varepsilon))^K} K! \{1 - \sigma\left(F - \frac{\varepsilon+1}{K\varepsilon+1}\right) + O(\sigma^2)\} \left(\prod_{i=1}^{K} x_{(i)}^{\varepsilon-1}\right) . \tag{2.12}$$

The major difference between models having small heterozygote advantage, as in (2.12), and slightly deleterious alleles, as in (2.10), is the replacement of σ in the former by $-2\sigma^2 \frac{L(K-L)}{K(K-1)}$ in the latter. If $\sigma \ll 1$ in both cases, we see from (2.8), for instance, that the homozygosity will be much more influenced by heterozygote advantage (or disadvantage) than by deleterious alleles. And the same conclusion would apply to any other statistic which treats allele frequencies symmetrically (*i.e.*, depends only on their order statistics).

The distribution of F has been studied by STEWART (1976) in the case $K = 3$, assuming the mutation parameter ε is known and that no selective differences operate. By conditioning on the observed value of $\prod_{i=1}^{K} x_{(i)}$, WATTERSON and PERLOW (1978) have found the distribution of F under the null hypothesis of no selection, free of the nuisance parameter ε, again in the case $K = 3$.

2.2 The infinitely many alleles model: Letting $K \to \infty$ and $\varepsilon \to 0$ in such a way that Θ and the σ_{ij} remain fixed, we find for instance that (2.8) remains true with $S_1 = \lim_{K\to\infty} \sum_1^K \sigma_{ii}/K$, $S_2 = \lim_{K\to\infty} \sum\sum_{i\neq j} \sigma_{ij}/[K(K-1)]$, and with

$$E(F|\text{neutrality}) = 1/(\Theta + 1)$$

and

$$\mathrm{Var}(F|\text{neutrality}) = 2\Theta/[(\Theta+3)(\Theta+2)(\Theta+1)^2] .$$

We shall, from now on, concentrate our attention on the neutral alleles case, the simple heterozygote advantage (or disadvantage) case (2.11), and the simple deleterious alleles case (2.9). In the latter case, suppose that the proportion of deleterious alleles converges to α as $K \to \infty$, that is

$$\lim_{K\to\infty} \frac{K-L}{K} = \alpha .$$

Then the correspondence between the alternative hypothesis models is that σ (for heterosis) is replaced in formulas by $-2\sigma^2\alpha(1-\alpha)$ in the deleterious case, at least in the leading selection term. Some examples will now be given.

In WATTERSON [1977, (3.1.11)], it was found that heterosis in the infinitely many alleles model led to a frequency spectrum in the population described by

$$\Phi(x) = \Theta x^{-1}(1-x)^{\Theta-1}\{1+\sigma x[2-(2+\Theta)x]/(1+\Theta)+O(\sigma^2)\} \text{ for } 0<x<1 ,$$

which may now be transcribed for the deleterious alleles model to read

$$\Phi(x) = \Theta x^{-1}(1-x)^{\Theta-1}\{1-2\sigma^2\alpha(1-\alpha)x[2-(2+\Theta)x]/(1+\Theta) + O(\sigma^3)\} ,$$
$$\text{for } 0<x<1 .$$

Thus the possibility of deleterious alleles tends to *decrease* the number of alleles having frequencies x below $2/(2+\Theta)$, and to *increase* those having frequencies x above $2/(2+\Theta)$, compared with the neutral model. One might have expected deleterious alleles to boost the number of low-frequency alleles.

The expected number of alleles, K_q say, whose frequencies are above some small threshold value q, was given in WATTERSON [1977, (3.2.1)] as

$$E(K_q) \doteq \int_q^1 \Theta x^{-1}(1-x)^{\Theta-1}dx + \sigma\,\Theta/(1+\Theta)^2 + O(\sigma^2)$$

for the heterosis model. For deleterious alleles, we obtain instead

$$E(K_q) \doteq \int_q^1 \Theta x^{-1}(1-x)^{\Theta-1}dx - 2\sigma^2\alpha(1-\alpha)\Theta/(1+\Theta)^2 + O(\sigma^3) .$$

Deleterious alleles tend to decrease the number of alleles of frequency above q, compared with a neutral allele model.

Section 3. Sample statistics.

We now discuss the effect of deleterious alleles on the composition of a sample of n genes, chosen at random from the population (2.10), or from the corresponding infinitely many alleles population.

A sample may contain a random number of alleles; we denote the number by k, and the numbers of genes of those various alleles by $n_1, n_2, \ldots, n_k$. Indeed, as it is the "unlabeled", or the "configuration" aspects of the sample which are of interest, we shall assume the allele numbers are arranged in decreasing order:

$$n_1 \geqslant n_2 \geqslant \ldots \geqslant n_k > 0 \ .$$

The *sample homozygosity* (calculated from observed allele frequencies rather than from observed homozygote frequencies) is

$$\hat{F} = \sum_{i=1}^{k} n_i^2/n^2 \ ,$$

and we shall denote by α_j the number of alleles that have j copies in the sample, so that $\hat{F} = \sum_{j=1}^{n} \alpha_j j^2/n^2$, $k = \sum_{j=1}^{n} \alpha_j$, and $n = \sum_{j=1}^{n} j\alpha_j$.

The probability of observing such a sample in the heterosis case was given by WATTERSON [1977, (4.1.4)] as

$$\Pr(k;n_1,n_2,\ldots,n_k) = \frac{n!}{\alpha_1!\alpha_2!\ldots\alpha_n!}\,[\Gamma(K\varepsilon)/\Gamma(n+K\varepsilon)]$$

$$[K!/(K-k)!\Gamma^k(\varepsilon)]\Big[\prod_{i=1}^{k}(\Gamma(n_i+\varepsilon)/n_i!)\Big]\{1+\sigma A_K+O(\sigma^2)\} \ , \qquad (3.1)$$

where

$$A_K \equiv [(1+\varepsilon)/(1+K\varepsilon)] - [n^2\hat{F}+n(2\varepsilon+1)+K\varepsilon(\varepsilon+1)]/(n+K\varepsilon)(n+1+K\varepsilon) \ .$$

In view of the remarks made in the previous section, (3.1) remains valid for the deleterious alleles model if we replace σ by $-2\sigma^2\dfrac{L(K-L)}{K(K-1)}$, and $O(\sigma^2)$ by $O(\sigma^3)$.

Thus $\hat{F}$ is clearly the appropriate statistic for testing for small departures from neutrality in any of the cases of heterozygote advantage, heterozygote disadvantage, or deleterious alleles.

In the infinitely many alleles version, (3.1) reduces to

$$\Pr(k;n_1,n_2,\ldots,n_k) = \frac{n!}{\alpha_1!\alpha_2!\ldots\alpha_n!n_1n_2\ldots n_k}\,[\Gamma(\Theta)/\Gamma(n+\Theta)]$$

$$\Theta^k\{1+\sigma A+0(\sigma^2)\} \ , \qquad (3.2)$$

where

$$A = n[1/(1+\Theta)-n\hat{F}/(n+\Theta)]/(n+1+\Theta) \ .$$

The deleterious alleles version of (3.2) is obtained by replacing σ by $-2\sigma^2\alpha(1-\alpha)$, and $O(\sigma^2)$ by $O(\sigma^3)$. Again, $\hat{F}$ is clearly a suitable test statistic for detecting departures from neutrality in either case. We tabulate $\hat{F}$'s distribution in *Section 4*, conditional on a given value of k (which removes dependence on the nuisance parameter Θ, assuming the neutral hypothesis is true).

Among the results obtained in WATTERSON (1977) that may be adapted to the deleterious alleles case, we quote the following two. Other results may be similarly adapted.

From (3.2), it may be shown [*cf.* WATTERSON 1977, (4.2.9)] that

$$E(k) = \Theta \sum_{i=0}^{n-1} (\Theta+i)^{-1} + \sigma\Theta n(n-1)/(1+\Theta)^2(n+\Theta)(n+1+\Theta) + O(\sigma^2) ,$$

which in the deleterious alleles context is replaced by

$$E(k) = \Theta \sum_{i=0}^{n-1} (\Theta+i)^{-1} - 2\sigma^2\alpha(1-\alpha)\Theta n(n-1)/(1+\Theta)^2(n+\Theta)(n+1+\Theta) + O(\sigma^3) .$$

Hence we may conclude that the presence of slightly deleterious alleles tends to reduce the number of different alleles observed in a sample, compared with the neutral case.

For the testing of neutrality using $\hat{F}$ conditional on a given k, WATTERSON [1977, (4.2.6)] found that the sample likelihood under heterozygote advantage is

$$\Pr(n_1,n_2,\ldots,n_k|k) = \frac{n!}{\alpha_1!\alpha_2!\ldots\alpha_n! n_1 n_2 \ldots n_k |S_n^{(k)}|}\{1-\sigma C+O(\sigma^2)\} \quad (3.3)$$

where $$C = \frac{n^2}{(n+\Theta)(n+1+\Theta)}\left[\hat{F} - \frac{1}{n} - \left(1-\frac{1}{n}\right) G(k)\right] ,$$

where $S_n^{(k)}$ is a Stirling number of the first kind, and where

$$\frac{1}{n} + \left(1-\frac{1}{n}\right) G(k) \equiv \frac{1}{n} + \left(1-\frac{1}{n}\right) \sum_{l=1}^{k} S_n^{(l)}/S_n^{(k)}$$

is the expected value of $\hat{F}$ under neutrality, $E(\hat{F}|k,\sigma=0)$. The corresponding likelihood for deleterious alleles is, as usual, obtained by replacing σ by $-2\sigma^2\alpha(1-\alpha)$, and $O(\sigma^2)$ by $O(\sigma^3)$, in (3.3). We see in either case that the likelihood ratio test, comparing the likelihood with selection with the likelihood when $\sigma = 0$, is most influenced by the quantity C, or equivalently, by the departure of $\hat{F}$ from its null hypothesis mean.

Under heterozygote advantage, the expected value of $\hat{F}$ is, from (3.3),

$$E(\hat{F}|k) = E(\hat{F}|k,\sigma=0) - \sigma n^2 \mathrm{Var}(\hat{F}|k,\sigma=0)/(n+\Theta)(n+1+\Theta) + O(\sigma^2) ,$$

which reads in the deleterious alleles context:

$$E(\hat{F}|k) = E(\hat{F}|k,\sigma=0) + 2\sigma^2\alpha(1-\alpha)n^2\mathrm{Var}(\hat{F}|k,\sigma=0)/(n+\Theta)(n+1+\Theta) + O(\sigma^3) ,$$

showing that deleterious alleles tend to increase the value of $\hat{F}$, conditional on k. The moments of $\hat{F}$, conditional on k and on $\sigma = 0$, were given explicitly in WATTERSON (1977) in terms of complicated functions of Stirling numbers of the first kind. See also Tables 2 and 3 below.

Section 4. Significance points for the $\hat{F}$ *statistic.*

Suppose that a sample of n genes contains k different alleles. The null hypothesis (neutral allele) distribution of the sample order statistics may be obtained from (3.2) as

$$\Pr(k;n_1,n_2,\ldots,n_k) = \frac{n!\Theta^k}{\alpha_1!\alpha_2!\ldots\alpha_n!n_1n_2\ldots n_k}\,\Gamma(\Theta)/\Gamma(n+\Theta)$$

and, conditional on k being given, from (3.3) as

$$\Pr(n_1,n_2,\ldots n_k|k) = \frac{n!}{\alpha_1!\alpha_2!\ldots\alpha_n!n_1n_2\ldots n_k}\Big/\left|S_n^{(k)}\right| \quad . \qquad (4.1)$$

EWENS (1972) was the first to obtain these distributions and to observe that (4.1) is free of the nuisance parameter Θ, the scaled mutation rate. The sample homozygosity, $\hat{F}$, conditional on given values of k and n, has a distribution that may be computed from (4.1) using much computer time or, more quickly, by simulation. Following EWENS (1972), KIRBY (1975), and STEWART (appendix to FUERST, CHAKRABORTY and NEI 1977), it is easy to generate samples having (4.1) as distribution, and hence to obtain the distribution, moments, etc. of sample statistics by simulation. In Figure 1, we show the frequency polygons of the distributions of $\hat{F}$ for $k = 2$, 5 and 10 combined with $n = 50$ and $n = 500$, obtained by simulation using 1000 samples each.

In Table 1, we give a more extensive tabulation of significance points for the $\hat{F}$ distribution, conservative in the sense that for 1%, 2.5%, 5%, 10% and 50%, no more than those proportions were observed at or below the tabulated $\hat{F}$ value; for the 90%, 95%, 97.5% and 99% points, no fewer than those proportions were

TABLE 1

Conservative % points for neutral $\hat{\mathrm{F}}$, *by simulation if* $\mathrm{k} > 2$

k	n	F_{min}	1	2.5	5	10	50	90	95	97.5	99	F_{max}
2	50	0.5000	0.5000	0.5000	0.5032	0.5200	0.7880	—	—	—	—	0.9608
	100	0.5000	0.5000	0.5008	0.5050	0.5288	0.8528	—	—	—	—	0.9802
	200	0.5000	0.5002	0.5018	0.5098	0.5392	0.8961	—	—	—	—	0.99005
	500	0.5000	0.5004	0.5032	0.5135	0.5525	0.9344	—	—	—	—	0.996008
3	50	0.33	0.34	0.35	0.37	0.41	0.62	0.89	—	—	—	0.9224
	100	0.33	0.34	0.37	0.40	0.44	0.67	0.96	0.96	—	—	0.9606
	200	0.33	0.34	0.37	0.40	0.46	0.68	0.95	0.98	—	—	0.98015
	500	0.33	0.36	0.39	0.44	0.48	0.76	0.98	0.99	0.99	—	0.992024
5	50	0.20	0.23	0.25	0.26	0.28	0.41	0.66	0.69	0.78	0.82	0.848
	100	0.20	0.24	0.26	0.28	0.30	0.44	0.73	0.80	0.84	0.89	0.922
	200	0.20	0.25	0.28	0.30	0.32	0.49	0.80	0.85	0.90	0.92	0.9605
	500	0.20	0.26	0.30	0.33	0.35	0.53	0.87	0.92	0.95	0.97	0.98408
7	50	0.14	0.18	0.18	0.20	0.21	0.29	0.46	0.54	0.57	0.64	0.7768
	100	0.14	0.19	0.20	0.22	0.24	0.34	0.57	0.66	0.72	0.81	0.8842
	200	0.14	0.20	0.22	0.23	0.25	0.39	0.65	0.74	0.79	0.83	0.94105
	500	0.14	0.21	0.23	0.25	0.28	0.43	0.73	0.81	0.86	0.91	0.976168
10	50	0.10	0.13	0.13	0.14	0.15	0.21	0.30	0.36	0.41	0.48	0.676
	100	0.10	0.14	0.15	0.16	0.17	0.25	0.40	0.45	0.50	0.54	0.829
	200	0.10	0.15	0.16	0.18	0.19	0.28	0.49	0.55	0.62	0.68	0.91225
	500	0.10	0.17	0.18	0.20	0.22	0.33	0.56	0.66	0.72	0.81	0.96436

— denotes significance level not possible.

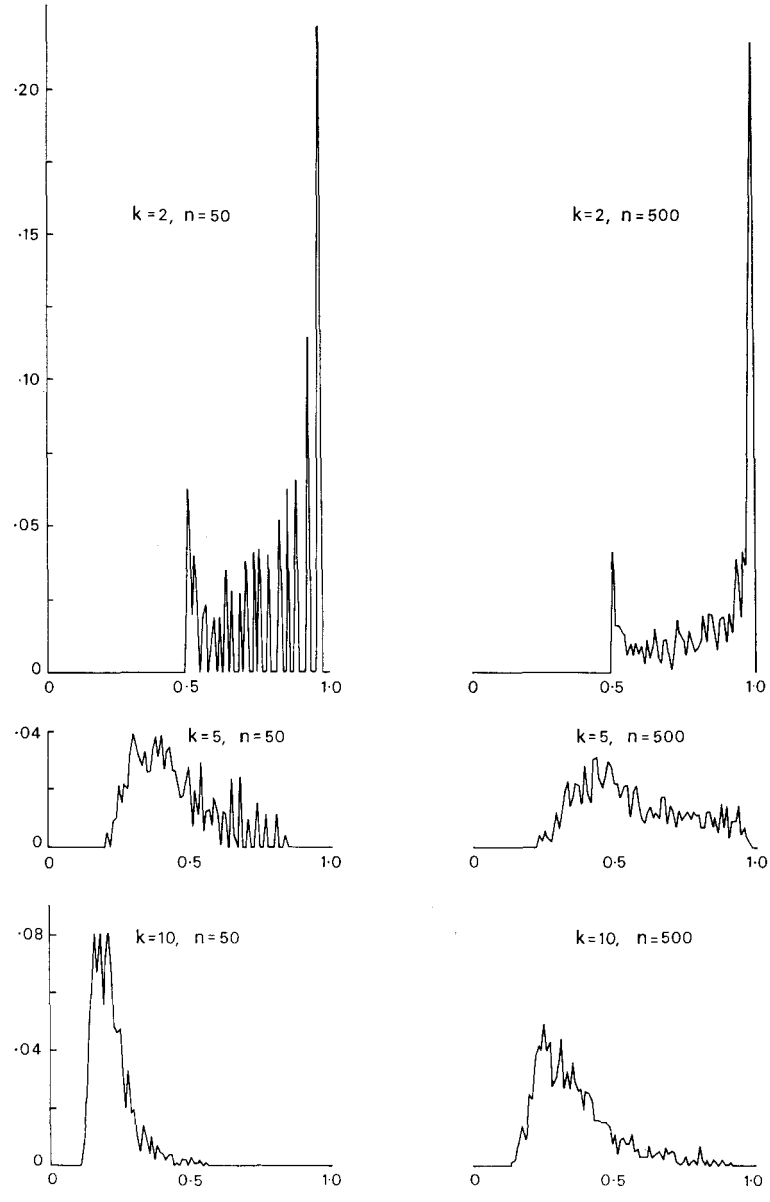

FIGURE 1.—Frequency polygons for the distribution of F.

observed below the tabulated $\hat{F}$ value. Again, the table is based on 1000 samples for each k, n combination except that the $k=2$ cases are by analytic computation, not simulation. We have also listed the lowest possible, and highest possible, $\hat{F}$ value. $\hat{F}$ is minimum in samples in which all alleles are equally represented (and then $\hat{F}=k^{-1}$); of course this may not be possible exactly if n is not divisible by k. The maximum $\hat{F}$ value is achieved when $k-1$ alleles have

TABLE 2

Mean of $\hat{F}$; *exact (simulation)*

k	n = 50	100	200	500
2	0.7812	0.8088	0.8306	0.8530
	(0.7786)	(0.8095)	(0.8264)	(0.8495)
3	0.6302	0.6714	0.7048	0.7401
	(0.6402)	(0.6804)	(0.6924)	(0.7379)
5	0.4388	0.4895	0.5322	0.5789
	(0.4396)	(0.4813)	(0.5275)	(0.5771)
7	0.3253	0.3764	0.4207	0.4704
	(0.3211)	(0.3788)	(0.4225)	(0.4746)
10	0.2242	0.2714	0.3134	0.3620
	(0.2227)	(0.2708)	(0.3156)	(0.3660)

relative frequency $1/n$ each, and one allele has relative frequency $1-(k-1)/n$. Then $\hat{F} = (k-1)n^{-2} + (1-(k-1)/n)^2$. For some (k,n) combinations, even the most extreme F values are not statistically significant, because even under the null hypothesis they were found to occur in frequencies in excess of the desired significance level.

In Tables 2 and 3 we exhibit the means and variances of $\hat{F}$ under the neutrality hypothesis; the exact values were obtained by computing (4.3.7) and (4.3.9) in WATTERSON (1977); the simulation values are based on 1000 samples in each case. Except possibly for the $k = 7$, $n = 50$ and $k = 10$, $n = 100$ variances, the simulation results agree very closely with the theoretical values.

TABLE 3

Variance of $\hat{F}$; *exact (simulation)*

k	n = 50	100	200	500
2	0.0265	0.0278	0.0281	0.0275
	(0.0261)	(0.0284)	(0.0287)	(0.0270)
3	0.0283	0.0325	0.0350	0.0366
	(0.0293)	(0.0334)	(0.0350)	(0.0368)
5	0.0189	0.0254	0.0306	0.0356
	(0.0182)	(0.0244)	(0.0289)	(0.0342)
7	0.0110	0.0169	0.0224	0.0286
	(0.0096)	(0.0175)	(0.0220)	(0.0282)
10	0.0048	0.0089	0.0133	0.0190
	(0.0047)	(0.0080)	(0.0131)	(0.0196)

Some numerical examples were quoted in WATTERSON (1977) for Drosophila samples and we repeat them here. In Table 4 we show the sample sizes, n, the numbers of alleles, k, the homozygosities, $\hat{F}$, the tail-probabilities of getting a more extreme $\hat{F}$ value, P (calculated exactly in WATTERSON 1977), the normal approximations to those tail probabilities, P_{norm}, and the simulation tail probabilities, P_{sim} (based on the actual k, n values, each simulation consisting of 1000 replicates). The simulation probabilities are consistent with those obtained by interpolating in Table 1.

The virtue of using simulated distributions is in respect of computer time. The simulated one-sided significance levels in Table 4 are in good agreement with the true values, where known, and indicate possible departures from neutrality in three of the four species. For the *simulans* sample, however, we had not previously obtained even the exact tail probability due to computer time limitations, whereas it may be simulated in approximately half a minute on a B6700, using 1000 samples.

We might interpret the data in Table 4 as perhaps indicating the presence of heterozygote advantage in *simulans*, and the presence of heterozygote disadvantage or of deleterious alleles in the *willistoni* and *equinoxalis* samples. However, as is usual in this work, there may be other reasons for samples to deviate from the neutral distribution, such as nonstationarity of the population composition, indistinguishability of different alleles, etc.

I thank MRS. M. WU for help with computing, and Professor M. NEI for pointing out the possibility of deleterious alleles influencing homozygosity.

LITERATURE CITED

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

FUERST, P. A., R. CHAKRABORTY and M. NEI, 1977 Statistical studies on protein polymorphism in natural populations, I. Distribution of single locus heterozygosity. Genetics **86**: 455–483.

KIMURA, M., 1956 *Stochastic Processes in Population Genetics*, Ph.D. thesis, University of Wisconsin, Madison.

TABLE 4

Drosophila sample statistics

Species	n	k	$\hat{F}$	P	$P_{\text{norm.}}$	$P_{\text{sim.}}$
willistoni	582	7	0.9230	0.00690	0.0048	0.009
tropicalis	298	7	0.6475	0.130	0.100	0.134
equinoxalis	376	5	0.9222	0.0355	0.0268	0.044
simulans	308	7	0.2356	?	0.0951	0.044

For *simulans*, left hand tail probabilities; for others, right tail probabilities.

KIRBY, K., 1975 A discussion of simulation results for various aspects of the neutral allele model. Theor. Pop. Biol. **7**: 277–287.

LI, W. H., 1977 Maintenance of genetic variability under mutation and selection pressures in a finite population. Proc. Natl. Acad. Sci. U.S. **74**: 2509–2513.

STEWART, F. M., 1976 Variability in the amount of heterozygosity maintained in neutral populations, Theor. Pop. Biol. **9**: 188–201.

WATTERSON, G. A., 1977 Heterosis or Neutrality? Genetics **35**: 789–814.

WATTERSON, G. A. and J. PERLOW, 1978 Homozygosity in three-allele populations. Theor. Pop. Biol. (In press).

WRIGHT, S., 1949 Adaptation and selection. pp. 365–389. In: *Genetics, Paleontology and Evolution*. Edited by G. L. JEPSON, G. G. SIMPSON and E. MAYR. Princeton University Press, Princeton.

Corresponding editor: W. J. EWENS