MAINTENANCE OF GENETIC VARIABILITY UNDER THE PRESSURE OF NEUTRAL AND DELETERIOUS MUTATIONS IN A FINITE POPULATION

WEN-HSIUNG LI

Center for Demographic and Population Genetics, University of Texas
Health Science Center at Houston, Houston, Texas

Manuscript received July 5, 1978
Revised copy received November 3, 1978

ABSTRACT

In order to assess the effect of deleterious mutations on various measures of genic variation, approximate formulas have been developed for the frequency spectrum, the mean number of alleles in a sample, and the mean homozygosity; in some particular cases, exact formulas have been obtained. The assumptions made are that two classes of mutations exist, neutral and deleterious, and that selection is strong enough to keep deleterious alleles in low frequencies, the mode of selection being either genic or recessive. The main findings are: (1) If the expected value ($\bar{q}$) of the sum of the frequencies of deleterious alleles is about 10% or less, then the presence of deleterious alleles causes only a minor reduction in the mean number of neutral alleles in a sample, as compared to the case of $\bar{q} = 0$. Also, the low- and intermediate-frequency parts of the frequency spectrum of neutral alleles are little affected by the presence of deleterious alleles, though the high-frequency part may be changed drastically. (2) The contribution of deleterious mutations to the expected total number of alleles in a sample can be quite large even if $\bar{q}$ is only 1 or 2%. (3) The mean homozygosity is roughly equal to $(1-2\bar{q})/(1+\theta_1)$, where θ_1 is twice the number of new neutral mutations occurring in each generation in the total population. Thus, deleterious mutations increase the mean heterozygosity by about $2\bar{q}/(1+\theta_1)$. The present results have been applied to study the controversial problem of how deleterious mutations may affect the testing of the neutral mutation hypothesis.

USING WRIGHT's (1949a) multiallelic distribution, I have recently developed formulas for the frequency spectrum, the mean number of alleles in a sample, and the mean and variance of homozygosity under mutation pressure, under either genic or recessive selection (LI 1977, 1978; see also WATTERSON 1978a). These formulas are general, but become computationally intractable when the intensity of selection is strong. Simpler formulas are therefore needed for the case of strong selection. In some particular cases, I have recently been able to reduce my formulas to simple forms by algebraic manipulations, but I have not been able to do so in general. I have, however, used a heuristic approach similar to that of WRIGHT (1966) to get approximate formulas that are useful for the study of strong selection. Recently, W. J. EWENS (personal communication) has pursued the same problem and has, somewhat earlier than I, obtained approximate formulas for the mean number of alleles in a sample and the mean homozygosity for the case of genic selection, using a different heuristic approach. My results for this case largely agree with his. The purpose of this communication is to present my new results mentioned above and to apply them to investigate the controversial problem of how deleterious mutations may affect the testing of the neutral mutation hypothesis (EWENS 1972; NEI 1975; OHTA 1976; WATTERSON 1978b,c).

In this study, I assume that there are two classes of mutations, neutral and deleterious, and that selection is sufficiently strong to keep deleterious alleles in low frequencies. Under certain circumstances, my approximate formulas for the frequency spectrum of neutral alleles are not very satisfactory. Nevertheless, they provide deep insight into the problem and enable us to compute fairly accurately both the mean number of neutral alleles in a sample and the mean homozygosity.

GENIC SELECTION

Consider a randomly mating population of effective size N. Let the number of possible allelic states at a locus be K, and let A_i denote the ith allele and x_i its frequency. We assume that there are two classes of alleles, *i.e.*, neutral alleles and deleterious alleles, and that the neutral class consists of the first I allelic states and the deleterious class the remaining $K-I$ states. Let s be the selection coefficient against deleterious alleles. We use the K-allele model: each gene, of whatever allelic type, has a mutation rate v per generation and the probability of A_i mutating to A_j is $v_1 = v/(K-1)$ for each $j \neq i$, $i = 1, \ldots, K$ (WRIGHT 1949b; KIMURA 1968a). Let $p = x_1 + \ldots + x_I$ be the sum of the frequencies of neutral alleles and $u_1 = I\,v_1$ the sum of the mutation rates to neutral alleles; similarly, let $q = x_{I+1} + \ldots + x_K$ and $u_2 = (K-I)v_1$. In addition, let $\alpha = 4Nv_1$, $\theta_1 = 4Nu_1 = I\alpha$, $\theta_2 = 4Nu_2 = (K-I)\alpha$, and $\theta_T = 4Nv = \theta_1 + \theta_2 - \alpha$. In this paper, we assume that s is at least one order larger than u_2 and that $S = 4Ns$ is sufficiently large so that the expected value ($\bar{q}$) of q at equilibrium is close to the equilibrium value in a population of infinite size, *i.e.*, $\bar{q} \approx \hat{q} = u_2/s = \theta_2/S$.

We first study the frequency spectrum, which is conventionally denoted by $\Phi(x)$ and has the meaning that $\Phi(x)dx$ represents the mean number of alleles whose frequency is between x and $x + dx$. (The frequency spectrum is also commonly known as the distribution of allele frequencies.) $\Phi(x)$ can be decomposed into the frequency spectrum of neutral alleles, $\Phi_1(x)$, and the frequency spectrum of deleterious alleles, $\Phi_2(x)$. We treat $\Phi_1(x)$ and $\Phi_2(x)$ separately.

First, let us consider $\Phi_2(x)$. To this end, we focus our attention on a particular deleterious allele, say A_i, $i > I$. It can be easily shown that the mean change of x_i per generation is approximately given by

$$M_i = v_1(1-x_i) - vx_i - sx_ip \ .$$

Following WRIGHT (1966), we replace p by $\hat{p} = 1 - u_2/s$, and obtain

$$M_i \approx v_1(1-x_i) - x_i(s+u_1-v_1) \ . \qquad (1)$$

My earlier computations for the case of $I=1$ suggest that this approximation introduces no serious errors as long as $S=4Ns$ is larger than 20 (see Tables 1 and 4 of LI 1978). Putting the variance of the change in x_i per generation equal to $x_i(1-x_i)/(2N)$, we find that the equilibrium probability density of x_i is given by

$$\phi(x_i) \approx \frac{\Gamma(S+\theta_1)}{\Gamma(S+\theta_1-\alpha)\Gamma(\alpha)} x_i^{\alpha-1}(1-x_i)^{S+\theta_1-\alpha-1} , \tag{2}$$

in which $\Gamma(\cdot)$ is the gamma function. Since there are $K-I$ allelic states in the second class,

$$\Phi_2(x) = (K-I)\phi(x) .$$

Letting K approach infinity, but keeping v and I/K constant, we obtain the following result for the infinite-allele model:

$$\Phi_2(x) \approx \theta_2 x^{-1}(1-x)^{S+\theta_1-1} . \tag{3}$$

(Note that, in the infinite-allele model, $\theta_1+\theta_2$ becomes equal to θ_T.) The expected number of deleterious alleles in a random sample of m individuals or $2m$ genes is equal to

$$\bar{k}_2 \approx \int_0^1 [1-(1-x)^{2m}]\Phi_2(x)dx \tag{4}$$

$$= \theta_2[(S+\theta_1)^{-1}+(S+\theta_1+1)^{-1}+\ldots+(S+\theta_1+2m-1)^{-1}]$$

$$\approx \theta_2 \log_e[(S+\theta_1+2m-0.5)/(S+\theta_1-0.5)]$$

$$= \theta_2 \log_e[1+2m/(S+\theta_1-0.5)] . \tag{5}$$

Using a different approach, EWENS (personal communication) has obtained a slightly different formula:

$$\bar{k}_2 \approx \theta_2 \log_e[1+2m/(S+\theta-0.5)] . \tag{6}$$

The contribution to mean homozygosity due to deleterious genes is given by

$$\bar{J}_2 = \int_0^1 x^2\Phi_2(x)dx$$

$$\approx \theta_2/[(S+\theta_1)(S+\theta_1+1)] , \tag{7}$$

which is negligibly small, if S is large. EWENS (personal communication) neglects this term by arguing that it is of order S^{-2}. The sample mean is expected to be slightly larger than the population mean given by (7), but the difference is negligible as long as m is larger than 25 (*cf.*, NEI and ROYCHOUDHURY 1974). The same comment applies to formulas (14a,b).

We now consider $\Phi_1(x)$. The mean change of x_i per generation for a neutral allele ($i \leq I$) is approximately given by

$$M_i = v_1(1-x_i) - vx_i + sx_iq . \tag{8}$$

Replacing q by u_2/s, we find

$$M_i \approx v_1(1-x_i) - (u_1-v_1)x_i$$

$$\phi(x_i) \approx \frac{\Gamma(\theta_1)}{\Gamma(\theta_1-\alpha)\Gamma(\alpha)} x_i^{\alpha-1}(1-x_i)^{\theta_1-\alpha-1} \; . \tag{9}$$

From (9), we obtain

$$\Phi_1(x) \approx \theta_1 x^{-1}(1-x)^{\theta_1-1} \tag{10}$$

for the model of infinite alleles. This is identical with the frequency spectrum for the case of neutral mutations with effective size N and mutation rate u_1 (KIMURA and CROW 1964). Thus, substituting u_2/s for q in equation (8) is equivalent to neglecting the class of deleterious mutations. As will be seen later, this approximation creates no serious errors when $\theta_1 \geq 1$. However, serious disagreements between (10) and the exact frequency spectrum occur at high allele frequencies when $\theta_1 < 1$. Fortunately, even in this case, $\Phi_1(x)$ in (10) agrees extremely well with the exact frequency spectrum at low and intermediate allele frequencies, and gives a quite accurate value for the mean number of neutral alleles in the population. Consequently,

$$\bar{k}_1 \approx \int_0^1 [1-(1-x)^{2m}]\Phi_1(x)dx \tag{11}$$

$$\approx \frac{\theta_1}{\theta_1} + \frac{\theta_1}{\theta_1+1} + \ldots + \frac{\theta_1}{\theta_1+2m-1} \tag{12a}$$

provides, in all cases, a close approximation to the mean number of neutral alleles in a sample (see examples and a more detailed expanation below). EWENS (personal communication) has obtained the same formula as (12a). This formula is expected to give an overestimate because it is obtained under the assumption that there exists no deleterious allele in the population.

Another estimate of $\bar{k}_1$ can be obtained as follows. We again neglect the class of deleterious mutations, but assume that the effective population size is $N\bar{p} \approx N(1-u_2/s)$ instead of N. From these two assumptions, we get

$$\bar{k}_1 \approx \frac{\theta_1'}{\theta_1'} + \frac{\theta_1'}{\theta_1'+1} + \ldots + \frac{\theta_1'}{\theta_1'+2m-1} \; , \tag{12b}$$

in which $\theta_1' = \theta_1(1-u_2/s)$. This is an underestimate because the actual effect of random drift on neutral genes is weaker than that created by an effective size of $N(1-u_2/s)$.

When using $\Phi_1(x)$ in (10) to compute $\bar{J}_1$, the contribution to mean homozygosity due to neutral genes, we must remember that this formula was obtained by assuming that the population is free of deleterious mutations and therefore every gene drawn from "the population" is neutral, *i.e.*, $\int_0^1 x\Phi_1(x)dx = 1$. Since the actual probability of a randomly drawn gene being neutral is $\bar{p}$ instead of 1, we need to multiply

$$P_c = \int_0^1 x^2\Phi_1(x)dx = \frac{1}{1+\theta_1} \tag{13}$$

by $\bar{p}^2 \approx (1-u_2/s)^2$ in order to obtain $\bar{J}_1$, namely,

$$\bar{J}_1 \approx \frac{(1-u_2/s)^2}{1+\theta_1}\ . \tag{14a}$$

This formula is identical with that of EWENS (personal communication). We may roughly regard P_c as the conditional probability that two genes randomly chosen from the population are of the same allelic type, given that they are neutral genes. As will be seen below, the actual frequency spectrum of neutral alleles is somewhat less dispersed over the whole allele frequency range than the approximate one given by (10); consequently, P_c is an underestimate. Because of this, formula (14a) tends to be an underestimate; however, it may become an overestimate when θ_1 is small and selection is weak, so that neutral alleles may temporarily become absent from the population and $\bar{p} < 1 - u_2/s$ (see examples below). An overestimate for P_c can be obtained by neglecting the class of deleterious mutations and assuming that the effective population size is $N(1 - u_2/s)$ instead of N. It is given by $P_c = 1/[1 + (1 - u_2/s)\theta_1]$, from which we get

$$\bar{J}_1 \approx (1-u_2/s)^2/[1+(1-u_2/s)\theta_1]\ . \tag{14b}$$

Since P_c is an overestimate, so is formula (14b).

Making use of these formulas, we can compute $\Phi(x) = \Phi_1(x) + \Phi_2(x)$, $\bar{k} = \bar{k}_1 + \bar{k}_2$, and $\bar{J} = \bar{J}_1 + \bar{J}_2$, in which $\bar{k}$ is the mean total number of alleles in a sample and $\bar{J}$ the mean homozygosity of the population.

Let us now examine the accuracy of the above formulas. This can be done by comparing these formulas with my earlier ones (LI 1977, 1978) for reasonably large $4Ns$ values. To a close approximation, my earlier results can be written as follows.

$$\Phi_1(x) = \theta_1 C_1 x^{-1}(1-x)^{\theta_T-1} \sum_{n=0}^{\infty} \frac{(-S)^n\Gamma(n+\theta_2)}{n!\Gamma(n+\theta_T)}(1-x)^n\ , \tag{15}$$

$$\Phi_2(x) = \theta_2 C_2 x^{-1}(1-x)^{\theta_T-1} \sum_{0}^{\infty} \frac{S^n\Gamma(n+\theta_1)}{n!\Gamma(n+\theta_T)}(1-x)^n\ , \tag{16}$$

$$\bar{k}_1 = \theta_1 C_1 \sum_{n=0}^{\infty} \frac{(-S)^n\Gamma(n+\theta_2)}{n!\Gamma(n+\theta_T)} \sum_{i=0}^{2m-1} (n+\theta_T+i)^{-1}\ , \tag{17}$$

$$\bar{k}_2 = \theta_2 C_2 \sum_{n=0}^{\infty} \frac{S^n\Gamma(n+\theta_1)}{n!\Gamma(n+\theta_T)} \sum_{i=0}^{2m-1} (n+\theta_T+i)^{-1}\ , \tag{18}$$

$$\bar{J}_1 = \theta_1 C_1 \sum_{0}^{\infty} (-S)^n\Gamma(n+\theta_2)/[n!\Gamma(n+\theta_T+2)]\ , \tag{19}$$

$$\bar{J}_2 = \theta_2 C_2 \sum_0^\infty S^n \Gamma(n+\theta_1)/[n!\Gamma(n+\theta_T+2)] \ , \tag{20}$$

$$C_1^{-1} = \sum_0^\infty (-S)^n \Gamma(n+\theta_2)/[n!\Gamma(n+\theta_T)] \ ,$$

$$C_2^{-1} = \sum_0^\infty S^n \Gamma(n+\theta_1)/[n!\Gamma(n+\theta_T)] \ .$$

When θ_1 and θ_2 are integers, some simplification of these formulas can be made. Given below are three examples.

For $\theta_1 = 1$ and $\theta_2 = r$, r a positive integer,

$$\Phi_1(x) = x^{-1} - x^{-1} e^{-S(1-x)} \sum_{i=0}^{r-1} S^i (1-x)^i / i! \ , \tag{21}$$

$$\Phi_2(x) = r x^{-1} e^{-Sx} \ , \tag{22}$$

$$\bar{J}_1 = \frac{1}{2} - rS^{-1} + r(r+1)/(2S^2) \ , \tag{23}$$

$$\bar{J}_2 = r/S^2 \ . \tag{24}$$

For $\theta_1 = 2$ and $\theta_2 = r$,

$$\Phi_1(x) = 2x^{-1}(1-x) - \frac{2r}{S-r} + \frac{2x^{-1}}{S-r} e^{-S(1-x)} \sum_0^{r=1} \frac{(r-i)}{i!} S^i (1-x)^i \ , \tag{25}$$

$$\Phi_2(x) = r x^{-1}(1-x) e^{-Sx} - r^2 (S-r)^{-1} e^{-Sx} \ , \tag{26}$$

$$\bar{J}_1 = \frac{1}{3} - \frac{r}{3(S-r)S^2} [2S^2 - 3(r+1)S + (r+1)(r+2)] \ , \tag{27}$$

$$\bar{J}_2 = r(S-2)/S^3 \ . \tag{28}$$

For $\theta_1 = 3$ and $\theta_2 = r$,

$$\begin{aligned} \Phi_1(x) &= 3x^{-1}(1-x)^2 - 3x^{-1}[(r+1)r - r(r-1)S + (r-1)(r-2)S^2/2]^{-1} \\ &\quad \times [r(r-1)Sx(1-x) - (r+1)rx(2-x) \\ &\quad + e^{-S(1-x)} \sum_0^{r=1} (r-i+1)(r-i)S^i(1-x)^i/i!] \ , \end{aligned} \tag{29}$$

$$\begin{aligned} \Phi_2(x) &= r x^{-1}(1-x)^2 e^{-Sx} - r^2 e^{-Sx}[(r+1)r - 2rS + S^2]^{-1} \\ &\quad \times [2S(1-x) - (r+1)(2-x)] \ , \end{aligned} \tag{30}$$

$$\begin{aligned} \bar{J}_1 &= \frac{1}{4} - \frac{r}{4S^3} [(r+1)r - r(r-1)S + (r-1)(r-2)S^2/2]^{-1} \\ &\quad \times [(r-1)S^4 - 5(r+1)S^3 + 4(r+1)(r+2)S^2 \\ &\quad - (r+1)(r+2)(r+3)] \ , \end{aligned} \tag{31}$$

$$\bar{J}_2 = r(S^2 - 4S + 6)/S^4 \ . \tag{32}$$

The corresponding formulas for $\bar{k}_1$ and $\bar{k}_2$ are more complicated and are not given here because it is less convenient to use them than to put $\Phi_1(x)$ and $\Phi_2(x)$ into formulas (4) and (11) and carry out numerical integrations. Formulas (21) to

(32) hold with a high degree of accuracy, though some approximations have been made to simplify them. The mathematical manipulations involved in the derivation are tedious, but the principle is rather simple and can be illustrated by the simplification of C_1 for the simplest case: $\theta_1 = \theta_2 = 1$. Namely,

$$\begin{aligned} C_1^{-1} &= \sum_{n=0}^{\infty} (-S)^n n!/[n!(n+1)!] \\ &= (-S)^{-1} \sum_{n=0}^{\infty} (-S)^{n+1}/(n+1)! \\ &= (-S)^{-1} [-1 + \sum_{n=0}^{\infty} (-S)^n/n!] \\ &= (1-e^{-S})/S \ . \end{aligned}$$

The same principle can be applied to derive formulas for any integral θ_1 and θ_2. Furthermore, using such formulas and the method of numerical interpolations, an approximate frequency spectrum of neutral alleles can be obtained for *arbitrary* (integral or nonintegral) θ_1 and θ_2; numerical computations show that the spectrum thus obtained is quite accurate when $\theta_1 \geq 2$, though it is not very satisfactory when $\theta_1 < 2$. (An example will be given below to illustrate this method

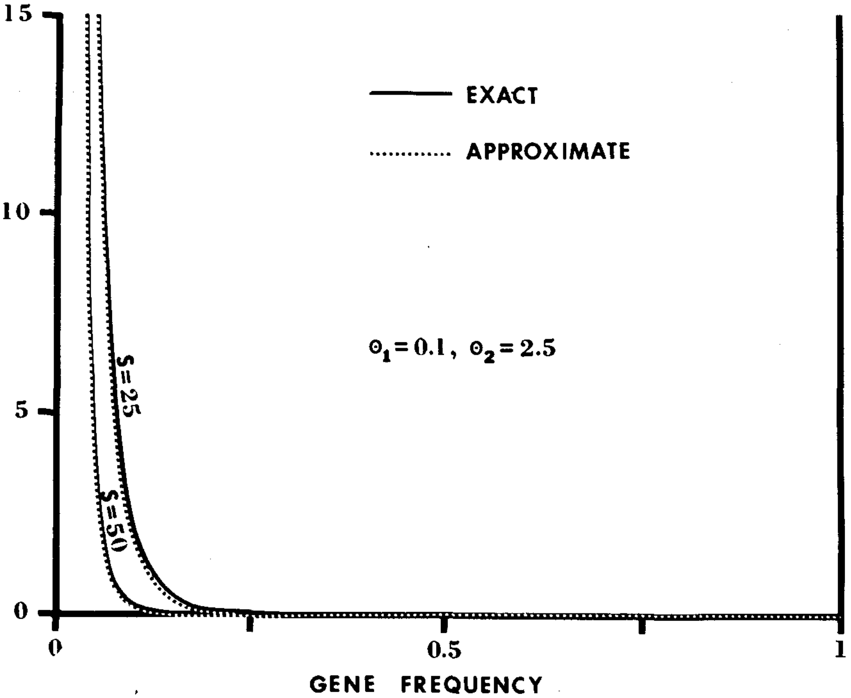

FIGURE 1.—Approximate and exact frequency spectra of deleterious alleles for two cases of genic selection: (1) $S = 25$, $\theta_1 = 0.1$, $\theta_2 = 2.5$ and (2) $S = 50$, $\theta_1 = 0.1$, $\theta_2 = 2.5$. The ordinate denotes $\Phi_2(x)$, which has the meaning that $\Phi_2(x)dx$ represents the expected number of deleterious alleles whose frequency is between x and $x + dx$. ——— Exact frequency spectrum. ······ Approximate frequency spectrum.

of interpolation.) For ease of discussion, we shall call formulas (15) to (32) the "exact" formulas, though terms of order e^{-S} have been neglected in some of these formulas.

Figure 1 shows the approximate (3) and the exact frequency spectrum (16) of deleterious alleles for two cases: (1) $S = 25, \theta_1 = 0.1, \theta_2 = 2.5$, and (2) $S = 50, \theta_1 = 0.1, \theta_2 = 2.5$. In both cases, the approximate frequency spectrum is almost indistinguishable from the exact one. Such close approximations also hold for larger θ_1 values. As an example, if $\theta_1 = 1$ and $\theta_2 = r$, formula (3) becomes $rx^{-1}(1-x)^S$ which is close to the exact frequency spectrum given by (22) as long as S is large and r/S is small. As another example, if $\theta_1 = 2$ and $\theta_2 = r$, formula (3) becomes $rx^{-1}(1-x)^{S+1}$, while the exact frequency spectrum is given by (26). The approximation is now even better. Thus, we may conclude that formula (3) holds well under the conditions specified in this paper. Numerical computations suggest that formula (3) gives an underestimate, unless θ_1 is large (*cf.*, Figure 1).

The comparisons between the approximate (10) and the exact frequency spectrum (15) of neutral alleles for the two cases given in Figure 1 are shown in Figure 2. Since the θ_1 value is the same for both cases, so is the approximate

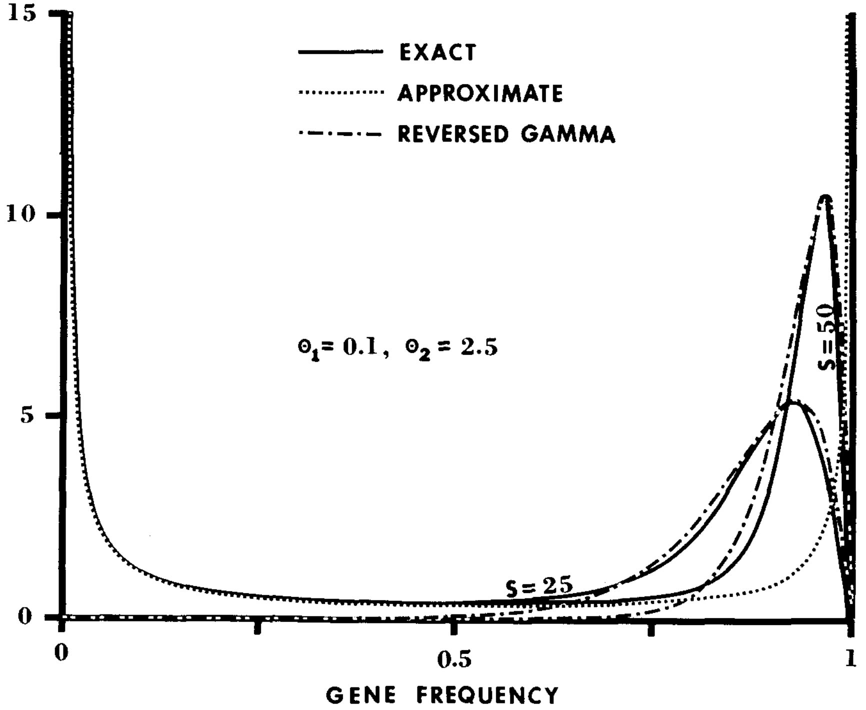

FIGURE 2.—Approximate and exact frequency spectra of neutral alleles for the same two cases shown in Figure 1. The ordinate denotes $\Phi_1(x)$, which has the meaning that $\Phi_1(x)dx$ represents the expected number of neutral alleles whose frequency is between x and $x+dx$. ——— Exact frequency spectrum. ...... Approximate frequency spectrum. —·—·—· Reversed gamma distribution. For detail, see text.

frequency spectrum (10) because it is determined by θ_1 alone (see the dotted line). It is seen from this figure that the approximate frequency spectrum agrees well with the exact one at low and intermediate allele frequencies, but deviates far from it at high allele frequencies. Such serious discrepancies are expected to arise whenever θ_1 is smaller than unity. The explanation is as follows: When $\theta_2 = 0$ (in the absence of deleterious mutations), $\Phi_1(x)$ is exactly given by (10) and is U-shaped, *i.e.*, $\Phi_1(x)$ has a peak at $x = 1$. As θ_2 increases, the probability of monomorphism decreases and the aforementioned peak becomes lower and moves inward; when $\theta_1 + \theta_2 = \theta_T$ becomes larger than one, $\Phi_1(x)$ becomes zero at $x = 1$, and the peak moves to somewhere below $x = 1$ (see LI 1978). Since the approximate frequency spectrum (10) does not change with θ_2, the approximation will become less and less satisfactory as θ_2 increases. Fortunately, the mean number of high-frequency alleles in the population computed by using (10) is only slightly larger than that computed by using (15). For example, the mean number of alleles whose frequency is higher than 0.6 is 0.923 for the curve with $S = 25$, 0.942 for the curve with $S = 50$ and 0.959 for the approximate frequency spectrum given in Figure 2. Since the mean number of high-frequency alleles in a sample of reasonable size should be roughly the same as that in the population, the discrepancies between the approximate and exact frequency spectra at high allele frequencies should introduce no serious errors into formula (12a), the approximate formula for the mean number of neutral alleles in a sample. That this is indeed the case can be illustrated by the following example: For the parameters specified in Figure 2 and $2m = 200$, formula (12a) gives $\bar{k}_1 = 1.57$ and formula (17) gives $\bar{k}_1 = 1.56$ for the case of $S = 25$, and 1.57 for the case of $S = 50$. More examples are given in Table 1.

Figure 3 shows that when $\theta_1 \geq 1$ no large discrepancies such as those of Figure 2 occur between the approximate and the exact frequency spectra of neutral alleles. For the case of $\theta_1 = 1$, the approximate frequency spectrum coincides with the exact one, except at the high-frequency end. For the case of $\theta_1 = 2.5$, some appreciable differences occur between the two frequency spectra at intermediate allele frequencies. The broken line is obtained by interpolation, using formulas (25) and (29). The procedure is as follows: First, we note that $\theta_2 = 2.5$ is not an integer. We therefore raise it to 3 and, at the same time, raise S to 30 so that the original value of $\theta_2/S = 2.5/25$ is maintained. (Alternatively, we may reduce θ_2 to 2 and S to 20.) Next, we note that $\theta_1 = 2.5$ is the average of $\theta_1 = 2$ and $\theta_1 = 3$. We therefore use the average of (25) and (29) with $r = \theta_2 = 3$ and $S = 30$ to approximate the frequency spectrum with $\theta_1 = 2.5$, $\theta_2 = 2.5$ and $S = 25$. It is seen from Figure 3 that the curve thus obtained gives an excellent fit to the exact frequency spectrum. Thus, for $\theta_1 \geq 2$ a close approximation to the frequency spectrum of neutral alleles can be obtained by interpolation, using formulas for integral θ_1 and θ_2; for $\theta_1 < 2$, however, this method is less successful.

The problem that remains to be solved is how to compute the high-frequency part of $\Phi_1(x)$ for $\theta_1 < 1$ when S is large, say $S > 100$, so that formula (15) becomes computationally intractable. This is an important case and a practical formula should be developed. Here, I present only two interesting properties.

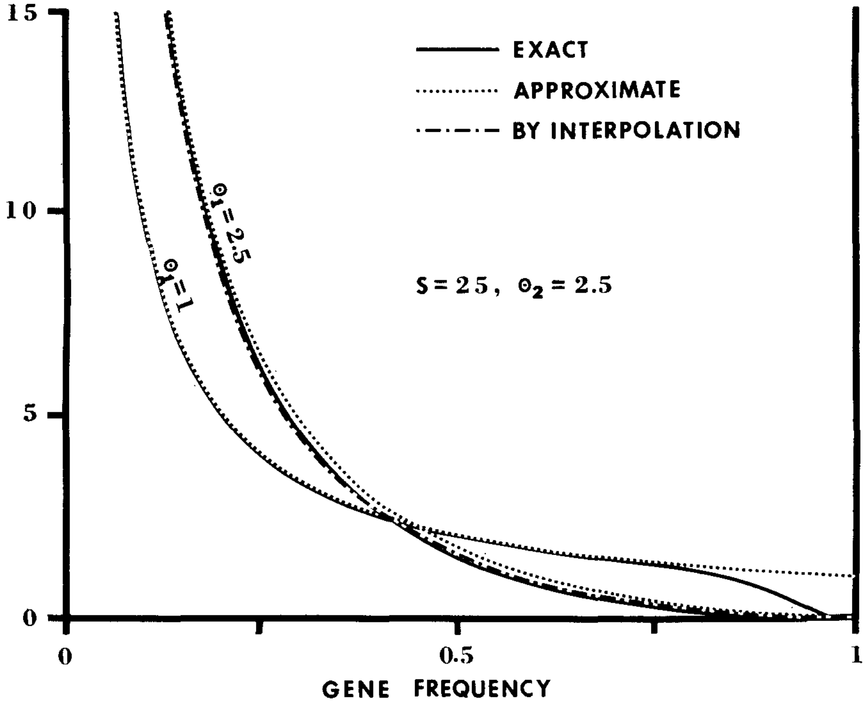

Figure 3.—Approximate and exact frequency spectra of neutral alleles for two cases of genic selection: (1) $\theta_1 = 1$, $\theta_2 = 2.5$, $S = 25$ and (2) $\theta_1 = 2.5$, $\theta_2 = 2.5$, $S = 25$. The ordinate denotes $\Phi_1(x)$, which has the meaning that $\Phi_1(x)dx$ represents the expected number of neutral alleles whose frequency is between x and $x + dx$. ——— Exact frequency spectrum. ······ Approximate frequency spectrum computed by using formula (10). —·—·—· Extrapolation by using formulas (25) and (29).

First, we note from Figure 2 that this part of $\Phi_1(x)$ resembles a reversed gamma distribution. The following is the theoretical basis for this resemblence. When θ_2/S and θ_1 are small, the distribution of q, the sum of the frequencies of deleterious alleles, can be approximated by the gamma distribution $\phi(q) = G(q; \theta_2, S)$ in which

$$G(y;\alpha,\beta) = \beta^{\alpha}\Gamma(\alpha)^{-1}e^{-\beta y}y^{\alpha-1} \ , \quad 0 \leq y \leq 1 \ ,$$

(Nei 1968). Since p, the sum of the frequencies of neutral alleles, is equal to $1 - q$, the distribution of p can be approximated by the reversed gamma distribution $\phi(p) = G(1-p; \theta_2, S)$. When θ_1 is very small, so that most of the time there is only one neutral allele in the population, $\Phi_1(x)$ should follow $\phi(p)$ closely except at frequencies near $x = 0$, where $\Phi_1(x)$ is very large. When θ_1 becomes larger but is still substantially smaller than one, the high-frequency part of $\Phi_1(x)$ should still resemble some reversed gamma distribution, though it will now have a lower peak and a longer tail, that is, the α and β values for this reversed gamma distribution are smaller than θ_2 and S, respectively. In

Figure 2, the high-frequency part of the curve with $S = 25$ follows roughly the reversed gamma distribution with $\alpha = 2$ and $\beta = 15$, while that of the curve with $S = 50$ follows roughly the reversed gamma distribution with $\alpha = 1.9$ and $\beta = 27$; these two distributions were obtained by trying some reasonable combinations of α and β values. Second, we note further that the peak of $\phi(q)$ should occur near $x = (\theta_2 - 1)/S$ because the peak of $G(y;\alpha,\beta)$ is at $y = (\alpha - 1)/\beta$. Therefore, the peak of $\Phi_1(x)$ should occur near $x = 1 - (\theta_2 - 1)/S$. For example, the peak of the curve with $S = 25$ occurs at $x = 0.93$, which is only slightly smaller than $1 - (\theta_2 - 1)/S = 0.94$, and that of the curve with $S = 50$ occurs at $x = 0.965$, which is again only slightly smaller than $1 - (\theta_2 - 1)/S = 0.970$.

In Table 1, we examine the accuracy of the approximate formulas for $\bar{k}_1$, $\bar{k}_2$, $\bar{J}_1$, and $\bar{J}_2$; the "exact" formulas for these quantities are given by (17), (18), (19) and (20), respectively. The parameters are specified in the table. We observe the following. (i) The accuracy of formula (12a) declines with increasing θ_2/S, but remains quite high as long as θ_2/S is not considerably larger than 0.1; if θ_1 is around one or smaller, formula (12a) is still fairly accurate even if θ_2/S is 0.2. This result indicates that if the sum of their frequencies is 0.1 or smaller, the presence of deleterious alleles causes no substantial reduction in the mean number of neutral alleles in a sample. Note that formula (12a) always gives an overestimate. Formula (12b) is less accurate than (12a), but provides a lower bound for the mean number of neutral alleles in a sample. (ii) Formula (5) is a close approximation to formula (18), and is somewhat better than formula (6), particularly when θ_2 is large. In most cases, formula (5) gives an underestimate—it becomes an overestimate only when θ_1 is very large. (iii) Formula (14a) gives a close approximation to formula (19) and is, on the whole, slightly better than formula (14b). Note that the latter always gives an overestimate while the former gives an underestimate except for the case of $\theta_1 = 0.1$ and $\theta_2 = 1$. It has been explained earlier why formula (14a) may become an overestimate when selection is weak and θ_1 is small. Interestingly, if S and θ_2 are raised to 50 and 2.5,

TABLE 1

Comparisons between the approximate and the exact formulas for $\bar{k}_1$, $\bar{k}_2$, $\bar{J}_1$ and $\bar{J}_2$[*]

			$\bar{k}_1$			$\bar{k}_2$			$\bar{J}_1$		$\bar{J}_2$	
θ_1	θ_2	(12a)	(12b)	(17)†	(5)	(6)	(18)†	(14a)	(14b)	(19)†	(7)	(20)†
0.1	1	1.57	1.54	1.57	2.42	2.37	2.45	0.820	0.824	0.818	0.0024	0.0028
0.1	2	1.57	1.52	1.56	4.83	4.66	4.99	0.730	0.737	0.732	0.0047	0.0056
0.1	4	1.57	1.46	1.55	9.67	8.99	9.83	0.582	0.592	0.573	0.0094	0.0116
1.0	1	5.88	5.67	5.83	2.38	2.33	2.40	0.451	0.463	0.453	0.0022	0.0025
1.0	2	5.88	5.45	5.77	4.75	4.58	4.80	0.405	0.426	0.408	0.0043	0.0050
1.0	4	5.88	5.01	5.65	9.50	8.86	9.60	0.320	0.356	0.325	0.0087	0.0100
5.0	1	19.07	18.38	18.85	2.22	2.18	2.23	0.150	0.157	0.154	0.0015	0.0017
5.0	2	19.07	17.68	18.64	4.43	4.29	4.44	0.135	0.147	0.141	0.0030	0.0033
5.0	4	19.07	16.24	18.20	8.86	8.33	8.81	0.107	0.128	0.119	0.0062	0.0063

[*] $S = 20$, $2m = 200$. The numbers in parentheses refer to formula numbers.

† Exact formulas.

but θ_1 remains equal to 0.1, then formulas (14a) and (19) both give $J_1 = 0.820$; note that the value of θ_2/S is the same as that of $1/20$. (Because of computational difficulty, no larger S has been tried.) This suggests that if S is large, formula (14a) will become an underestimate even if θ_1 is small and the exact value of $\bar{J}_1$ will be somewhere between the two values given by (14a) and (14b), probably closer to the former. (iv) Formula (7) is a close approximation to formula (20) and $\bar{J}_2$ is usually negligible. Note that the S value used in Table 1 is only 20. When S is larger, the agreement between the approximate and the "exact" formulas is expected to be even better. This has been borne out by extensive computations.

RECESSIVE SELECTION

Let the fitness of A_iA_j be $1-2s$ if $i, j > I$ and one if otherwise. This means that the deleterious alleles are completely recessive. We use the same notations as above and make the same assumption that $\bar{q}$ is close to $\hat{q} = \sqrt{u_2/(2s)} = \sqrt{\theta_2/(2S)}$, the equilibrium value in an infinite population.

We again focus our attention on the frequency of a particular deleterious allele, say A_i, $i > I$. It can be easily shown that the mean change of x_i per generation is given by

$$M_i = v_1(1-x_i) - vx_i - 2sx_i(1-q)q \ ,$$

approximately. As WRIGHT (1966) did, we replace q by $\hat{q}$ and find that

$$M_i \approx v_1(1-x_i) - x_i(\sqrt{2u_2s} + u_1 - v_1) \ . \qquad (33)$$

A comparison of (33) with (1) suggests that if we substitute $\sqrt{2u_2s}$ for s (or $\sqrt{2\theta_2S}$ for S) in the approximate formulas for genic selection, we will obtain the corresponding formulas for recessive selection. This is indeed the case and we have the following:

$$\Phi_1(x) \approx \theta_1 x^{-1}(1-x)^{\theta_1-1} \ , \qquad (34)$$

$$\Phi_2(x) \approx \theta_2 x^{-1}(1-x)^{\sqrt{2\theta_2S}+\theta_1-1} \ , \qquad (35)$$

$$\bar{k}_1 \approx \frac{\theta_1}{\theta_1} + \frac{\theta_1}{\theta_1+1} + \dots + \frac{\theta_1}{\theta_1+2m-1} \ , \qquad (36a)$$

$$\bar{k}_1 \approx \frac{\theta_1(1-\bar{q})}{\theta_1(1-\bar{q})} + \frac{\theta_1(1-\bar{q})}{\theta_1(1-\bar{q})+1} + \dots + \frac{\theta_1(1-\bar{q})}{\theta_1(1-\bar{q})+2m-1} \ , \qquad (36b)$$

$$\bar{k}_2 \approx \theta_2 \log_e[1 + 2m/(\sqrt{2\theta_2S} + \theta_1 - 0.5)] \ , \qquad (37)$$

$$\bar{J}_1 \approx (1-\bar{q})^2/(1+\theta_1) \ , \qquad (38a)$$

$$\bar{J}_1 \approx (1-\bar{q})^2/[1+\theta_1(1-\bar{q})] \ , \qquad (38b)$$

$$\bar{J}_2 \approx \theta_2/[(\sqrt{2\theta_2S} + \theta_1)(\sqrt{2\theta_2S} + \theta_1 + 1)] \ . \qquad (39)$$

In formulas (36b) and (38a,b), we have not replaced $\bar{q}$ by $\hat{q}$ because, when θ_2 is of order 1 or smaller, $\bar{q}$ is substantially smaller than $\hat{q}$ and the formula

$\bar{q} = \Gamma[\theta_2 + 1)/2]/[\sqrt{S}\,\Gamma(\theta_2/2)]$ should be used (NEI 1968). Numerical computations show that, as in the case of genic selection, formula (35) gives a close approximation to the frequency spectrum of deleterious alleles, and in what follows we shall be concerned only with the accuracy of formulas other than (35) and (37). The corresponding "exact" formulas have been given by LI (1977, 1978).

Figure 4 shows the approximate (34) and the exact frequency spectrum of neutral alleles for three cases: $\theta_1 = 0.1$, 0.36, and 1.5; in all cases, $S = 30$ and $\theta_2 = 0.75$. It is seen that in the first two cases the approximate frequency spectrum agrees well with the exact one at low and intermediate allele frequencies but deviates far from it at high frequencies, while in the third case there occur no large discrepancies. As in the case of genic selection, the approximate and the exact frequency spectrum give similar values for the mean number of neutral alleles whose frequency is higher than 0.01; for the above three cases, the former gives 1.50, 2.70, and 6.86, and the latter gives 1.47, 2.44, and 6.73. Consequently, formula (36a) holds fairly well; for the same three cases as above, it gives 1.73, 3.52, and 10.31 if $2m = 1000$, while the exact values are 1.72, 3.49, and 10.18.

Table 2 shows some comparisons between the approximate and the exact

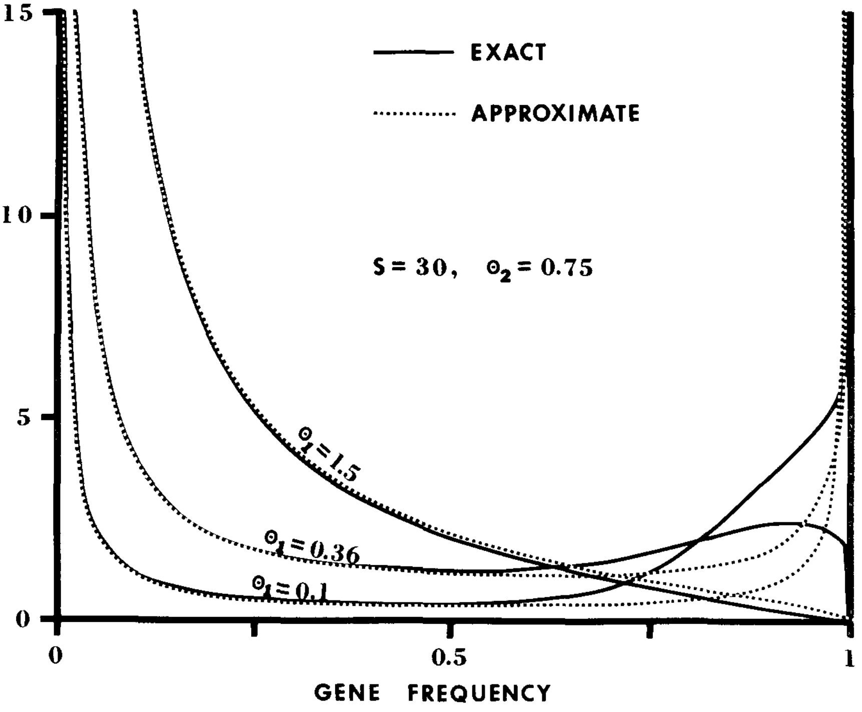

FIGURE 4.—Approximate and exact frequency spectra of neutral alleles for three cases of recessive selection: $\theta_1 = 0.1$, 0.36, 1.5; in all cases, $\theta_2 = 0.75$ and $S = 30$. The ordinate denotes $\Phi_1(x)$, which has the meaning that $\Phi_1(x)dx$ represents the expected number of neutral alleles whose frequency is between x and $x + dx$.——— Exact frequency spectrum. ······ Approximate frequency spectrum.

formulas for J_1 and J_2. It is seen that formula (38a) gives a quite accurate value for $\bar{J}_1$. The S value used is only 30 and formula (38a) tends to give an overestimate. If S becomes large, this formula will probably become an underestimate; unfortunately this is difficult to verify, because of computational difficulties. Formula (38b) is less accurate than formula (38a), but provides an upper bound for $\bar{J}_I$. Formula (39) tends to give an overestimate for $\bar{J}_2$, particularly when θ_2 is small. Fortunately, when S becomes larger than 50, $\bar{J}_2$ becomes less than 0.01 because it can be shown that $\bar{J}_2 < 1/(2S)$, and can therefore be neglected.

ESTIMATING θ FOR TESTING THE NEUTRAL MUTATION HYPOTHESIS

It has been controversial as to how one should estimate θ when testing the neutral mutation hypothesis. On the one hand, NEI (NEI and ROYCHOUDHURY 1974; NEI 1975), OHTA (1976) and some others advocate using the estimator given by

$$\theta_h = h/(1-h) \ , \tag{40}$$

in which h is the observed mean heterozygosity of the population for a large number of loci. On the other hand, EWENS (1972) and WATTERSON (1978b,c) contend that it is better to use the estimator defined through

$$k = \frac{\hat{\theta}_k}{\hat{\theta}_k} + \frac{\hat{\theta}_k}{\hat{\theta}_k + 1} + \ldots + \frac{\hat{\theta}_k}{\hat{\theta}_k + 2m - 1} \ , \tag{41}$$

in which k is the total number of alleles observed in a sample for a single locus. In my opinion, this controversy arises because of the failure to distinguish between the hypothesis of pan-neutrality (H_p), which postulates that all alleles in a population are neutral, and the hypothesis of neutral mutations (H_n), which postulates that the genic variation or the heterozygosity of a population is mainly due to neutral or almost neutral mutations. These two hypotheses may appear similar superficially, but are actually very different. In fact, H_p is much more stringent than H_n. The following example will make this point clear. Suppose

TABLE 2

*Comparisons between the approximate and the exact formulas for $\bar{J}_1$ and $\bar{J}_2$**

		$\bar{J}_1$			$\bar{J}_2$	
θ_1	θ_2	(38a)	(38b)	Exact	(39)	Exact
0.10	0.75	0.763	0.768	0.758	0.0141	0.0081
0.10	3.00	0.573	0.584	0.567	0.0152	0.0136
0.36	0.75	0.617	0.631	0.616	0.0132	0.0078
0.36	3.00	0.464	0.490	0.462	0.0147	0.0133
1.50	0.75	0.335	0.353	0.340	0.0099	0.0067
1.50	3.00	0.252	0.288	0.258	0.0126	0.0119

* $S = 30$.

that at a certain locus there are six alleles with the following frequencies: 0.60, 0.36, 0.01, 0.01, 0.01, 0.01. Then, for H_n to hold, it requires only that the first two alleles be neutral because the heterozygosity is mainly due to these two alleles. But, for H_p to hold, it requires that all six alleles be neutral; if any one of them is not neutral (say, one of the four low-frequency alleles is deleterious), then H_p is not true. In other words, H_n can be true even if the majority of the alleles are deleterious, but H_p is true only if every allele is neutral. This difference is of vital importance because the majority of mutations are deleterious and every natural population contains many deleterious genes. We note that what Kimura (1968a,b) proposed is not H_p but H_n. In Nei's and Ohta's approaches to testing H_n, one needs first to estimate θ. The θ value to be estimated is not θ_T, the total $4Nv$ value, but θ_1, the "neutral only" value, because H_n postulates that only neutral mutations are important to the polymorphism of a population. However, since H_n does not specify precisely what proportion of the mean heterozygosity is due to (almost) neutral mutations, only rough estimates of θ_1 can be obtained under this null hypothesis. In the method advocated by Nei (1975) and Ohta (1976), it is assumed that h is completely due to neutral mutations, ignoring the possibility that some (or many) of the rare alleles in the sample may be deleterious. This method has been criticized by Ewens (1972; personal communication) and Watterson (1978b,c), who argue that $\hat{\theta}_k$ is superior to $\hat{\theta}_h$ because k is a sufficient statistic for θ if all mutations (alleles) are neutral. This argument is valid if what is to be tested is H_p or what is to be estimated is θ_T. Here, however, the purpose of estimating θ is to test H_n, that is, what is to be estimated is θ_1. Since both estimating procedures ignore the possibility that some of the rare alleles may be deleterious, we need to examine the effect of deleterious mutations on $\bar{\theta}_h$ and $\bar{\theta}_k$, when comparing them as estimators of θ_1 under the null hypothsis of H_n. In the following, I shall use a numerical example to illustrate that the first approach is quite robust against the existence of rare deleterious alleles, whereas the second approach is not. I shall also discuss the controversial question of whether it is better to pool data from all loci studied or to treat each locus separately, the former approach being advocated by Nei (1975) and Ohta (1976) and the latter by Watterson (1978b,c).

Effect of deleterious mutations: Assume that the genome consists of a large number of identical loci; the problem of inhomogeneity among loci will be discussed later. Assume further that the parameter values for each locus are $\theta_1 = 1$, $\theta_2 = 10$, and $S = 500$, the mode of selection being genic. Using formulas (23) and (24), we find that the mean heterozygosity of the population is $\bar{H} = 0.52$. If the population were free of deleterious mutations, *i.e.*, $\theta_2 = 0$, the mean heterozygosity would decrease only slightly to $\bar{H} = \theta_1/(1 + \theta_1) = 0.5$. Thus, it is true that the mean heterozygosity is mainly due to neutral mutations or, in other words, the condition postulated by H_n is true. However, the condition postulated by H_p is far from being true because the population contains many deleterious alleles. If we follow the first approach and use gene frequency data from a large number of loci to compute h, the observed mean heterozygosity, the value obtained is expected to be close to 0.52, as long as the number of genes sampled for each locus is reasonably large (NEI and ROYCHOUDHURY 1974). Using $h = 0.52$ and equation (40), we find $\hat{\theta}_h = 1.08$, which is reasonably close to $\theta_1 = 1$. Next, let us consider the second approach. This approach uses single-locus data but, for ease of comparison, let us neglect the sampling error and assume $k = \bar{k}$. If $2m = 200$, then $\bar{k}_1 \approx 5.88$, $\bar{k}_2 \approx 3.36$ and $\bar{k} \approx 9.24$. Using $k \approx 9.24$ and equation (41), we find $\hat{\theta}_k \approx 1.78$, which is considerably larger than $\theta_1 = 1$. Thus, $\hat{\theta}_k$ is quite sensitive to the existence of rare deleterious alleles, but $\hat{\theta}_h$ is rather robust against such alleles.

Now see what we will get if we use these two estimated values to make predictions. First, consider the mean heterozygosity. If we assume that $\theta = \hat{\theta}_h = 1.08$ and all mutations are neutral, we will predict a heterozygosity of $1.08/(1 + 1.08) = 0.52$. As expected, this is equal to the mean heterozygosity. On the other hand, if we assume that $\theta = \hat{\theta}_k = 1.78$ and all mutations are neutral, we will predict a heterozygosity of 0.64, with a standard deviation equal to 0.16 (STEWART 1976). The difference between this predicted heterozygosity and the mean heterozygosity is 0.12, which is comparable to the standard deviation. This result suggests that if the discrepancy between the observed mean heterozygosity among loci and the value predicted by using $\theta = \hat{\theta}_k$ is used to test H_n, there will be a high probability of rejecting H_n when H_n is true, *i.e.*, the type I error will be high. Next, consider the frequency spectrum. We again assume that all mutations are neutral and $\theta = \hat{\theta}_h$ for the first approach, while $\theta = \hat{\theta}_k$ for the second approach. The frequency spectrum obtained by the first approach is represented by the curve with $\theta = 1.08$ in Figure 5, while that obtained by the second approach is represented by the curve with $\theta = 1.78$. It is seen that the former is very close to the solid line, the actual frequency spectrum, but the latter deviates far from it. Again we see that the type I error is high for the second approach. It is interesting to note that in the present example $\hat{\theta}_k$ is much smaller than $\theta_T = 11$, and the spectrum with $\theta = 1.78$ is very different from that with $\theta = 11$. If our purpose is to test H_p, we should try to get a more accurate estimate of θ_T so that the spectrum obtained will be closer to that with $\theta = 11$. For this purpose, it is better to use NEI's (1977) approach of estimating θ_T through the number of rare alleles than to use (41).

One might argue that in practice deleterious alleles may be usually too rare to appear in a sample, so that the second approach can also be applied to estimate θ when testing H_n. Let us see if this is true. In a recent review article, SIMMONS and CROW (1977) gave the following estimates for Drosophila populations: the rate of mutations causing mild deleterious effects is 6×10^{-5} per locus per generation, and these mutations reduce the fitness of heterozygotes on the average by about 0.02, the former being a minimum estimate. Using these two estimates and formula (5), we obtain $\bar{k}_2 \simeq 0.54$ if $2m = 200$, $\theta_1 \leq 2$ and $N = 10^4$; $\bar{k}_2$ increases slowly with increasing N. Note that this is a minimum estimate. Note also that, in addition to mildly deleterious mutations, there should be mutations that reduce the fitness of heterozygotes by a value, say, between 0.01 and 0.001, and there should be severely deleterious mutations such as recessive lethals. If

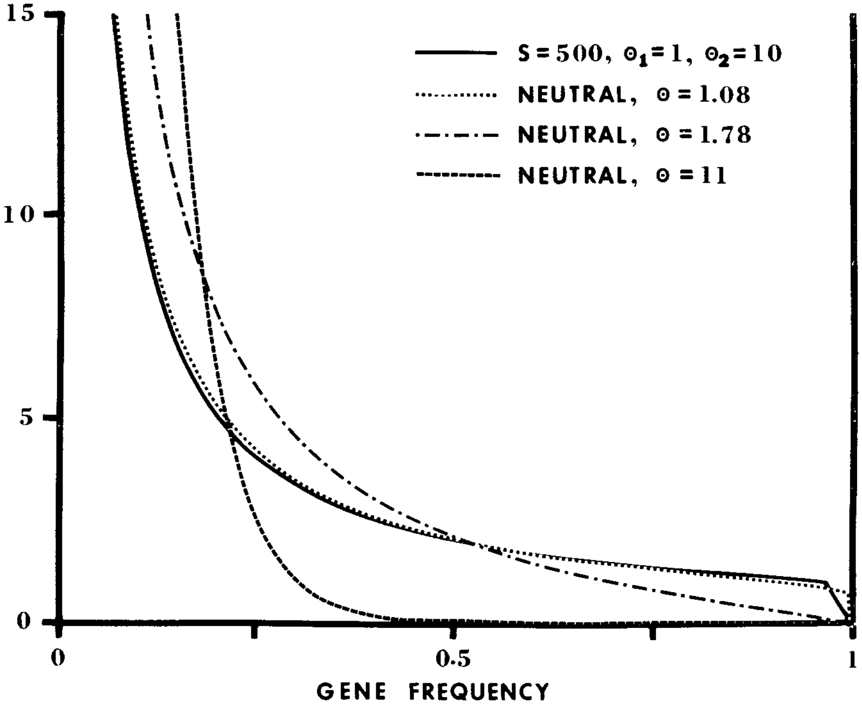

FIGURE 5.—Frequency spectra under various situations. The ordinate denotes $\Phi(x)$, which has the meaning that $\Phi(x)dx$ represents the expected number of alleles whose frequency is between x and $x + dx$. ——— Genic selection with $S = 500$, $\theta_1 = 1$, $\theta_2 = 10$. ······ Neutral mutations.

all these possibilities are taken into consideration, k_2 may easily become 1 or larger. Let us assume that this is true. Let us assume further that $\theta_1 \leq 0.2$. Then, in a sample of 200 genes, the expected number of different neutral alleles is smaller than 2.12 and, on the average, about one third of the total number of alleles in the sample is due to the presence of deleterious mutants. The assumption of $\theta_1 \leq 0.2$ is quite reasonable because, if $\theta_1 = 0.2$, the mean heterozygosity is greater than 0.16, which is roughly equal to or larger than the majority of mean heterozygosities observed in Drosophila species (*cf.*, AYALA *et al.* 1974 and data cited in NEI 1975). Although this assessment of the effects of deleterious mutations on k is admittedly rough and is based on data from Drosophila only, it suggests that the second approach of estimating θ can be very unfavorable for the testing of H_n.

Single-locus approach vs *pooling data*: A single-locus approach generally has the following three drawbacks: (i) The estimator of a parameter often has a large mean square error. This is true for $\hat{\theta}_k$ because k has a large variance (EWENS 1972). For example, if all mutations are neutral, then $\bar{k} = 1.57$ and $V(k) = 0.56$ for $\theta = 0.1$, and $k = 5.88$ and $V(k) = 4.24$ for $\theta = 1$, assuming $2m = 200$. (ii) The power of a single-locus test is generally low. This is true even if the hypothesis to be tested is H_p, not to mention the less stringent hypothesis H_n. As an example, let us consider WATTERSON's (1978a,c) homozygosity test. This seems to be the best single-locus test that has been devised for testing H_p. Yet, his results show that it is almost impossible to reject H_p if the sample contains fewer than three alleles (see Table 1 of WATTERSON 1978a). (iii) For testing H_n, we will not be able to use data from monomorphic loci, loci with low heterozygosity, because such loci contribute little to the observed average heterozygosity and therefore whether the alleles at such a locus are neutral or not is an irrelevant question under the null hypothesis of H_n. We note that the proportion of monomorphic loci is high in most of the species surveyed (*cf.*, FUERST, CHAKRABORTY and NEI 1977). Thus, a large amount of information will be wasted if we use a single-locus test. These three drawbacks can be overcome by pooling data from all loci studied. Of course, pooling data creates the problem of inhomogeneity because mutation rate varies from locus to locus and the mode and intensity of selection may also vary. However, the variation in mutation rate among loci may be taken into account by using a model of varying mutation rate as developed by NEI and his associates (*cf.*, NEI, CHAKRABORTY and FUERST 1976; FUERST, CHAKRABORTY and NEI 1977), though the problem of variation in the mode and intensity of selection among loci cannot be easily handled. At any rate, if the inhomogeneity among loci is not taken into consideration in a test, the type I error may become large.

In addition to the above, some other arguments have also been put forward by both sides. I discuss two of them here. Both are arguments against the first approach, namely, that $\hat{\theta}_h$ has a large mean square error and a large bias. These two criticisms can easily be refuted because they are based on single-locus estimation, but what NEI (1975) and OHTA (1976) propose is to estimate the average θ_1 value among loci by using data from a large number of loci. Obviously, if data from a large number of loci are used, the mean square error of $\hat{\theta}_h$ will become very small. This is also true for the bias for estimating the average θ_1 over the loci studied because it can be shown that the bias decreases quite rapidly as the number of loci used increases. Incidentally, even for a single-locus estimate, the bias of $\hat{\theta}_h$ may be a less serious problem than the sensitivity of $\hat{\theta}_k$ to the effect of deleterious mutations. For instance, for the example given in Figure 5, adding a bias of 40% to θ_h increases its mean value from 1.08 to 1.51, which is still considerably better than the estimate $\hat{\theta}_k = 1.78$. A bias of 40% is used in the above computation because EWENS (personal communication) argues that the simulation result of EWENS and GILLESPIE (1974) shows that for θ_T of order one the bias of $\hat{\theta}_h$ for a single-locus estimate is consistently 40% or more upwards, assuming that all mutations are neutral.

Several further remarks are in order. First, in the above I have compared $\hat{\theta}_h$ and $\hat{\theta}_k$ as two estimators of θ_1 under the null hypothesis of H_n and have argued that $\hat{\theta}_h$ is superior to $\hat{\theta}_k$. This is not to suggest that $\hat{\theta}_h$ is the best estimator that will ever be found, because it has some drawbacks, as mentioned above. In particular, if the number of loci used is not large, the estimate obtained by this method may not be very accurate. A more accurate approach is to use the number of alleles whose sample frequency is, say, higher than 0.01. This approach is, however, more complicated than that using h. In a future study, I shall examine whether the estimate obtained by the latter approach is close to that obtained by the former. If the answer is yes, then the latter is to be preferred to the former since it is much simpler. Second, although I have used a single-locus example to illustrate the difference between H_n and H_p, it should be borne in mind that H_n is postulated as a majority rule for the loci of a genome, and whether it is true or not can be decided only by studying a large number of loci. Third, it may be possible to find a statistic for testing H_n whose distribution is free of the nuisance parameter θ, but so far no such test statistics have been found. We note that although the testing procedures of EWENS (1972) and WATTERSON (1978a,b) do not require the estimation of θ, they are developed under the null hypothesis of H_p, not H_n. Fourth, since the neutral mutation hypothesis was proposed, terms such as the neutral alleles hypothesis (or theory), the neutrality hypothesis, etc., have appeared in the literature. These terms have been used sometimes as equivalents for the neutral mutation hypothesis and sometimes as equivalents for the hypothesis of pan-neutrality. This is very confusing. I suggest that a clear distinction always be made between the two hypotheses. Fifth, some authors (*e.g.*, WATTERSON 1978a,c) seem to think that "detecting selection among alleles" is equivalent to "testing the neutral mutation hypothesis," but it is not. To test the neutral mutation hypothesis is to detect selection among polymorphic alleles only, *i.e.*, among alleles whose frequencies are higher than 0.01, say. Detecting selection among alleles is equivalent to testing the hypothesis of pan-neutrality. Sixth, whether an allele is almost neutral or not is judged by whether random drift or selection plays a more important role in the population dynamics of this allele. KIMURA (1968a) proposed the definition $|2Ns| \ll 1$ but I (LI 1978) suggested that a better definition would be $|Ns| < 1$.

DISCUSSION

In obtaining the present theoretical results, it has been assumed that all deleterious mutations have the same selective disadvantage. This assumption may seem restrictive, but from the results we may draw the following general conclusions for the effects of deleterious mutations on the frequency spectrum, the mean number of alleles in a sample and the mean homozygosity:

First, if the expected value ($\bar{q}$) of the sum of the frequencies of deleterious alleles is around 10% or less, then the presence of such alleles causes no substantial reduction in the mean number of *neutral* alleles in a sample. Furthermore, the low and intermediate allele frequency parts of the frequency spectrum of neutral alleles are little affected by the presence of deleterious allels, though the high allele frequency part may be changed drastically.

Second, it can be shown that if S is much larger than $2m$, then formula (5), the formula for genic selection, reduces to

$$\begin{aligned}\bar{k}_2 &\simeq 2m\theta_2/S \\ &\simeq 2m\bar{q}\ . \end{aligned} \tag{42}$$

The same is true for formula (37), the formula for recessive selection, provided that $\sqrt{2\theta_2 S}$ is much larger than $2m$. Thus, a general principle emerges, namely, that the mean number of *severely deleterious* alleles in a sample is equal to the product of the sample size ($2m$) and the sum of the expected frequencies of such alleles ($\bar{q}$). (Here, by severely deleterious alleles, we mean alleles with large S.) We note that if $2m = 200$ and $\bar{q} = 0.01$, then formula (42) gives $\bar{k}_2 \simeq 2$, a large value. Thus, the contribution of deleterious mutations to $\bar{k}$ can be quite large even if $\bar{q}$ is only 1 or 2%.

Third, the presence of deleterious alleles reduces the mean homozygosity of a population from $1/(1+\theta_1)$ to about

$$\begin{aligned} \bar{J} \simeq \bar{J}_1 &\simeq (1-\bar{q})^2/(1+\theta_1) \\ &\simeq (1-2\bar{q})/(1+\theta_1) \ . \end{aligned} \tag{43}$$

Thus, the contribution of deleterious mutations to the mean heterozygosity of a population is roughly equal to $2\bar{q}/(1+\theta_1)$. An upper bound of the mean homozygosity is given by

$$\bar{J} = (1-\bar{q})^2/[1+(1-\bar{q})\theta_1] \ . \tag{44}$$

Of course, the second and third principles do not apply to the case where one or more of the deleterious alleles enjoy heterozygote advantage. These three principles are mainly for assessing the effects of mutations whose selective disadvantage is considerably large. The effects of mutations with very mild deleterious effect, say $S \leq 30$, can be studied by my earlier formulas (LI 1977, 1978).

I am indebted to W. J. EWENS for showing me his unpublished results and valuable comments. I also thank M. NEI, J. F. CROW and M. J. SIMMONS for discussions and suggestions. This study is supported by National Science Foundation grant DEB 77-09120 and Public Health Service grants GM 20293 and GM 19513.

LITERATURE CITED

AYALA, F. J., M. L. TRACEY, L. G. BARR, J. F. MCDONALD, and S. PEREZ-SALAS, 1974 Genetic variation in natural populations of five Drosophila species and the hypothesis of selective neutrality of protein polymorphisms. Genetics **77:** 343–384.

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

EWENS, W. J. and J. H. GILLESPIE, 1974 Some simulation results for the neutral allele model, with interpretations. Theoret. Pop. Biol. **6:** 35–57.

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., 1968a Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genet. Res. **11:** 247–269. ——, 1968b Evolutionary rate at the molecular level. Nature **217:** 624–626.

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

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. ——, 1978 Maintenance of genetic variability under the joint effect of mutation, selection, and random drift. Genetics **90**: 349–382.

NEI, M., 1968 The frequency distribution of lethal chromosomes in finite populations. Proc. Natl. Acad. Sci. U.S. **60**: 517–524. ——, 1975 *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam. ——, 1977 Estimation of mutation rate from rare protein variants. Am. J. Human Genet. **29**: 225–232.

NEI, M., R. CHAKRABORTY and P. A. FUERST, 1976 Infinite allele model with varying mutation rate. Proc. Natl. Acad. Sci. U.S. **73**: 4164–4168.

NEI, M. and A. K. ROYCHOUDHURY, 1974 Sampling variances of heterozygosity and genetic distances. Genetics **76**: 379–390.

OHTA, T., 1976 Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theoret. Pop. Biol. **10**: 254–275.

SIMMONS, M. J. and J. F. CROW, 1977 Mutations affecting fitness in Drosophila populations. Ann Rev. Genet. **11**: 49–78.

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

WATTERSON, G. A., 1978a The homozygosity test of neutrality. Genetics **88**: 405–417. ——, 1978b An analysis of multi-allelic data. Genetics **88**: 171–179. ——, 1978c The neutral alleles model, and some alternatives. Int. Stat. Inst. Conf. Proc., in press.

WRIGHT, S., 1949a Adaptation and selection. pp. 365–389. In: *Genetics, Paleontology and Evolution*. Edited by G. L. JEPSON, G. G. SIMPSON and E. MAYR. Princeton Univ. Press, Princeton, New Jersey. ——, 1949b Genetics of populations. Encyclopaedia Britannica, 14th ed. **10**: 111–112. ——, 1966 Polyallelic random drift in relation to evolution. Proc. Natl. Acad. Sci. U.S. **55**: 1074–1081.

Corresponding editor: B. S. WEIR