THE GENETIC STRUCTURE OF NATURAL POPULATIONS

by 45

RALPH DENNIS COOK

B. S., Northern Montana College, 1967

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics and Computer Science

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1969

Approved by:

Raja Nassar
Major Professor

TABLE OF CONTENTS

1. INTRODUCTION

The study of the genetic structure of natural populations is basic to the understanding of the mechanism of evolution. The investigation of this structure may lead to a better understanding of the processes by which a species evolves. Work in the field of population structure began shortly after the development of Mendelian heredity in 1900. The investigations of Fisher (1922) and Haldane (1924) were of major importance in laying the groundwork in this field. Both Fisher and Haldane restricted their treatments to simple situations. Haldane regarded the change in gene frequencies to be deterministic. This treatment is valid only if a population is infinitely large and is in an environment which remains constant or changes in deterministic ways. Fisher was the first to introduce the method of partial differential equations as a technique for studying the gene frequency distribution in a natural population. Fisher's method is equivalent to the method intrduced by Wright (1945) which makes use of the Fokker-Planck equation.

In this paper population structure will be dealt with under three different models. The first two models, the "island" and "continuum", to be presented were developed by Wright. The "stepping stone model" was developed by Kimura. Wright's island model will be treated extensively since it is believed to present the most favorable conditions for evolutionary advancement.

The probability distribution of gene frequency will be derived in two ways. The first, developed by Wright, is based on moments about the mean and will be presented in detail. The second, largely established by Kimura, is based on a consideration of the change in gene frequency as a stochastic process and will be presented in a somewhat less detailed manner.

While the contributions of other investigators are by no means insignificant, Wright and Kimura present the most extensive and complete coverage of this field. Hence, this paper will be primarily concerned with the investigations of these two men, and for the most part, restricted to the case of a single locus with two alleles.

2. GENERAL DISTRIBUTION FORM

There are two processes that cause changes in gene frequencies. The first is a directed process arising from systematic pressures (mutation, migration and selection); the second, a random process arising from random sampling of gametes in reproduction and random fluctuations in systematic pressures (Wright 1951). The systematic pressures lead to direct changes in gene frequency and the random processes to a gene frequency change indeterminate in direction but determinate in variance.

The random processes alone or coupled with directed processes operate to determine a probability distribution of gene frequencies. Such a distribution is stationary if the opposing forces of systematic pressures and random deviations are in balance.

The stationary distribution of a gene frequency may be viewed in several mathematically equivalent ways. For the purposes of the following discussions the distribution will be viewed as follows. Consider a panmictic population which consists of a large number of isolated or partially isolated groups. Assume that each group is the same size and subject to the same systematic pressures. Then the expected distribution of a particular gene at a given time among all groups is the distribution under consideration. This is commonly referred to as the "island" model of population structure.

The derivation of the general distribution form is based on the moments about the mean of the distribution. If a stationary distribution has been reached due to a balance between systematic pressures (Δq) and random deviations (δq) then the moments about the mean in one generation must equal those in the next generation. Let $f(q)$ and $f(\delta q)$ be the class frequencies of q and δq respectively. Thus, $\Sigma f(q)=1$ and $\Sigma f(\delta q)=1$. Assume Δq and δq are not correlated and that the mean of the random deviations is zero. Since the range of q is $0<q<1$ the range of δq must be $-q<\delta q<1-q$.

Now, the nth moment about the mean in one generation will be given by

$$\Sigma(q-\bar{q})^n f(q).$$

In the next generation the class frequencies of q will change by the amount $(\Delta q+\delta q)$ so that the nth moment in the next generation will be given by

$$\Sigma\Sigma\{(q-\bar{q}+\Delta q+\delta q)^n f(q)f(\delta q)\}.$$

If equilibrium has been reached the following relation will hold.

$$(2.1) \qquad \Sigma\Sigma\{(q-\bar{q}+\Delta q+\delta q)^n f(q)f(\delta q)\}=\Sigma(q-\bar{q})^n f(q).$$

Expanding the left-hand member and simplifying (2.1) becomes

$$(2.2) \qquad \Sigma\{(q-\bar{q})^{n-1}\Delta q f(q)\}+\frac{n+1}{2}\Sigma\{(q-\bar{q})^{n-2}\sigma^2_{\delta q}f(q)\}=0$$

Terms involving $(\Delta q)^2$, $(\delta q)^2$, $(\Delta q)(\delta q)^2$ and their higher powers were neglected in (2.2). Consequently, it is assumed that the systematic pressures and random deviations between generations are small, so that the loss from the neglected terms will be small.

As indicated above, the distribution of gene frequencies is discrete. However, in a group of N diploid individuals q can assume only the values 0, 1/2N, ..., (2N-1)/2N, 1, so that when N is large q may be approximated by a continuous random variable. If we let $\Phi(q)dq=f(q)$ and $\Delta q\Phi(q)dq=dx(q)$ then (2.2) becomes

$$\int_0^1(q-\overline{q})^n dx(q)+(n+1)/2\int_0^1(q-\overline{q})^{n+2}\sigma_{\delta q}^2\Phi(q)dq=0. \tag{2.3}$$

Integrating the first term by parts and letting n=1 we obtain x(0)=x(1). By definition the expression x(q) does not involve n, so that x(0)=x(1) for all values of n. Using this relation, in general, we have

$$\int_0^1(q-\overline{q})^{n-1}dx(q)=\{x(q)(q-\overline{q})^{n+1}\}_0^1-(n-1)\int_0^1 x(q)(q-\overline{q})^{n-2}dq$$

$$=(n-1)x(1)\int_0^1(q-\overline{q})^{n-2}dq-(n-1)\int_0^1 x(q)(q-\overline{q})^{n-2}dq.$$

Substituting the above expression in (2.3) and simplifying we obtain

$$\int_0^1(q-\overline{q})^{n-2}\{x(1)-x(q)+\tfrac{1}{2}\sigma_{\delta q}^2\Phi(q)\}dq=0.$$

This expression must be true whatever the value of n. This implies

$$x(q)-x(1)=\tfrac{1}{2}\sigma_{\delta q}^2\Phi(q). \tag{2.4}$$

Taking the differential of the logarithm of the left-hand member and substituting $\Delta q\Phi(q)dq$ for dx(q) (2.4) becomes

$$d\ \mathrm{Log}\{x(q)-x(1)\}=2\Delta q dq/\sigma_{\delta q}^2.$$

Integrating, we have

$$\mathrm{Log}\{x(q)-x(1)\}=2\int\Delta q dq/\sigma_{\delta q}^2+C.$$

Therefore,

$$x(q)-x(1)=C\ \exp\{2\int\Delta q/\sigma_{\delta q}^2 dq\}. \tag{2.5}$$

Equating (2.5) and (2.4) we have

(2.6) $$\Phi(q)=C_1/\sigma_{\delta q}^2 \exp\{2\int(\Delta q/\sigma_{\delta q}^2)dq\},$$

where C_1 is a constant such that

$$\int_0^1\Phi(q)dq=1.$$

Expression (2.6) gives not only the distribution of gene frequencies among groups at a given time but also gives the distribution of gene frequencies for one group over a long period of time.

Now consider the variance of the random deviations ($\sigma_{\delta q}^2$). A group of N monoecious diploid individuals may be regarded as the result of drawing 2N gametes at random from the preceeding generation. When considering only one locus with two alleles, the probability that any group should take on a particular value, $q_j=j/2N$, in the next generation is given by

$$\binom{2N}{j}p^{2N-j}q^j.$$

Let $\bar{q}$ be the mean of the entire population. Thus, $\delta q=q_j-\bar{q}$ and $E(\delta q)=0$. The variance of δq will be

$$\sigma_{\delta q}^2=q(1-q)2N_e,$$

where N_e is the effective size of a group (see Appendix A). (2.6) may now be written

(2.7) $$\Phi(q)=C/q(1-q) \exp\{4N_e\int(\Delta q/q(1-q))dq\}.$$

The constant C will depend on the type of systematic pressure (Δq) and will be determined for specific cases. It should be noted here that (2.7) is not the exact distribution form because certain terms were neglected and integration was used to approximate summation. However, it should be a good approximation when N is large, Δq and δq are small, and $0<q<1$. The last assumption is necessary since the integral substitution may produce some distortions at the terminal classes ($q=0$ or 1). While (2.7) is the desired distribution function, as will be seen later, it may not be in its most useful form.

Consider a population in which the groups are completely isolated and the effects of mutation and selection are ineffective. In this case, as an approximation, we may take $\Delta q=0$. Thus (2.7) becomes

$$\Phi(q)=C/q(1-q).$$

Here C is easily determined.

$$1/C=\int_{\beta}^{\alpha}dq/q(1-q)\acute{=}2\ \mathrm{Log}(2N-1).$$

where $\beta=1/2N$ and $\alpha=(2N-1)/2N$. This distribution is U-shaped with most of the groups having values of q near zero or one. It must be pointed out here that the distribution takes the above form when Δq is very small but not exactly zero. If Δq were zero the distribution would be nearly uniform since there would be no reason for one value of q to be favored over another. In this case random deviations would dominate the situation. The eventual fate of each group would be fixation at $q=0$ or 1.

The preceeding derivation of the probability distribution of gene frequencies is the work of Sewall Wright (1942). The underlying assumption is that the population has reached an equilibrium state;

or that the moments in one generation equal the corresponding moments in the next generation. This approach leaves no indication of the form of the distribution before equilibrium has been reached.

Kimura (1955) considered the process of change in gene frequency as a chance event evolving in time, that is, as a stochastic process. He reasoned that for a natural population to assume a dominant role in evolution it should consist of a large number of individuals so that gene frequencies may be regarded as continuous random variables. Also the changes in gene frequency in such a population must be very slow. Accordingly, the process of change in gene frequency may be regarded as a continuous stochastic process. Assuming that the probability distribution at a given time t depends only on the gene frequencies at a preceeding time t_o, the change in gene frequency may be considered as being Markovian. Using the above assumption a brief sketch of Kimura's derivation will now be presented.

Consider two alleles with respective frequencies p and 1-p. Let $\Phi(p',p;t)$ be the conditional probability that the frequency of allele A is p at time t, given that the initial frequency was p' at time t=0. Also let $g(\delta p,p;\delta t,t)$ be the probability density that the gene frequency changes from p to $p+\delta p$ in the time interval $(t,t+\delta t)$. Now the process of the change in gene frequency may be represented as

$$\Phi(p',p;t+\delta t)=\int\Phi(p',p-\delta p;t)g(\delta p,p-\delta p;\delta t,t)d(\delta p).$$

This expression is a direct result of the assumption that the process is Markovian. Essentially it means that the probability that the gene frequency is p at time $t+\delta t$ is the total of all the probabilities of ways in which the gene frequency is $(p-\delta p)$ at time t and all the ways the gene frequency may increase by the amount δp in the time interval $(t,t+\delta t)$. The integral is taken over all values of δp, where δp may assume any value such that $0<p-\delta p<1$.

Expanding the right side by Taylor's expansion we have

$$\Phi g=\Phi g-\delta p\,\frac{\partial(\Phi g)}{\partial p}+\frac{(\delta p)^2}{2!}\,\frac{\partial^2(\Phi g)}{\partial p^2}-\frac{(\delta p^3)}{3!}\,\frac{\partial^3(\Phi g)}{\partial p^3}+\ldots$$

where $\Phi=\Phi(p',p;t)$ and $g=g(\delta p,p;\delta t,t)$.

Therefore, $$\Phi=\Phi\int gd(\delta p)-\frac{\partial}{\partial p}\{\Phi\int(\delta p)gd(\delta p)+$$

$$+\frac{\partial^2}{2\partial p^2}\{\Phi\int(\delta p)^2gd(\delta p)-\ldots,$$

assuming that the order in which the various operations are performed may be changed. Now, noting that $\int gd(\delta p)=1$, the above expression may be written in the form

$$\frac{\Phi(p',p;t+\delta t)-\Phi(p',p;t)}{\delta t}=-\frac{\partial}{\partial p}\{\Phi\frac{1}{\delta t}\int(\delta p)gd(\delta p)$$

$$+\tfrac{1}{2}\frac{\partial^2}{\partial p^2}\{\Phi\frac{1}{\delta t}\int(\delta p)^2gd(\delta p)\}-\ldots\ .$$

Let $$\lim_{\delta t\to o}\frac{1}{\delta t}\int(\delta p)gd(\delta p)=M(p,t)$$

$$\lim_{\delta t\to o}\frac{1}{\delta t}\int(\delta p)^2gd(\delta p)=V(p,t)$$

and assume $$\lim_{\delta t\to o}\frac{1}{\delta t}\int(\delta p)^ngd(\delta p)=0 \text{ for } n\geqslant 3.$$

Therefore, taking the limit of the above expression we have

$$\frac{\partial\Phi}{\partial t}=\tfrac{1}{2}\frac{\partial^2}{\partial p^2}(V(p,t)\Phi)-\frac{\partial}{\partial p}(M(p,t)\Phi).$$

Since all quantities which cause gene frequency to change are measured with one generation as a time unit, Kimura suggested replacing $V(p,t)$

and M(p,t) by $V_{\delta p}$ and $M_{\delta p}$ the variance and mean of change in gene frequency per generation. Thus,

$$\partial\Phi/\partial t=\tfrac{1}{2}\partial^2(V_{\delta p}\Phi)/\partial p^2-\partial(M_{\partial p}\Phi)/\partial p. \tag{2.8}$$

Kimura (1955) showed that the first and second terms on the right side of (2.8) give the rate of change in the probability distribution due to random fluctuations and systematic pressures respectively. Equation (2.8) is referred to as the Kolmogrov forward equation by mathematicians and the Fokker-Planck equation by physicists. Because (2.8) is dependent on time t it may be used to study the probability distribution of gene frequencies before a state of equilibrium has been reached. This aspect of the distribution will be expanded later for specific examples.

3. RANDOM GENETIC DRIFT

Random genetic drift is the process by which gene frequencies change from generation to generation due to random sampling of gametes in reproduction. Previously, genetic drift was treated in conjunction with systematic pressures. However, in finite populations with little or no systematic pressures random genetic drift becomes a dominating factor.

The problem of drift was first treated mathematically by R. A. Fisher using differential equations. He termed this type of change in gene frequency the Hagedoorn Effect. Fisher's general approach to the problem was adequate, however, his conclusions were incorrect. The correct solution for the case of a single locus with two alleles was first obtained by Wright with his method of path coefficients. Due to the random sampling of gametes in each generation, the ultimate fate of a gene in a population subjected to only random drift will be complete fixation or loss. The process by which a finite population reaches complete fixation is known as the "decay" of variability because the population gradually loses its capacity to change genetically. Both Fisher and Wright in their treatments assumed that a steady state of decay had been attained, but little was known about the process by which a steady state was reached. Kimura (1954) obtained a complete solution to this problem for the case of a single locus with two alleles. He (Kimura 1955a, 1955b) later extended his treatment to an exact solution for the tri-allelic case and an approximation to the multi-allelic case. A general treatment of Kimura's solution will be presented here.

Consider a random mating population of N diploid individuals and

and assume that the effects of systematic pressures are so small that they may be neglected. Furthermore, let N be small enough to allow gene frequency to change from generation to generation due to random genetic drift. Using Kimura's notation, the mean and variance of the change in gene frequency per generation for a single locus with two alleles will be given by $M_{\delta p}=0$ and $V_{\delta p}=p(1-p)/2N$, where N is the variance effective number. Substituting these into (2.8) we obtain the partial differential equation (Kimura 1954).

$$\frac{\partial\Phi}{\partial t} = \frac{1}{4N}\frac{\partial^2}{\partial p^2}\{p(1-p)\phi\},\ 0<p<1.$$

Kimura (1954) gave the solution of the above equation in the form

$$\phi(p',p;t)=\Sigma 6p'(1-p')\ \exp(-t/2N)+30p'(1-p')(1-2p')(1-2p)\ \exp(-3t/2N)+...$$

For $t>0$ the series is uniformly convergent. This can be seen by noting that the exponential term rapidly approaches zero.

Figures 3.1a and 3.1b show the change in the probability distribution of gene frequencies for p=.5 and .1 respectively. The area under the curves represents the probability that both genes still exist in the population.

It may be seen from these figures that this probability gradually decreases with time. The fixed classes (p=0 or 1) gradually "absorb" probability until the curves become horizontal; at such time a steady state of decay has been reached. This rate is given as 1/2N per generation.

Kimura (1956) obtained the exact solution for the case of a tri-allelic locus by an extension of the method used in the previous case. It was

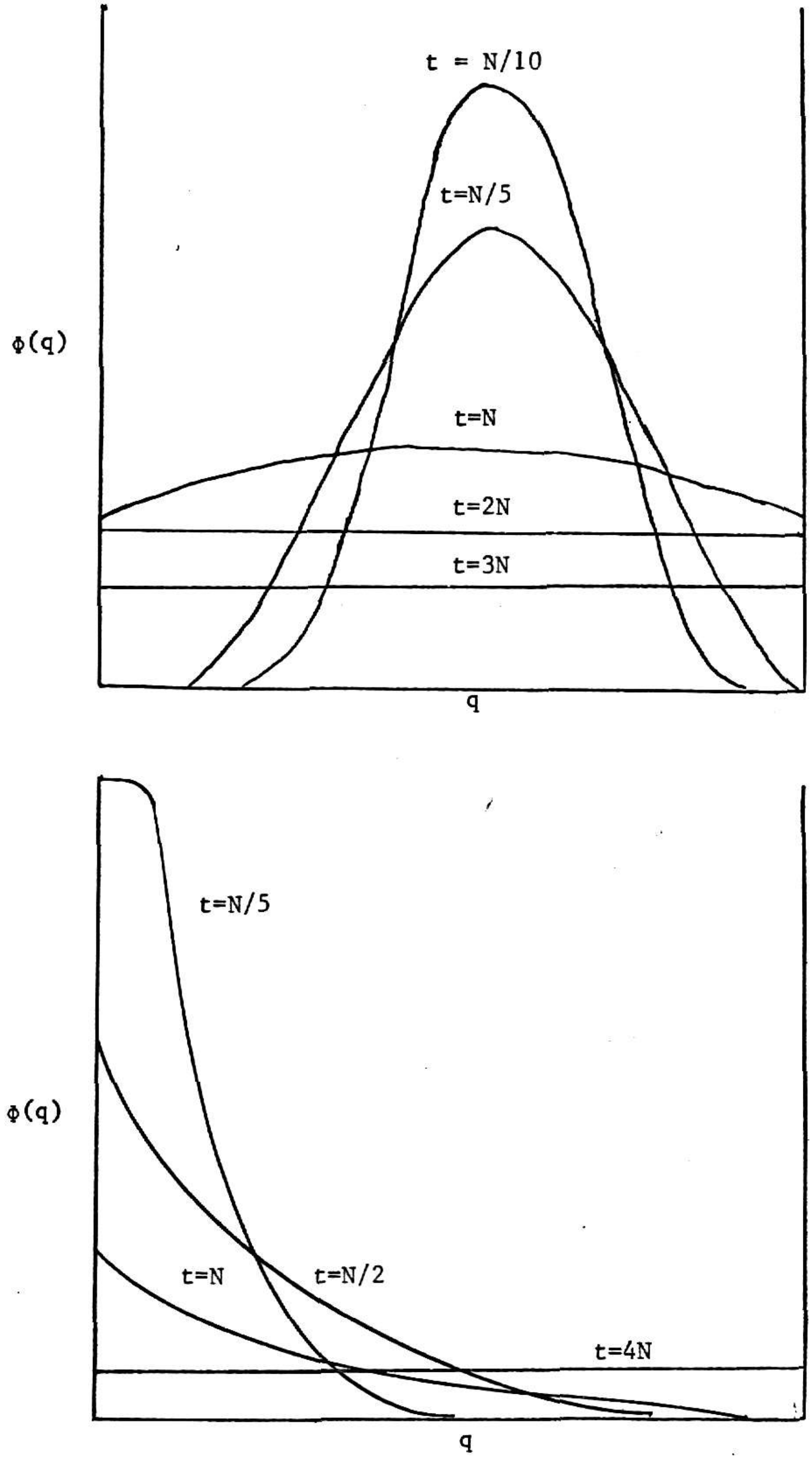


Fig. 3.1a and 3.1b. The change in the probability distribution of gene frequencies due to random genetic drift.

found that the final rate of decay is 3/2N per generation as opposed to 1/2N for a pair of alleles.

To my knowledge the exact solution for the general case of N alleles has not been derived. However, Kimura has obtained the following approximation. Consider a population which contains N alleles, A_1, A_2, ..., A_n with respective frequencies p_1, p_2, ..., p_n. The probability density that the population contains K alleles with respective frequencies x_1, ..., x_k in the tth generation is given asymptotically by

$$\Phi_{1,2\ldots k}(p_1\cdots p_k;x_1\ldots x_k;t)\sim(2k-1)!\left(\prod_{i=1}^{k}p_j\right)\exp\{-k(k-1)t/4N\},$$

where K<N. This expression depends on the assumption that the population size N is much larger than the number of alleles. The results obtained by Kimura indicate that as the number of alleles in a population increases, the rate at which alleles are eliminated from the population also increases. Thus, random drift may be effective in keeping the number of alleles in a population relatively small.

4. DISTRIBUTION UNDER MUTATION PRESSURE

Let the rate of mutation of a given gene to its allele be u and let v be the rate of reverse mutation. Assume that u and v are the same for all groups. The amount of change in the frequency of the given gene per generation is given by

$$\Delta q = up(\text{gain}) - vq(\text{loss})$$

$$= u(1-q) - v(q). \tag{4.1}$$

From the above expression it can be seen that the increase or decrease of q depends upon the relative size of the gain or loss per generation. If, in any given generation, the gain is larger than the loss the value of q will increase. However, as q increases so must its loss increase, so that eventually the amount of loss will balance the gain. When the loss does balance the gain we will have up=vq or $\Delta q=0$. Under such a condition the equilibrium points are determined;

$$\hat{q}=u/(u+v) \text{ and } \hat{p}=v/(u+v).$$

Because of the random deviations the equilibrium points will never be continually realized in all groups. One point should be repeated here; it is a balance between the systematic pressures and random deviations that will lead to a stationary distribution of a gene frequency.

Assuming that migration and selection are ineffective, the distribution function of q under recurrent mutation pressure is obtained by substituting (4.1) in the general distribution function (2.7). Thus,

$$\Phi(q)=C/q(1-q)\ \exp\{4N\int(u(1-q)-v(q))/q(1-q)dq\}$$

$$=C/q(1-q)\ \exp\{4Nu\ \mathrm{Log}\ q+4Nv\ \mathrm{Log}\ (1-q)\}.$$

Put 4Nu=U and 4Nv=V, then $\Phi(q)$ becomes

$$\Phi(q)=Cq^{U-1}(1-q)^{V-1},$$

which is in the form of a simple Beta distribution. Hence, the constant C is determined.

$$C=\Gamma(U+V)/\Gamma(U)\Gamma(V).$$

Because of the available form of the distribution the mean and variance are easily found to be;

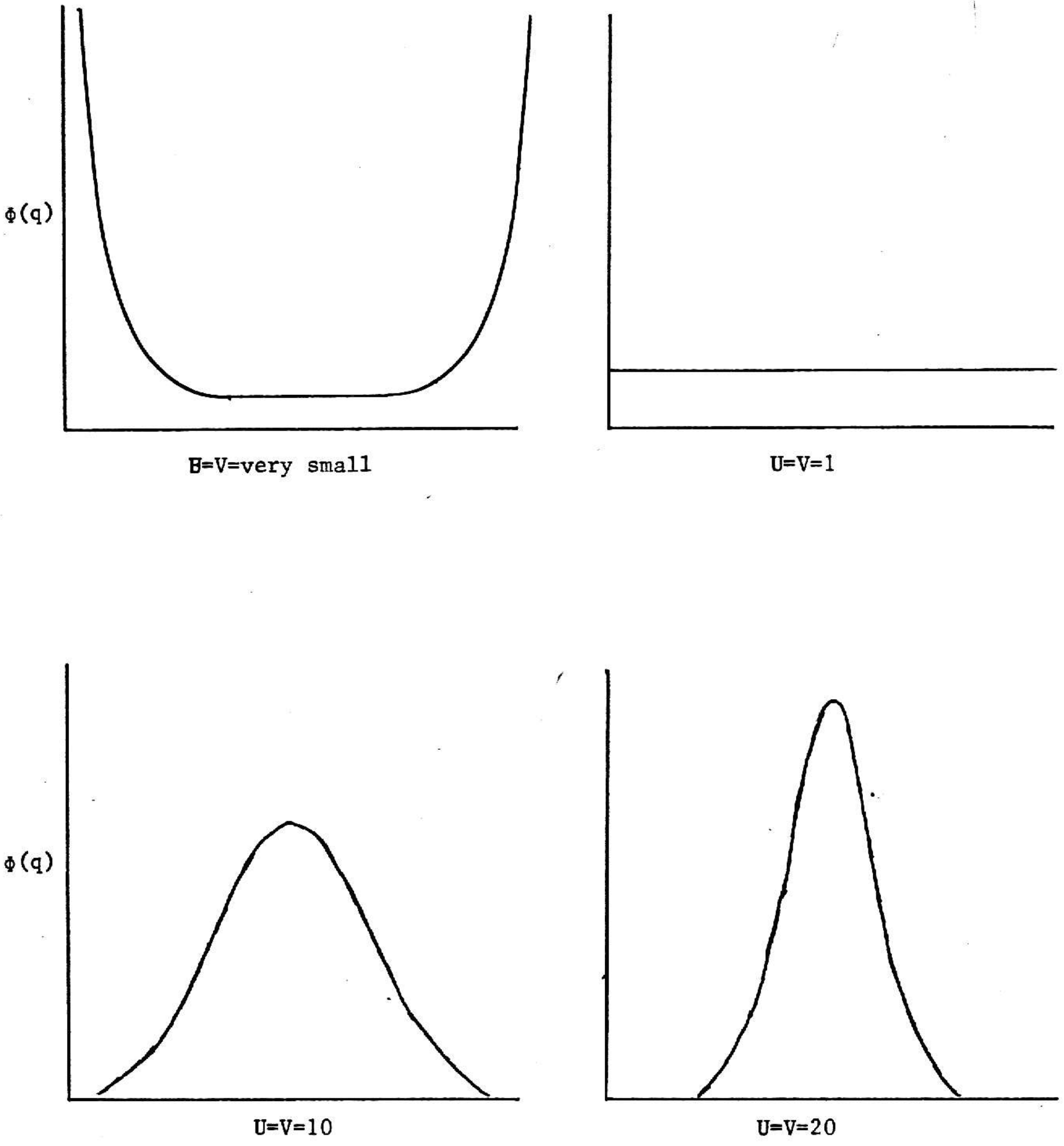
$$\bar{q}=U/(U+V)=u/(u+v)$$

and

$$\sigma_q^2=UV/(U+V)^2(U+V+1)=\bar{q}(1-\bar{q})/\{4N(u+v)+1\}.$$

The mean of the distribution and the equilibrium value of q are equal. Accordingly, the distribution of q under recurrent mutation pressure will vary around the equilibrium value, as would be expected.

Figure 4.1 gives the form of the distribution for various effective sizes. It can be readily seen that the relative population size depends not only on N but also on the mutation rate. The larger the mutation rate becomes the smaller N may be for the population to be considered large (U=V>2). Various other conclusions are immediately available from examination of the figure.



B=V=very small

U=V=1

U=V=10

U=V=20

Fig. 4.1. Distribution of q under recurrent mutation pressure. The symmetry of the distribution is caused by the assumption u=v.

5. DISTRIBUTION UNDER MIGRATION PRESSURE

Assume that the effects of mutation and selection on each group are totally ineffective. Further assume that each group exchanges a proportion m of individuals with a random sample of equal size from the entire population every generation. This assumption is actually artificial. Any given group may actually exchange a higher proportion of individuals with neighboring groups than with groups that are distantly removed. As a result, the immigrants will not constitute a random sample from the entire population. This problem will be dealt with more extensively later.

If q is the frequency of a given gene in one generation then the frequency of the same gene in the next generation will be given by

$$q'=(1-m)q+m\bar{q}'=q-m(q-\bar{q}),$$

where $\bar{q}$ is the frequency in the entire population.

Therefore,

$$\Delta q=q'-q=-m(q-\bar{q})=m\bar{q}-mq.$$

Before proceeding to the general distribution form the variance must be considered more extensively than it was in previous sections. If the migration scheme hypothesized above is continued for an extended period the immigrants will tend to move the group gene frequencies toward the average of the total population. In one generation the deviation of a group frequency from the population average is $(q-\bar{q})$; in the next generation after migration the deviation will be

$$q'-\bar{q}=(q-m(q-\bar{q}))-\bar{q}=(1-m)(q-\bar{q}).$$

The variance of q among all groups in the next generation will be

$$\sigma^2_{q'}=E\{(q'-\bar{q})^2\}=(1-m)^2E\{(q-\bar{q})^2\}=(1-m)^2\sigma^2_q,$$

where (σ^2_q) is the variance of q in the first generation. It was assumed, in the derivation of the general distribution function, that the moments about the mean remained constant from generation to generation. Therefore, the variance of q must remain constant. For this to be true the decrease in variance due to immigrants must be compensated for by the sampling variance of the new gene frequency. Now, for groups of effective size N the sampling variance of each group in the next generation will be

$$(q+\Delta q)(1-q-\Delta q)/2N.$$

The average value of this sampling variance for all groups is

$$\overline{\sigma^2_{\delta q}}=\frac{1}{2N}\int_0^1\{q-m(q-\bar{q})\}\{1-q+m(q-\bar{q})\}\Phi(q)dq.$$

After integration we obtain

(5.1) $$\overline{\sigma^2_{\delta q}}=\{\bar{q}(1-\bar{q})-(1-m)^2\sigma^2_q\}/2N.$$

which is the contribution of random deviations to the variance. Therefore, the variance in the next generation will become (Wright 1953)

(5.2) $$\sigma^2_{q'}=\sigma^2_q=\{\bar{q}(1-\bar{q})\}/\{2N-(2N-1)(1-m)^2\}.$$

If the percentage of immigrants m is small we may take q(1-q)/2N as an approximation of (5.1). Therefore,

$$\Phi(q)=C/q(1-q)\ \exp\{4N\int\{m\bar{q}-mq\}/q(1-q)dq\}$$

$$=C/q(1-q)\ \exp\{4Nm\bar{q}\ \mathrm{Log}\ q+4Nm\bar{p}\ \mathrm{Log}\ (1-q)\}.$$

Let U=4Nmq and V=4Nmp then,

$$\Phi(q)=Cq^{U-1}(1-q)^{V-1} \tag{5.3}$$

where $C=\Gamma(U+V)/\Gamma(U)\Gamma(V)$.

From (5.3) the variance of q is

$$\sigma_q^2=\bar{q}(1-\bar{q})/(4Nm+1).$$

This is a good approximation of (5.2) when m is small.

Let us consider now the probability distribution of gene frequencies before a state of equilibrium has been reached. Let A and a be a pair of alleles in the respective proportions p and 1-p from a random mating population of effective size N. Suppose that the population exchanges a proportion m of its individuals with a random sample from some larger area each generation. Then the mean and variance of the rate of change per generation may be given in Kimura's notation by $M_{\delta p}=m(p-p)$ and $V_{\delta p}=p(1-p)/2N$, where p is the frequency of gene A in the immigrants. Substituting $M_{\delta p}$ and $V_{\delta p}$ in (2.8) and solving the resulting differential equation, Kimura and Crow (1956) obtained

$$\Phi(p',p;t)=\Sigma X_i(p)\ \exp\{-i(m+\frac{i-1}{4N})t\}$$

as a solution. In this expression

$$X_i(p)=p^{Б-1}(1-p)^{(A-B)-1}F(A+i-1,-i,A-B,1-X)$$

$$\cdot\ F(A+i-1,-i,A-B,1-p)\ \frac{\Gamma(A-B+i)\Gamma(A+2i)\Gamma(A+i-1)}{i!\Gamma^2(A-B)\Gamma(B+i)\Gamma(A+2i-1)}$$

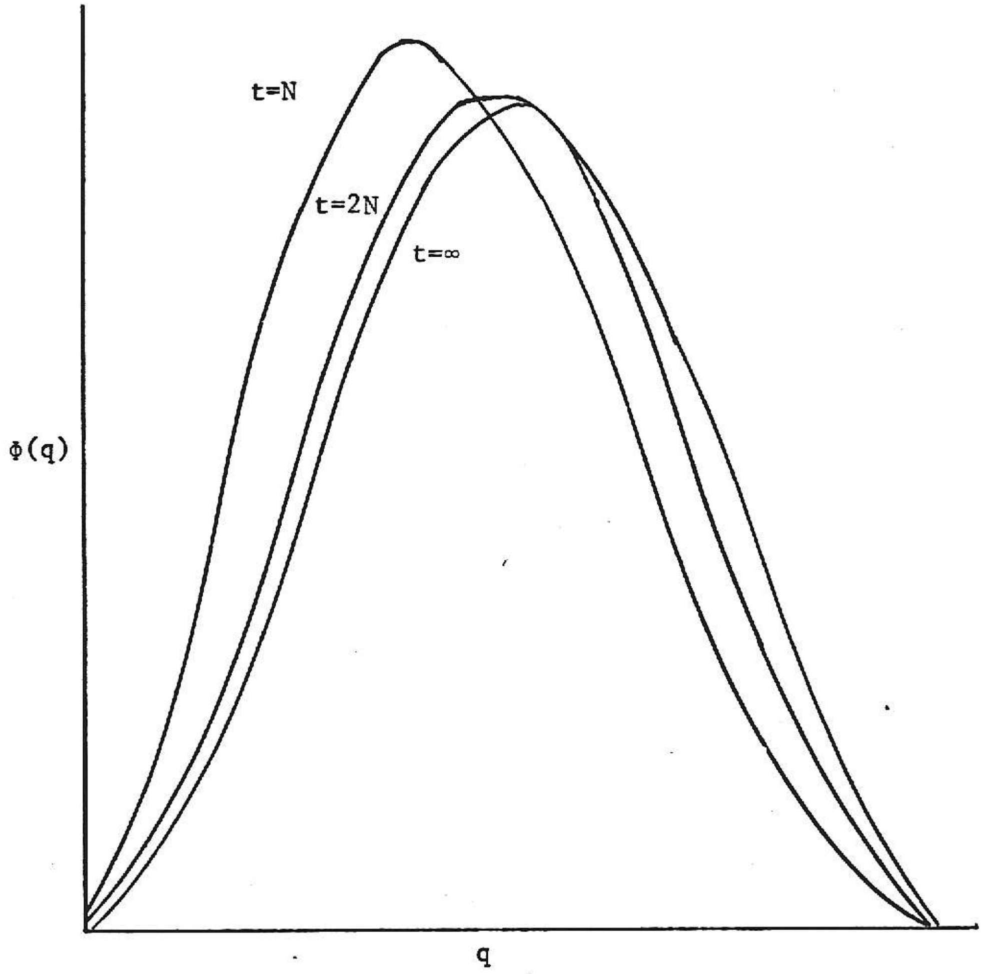


Fig. 5.1. Asymptotic behavior of the distribution assuming a finite population, m=0.5 and p=0.2.

$A=4Nm$ and $B=4Nm\bar{x}$. F denotes the hypergeometric function. Figure (5.1) shows the form of this distribution for three values of t.

As t becomes large the distribution curve gradually approaches the steady state gene frequency distribution that was previously derived under Wright's island model. That is

$$\lim_{t\to\infty} \Phi(p',p;t)=\Phi(p)$$

where $\Phi(p)$ is equation (5.3).

It should be noted here that mutation pressure may be included in the previous discussion by putting $m=m+u+v$ and $mp=m\bar{x}+v$ where u and v are the mutation rates as previously defined.

6. DISTRIBUTION UNDER SELECTION PRESSURE

The primary assumption when considering selection is that some genotypes in a population have a distinct advantage over other genotypes. Consider one locus with two alleles A and a. There will be three possible genotypes with respect to the locus; AA, Aa, and aa. Let w denote the fitness of a genotype. That is, if w=1 for genotype AA and w=(1-s) for genotype aa then, in any given generation, for every one individual of the AA type surviving there will be (1-s) individuals of the aa type. The number s is such that $0<s<1$ and is usually known as the coefficient of selection.

To determine a general expression for selection pressure let the fitness of the three genotypes AA, Aa and aa be 1, $(1-s_1)$ and $(1-s_2)$ respectively, where s_i is constant from generation to generation. Assume that the three genotypes are produced in the proportion p^2, $2pq$ and q^2. The average fitness for the entire population is given by

$$\overline{w}=p^2(1)+2pq(1-s_1)+q^2(1-s_2)=1-2s_1pq-s_2q^2.$$

The change in frequency of the recessive gene a per generation is

$$\Delta q=\{pq(1-s_1)+q^2(1-s_2)\}/\overline{w}-q=pq\{-s_1+(2s_1-s_2)q\}/\overline{w}.$$

Noting that $$d\overline{w}/dq=2\{-s_1+(2s_1-s_2)q\}$$

Δq may be written in the form (Wright 1937)

$$\Delta q=\frac{q(1-q)}{2w}\frac{d\overline{w}}{dq}\,. \tag{6.1}$$

Equation (6.1) assumes that the fitness of a genotype is independent of selection or that the coefficients of selection remain constant regardless of the frequency of a genotype.

The distribution function $\Phi(q)$ under selection pressure alone becomes (Wright 1937)

$$\Phi(q)=(C/q(1-q))\exp\{4N\int \frac{q(1-q)}{2w}\frac{d\bar{w}}{dq}\ dq/q(1-q)\ \}$$

$$=C\bar{w}^{2N}/q(1-q).$$

The final form of $\Phi(q)$ will depend on the type of selection that is involved. As an example assume that a is a recessive lethal gene. In this case s=1 and

$$\bar{w} = 1-2s_1pq-s_2q^2= 1-2s_1pq$$

Thus, $$\Phi(q) = C(1-2s_1pq)^{2N}/q(1-q).$$

Under most types of selection a stable gene frequency equilibrium point will exist and may be determined by putting $\Delta q=0$. Consider the case of selection favoring the genotype Aa. For convenience let the three genotypes AA, Aa and aa have fitnesses of $(1-s_1)$, 1 and $(1-s_2)$ respectively, where $0<s<1$. s_1 and s_2 are assumed to be constant. Then

$$\bar{w}=1-s_1p^2-s_2q^2$$

and $$\Delta q=pq(s_1p-s_2q)/(1-s_1p^2-s_2q^2).$$

Putting $\Delta q=0$ and solving for p we obtain

$$\hat{p}=s_2/(s_1+s_2) \text{ and } \hat{q}=s_1/(s_1+s_2).$$

The equilibrium value p and q are independent of the initial gene frequencies and since s_1 and s_2 are constants this is a stable equilibrium condition.

Let us assume that selection is against the heterozygote. In this case the fitness of AA, Aa and aa will be 1, (1-s) and 1 respectively. Again assume that the population is random mating and that the initial proportions of the three genotypes are p^2, 2pq and q^2 before selection. Proceeding as before we obtain

$$\bar{w}=1-2spq$$

and

$$\Delta q=spq(2q-1)/(1-2spq).$$

Assuming that s is very small Δq may be expressed in the form

$$\Delta q\simeq 2spq(1-\tfrac{1}{2}).$$

From inspection of the above expression it can be seen that $\Delta q=0$ when $q=\frac{1}{2}$, so that $q=\frac{1}{2}$ is the equilibrium point. However, this is an unstable equilibrium point because if q is greater than ½, Δq will be positive and the frequency of q will increase. A similar condition exists when q is less than ½. In the absence of other systematic pressures the ultimate fate of this type of selection in a large panmictic population is fixation at either q=0 or 1. The distribution functions for the above two cases may be obtained quite simply by substituting the given expression of Δq in the general form of $\phi(q)$.

During the previous discussion of selection pressure it was assumed that the coefficient of selection (s) remained constant between generations. However, random fluctuations of selection intensity may be an important factor in causing random fluctuations in gene frequency. To isolate the effect of random fluctuations of selection intensity, Kimura (1955) proposed a population large enough to allow the effects of random sampling to be neglected. Considering only a pair of alleles A and a such that if p is the frequency of A in generation t then the frequency of A in generation (t+1) will be given by p+sp(1-p). Also assume that s=0 over a large number of generations and that the variance of $s(V_s)$ is constant. Under the above assumptions $M_{\delta p}=0$ and $V_{\delta p}=V_{s}p^{2}(1-p)$. Substituting these expressions into (2.8) and solving the partial differential equation assuming that the initial condition is a fixed gene frequency we obtain (Kimura 1955)

$$\Phi(p',p;t)=(2\pi V_s t)^{-\frac{1}{2}}\exp\left[-\frac{V_s}{8}t-\frac{\{\mathrm{Log}\,\frac{p(1-p')}{(1-p)p'}\}^2}{2V_s t}\right]\frac{(p'(1-p'))^{\frac{1}{2}}}{(p(1-p))^{3/2}}$$

Figure (6.1) shows the process of change in the gene frequency distribution. As t increases the distribution becomes U-shaped. The gene frequency accumulates near fixation and loss but never becomes lost or fixed completely.

If the genes are not neutral, that is $\bar{s}=0$, then $M_{\delta p}=0$ should be replaced by $M_{\delta p}=\bar{s}p(1-p)$ in the partial differential equation. However, to my knowledge, no exact solution for this form has been found.

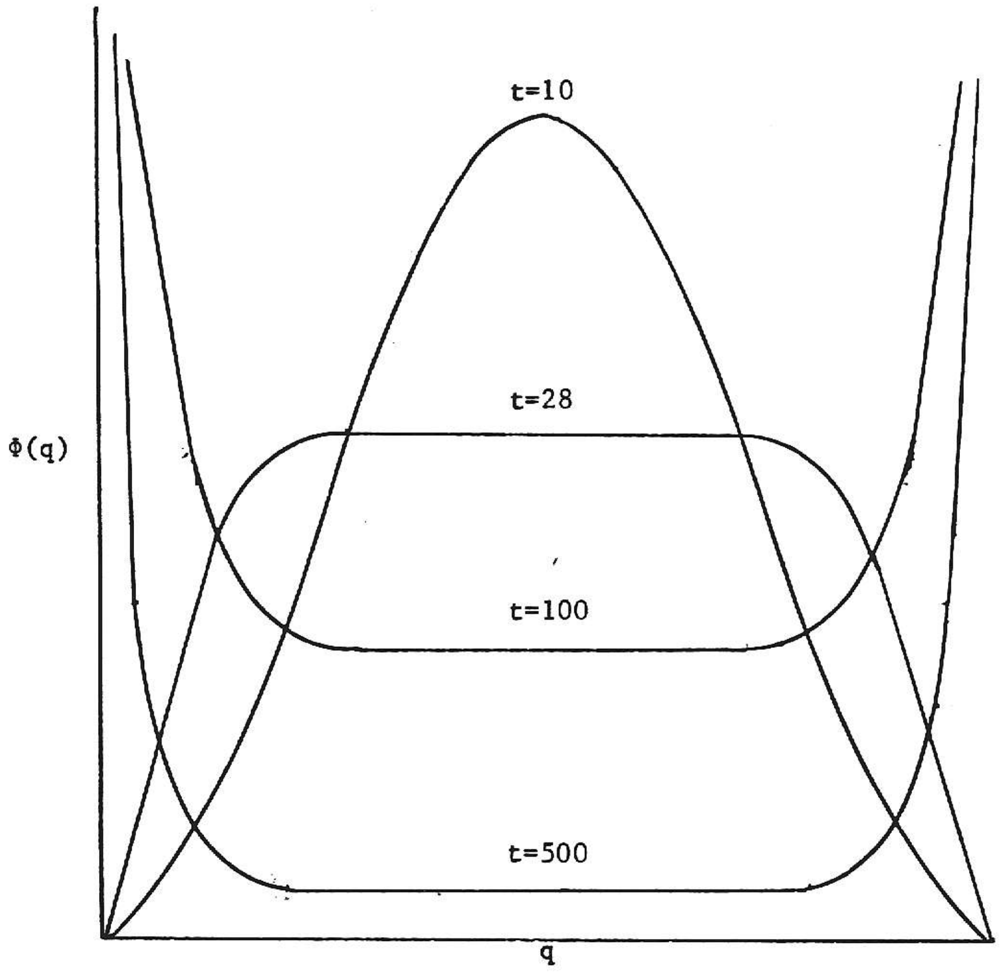


Fig. 6.1. The process of change in the gene frequency distribution assuming random fluctuations of selection intensity with s=0.

7. DISTRIBUTION UNDER JOINT PRESSURES

The previous discussion of distribution form is unrealistic in the sense that a natural population would rarely be subjected to only one type of pressure. More realistic would be the simultaneous consideration of two or more types of pressures. Consider the combined effects of mutation and migration under the same general assumptions that were previously given. The changes in gene frequency per generation under the influence of mutation and migration were given previously, so that the change in frequency from these joint effects may be given as the sum of the individual changes. Thus,

$$\Delta q=up-vq+m\bar{q}-mq=up-vq+m\bar{q}p-m\bar{p}q$$

$$=(u+m\bar{q})p-(V+m\bar{p})q.$$

Substituting q in the general expression for $\phi(q)$ we have

$$\phi(q)=C/q(1-q)\ \exp\{4N\textstyle\int\{(u+m\bar{q})(1-q)-(v+m\bar{p})q\}/q(1-q)dq\}$$

$$=C/q(1-q)\ \exp\{4N(u+m\bar{q})\ \mathrm{Log}\ q+4N(v+m\bar{p})\ \mathrm{Log}\ (1-q)\}.$$

Put $U=4N(u+mq)$ and $V=4N(v+mp)$

then, $$\phi(q)=Cq^{U-1}(1-q)^{V-1}.$$

The distribution under the joint systematic pressures is a Beta distribution with U and V equal to the sums of the corresponding terms from the Beta distributions under the single systematic pressures. The mean and variance of the above distribution are given by

$$\bar{q}=V/(u+v)$$

and

$$\sigma_q^2=\bar{q}(1-\bar{q})/\{4N(m+u+v)+1\}.$$

If conditions are the same in all groups the mean and variance given above are not only the mean and variance among all groups at any given time but also they are the variance for one group over a long period of time.

Following the above procedure the distribution under the joint effects of mutation and selection will be found to be (Wright 1942)

$$\Phi(q)=C\bar{w}^{2N}q^{U-1}(1-q)^{V-1} \tag{7.1}$$

where $U=4Nu$ and $V=4Nv$.

This form of the distribution function is very general. If U and V are given the values

$$U=4N(u+m\bar{q})$$

and

$$V=4N(v+m\bar{p})$$

then (7.1) becomes the distribution function under the combined effects of mutation, migration and selection. By letting U and V assume the appropriate values one can obtain the distribution function under migration and selection. Assuming that there is no selection ($w=1$) (7.1) will assume the distribution form previously given for mutation, migration and the joint effects of mutation and migration. Also, by letting $U=V=0$ (no mutation or migration effects) (7.1) becomes the distribution function under selection pressure. Thus equation (7.1) is as general as (2.7) and may be preferred because the various distributions are more readily obtainable.

In the derivation of the previous distribution functions, it was assumed while not explicitly stated, that the effects of systematic pressures on one locus is independent of all other loci. A more general treatment requires that the interactions between different loci be taken into account. In general, the fitness of a genotype will depend on a particular combination of genes rather than on the presence of a single gene. The joint distribution function for k loci can be written as the product of k distributions of the form given in (7.1). Thus,

$$(7.2) \qquad \Phi(q_1,q_2,\ldots q_k)= C_w^{-2N} \prod_{i=1}^{k} q_i^{U_i-1}(1-q_i)^{V_i-1}$$

where $U_i = 4N(m_i\bar{q}_i+v_i)$

and $V_i = 4N(m_i\bar{p}_i+v_i)$.

The average fitness over all genotypes is represented by $\bar{w}$. For k pairs of genes the joint distribution (7.2) will be represented by a surface in k+1 dimensional space.

8. ISOLATION BY DISTANCE

Wright (1942) pointed out that the island model of population structure is not likely to be realized in nature; namely because of the unrealistic assumption that the immigrants to a group comprise a random sample from the entire population. As indicated previously, the immigrants are not likely to be representative of the entire population but rather of neighboring groups. To take this into account, Wright (1942) proposed the continuum model of population structure which assumes that the population is distributed uniformly over a wide area and that mating individuals are limited to a "neighborhood" of limited distance. Thus, the farther two individuals are apart the less chance they will have of mating. Wright (1942) examined the above model in terms of departures from random mating by using the average inbreeding coefficient relative to the total population. This coefficient (F) has been defined as "the correlation between uniting gametes with respect to the gene complex as an additive system". In order to give the coefficient (F) a more concrete meaning recall that the variance of q under migıation pressure was given as

$$\sigma_q^2=\bar{q}(1-\bar{q})/(4Nm+1).$$

Dobzhansky and Wright (1941) gave

$$F=1/(4Nm+1)$$

as a good approximation to F when m is small. As is expected, under the island model, the smaller the proportion of immigrants to any group the larger the value of F. Thus a group is completely inbred (F=1) when m=0. For a more complete discussion of the inbreeding coefficient see Appendix C.

The development of Wright's continuum model is based on two important theorems from the theory of path coefficients. Only that portion of the theory of path coefficients which is necessary to the development of Wright's model will be presented here.

THEOREM 8.1. The correlation between two variables is the sum of the products of the chains of individual path coefficients along all the paths by which they are connected.

THEOREM 8.2. Let an effect x be produced by two correlated causes A and B. Also let a and b be the path coefficients from causes A and B to effect x respectively. Then

$$a^2+b^2+2abr=1$$

where r is the correlation coefficient between causes A and B.

For an elementary discussion of path coefficients and proof of the above two theorems see Appendix B.

Before developing Wright's continuum model based on the inbreeding coefficient F, some relationships between path coefficients and F must be established.

A zygote may be considered as being equally and linearly determined by each of its gametes. Consequently the two path coefficients from the gametes to the zygote must be equal (Fig. 8.1). Let a be the path coefficient between a gamete and a zygote and let F be the correlation between uniting gametes. The primes represent corresponding values in the preceeding generation. By Theorem 8.2 we have the following relation.

$$2a'^2+2a'^2F=1$$

or

(8.1) $$a'=(1/2(1+F'))^{\frac{1}{2}}$$

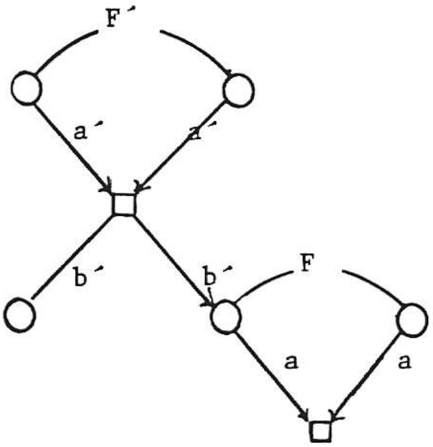


Fig. 8.1. Path for one generation.

Now consider the paths from the zygote to its gametes in Fig. 8.1. Since the zygote produces gametes by the random process of segregation the two gametes should be produced in equal frequency. Thus, the path coefficients from a zygote to each of its gametes are equal and the gametes are not correlated. As a result, by Theorem 8.1, the correlation between a zygote and one of its gametes is the path coefficient from the zygote to the gamete. This correlation must be the same as that between the zygote and one of its gametes from the preceeding generation. Therefore,

$$b=a'+a'F.$$

Substituting the expression for a´given in (8.1) we have

(8.2) $$b=\{(1+F')/2\}^{\frac{1}{2}}.$$

The zygotic generation path coefficient (the path coefficient from a zygote to a zygote) is given by

(8.3) $$ba=\tfrac{1}{2}\{(1+F')/(1+F)\}^{\frac{1}{2}}$$

and the gametic generation path coefficient (the path coefficient from a gamete to a gamete) is given by

$$a'b=\{2(1+F')\}^{-\frac{1}{2}}\{(1+F')/2\}^{\frac{1}{2}} = \tfrac{1}{2}. \tag{8.4}$$

It should be noted here that expressions (8.3) and (8.4) are not correlation coefficients. That is, $a'b$ is not the correlation between a preceeding generation gamete and a present generation gamete. This correlation is given by

$$r=a'b+F'a'b=b^2=(1+F')/2.$$

Let r be the correlation coefficient between mating individuals. Using Theorem 8.1 in reference to Fig. 8.2 it can be easily seen that we have the following relationship between F and r.

$$F=brb. \tag{8.5}$$

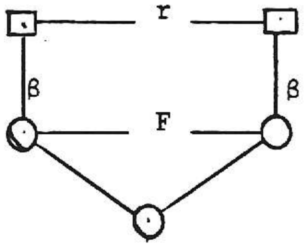


Fig. 8.2. Correlation between uniting gametes in terms of correlation between mates.

To proceed with the development of Wright's model, consider a "neighborhood" of N monoecious diploid individuals whose gametes unite at random in every generation, including the possibility of self fertilization. Now,

$$\text{pr(2 gametes of the same individual unite)} = 1/N$$

and

$$\text{pr(2 gametes from different individuals unite)} = (N-1)/N.$$

In the first case the correlation between mating individuals is $r=1$. If two gametes from different individuals unite there will be four paths in going from z_1 to z_2 (Fig. 8.3). r' denotes the correlation between mating individuals in the preceeding generation. From Theorem 8.1

$$r= 4r'a'^2b'^2.$$

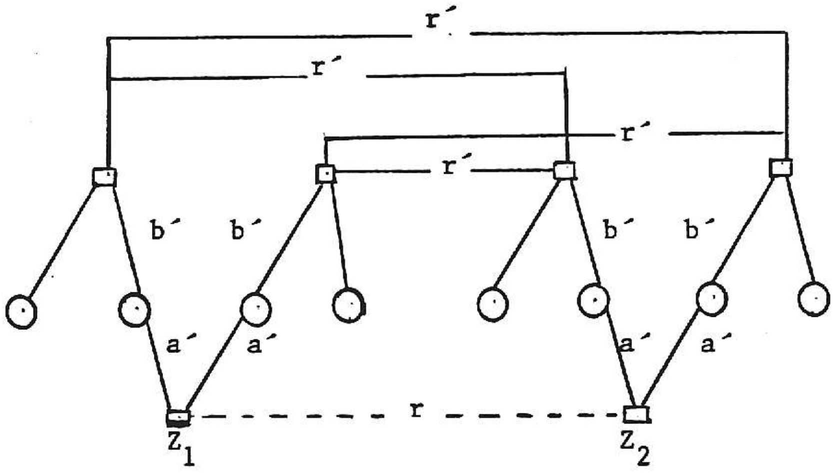


Fig. 8.3. Correlation between remote relatives.

Thus, the average value for the entire group will be

$$\bar{r}= \frac{1}{N} +(N-1)/N\cdot 4a'^2b'^2r'.$$

Simplifying by use of (8.5) we have

(8.6) $$F= \frac{1}{N} b^2+(N-1)/N\cdot 4b^2a'^2b'^2r'.$$

From (8.2), (8.4), (8.5) and (8.6) we have the fundamental relation

(8.7) $$F= \frac{1}{N} b^2+(\frac{N-1}{N})F'.$$

Now let the size of a neighborhood be N, that is, the two parents of any individual are from a neighborhood consisting of N individuals. Assume that the distribution of individuals over a neighborhood is uniform so that N is directly proportional to the size of the neighborhood. Thus the four grandparents of any individual may be considered as drawn at random from an area of size 2N. In general we may think of the ancestors of any generation K as being drawn at random from an area of size KN.

Let F be the correlation between uniting gametes for a neighborhood of size N and F_K be that correlation for an area of size KN (a group of KN individuals). Let the primes denote the correlation from previous generations. Using the above notation, (8.7) may be written as

$$F_1 = \frac{1}{N} b^2 + (\frac{N-1}{N}) F_2', \tag{8.8}$$

where F_2' denotes the correlation for a group of 2N individuals in the previous generation. It should be noted here that expressions (8.7) and (8.8) are identical, only the notation has changed. From (8.2) and (8.8) we have

$$F_1 = \frac{1}{N}(\frac{1+F_1'}{2}) + \frac{N-1}{N} F_2'.$$

Similarly,

$$F_2' = \frac{1}{2N}(\frac{1+F_1''}{2}) + \frac{2N-1}{2N} F_3'',$$

$$F_3'' = \frac{1}{3N}(\frac{1+F_1'''}{2}) + \frac{3N-1}{3N} F_4''', \ldots .$$

If the population has reached an equilibrium state the primes may be dropped because F will not change from generation to generation. Furthermore, let F_1=F. Then

$$(8.9) \qquad F= \frac{1+F}{2N}\left[1+ \frac{1}{2}\left(\frac{N-1}{N}\right)+ \frac{1}{3}\left(\frac{N-1}{N}\right)\left(\frac{2N-1}{2N}\right)+ \frac{1}{4}\left(\frac{N-1}{N}\right)\left(\frac{2N-1}{2N}\right)\left(\frac{3N-1}{3N}\right)+\ldots\right].$$

This expression will hold for some finite group with random mating neighborhood of basic size N. In general, the value of F for an area K times the basic neighborhood is obtained by summing the first K-1 terms of (8.9) assuming $F_i=0$, $i>K$. When the population size becomes very large the value of F must approach unity. This implies that the sum of the series within the brackets of (8.9) is N.

Let t_i represent the ith term in the series within the brackets of (8.9), then

$$F= \frac{(1+F)}{2N}\,\Sigma t_i$$

or

$$(8.10) \qquad F=\Sigma t_i/(2N-\Sigma t_i).$$

In general, the kth term in the series (t_i) may be written in the form

$$t_k= \frac{1}{k}\prod_{i=1}^{k}\left(1- \frac{1}{ik}\right) = \frac{\{(k-1)N-1\}}{kN}\,t_{k-1}.$$

From this relation it may be shown that the sum of the first k-1 terms of (8.9) is

$$\Sigma t_i=N(1-kt_k)$$

so that (8.10) may be written as

$$F=(1-kt_k)/(1+kt_k).$$

Figure 8.4 (Wright 1951) shows some relationships between F and selected values for K and N.

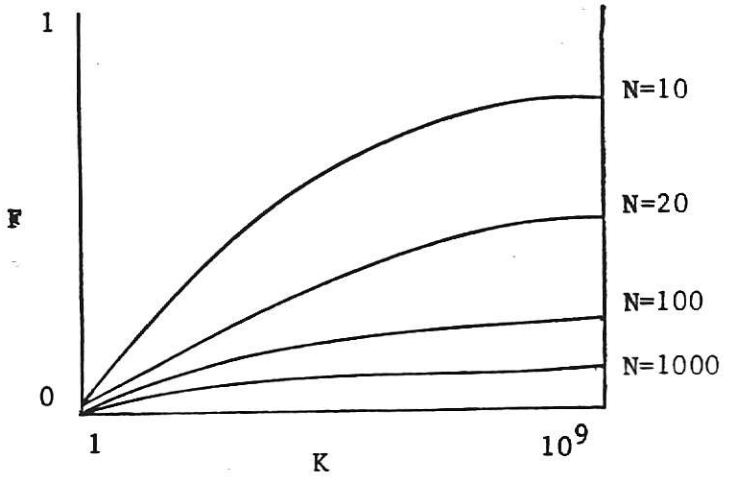


Fig. 8.4. Relationships between F and selected values for K and N.

The relationship between F and the variance of gene frequencies under the island model (assuming only migration pressure) is presented in Fig. (8.5) (Wright 1951).

Before developing a formula for the distribution of gene frequencies under the continuum model a random breeding unit within the population must be defined. Consider a continuous population of size N which is divided into H subpopulations each of size N_s. Let each subpopulation be composed of K random breeding groups each of size N_g. The inbreeding coefficient F will be zero relative to the random breeding groups, F_s relative to the subpopulations and F_t relative to the total population.

Given a single locus with two alleles, A and a, the proportion of heterozygotes in a population with inbreeding coefficient F has been shown to be (Wright's equilibrium law, Appendix C)

$$P=2q(1-q)(1-F_t). \quad (8.11)$$

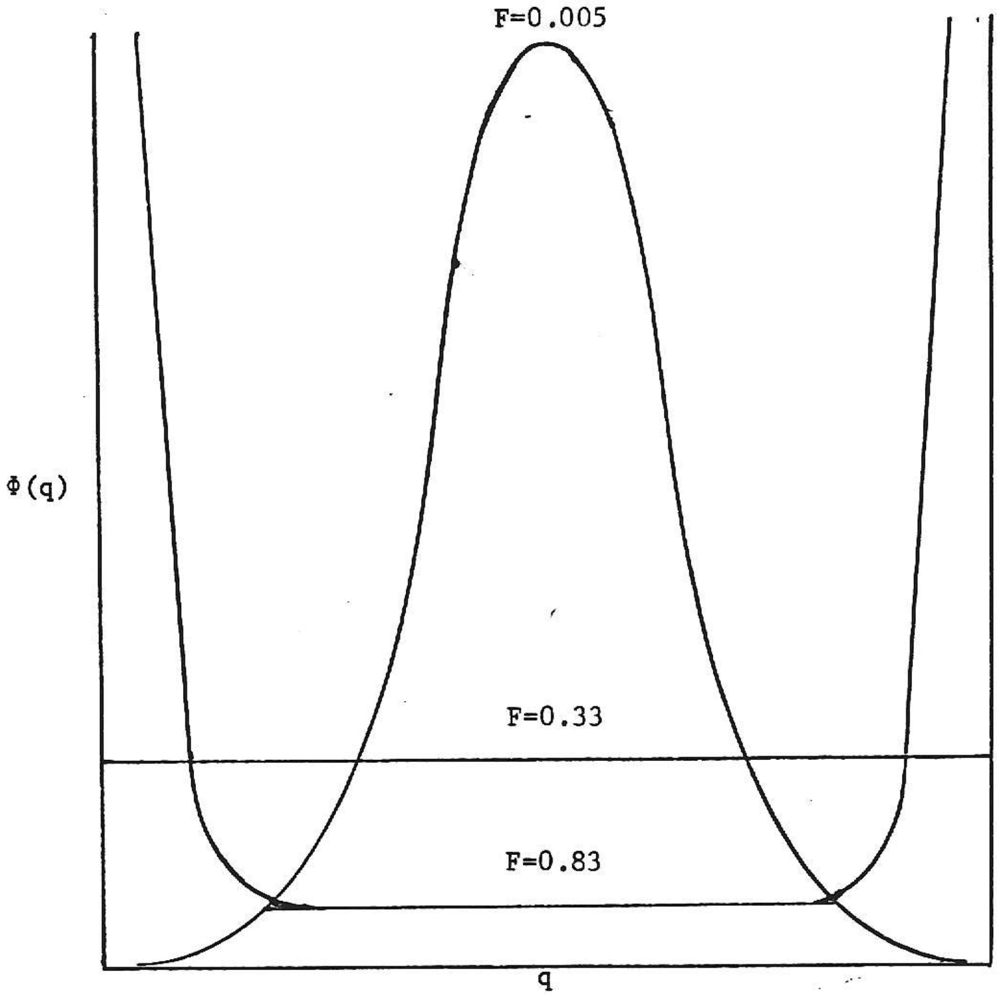


Fig. 8.5. The relationship between F and the variance of gene frequencies under the island model.

If F_t remains constant from generation to generation the proportion of heterozygotes will also remain constant. Under the population hypothesised above P may also be represented as the average heterozygosis of the subpopulations.

$$P= \sum_{i=1}^{H} \{2q_i(1-q_i)(1-F_s)\}/H$$

$$= 2(1-F_s)(\overline{q}-\Sigma q_i^2/H). \tag{8.12}$$

The variance of q among subpopulations is given by

$$\sigma_q^2 = \frac{1}{H} \Sigma q_i^2-\overline{q}^2. \tag{8.13}$$

Substituting (8.13) in (8.12) we have

$$P = 2(1-F_s)\{\overline{q}(1-\overline{q})-\sigma_q^2\}.$$

From (8.11)

$$2\overline{q}(1-\overline{q})(1-F_t) = 2(1-F_s)\{\overline{q}(1-\overline{q})-\sigma_q^2\}$$

or

$$\sigma_q^2 = \{\overline{q}(1-\overline{q})(F_t-F_s)\}/(1-F_s). \tag{8.14}$$

Equation (8.14) was derived under the continuum model. However, equations (8.11), (8.12) and (8.13) also apply to the island model if H is considered as the number of groups rather than the number of subpopulations. Thus (8.14) is applicable to either model. Recall that the variance of q under the island model for the joint effects of mutation and migration was given by

$$\sigma_q^2 = \overline{q}(1-\overline{q})/\{4N(m+u+v)+1\}.$$

Equating the two expressions for σ_q^2 we have

$$m+v+u=(1-F_t)/(F_t-F_s)4N.$$

The mean of $\Phi(q)$ under the island model for mutation and migration pressure is

$$\bar{q}=v/(u+v).$$

Thus $$m+v/\bar{q}=(1-F_t)/(F_t-F_s)4N.$$

The distribution of gene frequencies under the continuum model, assuming only mutation and migration pressure, may be found by substituting the above expression in $\Phi(q)$ derived under the island model with corresponding assumptions. The distribution is given approximately by

(8.15) $$\Phi(q)=Cq^{U\bar{q}-1}(1-q)^{U(1-\bar{q})-1}$$

where $$U=(1-F_t)/(F_t-F_s)$$

and $$C=\Gamma(U\bar{q})\Gamma(U(1-\bar{q}))/\Gamma(U).$$

The distribution under migration, mutation and selection may be obtained by multiplying (8.15) by $(\bar{w}^{2N})$ where $\bar{w}$ is the average fitness for the type of selection involved.

As can be seen from the previous discussions, the island and continuum models are similar in many respects. The main difference between the two models is that neighboring groups should be similar under the continuum model but uncorrelated under the island model.

9. STEPPING STONE MODEL OF POPULATION STRUCTURE

Kimura and Weiss (1964) proposed a third model of population structure. Their stepping stone model is actually a rough combination of the two models by Wright which were discussed earlier. Under the island model of population structure Wright postulated a natural population consisting of many partially isolated groups each of which exchange a proportion m with a random sample of the entire population every generation. It was this assumption that led Wright to postulate his continuum model of population structure. The population hypothesised by the stepping stone model consists of many groups or colonies exchanging a given proportion of individuals with adjacent groups and a given proportion with the entire population every generation. Essentially the stepping stone model consists of the island model with the added assumption that each group exchanges a given proportion of individuals with adjacent groups every generation. Wright analized his continuum model by considering the correlation between uniting gametes; Kimura and Weiss were interested in the correlation of gene frequencies between two colonies which are a given distance apart. The two methods of considering correlation are actually equivalent.

One Dimensional Case

The simplest situation under the stepping stone model is constructed by considering an infinite array of groups with their position represented by integers on a line. Consider a single locus with two alleles A and a. Assume that individuals in any given group can migrate at most "one step" in either direction each generation and that the gene frequency in each group may change by mutation, migration and random sampling of gametes.

Selection will be dealt with later. Kimura and Weiss, in order to simplify the treatment, further assumed the generation time to be discrete. Let p_i be the frequency of gene A in the ith colony in any given generation. Then the frequency in the next generation will be given by

$$(9.1) \qquad p_i'=(1-m+n)p_i+vq_i-up_i+\frac{m}{2}(p_{i-1}+p_{i+1})+n\bar{p}+\varepsilon_i,$$

where m is the rate of migration from a group to adjacent groups such that a proportion m/2 is exchanged between adjacent groups each generation. The proportion of individuals from each group exchanged with a random sample of the entire population every generation is represented by n. Also u and v are the mutation rates as previously defined and ε_i is the change in the frequency of p_i each generation due to random sampling of gametes. As previously shown, ε_i follows a binomial distribution with mean and variance given by

$$E_\delta(\varepsilon_i)=0$$

and

$$E_\delta(\varepsilon_i^2)=p_i(1-p_i)/2N_e$$

where N_e is the effective size of a group and E_δ is the expectation with respect to a single group.

After simplification (9.1) may be written in the form

$$(9.2) \qquad p_i'=(1-m-v-u-n)p_i+\frac{m}{2}(p_{i-1}+p_{i+1})+v+n\bar{p}+\varepsilon_i.$$

Kimura and Weiss have stated that the effect of exchanging a given proportion n of each group with a random sample from the entire population is formally equivalent to mutation. Without loss of much generality (9.2) may be written in the form

(9.3) $$p_i'=(1-m-n)p_i+\frac{m}{2}(p_{i-1}+p_{i+1})+n\overline{p}+\varepsilon_i.$$

Let d_i denote the deviation of the gene frequency in the ith group from the population mean, that is $d_i=p_i-\overline{p}$. Then (9.3) becomes

$$d_i'+\overline{p}=(1-m-n)(d_i+\overline{p})+\frac{m}{2}(d_{i-1}+d_{i+1}+2\overline{p})+n\overline{p}+\varepsilon_i,$$

or

(9.4) $$d_i'=\alpha d_i+\beta(d_{i-1}+d_{i+1})+\varepsilon_i$$

where $\alpha=1-m-n$ and $\beta=m/2$.

Let V be the variance of gene frequency among groups.

$$V=E_\Phi(d_i^2).$$

Let r_k be the correlation coefficient of the gene frequencies between two groups which are k steps apart.

$$r_k=E_\Phi(d_i d_{i+k})/V.$$

In the above two expressions E_Φ is the expectation with respect to gene frequency among all groups.

To obtain an expression for the variance of gene frequency among colonies begin by squaring both sides of (9.4) to obtain

$$d_i'^2=\alpha p_i^2+\beta^2(d_{i-1}^2+2d_{i-1}d_{i+1}+d_{i+1}^2)+\varepsilon_i^2+\varepsilon_i\alpha d_i+\beta\varepsilon_i(d_{i-1}+d_{i+1})+\beta\alpha(d_{i-1}d_i+d_i+1d_i).$$

Taking the expectation of the above expression assuming

$$E(\varepsilon_i d_i)=0 \text{ and } E(\varepsilon_i\varepsilon_{i+k})=0$$

we have

$$V'=E_{\Phi}(d_i'^2)=\alpha^2V+2V\beta^2(1+r_2)+4V\alpha\beta r_1+E_{\Phi}\{p_i(1-p_i)/2N_e\}$$

$$=\alpha^2V+4\alpha\beta Vr_1+2\beta^2V(1+r_2)-V/2N_e+\overline{p}(1-\overline{p})/2N_e.$$

If the population is in an equilibrium state the variance will not change from generation to generation, hence, $V'=V$ and the above may be written in the form

$$(9.5)\qquad V_p\{1-\alpha^2-4\alpha\beta r_1-2\beta^2(1+r_2)+½N_e\}=\overline{p}(1-\overline{p})/2N_e.$$

A more definite form of the variance will be found when the expressions for r_1 and r_2 have been obtained.

An expression for the correlation coefficient between two groups which are k steps apart can be obtained by finding the expected value of the product (d_id_{i+k}): d_i' and d_{i+k}' are equal to corresponding expressions in the form of (9.4). Thus,

$$d_i'd_{i+k}'=\alpha^2d_id_{i+k}+\beta^2(d_{i-1}d_{i+k-1}+d_{i+1}d_{i+k-1}+d_{i-1}d_{i+k+1}+d_{i+1}d_{i+k+1})$$

$$+\beta(d_{i+k-1}+d_{i+k+1})\alpha d_i+\beta(d_{i-1}+d_{i+1})\alpha d_{i+k}.$$

Taking expectations and dividing both sides by V we have

$$r_k'=\alpha^2r_k+\beta^2(2r_k+r_{k-2}+r_{k+2})+2\alpha\beta(r_{k-1}+r_{k+1}).$$

If the population is at equilibrium then $r_k'=r_k$ and

$$(9.6)\qquad (2\beta^2+\alpha^2-1)r_k+\beta^2(r_{k+2}+r_{k-2})+2\alpha\beta(r_{k-1}+r_{k+1})=0.$$

This expression (9.6) holds for $k>1$. If $k=1$ then r_{-1} should be replaced by r_1 because no distinction is made between correlation in positive and negative directions. Thus, if $k=1$ (9.6) becomes

$$(9.7)\qquad (\alpha^2+2\beta^2-1)r_1+\beta^2(r_3+r_1)+2\alpha\beta(r_2+1)=0,$$

noting that $r_0=1$.

To solve the difference equation (9.6) let $r_k=\lambda^k$ and substitute this in (9.6). Accordingly,

$$\beta^2\lambda^4+2\alpha\beta\lambda^3+(\alpha^2+2\beta^2-1)\lambda^2+2\alpha\beta\lambda+\beta^2=0.$$

The above 4th order equation in λ has the following four roots:

$$\lambda_1=\frac{1}{2\beta}\{(1-\alpha)+\{(1-\alpha)^2-(2\beta)^2\}^{\frac{1}{2}}\}$$

$$\lambda_2=\frac{1}{2\beta}\{(1-\alpha)-\{(1-\alpha)^2-(2\beta)^2\}^{\frac{1}{2}}\}$$

$$\lambda_3=-\frac{1}{2\beta}\{(1+\alpha)+\{(1+\alpha)^2-(2\beta)^2\}^{\frac{1}{2}}\}$$

$$\lambda_4=-\frac{1}{2\beta}\{(1+\alpha)-\{(1+\alpha)^2-(2\beta)^2\}^{\frac{1}{2}}\}$$

where $\lambda_1>1$, $0<\lambda_2<1$, $\lambda_3<-1$ and $\lambda_4<0$. The required general solution to (9.6) should then be a linear combination of the four roots or

$$r_k=\sum_{i=1}^{4}C_i\lambda_i^k$$

where the C's are constants. These constants may be determined in the following manners. In order that

$$\lim_{k\to\infty} r_k=0$$

we must have $C_1=C_3=0$, because both λ_1 and λ_3 are greater than one in absolute value. Also r_0 must equal one so that

$$r_0=C_2\lambda_2^0+C_4\lambda_4^0=C_2+C_4=1.$$

From the above requirements plus the requirement that r must satisfy (9.7), C_2 and C_4 are determined.

$$C_2=R_1/(R_1+R_2) \text{ and } C_4=R_2/(R_1+R_2)$$

where
$$R_1=\{(1+\alpha)^2-(2\beta)^2\}^{\frac{1}{2}}$$

and
$$R_2=\{(1-\alpha)^2-(2\beta)^2\}^{\frac{1}{2}}.$$

Let $r(k)=r_k$. Then

$$r(k)=C_2\lambda_2^k+C_4\lambda_4^k \tag{9.8}$$

is the correlation of gene frequencies between groups which are k steps apart.

Solving (9.8) for r(1) and r(2) and substituting these in (9.5) we obtain

$$V=\overline{p}(1-\overline{p})/(1+2N_eC_0)$$

where
$$C_0=2R_1R_2/(R_1+R_2).$$

Let m=0 so that there is no migration between neighboring groups. Under this assumption the stepping stone model reduces approximately to the island model of population structure and the variance becomes

$$V=\overline{p}(1-\overline{p})/\{1-2N_e(2n-n^2)\}.$$

This expression for the variance agrees with Wright's formula for variance under the island model except for the term n^2 in the denominator. In most cases the difference between the two expressions should be negligible since n^2 will be very small.

If m is much larger than n then R_1 and R_2 can be approximated by

$$R_1=2(1-m)^{\frac{1}{2}} \text{ and } R_2=(2mn)^{\frac{1}{2}}.$$

In this case the variance reduces to

$$V=\overline{p}(1-\overline{p})/\{1+4N_e(2mn)^{\frac{1}{2}}\}.$$

Furthermore, Kimura and Weiss gave

$$r(k)= \exp\{-(2n/m)^{\frac{1}{2}}k\} \tag{9.9}$$

as a close approximation to (9.8).

In the previous treatment of the dimensional stepping stone model it was assumed that migration was restricted to one step per generation. Suppose in any generation individuals can migrate p steps per generation. In this case Kimura and Weiss suggested using the variance of migration distance per generation in place of m. Therefore, (9.9) may be approximated by

$$r(k)= \exp\{-\{(2n)^{\frac{1}{2}}/\sigma_m\}k\}$$

where $$\sigma_m^2=\sum_{j=1}^{p} j^2 m_j$$

and $m_j/2$ is that proportion of individuals exchanged between groups which are k steps apart.

The probability distribution of gene frequencies under the stepping stone model is of the same form as the distribution under the island model which was given previously as

$$\Phi(q)=\Gamma(U+V)/\Gamma(U)\Gamma(V)\cdot p^{U-1}(1-p)^{V-1}$$

where U=4Nmp and V=4Nmp. Under the island model m corresponds to n under the stepping stone model. To distinguish between the two values of m let m_1 be the proportion exchanged between adjacent groups under the stepping stone model. The value m under the island model is the proportion of individuals from each group exchanged with a random sample from the entire population. Accordingly, under the stepping stone model we may take

$$m=m_1+n$$

to be that proportion migrating to each group. However, now the immigrants do not constitute a random sample from the entire population; the immigrants between adjacent groups being correlated by the amount r(1). Consider two groups that have r(1)=1 and exchange a proportion m every generation. Since the correlation is one the immigrants will not cause the gene frequency to change in the two groups so that this case is formally equivalent to having m=0. In the same manner, if two groups are correlated by the amount r(1) and exchange a proportion m, this is equivalent to assuming that the stepping stone model, we may take

$$m=n+m_1(1-r(1))$$

to obtain the approximate gene frequency distribution. When n is very small the variance of p may be approximated by

$$V=\overline{p}(1-\overline{p})/\{1+4N_e m_1(1-r(1))\}.$$

It can easily be seen that the variance is dependent on the size of the correlation between adjacent groups. The variance will be minimum when

$r(1)=-1$ and will increase with $r(1)$, becoming maximum when $r(1)=1$. Kimura and Weiss gave

$$r(1)=1-(2n/m_1)^{\frac{1}{2}}$$

as an approximation to $r(1)$ when $n<m_1<1$.

Two and Three Dimensional Cases

In the two dimensional case, assume that the population consists of an array of groups in a square lattice with the groups located at the intersections of the lines. Each group occupies a point which may be denoted by a pair of integers (k_1,k_2). Assume that each group exchanges a given proportion m_x of its individuals each generation with adjacent groups in the horizontal direction and a given proportion m_y with adjacent groups in the vertical direction. Also, that the effective number of each group remains constant from generation to generation. Now the proportion of individuals which migrates to the four neighboring groups is given by

$$m=m_x+m_y.$$

If we let n denote the proportion exchanged each generation with a random sample of the entire population, the entire proportion of individuals exchanged in each generation for a single group is given by

$$m=m_x+m_y+n.$$

The computations involved in deriving the general expression for the correlation coefficient in the two and three dimensional cases are very complicated. Only the general form of the correlation coefficient will be presented in these cases.

When the population is at equilibrium, Kimura and Weiss (1964) gave

$$r(k_1,k_2)= \frac{A_1(k_1,k_2)+A_2(k_1,k_2)}{A_1(0,0)+A_2(0,0)}$$

as the correlation of gene frequencies between groups which are k_1 steps apart in the horizontal direction and k_2 steps apart in the vertical direction. In the above expression

$$A_1(k_1k_2)= \frac{1}{8\pi^2} \int_0^{2\pi}\int_0^{2\pi} \frac{\cos(k_1\theta_1)\cos(k_2\theta_2)}{n+m_x(1-\cos(\theta_1))+m_y(1-\cos(\theta_2))}\, d\theta_1 d\theta_2$$

and

$$A_2(k_1k_2)= \frac{1}{8\pi^2} \int_0^{2\pi}\int_0^{2\pi} \frac{\cos(k_1\theta_1)\cos(k_2\theta_2)}{2-n-m_x(1-\cos(\theta_1))-m_y(1-\cos(\theta_2))}\, d\theta_1 d\theta_2.$$

In the three dimensional case consider the same general assumptions as previously presented except now the population is represented as a cubic lattice with each group occupying a point denoted by three integers (k_1, k_2, k_3). In this case migration will be in three directions with respective rates m_x, m_y and m_z.

The correlation of gene frequencies between groups which are k_1, k_2 and k_3 steps apart in the x, y and z directions respectively is given by

$$r(k_1k_2k_3)= \frac{A_1(k_1k_2k_3)+A_2(k_1k_2k_3)}{A_1(0,0,0)+A_2(0,0,0)}$$

where

$$A_1(k_1k_2k_3)= \frac{1}{2\pi^3} \int_0^{\pi}\int_0^{\pi}\int_0^{\pi} \frac{\cos(k_1\theta_1)\cos(k_2\theta_2)\cos(k_3\theta_3)d\theta_1 d\theta_2 d\theta_3}{n+m_x(1-\cos(\theta_1))+m_y(1-\cos(\theta_2))+m_z(1-\cos(\theta_3))}$$

and

$$A_2(k_1k_2k_3)= \frac{1}{2\pi^3} \int_0^{\pi}\int_0^{\pi}\int_0^{\pi} \frac{\cos(k_1\theta_1)\cos(k_2\theta_2)\cos(k_3\theta_3)d\theta_1 d\theta_2 d\theta_3}{2-n-m_x(1-\cos(\theta_1))-m_y(1-\cos(\theta_2))-m_z(1-\cos(\theta_3))}\ .$$

10. DISCUSSION

The concept of population structure is of major importance in evolution. A change in gene frequency may be considered as a basic elementary step in an evolutionary process. As a consequence any pressure which causes gene frequency to change can be regarded as an essential factor in evolution at the population level. Generally speaking, the relative size and form of a natural population determines which pressures are most important in changing gene frequency.

In large random mating populations the effects of random drift are negligible and the counteracting systematic pressures assume a dominate role. If the systematic pressures remain constant over a large number of generations they will determine certain equilibrium conditions for each locus involved. When all genes reach their equilibrium points there will be very little, if any, further genetic change in a population under constant environmental conditions. Only when environmental conditions change will the selective values of various genotypes change. At such time new equilibrium conditions will be determined and the gene frequencies will begin to move in a new direction. Depending on the extent of the environmental change, a natural population may approach extinction if the required genotypes are not readily available.

In small completely isolated populations random drift assumes a dominate role. In these populations the genes and thus genotypes will tend to become either fixed or lost. Because most genotypes will consist of gene combinations fixed at random they may not be the most favorable for the evolution of the population. Populations of this type will be largely nonadaptive and their ultimate fate is probably extinction.

Populations of intermediate size present the most favorable conditions for evolution of a species. Selection will be effective to a certain extent in determining the favorable genotypes while random drift will permit a wide variety of genotypes to exist. The existence of a wide variety of genotypes will permit selection to act in many ways if environmental conditions should suddenly change.

Along with the problem of continued existence of a species the problem of the formation of a new species has attracted much attention. Wright recognized two distinct ways in which species formation may occur: (a) the transformation of a population as a whole into a completely new species (b) the division of a population into one or more new species. Wright advocated the hypothesis that the most favorable population for the formation of a new species or continued existence of an established one is a population which is subdivided into a number of partially isolated groups and has a balance between all pressures which cause gene frequency to change. Such a population has several advantages over a random mating population of comparable size. A subdivided population will tend to maintain more alleles at each locus as opposed to the establishment of one type of allele at each locus in a panmictic population. Because of the local variability and number of alleles at each locus a subdivided population may maintain more genotypes that are of an optimal type than will a panmictic population. A subdivided population will be more suitable than a random mating population to an area in which the conditions are not uniform. Perhaps the most important asset of the subdivided population is that it may evolve continuously without change in the environmental conditions. If the environmental conditions do change this population will be able to adapt itself to the new conditions quite readily.

The importance of the three models discussed can be seen from the previous evolutionary implications. The island and continuum models, established by Wright, are considerations of two extreme cases. The actual form of a population may be a mixture between the two models. It is likely that when considering a population over a large area the form will approach the island type; but when considering a population over a relatively small area the form will approach the continuum type. When the continuum model was considered it was assumed that the population was uniformly distributed over a given area. Actually, the distribution of a natural population will be anything but uniform. The differential density of populations from area to area is an important factor when considering the distribution of gene frequencies. If the distribution of a population is very irregular the variance of the distribution of gene frequencies may be much larger than previously indicated.

Kimura's stepping stone model is a rough combination of Wright's models. A natural population, without restricting assumptions placed on it, may more nearly conform to Kimura's model than either of Wright's.

ACKNOWLEDGEMENT

The writer would like to thank Dr. Raja Nassar, Department of Statistics, Kansas State University, for his helpful suggestions and advice during the preperation of this report.

APPENDIX A
Effective Population Number

In a natural population the total number of individuals in a given generation may be large, while the total number of individuals contributing to the next generation may be quite small. Not every individual will reach maturity and mate. Some mating individuals may not leave progeny that survive to sexual maturity and mate in the next generation. Regardless of their genotypes, those individuals that do not mate will not contribute to the genetic composition of the next generation. Even if all capable individuals do mate and contribute to the next generation the expected number of progeny may not be the same for all individuals. Thus, there is a need for some standard measure of population size.

An ideal population is defined as being a population of N breeding individuals, half females and half males, mating at random with the variance of the random deviations of gene frequencies given by

$$\sigma^2_{\delta q}=q(1-q)/2N.$$

Furthermore, in an ideal population each individual has an equal expectation of progeny. Under this assumption the distribution of the number of progeny per parent approaches the Poisson distribution. Kimura and Crow (1962) indicated that in most natural populations the distribution of progeny has a variance greater than the corresponding Poisson value. An ideal population is the standard measure with which other populations are compared. The actual size of a population is reduced to a number equivalent to that in an ideal population.

The effective number of a population is defined as the size of an idealized population that would have the same amount of inbreeding or of random gene frequency drift as the population under consideration. Any natural population will actually have two effective numbers; an inbreeding effective number and a variance effective number. Under many conditions these numbers will be identical or very similar. As will be seen the inbreeding effective number depends primarily on the number in the parent generation whereas the variance effective number depends on the number in the progeny generation.

The concept of an effective population number has been discussed by Wright, Haldane, Morton and Crow. The following treatment is primarily that presented by Kimura and Crow (1962).

Inbreeding Effective Number

Consider a monoecious diploid population in which mating is entirely at random. The inbreeding effect may be found as follows. Let p_t be the probability that a pair of homologus genes in an individual in generation t come from the same parent in generation t-1. Also let f_t be the inbreeding coefficient in generation t or equivalently f_t is the probability tnat a pair of homologus genes came from a common ancestor. That ancestor may have been in the (t-1)st generation or in some preceeding generation. Two gametes coming from the same individual in the (t-1)st generation have a probability ½ of carrying the same gene. Thus

$$\tfrac{1}{2}p_t = pr \text{ (two gametes are from the same parent and carry the same gene).}$$

If the genes in the uniting gametes are different they have, by definition, a probability of f_{t-1} of being from a common ancestor. This holds whether or not the genes come from the same individual in the (t-1)st generation. Thus

$\frac{1}{2}p_t f_{t-1}$ = pr (two gametes are from the same parent, carry different genes and come from a common ancestor).

and

$(1-p_t)f_{t-1}$ = pr (two gametes are from different parents and are derived from a common ancestor).

The inbreeding effect in the tth generation is given by the sum of the previous three probabilities.

$$f_t = \tfrac{1}{2}p_t + \tfrac{1}{2}p_t f_{t-1} + (1-p_t)f_{t-1}.$$

If a gamete is equally likely to have come from any of the potential parents in the previous generation we have

$$p_t = 1/N_{t-1}.$$

where N_{t-1} is the number of contributing individuals in the (t-1)st generation; the inbreeding effective number is defined simply to be

$$N_e = 1/p_t.$$

When a gamete is not equally likely to have come from any contributing individual p_t may be determined as follows. Consider a population of N monoecious diploid individuals each of which contributes a variable number (k) of gametes to the next generation. The mean gamete number will be given by $\overline{k}=\Sigma k/N_{t-1}$ and the variance by $V_k=\Sigma k^2/N_{t-1}-\overline{k}^2$, where the summation is over the N_{t-1} contributing individuals in the (t-1)st generation. The number of ways two gametes from one parent may be selected is k(k-1)/2. The total number of ways two gametes from the same parent may be selected is $k(k-1)/2$. The total number of pairs of gametes is $N_{t-1}\overline{k}(N_{t-1}\overline{k}-1)/2$. Therefore

$$p_t=\Sigma k(k-1)/N_{t-1}\overline{k}(N_{t-1}\overline{k}-1). \tag{A.1}$$

From the definitions for the mean and variance we have $\Sigma k^2=N_{t-1}V_k+N_{t-1}\overline{k}^2$ and $\Sigma k=N_{t-1}\overline{k}$. Substituting these expressions in (A.1) the inbreeding effective number becomes

$$N_e=1/p_t=(N_{t-1}\overline{k}-1)/(\overline{k}-1+V_k/\overline{k}). \tag{A.2}$$

As previously stated, in an ideal population each gamete has a probability $1/N_{t-1}$ of coming from any particular individual in the preceeding generation. The probability that k randomly chosen gametes will come from a given individual has a binomial distribution with mean $\overline{k}$ and variance $V_k=N_{t-1}\overline{k}(1/N_{t-1})(1-1/N_{t-1})=\overline{k}(1-1/N_{t-1})$. Substituting these expressions in (A.2) we have $N_e=N_{t-1}$ as expected under an ideal population.

Variance Effective Number

As before, assume that the population consists of N_{t-1} individuals each contribution a variable number (k) of gametes to the next generation. Consider a single locus with two alleles A and a, such that the frequency of A is $p=p_1+p_2$ where p_1 is the frequency of the homozygote and p_2 is that proportion of p furnished by the heterozygote. The number of homozygotes and heterozygotes in the (t-1)st generation will be $n_1=N_{t-1}p_1$ and $n_2=2N_{t-1}p_2$ respectively.

The number of A genes contributed to the next generation will be

$$\sum_{i=1}^{n_1} k_i+\sum_{j=1}^{n_2} L_j$$

where L_j is the number of A genes contributed by the jth heterozygote. The difference in the number of A genes from the (t-1)st generation to the tth generation may be given by

$$N_{t-1}\bar{k}\delta p=\Sigma^{n_1}k_i+\Sigma^{n_2}L_i-N_{t-1}\bar{k}(p_1+p_2) \tag{A.3}$$

$p_1$and p_2 may be written as

$$p_1 = \Sigma^{n_1}1/N_{t-1} = \Sigma^{n_1}\bar{k}/N_{t-1}\bar{k}$$

and

$$p_2 = \Sigma^{n_2}\bar{k}/2N_{t-1}\bar{k}.$$

Accordingly, (A.3) may be written in the form

$$N_{t-1}\bar{k}\delta p = \Sigma^{n_1}(k_i-\bar{k})+\tfrac{1}{2}\Sigma^{n_2}(k_i-\bar{k})+\Sigma^{n_2}(L_i-k/2). \tag{A.4}$$

Letting E denote expectation the variance is given by $V_{\delta p} = E(\delta p)^2$ since $E(\delta p) = 0$. Squaring both sides of (A.4) and taking expectations we obtain

$$(N_{t-1}\bar{k})^2V_{\delta p} = E\{\Sigma^{n_1}(k-\bar{k})+\tfrac{1}{2}\Sigma^{n_2}(k-\bar{k})\}^2+E\{\Sigma^{n_2}(L-k/2)\}^2. \tag{A.5}$$

The terms in the two expectations of (A.5) are assumed to be uncorrelated. Therefore, the expectation of the cross product is zero. Now, performing the indicated operations and simplifying (A.5) becomes

$$(N_{t-1}\bar{k})^2V_{\delta p} = V_k/N_{t-1}\{N_{t-1}(n_1+n_2/4)-(n_1+n_2/2)^2\}+ \frac{n_1\bar{k}}{4}. \tag{A.6}$$

Let α be a measure of the departure from Hardy-Weinberg proportions so that $p_1 = p(1-p)(1-\alpha)$. Then the following relations hold

$$n_1+n_2/2 = N_{t-1}p$$

$$n_2 = 2N_{t-1}p(1-p)(1-\alpha_{t-1})$$

$$n_1+n_2/4 = N_{t-1}(p-p_1/4) = N_{t-1}\ p-p(1-p)(1-\alpha_{t-1}/4)\ .$$

Substituting the above relations into (A.6) and simplifying we obtain

$$\frac{2N_{t-1}\bar{k}}{p(1-p)}\,V_{\delta p} = \frac{N_{t-1}}{N_{t-1}-1}\ \frac{V_k}{k}\,(1-\alpha_{t-1})+(1-\alpha_{t-1}).$$

Since $V_{\delta p}$ has the value $p(1-p)/2N_e$ in an ideal population we obtain (Kimura and Crow 1962).

$$N_e = 2N_t/(1-\alpha_{t-1}+(1+\alpha_{t-1})S_k^2/\bar{k})$$

where $S_k^2 = \Sigma(k-\bar{k})^2/(N_{t-1}-1)$, for the variance effective number.

In the preceeding brief derivations a monoecious diploid population was assumed. Analagous results may be obtained assuming separate sexes. Under the assumption of separate sexes the inbreeding and variance effective numbers become respectively

$$N_e = 1/p_t = (N_{t-2}\bar{k}-2)/(\bar{k}-1+V_k/\bar{k})$$

and

$$N_e = 2N_t/(1-\alpha_{t-1}+(1+\alpha_{t-1})S_k^2/\bar{k})\,.$$

The above variance effective number was derived assuming the progeny distributions are the same for both sexes of parents.

APPENDIX B
Method of Path Coefficients

The method of path coefficients was developed primarily by Sewall Wright. It has proved to be an effective technique for many problems in theoretical genetics and statistical analysis of cause and effect in a system of correlated variables.

Suppose the variable X_0 is determined completely and linearly by the variables $X_1, X_2, \ldots, X_n$. Also suppose that all variables are measured from their respective means. Using linear regression methods the following relationship may be realized.

(B.1) $$X_0 = \beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n,$$

where the β's are the respective regression coefficients. Let $x_i = X_i/\sigma_{ii}$, then (B.1) becomes

$$x_0\sigma_{00} = \beta_1 x_1\sigma_{11}+\beta_2 x_2\sigma_{22}+\ldots+\beta_n x_n\sigma_{nn}$$

or

$$x_0 = \beta_1 \frac{\sigma_{11}}{\sigma_{00}} x_1+\beta_2 \frac{\sigma_{22}}{\sigma_{00}} x_2+\ldots+\beta_n \frac{\sigma_{nn}}{\sigma_{00}} x_n.$$

Put $\theta_i = \beta_i\sigma_{ii}/\sigma_{00}$, then

(B.2) $$x_0 = \theta_1 x_1+\theta_1 x_2+\ldots+\theta_n x_n.$$

The θ's are defined as path coefficients and are equivalent to standard partial regression coefficients. A standard partial regression coefficient of x_0 on x_i will be a path coefficient when all variables included in the regression equation are causes of x_o and all relevant causes are included.

Let $V(x_i)$ denote the variance of the ith variable and r_{ij} the correlation between the ith and jth variables. Squaring both sides of (B.2) and taking expectations we have

$$E(x_0^2) = \theta_1^2 E(x_1^2)+\ldots+\theta_n^2 E(x_n^2)+\sum_{i\neq j} \theta_i \theta_j E(x_i x_j).$$

Since all variables are standardized, $E(x_i^2) = V(x_i) = 1$ and $E(x_i x_j) = r_{ij}$. Thus,

$$\theta_1^2+\theta_2^2+\ldots+\theta_n^2+\sum_{i\neq j} \theta_i \theta_j r_{ij} = 1 \tag{B.3}$$

and if all variables are independent

$$\theta_1^2+\theta_2^2+\ldots+\theta_n^2 = 1.$$

The quantities θ_1^2, θ_2^2, ... were termed coefficients of determination by Wright. If x_2, ..., x_n are kept constant such that x_1 maintains the same amount of variability as before, then the variability in x_0 would be θ_1^2 or equivalently the variance in X_0 would be $\theta_{x_0}^2 \theta_1^2$. In this way θ_1^2 measures the proportion of variability in X_0 which is directly attributable to X_i.

Consider two variables X and Y which are the result of n_1 common causes (A_i), n_2 causes (B_i) which effect only variable X and n_3 causes (C_i) which effect only variable Y. Let n_4 of these causes be correlated and let $n = n_1+n_2+n_3$. Considering all variables in standardized form, the X and Y variables may be represented by

$$X = b_1B_1+b_2B_2+\ldots+b_{n_2}+a_1A_1+a_2A_2+\ldots+a_{n_1}A_{n_1}$$

$$Y = c_1C_1+c_2C_2+\ldots+c_{n_3}C_{n_3}+a_1'A_1+a_2'A_2+\ldots+a_{n_1}A_{n_1},$$

where the small case letters denote the corresponding path coefficients.

Multiplying the expressions for X and Y given above we obtain

(B.4) $$XY = \sum^{n_3}\sum^{n_2} c_i C_i b_j B_j + \sum^{n_3}\sum^{n_1} c_i C_i a_j A_j + \sum^{n_1}\sum^{n_2} a'_i A_i b_i B_j + \sum^{n_1}\sum^{n_1} a'_i A_i a_i A_j$$

Since all variables are standardized $E(XY) = r_{xy}$, $E(C_iB_j) = r_{C_iB_j}$, $E(C_iA_j) = r_{C_iA_j}$, $E(A_iB_j) = r_{A_iB_j}$ and $E(A_iA_j) = r_{A_iA_j}$. Thus, taking the expectation of (B.4) we obtain

(B.5) $$r_{xy} = \Sigma\Sigma c_i b_i r_{C_iB_j} + \Sigma\Sigma c_i a_j r_{C_iA_j} + \Sigma\Sigma a'_i b_j r_{A_1B_j} + \Sigma\Sigma a'_i a_j r_{A_iB_j}.$$

Now, $(n-n_4)$ of the above correlation coefficients will be zero. Accordingly, (B.5) may be written in the form

(B.6) $$r_{xy} = a'_1 a_1 + a'_2 a_2 + \ldots + a'_{n_1} a_{n_1} + \sum_{i=1}^{n_4} \rho_i r_i ,$$

where ρ_i denotes the product of the path coefficients for the ith correlation coefficient.

Rather than work directly with standardized equations Wright advocated applying the method of path coefficients directly to a path diagram in which the dependent variables are represented as additively and completely determined by others.

It should be noted that expression (B.3) is a generalization of Theorem 2.2 and expression (B.6) is equivalent to Theorem 2.1.

APPENDIX C
Inbreeding Coefficient F

Inbreeding is a genetic term referring to a mating system in which mated individuals are more closely related than with random members of the population as a whole. The inbreeding coefficient F has been defined as the probability that two genes at a given locus are identical. In this definition the term identical refers to two genes being identical because they are copies arising in the reproductive process of one gene occuring in some previous generation. In general, inbreeding in natural populations will lead to a correlation between uniting gametes.

Consider a single locus with two alleles A and a which occur in the proportions p and q respectively. Under random mating the probability that two A gametes will unite is p^2. If a certain degree of inbreeding exists in the population then the probability that two A gametes will unite will be greater than p^2. Let $p^2+\varepsilon$ be this probability where $0<\varepsilon<1-p^2$. Let g represent the gametes from one sex and g' the gametes from the other sex, such that $g=g'=1$ for A gametes and $g=g'=0$ for a gametes. From Table 1 we have $g=g'=p$, $\sigma_g^2=\sigma_{g'}^2=pq$ and $\sigma_{gg'}^2=\varepsilon^2$. Therefore, the correlation between uniting gametes is, by definition,

$$F = \sigma_{gg'}/\sigma_g\sigma_{g'} = \varepsilon/pq$$

or

$$\varepsilon = Fpq.$$

With a given degree of inbreeding the zygotic proportion will be (Wright 1922)

$$\begin{array}{llll} & AA & Aa & aa \\ (C.1) & p^2+Fpq & 2pq(1-F) & q^2+Fpq. \end{array}$$

If the inbreeding coefficient F remains constant from generation to generation

Table 1. Correlations between uniting gametes.

q' q	1	0	total
1	$p^2+\varepsilon$	$pq-\varepsilon$	p
0	$pq-\varepsilon$	$q^2+\varepsilon$	q
	p	q	1

the zygotic proportions will also remain constant. Expression (C.1) is a generalization of the Hardy-Weinbery Law in which a random mating population (F=0) was assumed and is generally referred to as Wright's equilibrium law.

BIBLIOGRAPHY

Crow, J. F. and Kimura, M., 1956. Some Genetic Problems in Natural Populations. Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability 4:1.

Crow, J.F. and Kimura, M., 1963. The Measurement of Effective Population Number. Evolution 17:279.

Crow, J. F. and Morton, N. E., 1955. Measurement of Gene Frequency in Small Populations. Evolution 9:202.

Dobzhansky, T. and Wright, S., 1941. Genetics of Natural Populations. Genetics 26:23.

Fisher, R. A., 1922. On The Dominance Ratio. Proceedings of the Royal Society, Edinburgh 42:321.

Feller, W., 1951. Diffusion Processes in Genetics. Proceedings Second Berkeley Symposium on Mathematical Statistics and Probability 2:227.

Haldane, J. B. S., 1924. A Mathematical Theory of Natural and Artificial selection. Proceedings of the Cambridge Philosophical Society 23:19

Kimura, Motoo, 1951. Effect of Random Fluctuation of Selective Value on the Distribution of Gene Frequencies in Natural Populations. National Institute of Genetics Annual Report (Japan) 1:45.

Kimura, Motoo, 1952. On the Process of Decay of Variability Due to Random Extinction of Alleles. National Institute of Genetics Annual Report (Japan) 2:60.

Kimura, Motoo, 1954. Process Leading to Quasi-fixation of Genes in Natural Populations Due to Random Fluctuation of Selection Intensities. Genetics 39:280.

Kimura, Motoo, 1955a. Solution of a Process of Random Genetic Drift with a Continuous Model. Proceedings of National Academy of Science 41:144.

Kimura, Motoo, 1955b. Random Genetic Drift in a Multi-allelic Locus. Evolution 9:419.

Kimura, Motoo, 1955c. Stochastic Processes and Distirbution of Gene Frequencies under Natural Selection. Cold Springs Harbor Symposium 20:33.

Kimura, Motoo, 1956. Random Genetic Drift in a Tri-allelic Locus; Exact Solution with a Continuous Model. Biometrics 12:57.

Kimura, Motoo and Weiss, G. H., 1964. The Stepping Stone Model of Population Structure and the Decrease of Genetic Correlation with Distance. Genetics 49:561.

Li, C. C., 1955. Population Genetics. Chicago: University of Chicago Press.

Weiss, G. H. and Kimura M., 1965. A Mathematical Analysis of the Stepping Stone Model of Genetic Correlation. Journal of Applied Probability 2:129.

Wright, S., 1931. Evolution in Mendelian Populations. Genetics 16:97.

Wright, S., 1937. The Distribution of Gene Frequencies in Populations. Proceedings National Academy of Science 23:307.

Wright, S., 1942. Isolation by Distance. Genetics 28:114.

Wright, S., 1945. The Differential Equation of the Distribution of Gene Frequencies. Proceedings National Academy of Science 31:382.

Wright, S., 1951. The Genetical Structure of Populations. Annals of Eugenics 15:323.

Wright, S., 1952. The Theoretical Variance Within and Among Subdivisions of a Population That is in a Steady State. Genetics 37:312.

THE GENETIC STRUCTURE OF NATURAL POPULATIONS

by

RALPH DENNIS COOK

B. S., Northern Montana College, 1967

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics and Computer Science

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1969

The genetic structure of natural populations is considered primarily under three models. The "island" and "continuum" models were developed by Wright and represent two extreme forms of population structure. The "stepping stone" model was proposed by Kimura and Weiss. This model is a rough combination of Wright's two models and probably represents a natural population better than either of them. Wright analized his continuum model by considering the correlation between uniting gametes; Kimura and Weiss, through their stepping stone model, examined the correlation of gene frequencies between two colonies which are a given distance apart. These two methods of considering correlation appear to be equivalent.

The distribution of gene frequencies is considered in two ways. First, the probability distribution is derived under the island model assuming a balance between random deviations and systematic pressures. This steady state distribution is also shown to be applicable to the continuum and stepping stone models. The effects of migration, mutation and selection are considered, both jointly and independently, on the steady state distribution. Second, the distribution of gene frequencies before a steady state has been reached is considered by treating the change in gene frequency as a stochastic process. Kimura has shown that this treatment of change in gene frequency leads to the Fokker-Planck equation. Using this equation the effects of random sampling of gametes and random fluctuations of selection pressures are isolated.

Consideration of the effective population number, the method of path coefficients and the inbreeding coefficient is necessary to the development of the three models presented. Brief treatments of these three topics are presented in the appendices.