GENETIC DRIFT--A STOCHASTIC PROCESS

by

CLEMENT JOHN MAURATH

A. B., Fort Hays Kansas State College, 1965

A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Approved by:

H C Fryer

Major Professor

TABLE OF CONTENTS

INTRODUCTION

From the standpoint of population genetics, one of the most elementary steps in evolution is the change in gene frequency, especially the change due to natural selection. Since there exist various factors which introduce an element of indeterminacy into the process, it is not difficult to imagine that the process is continuous. One of these factors, random sampling of gametes due to finite population size, is of special interest. There are also systematic pressures that affect gene frequency. Among these are selection, migration, and mutation. The change due to selection is controlled by the amount of selection, or selection intensity. It is also found that there exists a fluctuation of these selection intensities from generation to generation. These two points of interest, random sampling of gametes and fluctuation of selection intensities, cause a phenomenon known as genetic drift.

Genetic drift due to random sampling of gametes will cause the gene in question to become either completely fixed or completely lost from the population and will approach one of these limits asymptotically. In reaching one or the other of these limits, the gene frequency varies as a stochastic process (see Fig. 1).

Genetic drift due to fluctuation of selection intensities also approaches either fixation or loss asymptotically. But for this case the gene frequency will become fixed before it reaches complete homozygosity (see Fig. 2). Thus if we have a pair of alleles A_1 and A_2 and genotypes A_1A_1, A_1A_2, and A_2A_2, after a

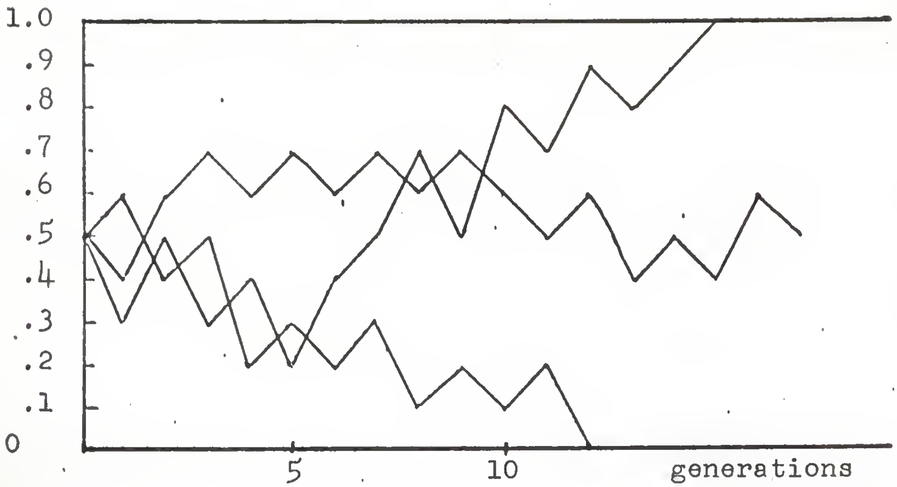

Fig. 1. Three examples of genetic drift due to random sampling of gametes in finite populations. Original gene frequency is .5.

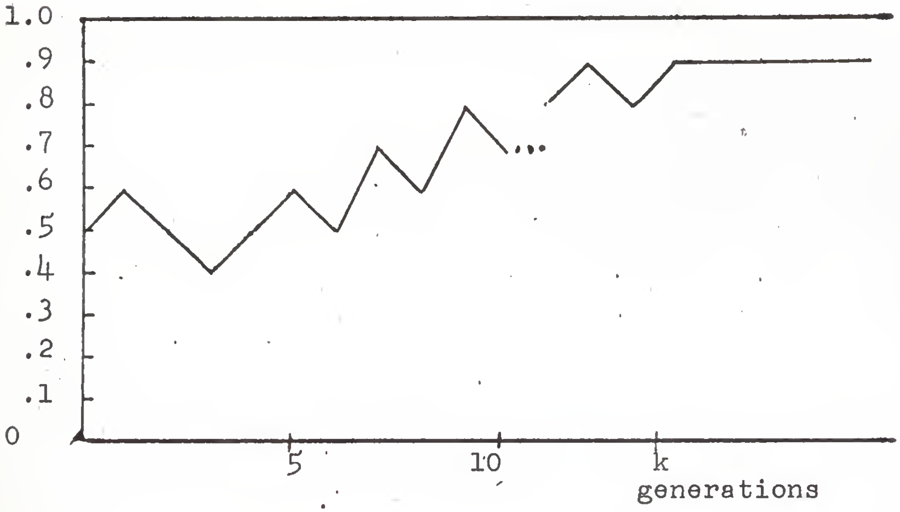

Fig. 2. Example of genetic drift due to fluctuation of selection intensity, reaching fixation at about .9 after k generations.

certain number of generations, the genotypes will become fixed as A_1A_1 and A_2A_2. On the other hand, drift due to random sampling of gametes will produce all A_1A_1 or all A_2A_2. In both cases all heterozygotes will be lost.

Mathematical treatments will be presented for these cases of genetic drift.

HISTORICAL BACKGROUND

Hagedoorn Effect

Appearing in 1921 was some of the first mathematics dealing with genetic drift. Fisher (1921) proposed the following: If p is the proportion of any gene, and q is the frequency of its allelomorph, then in N individuals of any generation we have 2Np genes scattered at random. Let $\cos \theta = 1 - 2p$, where $0 < \theta < \pi$. For a second generation of N individuals formed at random, the standard deviation of p will be

$$\sigma p = \sqrt{\frac{pq}{2N}} \; ,$$

thus

$$\sigma\theta = \sqrt{\frac{pq}{2N}} \; \frac{d\theta}{dp} = \frac{1}{\sqrt{2N}} \; .$$

Since this is independent of θ, Fisher calculated the changes in the distribution of θ, in the absence of selection. If $y(\theta)d\theta$ represents the distribution of θ in any one generation, the distribution in the next generation will be

$$y + \Delta y = \int_0^{\pi} \frac{1}{\sqrt{2\pi\sigma}} e^{-\delta\theta^2/2\sigma^2} (y + y'\delta\theta + \frac{\delta\theta^2}{2!} y'' + \ldots)$$

$$= y + \frac{\sigma^2}{2} y'' + \ldots \qquad (1)$$

since $\sigma_\theta^2 = 1/2N$ is very small. The rate of change of $y(\theta)$ is given by

$$\frac{\partial y}{\partial T} = \frac{1}{4N} \frac{\partial^2 y}{\partial \theta^2} \ . \qquad (2)$$

Since no distinction has been drawn between the gene and its allelomorph, the above solutions are symmetric. The stationary case is $y = A/\pi$, where A is the number of factors present (unfixed loci).

Fisher explains that when y is increasing,

$$y = A_0 e^{kT} \frac{p}{2 \sinh \frac{1}{2} p\pi} \cosh p(\theta - \frac{\pi}{2}) \qquad (3)$$

and when y is decreasing,

$$y = A_0 e^{-kT} \frac{p}{2 \sin \frac{1}{2} p\pi} \ . \cos p(\theta - \frac{\pi}{2}) \qquad (4)$$

where

$$k = \frac{p^2}{4N} \ . \qquad (5)$$

Gene Extinction Due to Drift. Fisher (1921) represents by e^{-h} the chance that a particular gene borne by a single individual will not be represented in the next generation. The chance of extinction for a factor of which b genes are in existence will be e^{-bh}. When θ is near zero, p, which is always equal to $\sin^2 \frac{\theta}{2}$, will be nearly equal to $1/4\ \theta^2$. Let $t = \sin 1/2\ \theta$, then the number of genes in existence is $2Nt^2$ and the chance of their extinction in one generation is e^{-2Nht^2}.

This chance is negligible except when t is very small and may be equated to $1/2\ \theta$; hence the number of genes exterminated in any one generation is

$$2 \int_0 y e^{-2Nht^2}\, d\theta = 4 \int_0 y e^{-2Nht^2}\, dt\ . \qquad (6)$$

In the stationary case, $y = A/\pi$ and the number of genes exterminated will be

$$\frac{A}{\pi}\ \frac{2\sqrt{2\pi}}{\sqrt{4hN}} = A\sqrt{\frac{2}{\pi hN}}\ ;$$

if new mutations occur at rate $N\mu$, then gene frequency equilibrium will occur at

$$A = \sqrt{\frac{\pi h}{2}}\, N^{3/2}\ \mu\ .$$

In the absence of mutation, there is extinction, and the number of factors diminishes. Considering equation (4) when θ

is small, one gets

$$\cos p(\theta - \frac{\pi}{2}) = \cos \frac{1}{2} p\pi + 2p \sin \frac{1}{2} p\pi$$

$$\cdot t - 2p^2 \cos \frac{1}{2} p\pi \cdot t^2 \ldots$$

so the rate of extinction is:

$$A_0 e^{-kT} \frac{p}{2 \sin \frac{1}{2} p\pi} \sqrt{\frac{2\pi}{hN}} \left[\cos \frac{1}{2} p\pi + 2p \sin \frac{1}{2} p\pi \sqrt{\frac{2}{2hN}} \right]$$

giving

$$k = p \sqrt{\frac{2\pi}{hN}} \left[\frac{1}{2} \cot \frac{1}{2} p\pi + \frac{p}{2\pi hN}\right].$$

Equating this to (5) gives:

$$\frac{p^2}{N}\left[\frac{1}{4} - \frac{1}{h}\right] = \sqrt{\frac{2\pi}{hN}} \frac{p}{2} \cot \frac{1}{2} p\pi$$

when $\cot \frac{1}{2} p\pi$ is of the order of $\frac{1}{\sqrt{N}}$, so that p is near 1, $1 = 1/4N$.

Hagedoorn (1921) was one of the first to indicate this random effect and it was so named by Fisher, "The Hagedoorn Effect". Fisher's value of 1/4N was later disproved by Wright which will be discussed in the following section.

Sewall Wright Effect

In a paper in 1921, Wright gave a general method for determining the decrease in heterozygosis. He stated that for two alleles per locus the rate of loss per generation is 1/2N in the case of a breeding population of N individuals either equally divided between males and females or composed of monoecious individuals. This is different from the result given by Fisher above and will be explained later in this section.

Wright expanded on the subject in 1931 and gave these results.

Consider a population in which there are N_m breeding males and N_f breeding females, and random mating. The proportion of matings between full brother and sister will be $\frac{1}{(N_m N_f)}$, that between half brother and sister $\frac{(N_m + N_f - 2)}{(N_m N_f)}$, and that between less closely related individuals $\frac{(N_m - 1)(N_f - 1)}{(N_m N_f)}$. The correlation between mated individuals may be written, giving due weight to these three possibilities.

$$M = a'^2 b'^2 \left[\frac{1}{N_m N_f} (2 + 2M') + \frac{N_m + N_f - 2}{N_m N_f} (1 + 2M') + \frac{(N_m - 1)(N_f - 1)}{N_m N_f} 4M' \right] , \qquad (7)$$

where $a = \sqrt{\dfrac{1}{2(1 + F)}}$ is the path coefficient, gamete to fertilized egg, $b = \sqrt{1/2(1 + F')}$ is the path coefficient, zygote to gamete, and F is the correlation between uniting egg and sperm, and where primes are used to indicate the number of generations preceding the one in question. The proportional change in heterozygosis is given by:

$$F = F' + \frac{N_m + N_f}{8N_mN_f} (1 + 2F' + F'') .$$

The proportion of heterozygosis

$$P = P' - \frac{N_m + N_f}{8N_mN_f} (2P' - P'') .$$

It is to be expected that the proportional change per generation will reach approximate constancy. This rate was found by equating P/P' to P/P" to be:

$$-\frac{\triangle P}{P'} = \frac{1}{2}\left(1 + \frac{N_m + N_f}{4N_mN_f}\right) - \frac{1}{2}\sqrt{1 + \left(\frac{N_m + N_f}{4N_mN_f}\right)^2} .$$

This gives for small populations

$$\left(\frac{1}{8N_m} + \frac{1}{8N_f}\right)\left(1 - \frac{1}{8N_m} - \frac{1}{8N_f}\right)$$

as a close approximation, and for large populations

$$\frac{1}{8N_m} + \frac{1}{8N_f} .$$

For the simplest case of mating brother with sister or $N_m = N_f = 1$, the rate of loss of heterozygosis is $1/4(3 - \sqrt{5})$, or 19.1 per cent per generation. For the case $N_m = 1$ and $N_f \doteq \infty$, the rate of loss is about 11 per cent per generation. For a more useful case in which there is a relatively limited number of males but unlimited number of females, the rate becomes $1/8\ N_m$. An especially important case is the population which is equally divided, or $N_m = N_f = 1/2\ N$. In this case the rate is $1/(2N + 1)$, or approximately $1/2\ N$.

If only random mating cases are considered, then gametes have a chance $1/N$ of coming from the same individual and $(N - 1)/N$ of coming from different individuals. The correlation between uniting gametes may then be written

$$F = \frac{1}{N} b^2 + \left(\frac{N - 1}{N}\right) 4b^2 a'^2 F'$$

and

$$P = \frac{(2N - 1)}{2N} P' .$$

This result does not differ appreciably from that of the preceding case. The rate of loss is exactly $1/2\ N$ instead of

$$\frac{1}{2N + 1} .$$

The simplest case is continued self-fertilization in which $N = 1$ and the formula gives 50 per cent loss per generation, as would be expected.

In order to determine generally the distribution of gene frequencies, Wright (1931) considers the way in which genes A_1 with frequency q are distributed after one generation of random mating. In a population of N breeding individuals, each of the specified genes will have 2Nq representatives among the zygotes and their allelomorphs 2N(1 - q). A random sample of the same size will be distributed according to the expression

$$\left[(1 - q)A_2 + qA_1\right]^{2N} . \tag{8}$$

The contribution of this sample to the frequency class with an allelomorphic ratio $q_1:(1 - q_1)$ will be in proportion to the $2Nq_1{}^{th}$ term of the above expression and to the number of genes included in the contributing class (f). The sum of contributions from all such classes should give the $2Nq_1{}^{th}$ term an absolute frequency which is smaller than its value in the preceding generation (f_1) by the amount 1/(2N + 1), as given above. Thus the following equation is given to solve for f as a function of q.

$$f_1(1 - \frac{1}{2N + 1}) = \frac{(2N)!}{(2Nq_1)!(2N(1 - q_1))!} \sum q^{2Nq_1}(1 - q)^{2N(1-q_1)} f$$

Let f = $\phi(q)/2N = \phi(q)dq$, and replacing summation by integration, the result is:

$$\frac{\phi(q_1)}{2N + 1} = \frac{(2N)!}{(2Nq_1)!(2N(1 - q_1))!} \int_0^1 q^{2Nq_1}(1 - q)^{2N(1-q_1)}\phi(q)\,dq \tag{9}$$

The cases of 2 and 3 monoecious individuals as worked out by simple algebra suggests an approach to a uniform distribution. As a trial, let $\phi(q) = C$. This is a solution since

$$\frac{C}{2N + 1} = \frac{C(2N)!}{(2Nq_1)!(2N(1 - q_1))!} \quad \frac{\Gamma(2Nq_1 + 1)\ \Gamma(2N - 2Nq_1 + 1)}{\Gamma(2N + 2)} .$$

It would appear that after a cross mating the gene frequencies will spread out from 50 per cent toward fixation or loss until a practically uniform distribution is reached. The frequencies of all classes will then decrease at a rate of about 1/2N as 1/4N of the genes become fixed and 1/4N become lost per generation if $q = 1/2$ initially.

Wright (1931) points out that we must examine the terminal points before fully accepting this solution. The amount of fixation at the extremes, if N is large, can be found directly from the Poisson series. The contribution to the zero class when the mean number in the sample is $e^{-m}(m = 1, 2, 3, \ldots)$ is:

$$(e^{-1} + e^{-2} + e^{-3} + \ldots)f = \frac{e^{-1}}{1 - e^{-1}} f = .582\ f.$$

This is larger than 1/2 f as stated above, but is attributed to the distortion near the ends due to the element of approximation involved in using integration for summation.

If mutation is occurring, however low the rate, the decline in heterozygosis cannot go on indefinitely. There will come a time when the chance elimination of genes will be exactly balanced by new genes arising by mutation. The equation to be solved is equation (9). By trial and error, Wright (1931) finds $\emptyset(q) = C_1q^{-1} + C_2(1 - q)^{-1}$ as a solution. The terminal condition, reduction of the class of fixed genes by an occasional mutation contributing to the class $q = (2N - 1)/2N$, necessarily involves the appearance of new genes contributing to the class $q = 1/2N$. This means that only the symmetrical solution $\emptyset(q) = Cq^{-1}(1 - q)^{-1}$ can be accepted as descriptive of the distribution of the entire array of genes at equilibrium, provided there is no selection, migration, or recurrence of the same mutation. Thus letting

$$f = \frac{C}{2N}\, q^{-1}(1 - q)^{-1} \quad \text{and} \quad \sum f = 1,$$

$$C = \frac{1}{2\left[.577 + \log(2N - 1)\right]} \cong \frac{1}{2 \log 3.6N} \,. \qquad (10)$$

Before attainment of equilibrium with respect to heterozygosis the distribution will pass through phases of approximately the form $\emptyset(q) = C_1q^{-1}(1 - q)^{-1} + C_3$, in which the term C_1 gradually displaces C_3 as the number of temporarily fixed genes approaches equilibrium with mutation. As the chance of complete fixation increases, the chance of mutation must be taken into account. The distribution passes through phases of the type $C_2(1 - q)^{-1} + C_3$, C_2 gradually displacing C_3, relatively, but itself declining as the chance of complete loss increases.

If there is reverse mutation, but at a very slow rate, a term C_1q^{-1} must be added to the formula, and an equilibrium will be reached in the form $Cq^{-1}(1 - q)^{-1}$. Thus in the long run, the gene will be completely fixed or completely absent from the population with frequencies proportioned to the mutation rates to and from the gene respectively. Occasionally these conditions will not be quite complete and at extremely rare intervals the gene will drift from one state to the other.

The turnover among genes in equilibrium in the distribution $Cq^{-1}(1 - q)^{-1}$ can be determined by consideration of the variance of q and independently by application of the Poisson law. Let $\sigma q^2 = \dfrac{\sum(q - 1/2)^2 f}{\sum f}$ be the variance of q, excluding the terminal classes. This variance is increased in the following generation by the spreading out of each frequency class as a result of random sampling. The variance from the spreading of a single class is $q(1 - q)/2N$ and the average is thus

$$\Delta\sigma q^2 = \frac{\sum q(1 - q)f}{2N\sum f} = \frac{1}{2N}\left(\frac{1}{4} - \sigma^2 q\right) = \frac{2N - 1}{(2N)^2} C ,$$

where C is as in (10). The sum $\sigma q^2 + \Delta\sigma q^2$ includes the newly fixed factors whose contribution is 1/4 k where k is the rate of fixation, plus loss, but excludes mutation. The contribution of the new mutations to the variance is

$\dfrac{k(N - 1)^2}{(2N)^2}$; therefore

$$\sigma q^2 + \Delta\sigma q^2 - \frac{1}{4} k + k \left(\frac{N - 1}{2N}\right)^2 = \sigma q^2$$

$$K = C = \frac{1}{2 \log 3.6N} .$$

The proportion exchanged at each extreme is thus about half the corresponding subterminal class when N is large. This compares with the proportion as determined by the Poisson law, which is .46 times the subterminal class instead of .50.

Referring to Fisher's equations, (1) and (2), Wright made the following remarks. He claims that equation (2) gives the wrong solution, and he also points out a comparison of the equations. He states that in a breeding population of one million with one mutation per 1000 individuals, Fisher's formula $\sqrt{\pi/2}\ N^{3/2}\ \mu$ gives 1,250,000 unfixed factors with a turnover of .08 per cent, while his formula $2N\mu \log 3.6N$, gives 30,000 unfixed factors and a turnover of 3.3 per cent.

Fisher yielded to Wright, and Wright (1931) printed a note from Fisher to this effect. Equation (2) should have read

$$\frac{\partial y}{\partial T} = \frac{1}{4N} \frac{\partial}{\partial \theta} (y \cot \theta) + \frac{1}{4N} \frac{\partial^2 y}{\partial \theta^2}$$

and with this he agrees with Wright's value of 1/2N.

MODERN APPROACH TO GENETIC DRIFT

Kimura's Treatment of Random Genetic Drift Due to Random Sampling of Gametes

As was noted before, Fisher (1921) and Wright (1931) gave solutions to this problem. Fisher used differential equations and Wright used differential equations and path coefficients.

Kimura (1955c) states that in these works it was assumed that a state of steady decay had been reached. Nothing was known about the complete solution which might show how the process finally leads to the state of steady decay. Kimura showed that the process approaches asymptotically the state of steady decay by finding the moments of the distribution and using the Fokker-Planck equation.

Again considering a finite random mating population of N diploid parents, where A_1 and A_2 are a pair of alleles with frequencies x and 1 - x, respectively, when there is no selection, mutation, or migration, Kimura (1955c) states that an adequate description of the change in gene frequencies is to give the frequency distribution f(x, t) at the t^{th} generation, where x takes on a series of discrete values: 0, 1/2N, 2/2N, . . ., 1 - 1/2N, 1. Without serious error, x can be considered continuous for large N.

First of all, Kimura (1954) gave as the n^{th} moment of the distribution about zero:

$$\mu_n'(t) = p - 3pq \frac{n-1}{n+1} (1-\lambda_1)^t - 5pq(p-q) \frac{(n-2)(n-1)}{(n+1)(n+2)} (1-\lambda_2)^t$$

$$- 7pq(-5pq + 1) \frac{(n-3)(n-2)(n-1)}{(n+1)(n+2)(n+3)} (1 - \lambda_3)^t$$

$$-9pq(14pq^2 - 7pq + p - q) \frac{(n-4)(n-3)(n-2)(n-1)}{(n+1)(n+2)(n+3)(n+4)} (1-\lambda_4)^t$$

$$+ \alpha \left[(1 - \lambda_5)^t\right] \qquad (11)$$

where $q = 1 - p$ and $\lambda_i = \dfrac{i(i + 1)}{4N}$, $i = 1, 2, ...$

Using a more sophisticated method, Kimura (1955a) presented the following: Let x_t be the gene frequency in the t^{th} generation, and let δx_t be the amount of change due to random sampling of gametes per generation, such that

$$x_{t+1} = x_t + \delta x_t \; . \qquad (12)$$

Let $\mu_n'^{(t+1)} = E(x_{t+1}^n)$ be the n^{th} moment of the distribution about zero in the $(t + 1)^{th}$ generation. He then writes $E(x_{t+1}^n)$ in terms of $(x_t + \delta x_t)^n$. This is done in two steps; first, taking expectation for the random change δx_t, which will be denoted by E_δ, and, second, taking the expectation for the existing distribution, denoted by E_*.

Note that $E_\delta(\delta x_t) = 0$, $E_\delta(\delta x_t)^2 = x_t(1 - x_t)/2N$, etc., so

$$\mu_n{}'^{(t+1)} = E(x_t + \delta x_t)^n$$

$$= E_*\left[x_t^n + \binom{n}{1} x_t^{n-1} E_\delta(\delta x_t) + \binom{n}{2} x_t^{n-2} E_\delta(\delta x_t)^2 + \ldots\right]$$

$$= E_*\left[x_t^n + \frac{n(n-1)}{2} x_t^{n-2} \frac{x_t(1-x_t)}{2N} + \alpha\left(\frac{1}{N^2}\right)\right] \quad , \qquad (13)$$

assuming that N is large enough so that terms of $1/N^2$ and higher can be omitted without serious error. The equation is then:

$$= \left[1 - \frac{n(n-1)}{4N}\right] \mu_n{}'^{t} + \frac{n(n-1)}{4N} \mu'^{(t)}_{n-1} \; . \qquad (14)$$

For large N the moments change very slowly so equation (14) is replaced by the system of differential equations.

$$\frac{d\mu'^{(t)}_n}{dt} = - \frac{n(n-1)}{4N} \left[\mu'^{(t)}_n - \mu'^{(t)}_{n-1}\right] \; , \; (n = 1, 2, 3, \ldots) \; . \qquad (15)$$

If the population starts from gene frequency p $(0 < p < 1)$, $\mu'^{(0)}_n = p^n$ and the nth moment is a solution of (15).

$$\mu'^{(t)}_n = p + \sum_{i=1}^{\infty} (2i+1)pq(-1)^i F(1-i, i+2, 2, p)$$

$$\times \frac{(n-1) \ldots (n-i)}{(n+1) \ldots (n+i)} e^{-\left[i(i+1)/4N\right] t} \qquad (16)$$

where F(1 - i, i + 2, 2, p) is the hypergeometric function. For finite n the series is finite.

He next derives the probability $f(1, t)$ of the gene A_1 becoming fixed in the population by the t^{th} generation. Note that

$$f(1, t) = \lim_{n \to \infty} \sum_{x=0}^{1} x^n f(x, t) = \lim_{x \to \infty} \mu'^{(t)}_n .$$

He now has an infinite series

$$f(1,t) = p + \sum_{i=1}^{\infty} (2i+1)pq(-1)^i F(1-i,i+2,2,p)e^{-[i(i+1)/4N]t} \tag{17}$$

whose convergence must be examined. At this point he introduces the Gegenbauer polynomial $T^1_{i-1}(z)$ which is related to the hypergeometric function by

$$T^1_{i-1}(z) = \frac{i(i+1)}{2} F(i+2, 1-i, 2, \frac{1-z}{2}) .$$

Using this relation and putting $p = \frac{(1-r)}{2}$, where $(-1 < r < 1)$, he obtains:

$$f(1,t) = p + \sum_{i=1}^{\infty} (-1)^i \frac{(2i+1)}{2i(i+1)} (1-r^2)T^1_{i-1}(r)e^{-[i(i+1)/4N]t} . \tag{18}$$

Using the recurrence relation,

$$(2i+1)(1-r^2)T^1_{i-1}(r) = i(i+1)P_{i-1}(r) - i(i+1)P_{i+1}(r) \tag{19}$$

the above formula becomes:

$$f(1,t) = p + \sum_{i=1}^{\infty} \frac{(-1)^i}{2} \left[P_{i-1}(r) - P_{i+1}(r)\right] e^{-\left[i(i+1)/4N\right] t} \qquad (20)$$

where $P_n(r)$ represents a Legendre polynomial. For $t = 0$, the partial sum of the first n terms of equation (20) is $(-1)^{n-1} \frac{(P_{n-1} - P_n)}{2}$ $(n \geq 3)$.

To obtain the probability of gene A_2 being fixed, f(0, t) is obtained by replacing p with q and r with -r.

In the notation of equation (11) he has

$$\begin{aligned} f(1,t) = p &- 3pq(1-\lambda_1)^t - 5pq(p-q)(1-\lambda_2)^t \\ &- 7pq(-5pq+1)(1-\lambda_3)^t - 9pq(14pq^2-7pq+p-q)(1-\lambda_4)^t \\ &+ \alpha \left[(1 - \lambda_3)^t\right] \end{aligned} \qquad (21)$$

and again f(0, t), the probability of complete loss, is found by replacing p by q.

He now has the probability for the fixed classes and he makes this statement

$$f(1,t)+f(0,t) = 1 - \sum_{j=0}^{\infty} \left[P_{2j}(r) - P_{2j+2}(r)\right] e^{-\left[(2j+1)(2j+2)/4N\right] t} \qquad (22)$$

which is 0 when t = 0 and tends to 1 when $t \to \infty$.

He then considers the probability distribution of unfixed classes. The variance of the rate of change in gene frequency due to random sampling of gametes is $V_{\delta x} = \frac{x(1 - x)}{2N}$. So if $\phi(x, t)$ is the relative probability that the frequency of the gene in the population will take any value between x and x + dx (0 < x < 1) in the tth generation, $\phi(x, t)$ satisfies the partial differential equation derived from equation (49). (See Appendix.)

$$\frac{\partial \phi}{\partial t} = \frac{1}{4N} \frac{\partial^2}{\partial x^2} \left[x(1 - x)\phi\right] \tag{23}$$

To solve this equation he uses $\phi \propto X_i(x) e^{-\lambda_i t}$ or $X_i(x) e^{-\left[i(i+1)/4N\right] t}$ and this gives the hypergeometric equation

$$x(1-x) \frac{d^2 X_i}{dx^2} + 2(1-2x) \frac{dX_i}{dx} - (1-i)(i+2)X_i = 0$$

or using x = (1 - z)/2 such that z = 1 - 2x gives Gegenbauer equation

$$(z^2-1) \frac{d^2 X_i}{dz^2} + 4z \frac{dX_i}{dz} - (i-1)(i+2)X_i = 0 \tag{24}$$

Looking at equations (16) and (17), he derives the moment formula

$$\mu'^{(t)}_n - 1^n f(1, t) = \int_0^1 x^n \phi(x, t)dx$$

which suggests a solution of equation (23) of the form

$$\sum_{i=1}^{\infty} C_i X_i(x) e^{-\left[i(i+1)/4N\right] t} . \tag{25}$$

Comparing this with equation (24), it was found that a solution for (24) is the Gegenbauer polynomial $X_i = T^1_{i-1}(z)$. Thus

$$\phi(x,t) = \sum_{i=1}^{\infty} C_i T'_{i-1}(t)\ e^{-\left[i(i+1)/4N\right]t} \qquad (26)$$

He gives this method for solving for the C_i. Since initial gene frequency of population is p, then

$$\delta(x - p) = \sum_{i=1}^{\infty} C_i T'_{i-1}(z)$$

where $\delta(x)$ represents the delta function. Multiply both sides by $(1 - z^2)T'_{i-1}(z)$ and using orthogonal property,

$$\int_{-1}^{1} (1-z^2)T'_m(z)T'_{i-1}(z)dz = \delta_{m,i-1} \frac{2(i+1)i}{(2i+1)}\ , \qquad (27)$$

where m in Kronecker's notation represents zero or a positive integer; thus

$$2\left[1 - (1-2p)^2\right]\ T'_{i-1}(1-2p) = C_i \frac{2(i+1)i}{(2i+1)}$$

$$C_i = 4pq \frac{(2i + 1)}{9(i + 1)} T'_{i-1}(1 - 2p)\ . \qquad (28)$$

Some of these values are given by $C_1 = 6\ pq$,

$$C_2 = -30pq(p - 2),\ C_3 = 84\ pq(-5pq + 1).$$

The formal solution is

$$\phi(x,t) = \sum_{i=1}^{\infty} \frac{(2i+1)(1-r^2)}{i(i+1)} T'_{i-1}(r)\ T'_{i-1}(z)\ e^{-\left[i(i+1)/4N\right]t} \qquad (29)$$

or in terms of hypergeometric function,

$$\emptyset(x,t) = \sum_{i=1}^{\infty} pqi(i+1)(2i+1)F(1-i,i+2,2,p)F(1-i,i+2,2,x)$$
$$e^{-[i(i+1)/4N]\,t} . \qquad (30)$$

By noting that $\frac{dP_i(z)}{dt} = T'_{i-1}(z)$ and $P_n(1) = 1$, he gives the possibility that both A_1 and A_2 coexist in the tth generation.

$$\Omega_t = \int_0^1 \emptyset(x,t)dx = \int_{-1}^1 \emptyset(x,t)\frac{dz}{2}$$

$$= \sum_{m=1}^{\infty} \frac{(4m-1)(1-r^2)}{(2m-1)2m} T'_{2m-2}(r)\ e^{-[(2m-1)2m/4N]\,t} \qquad (31)$$

for $t > 0$, the series is convergent and as $t \to \infty$ Ω_t becomes zero. He then gives the proof that when $t = 0$, the series converges to 1. If $\Omega_{0,n}$ is the partial sum of first n terms, then by a recurrence relation

$$\frac{(4m-1)(1-r^2)T'_{2m-2}(r)}{(2m-1)2m} = P_{2m-2}(r) - P_{2m}(r)$$

$$\Omega_{0,n} = 1 - P_{2n}(r).$$

Using $P_n(z) = \frac{1}{\pi}\int_0^{\pi} \left[z + \sqrt{z^2-1}\cos t\right]^n dt$

he shows that for $|r| < 1$, $P_{2n}(r) \to 0$ as $n \to \infty$.

$$|P_{2n}(r)| \le \frac{1}{\pi}\int_0^{\pi} \left|r + \sqrt{r^2-1}\cos t\right|^{2n} dt$$

$$= \frac{1}{\pi}\int_0^{\pi} \left[r^2 + (1 - r^2)\cos^2 t\right]^n dt \rightarrow 0 \text{ as } n \rightarrow \infty .$$

Also

$$\Omega_t = \sum_{j=0}^{\infty} \left[P_{2j}(r) - P_{2j+2}(r)\right] \quad e^{-\left[(2j+1)(2j+2)/4N\right] t}$$

from equation (31) which says $f(1,t) + \Omega_t + f(0,t) = 1$ from equation (22). For $t > 0$ the series is seen to be convergent and as $t \rightarrow \infty$, $\Omega_t \rightarrow 0$, giving the asymptotic formula

$$\Omega_t \sim 6pq\ e^{-(1/2N)t} \qquad (31.1)$$

and for $t = 0$, Ω_t converges to 1.

Finally, from equation (29) we have the probability of heterozygosis,

$$H_t = \int_0^1 2x(1 - x)\phi(x,\ t)dx$$

$$= \sum_{i=1}^{\infty} pq\ \frac{(2i+1)}{i(i+1)}\ T^{!}_{i-1}(1-2p) \int_{-1}^{1} (1-z^2)T^{!}_{i-1}(z)e^{-\left[i(i+1)/4N\right] t}dz.$$

From equation (27) where $m = 0$, the integral above is zero except when $i = 1$. Hence

$$H_t = pq \cdot \frac{3}{2} \cdot 1 \cdot \frac{4}{3} \cdot e^{-(1/2N)t} = 2pqe^{-(1/2N)t} = H_0e^{-(1/2N)t} \qquad (32)$$

and this shows that heterozygosis decreases at the rate 1/2N per generation. This is the exact result of Wright and Fisher's corrected result as given previously.

Kimura gives a short proof that this is valid. If p is the frequency of A_1 and $qp(1 - p)$ is the frequency of heterozygotes, then if $p + \delta p$ is the change in p for one generation, $1 - p - \delta p$ is the change in $1 - p$ for that generation. The expected value of the heterozygotes is

$$E\left[2(p + \delta p)(1 - p - \delta p)\right] = 2p(1 - p) - 2E(\delta p)^2$$

$$= 2p(1 - p) - 2\frac{p(1 - p)}{2N} = \left[1 - \frac{1}{2N}\right] 2p(1 - p)$$

as was to be shown.

Again going back to the notation of equation (11) he writes:

$$\Omega_t = 6pq\, e^{-(1/2N)t} + 14\, pq(-5pq+1)e^{-(6/2N)t} + \alpha\left[e^{-(15/2N)t}\right] \tag{33}$$

and also the variance of the distribution in the t^{th} generation is from equations (21) and (25),

$$V_t = pq - pq\, e^{-(1/2N)t}\,. \tag{34}$$

This says that the variance approaches its limiting value pq at the rate 1/2N per generation.

Kimura (1955b) also considers the case where N is changing gradually from generation to generation in a deterministic way such that N_t can be represented as a continuous function of t. In this case equation (31.1) becomes:

$$\Omega_t \sim 6pq\, e^{-\int_0^t dt/2N_t} \tag{34.1}$$

and equation (32) becomes:

$$H_t = 2pq\ e^{-\int_0^t dt/2N_t} \qquad (34.2)$$

Thus a necessary and sufficient condition that for a growing population, H_t and Ω_t to vanish at the limit when t becomes ∞ is that the integral $\int_0^\infty \frac{dt}{N_t}$ diverges, i.e., N_t must be at most of the order of T at the limit. If the population increases more rapidly, heterozygosis cannot be eliminated entirely.

On the other hand, if N changes stochastically around its mean $\overline{N}$ with sufficiently small deviations compared with $\overline{N}$ and if these deviations are mutually independent, then N in equations (31.1) and (32) should be replaced by $\overline{N} - \frac{V_N}{\overline{N}}$, where V_N is the variance of N.

Random Genetic Drift Due to Random Fluctuation of Selection Intensities

Kimura (1954) considers a pair of alleles lacking dominance and the process of change of their frequencies when their selection coefficients fluctuate fortuitously from generation to generation around a mean zero.

Consider a large random mating population where the effect of random sampling of gametes is negligible with alleles A_1 and A_2. If x is the relative frequency of the gene A_1 in the population and s is the selection coefficient of A_1, then the rate of change of gene frequency due to selection is approximately

$$\delta x = sx(1 - x)$$

per generation, when s is small. If there is random fluctuation in the selection intensity, s and δx are random variables.

Let the mean of s be $\bar{s}$ and its variance V_s. Then the mean of δx is

$$M_{\delta x} = \bar{s}\ x(1 - x)$$

and the variance

$$V_{\delta x} = V_s\ x^2(1 - x)^2.$$

Thus we would expect a certain irregularity in the process of change in gene frequency from generation to generation.

When the rate of change is small, this process may be treated as a continuous Markov process. If x is the gene frequency at the t^{th} generation and the function $\emptyset(x, p; t)$ denotes the density of the conditional probability that the frequency lies between x and x + dx at the t^{th} generation given that the initial gene frequency was p at t = 0, we have

$$\frac{\partial \emptyset(x,p;t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[V_{\delta x}\ \emptyset(x,p;t)\right] - \frac{\partial}{\partial x}\left[M_{\delta x}\ \emptyset(x,p;t)\right] . \quad (35)$$

This equation is known as "Kolmogorov's forward differential equation" and also as the "Fokker-Planck equation". Wright (1945) was the first to apply this equation to population genetics. (See Appendix.)

The left-hand side of this equation represents the rate of change of the relative probability of any class per generation and can be written as the two terms on the right. Of the terms on the right-hand side, the first is due to random fluctuation and the second is due to the directed change.

Making the substitutions for $M_{\delta x}$ and $V_{\delta x}$, equation (35) becomes:

$$\frac{\partial \phi}{\partial t} = \frac{V_s}{2} \frac{\partial^2}{\partial x^2} \left[x^2(1-x)^2\phi\right] - \bar{s} \frac{\partial}{\partial x} \left[x(1-x)\phi\right], \quad (0 < x < 1) \qquad (36)$$

Let

$$\phi = 2x^{(1+s_1/2)-2} \, (1-x)^{(1+s_1/2)-2} \, e^{-\lambda t_1}U$$

and

$$x = \frac{1}{2} \left[1 + \tanh\,(\theta/2)\right]$$

where $t_1 = (tV_s)/2$ and $s_1 = (2\bar{s})/V_s$.

This reduces (36) to

$$\frac{d^2U}{d\theta^2} + \left[\lambda - \frac{1+s_1{}^2}{4} - \frac{s_1}{2} \tanh\,(\frac{\theta}{2})\right]U = 0, \quad (-\infty < \theta < \infty).$$

Kimura (1955) gives the following two independent solutions.

$$U_+ = \left[\frac{e^\theta}{1 + e^\theta}\right]^a \left[\frac{1}{1 + e^\theta}\right]^b F(a+b,\ a+b+1,\ 1+2a,\ \frac{e^\theta}{1+e^\theta})$$

$$U_- = \left[\frac{e^\theta}{1 + e^\theta}\right]^{-a} \left[\frac{1}{1 + e^\theta}\right]^{b} F(-a+b,\ a+b+1,\ 1-2a,\ \frac{e^\theta}{1+e^\theta})$$

where

$$a = \sqrt{\left(\frac{1 - s_1}{2}\right)^2 - \lambda} \quad \text{and} \quad b = \sqrt{\left(\frac{1 + s_1}{2}\right)^2 - \lambda}\ .$$

If the gene A is randomly selected such that the mean value of its selection coefficient is zero if taken over very long periods of time, then

$$M_{\delta x} = 0,$$

$$V_{\delta x} = V_s x^2(1 - x)^2 ,$$

where V_s is the variance of s; thus equation (35) reduces to

$$\frac{\partial \phi}{\partial t} = \frac{V_s}{2} \frac{\partial^2}{\partial x^2} \left[x^2(1 - x)^2 \phi\right]. \qquad (37)$$

This equation has singularities at the boundaries so that no arbitrary conditions can be imposed there, but he shows that if an initial condition $\phi(x, p; 0)$ is given, a continuous stochastic process satisfying equation (37) can be uniquely determined.

Still considering the case of no dominance, Kimura (1952) makes the transformation

$$z = \log \left(\frac{x}{1 - x}\right).$$

Then the rate of change of the value of z per generation becomes $\delta z = s$. If the gene frequency in the population is measured by the z scale, it changes continuously from $-\infty$ to ∞ as x changes from 0 to 1. Thus the distribution of z is approximately normal. The mean and variance of δz are equal respectively to the mean $\bar{s}$ and variance V_s of the selection coefficient s.

It follows that by using the same transformation he was able to solve equation (37). Let

$$u = 1/2\ e^{(V_s/8)t}\ x^{3/2}\ (1 - x)^{3/2}\ \phi$$

and

$$z = \log\left(\frac{x}{1 - x}\right) .$$

The result is

$$\frac{\partial u}{\partial t} = \frac{V_s}{2}\ \frac{\partial^2 u}{\partial z^2} . \tag{38}$$

This equation is also known as the heat conduction equation and it is already established that there is a unique solution which is continuous for $-\infty$ to $+\infty$ when $t \geq 0$ and reduces to $u(z, 0)$ when $t = 0$.

$$u(z,t) = \frac{1}{\sqrt{2\pi V_s t}} \int_{-\infty}^{\infty} e^{-(z-\beta)^2/2V_s t}\ u(\beta,0)\,d\beta$$

Then if the initial distribution of gene frequencies $\phi(x,p;0)$

is given and after making substitutions, we have the unique solution which satisfies (37) and is continuous between 0 and 1.

$$\phi(x,p;t) = \frac{1}{\sqrt{2\pi V_s t}} \frac{e^{-(V_s/8)t}}{[x(1-x)]^{3/2}} \int_0^1 \exp\left[-\frac{\left[\log\frac{x(1-y)}{(1-x)y}\right]^2}{2\,V_s t}\right] \times \sqrt{y(1-y)}\ \phi(y,0)\,dy\ . \qquad (39)$$

On the other hand, if the initial condition is not a continuous distribution $\phi(x, p; 0)$, but is a given gene frequency x_0, then the relative probability that the gene frequency in the t^{th} generation will be between x and x + dx is

$$\phi(x,p;t) = \frac{1}{\sqrt{2\pi V_s t}} \exp\left[-\frac{V_s t}{8} - \frac{\left[\log\frac{x(1-x_0)}{(1-x)x_0}\right]^2}{2\,V_s t}\right] \times \frac{[x_0(1-x_0)]^{1/2}}{[x(1-x)]^{3/2}}\ . \qquad (40)$$

If x = .5, the distribution curve becomes unimodal if the number of generations is less than $4/3V_s$, but becomes bimodal if it exceeds this value. (See Fig. 3.)

The mean of the distribution is always

$$x_0 = \int_0^1 x\phi(x,\ t)\,dx \qquad (41)$$

but the variance

$$V_t = \int_0^1 (x - x_0)^2\ \phi(x,\ t)\,dx \qquad (42)$$

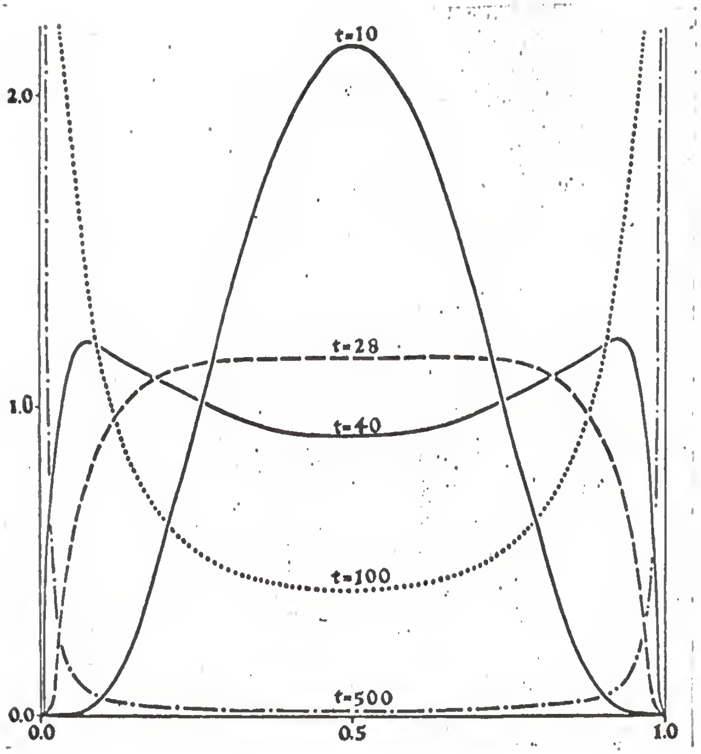

Fig. 3. Illustration of the process of change in the distribution of gene frequencies with random fluctuation in the selection intensities. It is assumed that there is no dominance, the initial gene frequency is .5, and the variance of the selection coefficient is .0483.

increases in successive generations.

It can be represented asymptotically for large t

$$V_t = x_0(1 - x_0) \sqrt{\frac{\pi x_0(1-x_0)}{2V_s t}}\, e^{-V_s t/8} + \propto \left[\frac{e^{-V_s t}}{t\sqrt{t}}\right] \quad . \qquad (43)$$

Thus as t becomes very large, V_t is very close to $V_s/8$.

A highly complicated treatment of the terminal parts of the distribution is given in Kimura (1954), pages 286-289.

Comparison of Two Methods

Wright has repeatedly emphasized the evolutionary significance of random drift in a natural population which is subdivided into many partially isolated subgroups. His theory is accepted by many evolutionists. On the other hand, Fisher and Ford (1947) emphasized the prevalence of drift due to fluctuation of selection intensities and challenged the theory of Wright by denying any significance of random drift due to small population numbers in evolution. This led to experimental studies by members of the school of Fisher and Ford (Sheppard, 1951).

In spite of all of the experimental studies, no mathematical analysis was made. This prompted Kimura to make the studies as mentioned before. With his results he makes the following comparison.

$$\frac{1}{2N} \cong \frac{V_s}{8} \qquad (44)$$

Although this is a rather restricted formula, it could be used to calculate N or V_s if one or the other is known.

The effect produced by the random fluctuation in natural selection is stated as being of relatively little importance for small populations. However, in large populations it has a remarkable effect that in the case of no dominance, the distribution curve is modified markedly in the parts where the frequency of either allele is low.

Another comparison between drift due to random mating in small populations and due to fluctuation of selection intensities is that when due to finite size, the gene in question may indeed be lost, while if due to the latter case the gene may reach an equilibrium near the fixation point, called quasi-fixation. This is the asymptotic case as noted before.

Fixation of Mutant Gene

In large natural populations, gene mutations may be occurring in each generation. While most of the genes are deleterious, some turn out to be advantageous. These advantageous mutant genes have a tendency to increase their frequencies in later generations, and thus have a chance for establishment even in large populations. Wright (1931, 1942) studied this problem and gave some solutions. Kimura (1957) and Robertson (1960) presented solutions under general conditions for the probability that a mutant gene would become fixed in a population.

Equation (49) takes the form

$$\frac{\partial u}{\partial t} = \frac{p(1-p)}{4N}\frac{\partial^2 u}{\partial p^2} + sp(1-p)\left[h + (1-2h)p\right]\frac{\partial u}{\partial p} \qquad (44.1)$$

where the selective advantage of mutant homozygote is s and that of heterozygote is sh. The solution, u(p, t), is the probability that the mutant gene reaches fixation by the t^{th} generation, given that its initial frequency is p. This probability is equivalent to that of equation (17).

Kimura (1957) defines the probability of ultimate fixation by

$$u(p) = \lim_{t \to \infty} u(p, t).$$

For the neutral mutant gene, u(p) = p. If v is the initial number of mutant genes, $u(p) = \frac{v}{2N}$ and the probability of fixation per mutant gene is $\frac{1}{2N}$.

For the general case Kimura (1957) sets $\frac{\partial u}{\partial t} = 0$, and obtains

$$u(p) = \frac{\int_0^p e^{-2N_s(2h-1)x(1-x)-2N_{sx}}\,dx}{\int_0^1 e^{-2N_s(2h-1)x(1-x)-2N_{sx}}\,dx} \qquad (44.2)$$

where 2h - 1 is the measure of dominance.

The following are more simplified equations for ultimate frequency at fixation.

For additive:

$$u(p) = \frac{\int_0^p e^{-2N_s x}\, dx}{\int_0^1 e^{-2N_s x}\, dx} = \frac{1 - e^{-2N_s x}}{1 - e^{-2N_s}} .$$

For recessive:

$$u(p) = \frac{\int_0^p e^{-2N_s x^2}\, dx}{\int_0^1 e^{-2N_s x^2}\, dx} .$$

For dominance:

$$u(p) = \frac{\int_0^p e^{2N_s x(x-2)}\, dx}{\int_0^1 e^{2N_s x(x-2)}\, dx} .$$

By expanding each of these by the Taylor series and looking only at the first two terms, since others will be very small, for small N_s one obtains:

for additive:

$$u(p) = p + p(1 - p)N_s;$$

for recessive:

$$u(p) = p + \frac{2}{3} p(1 - p^2)N_s;$$

and for dominance:

$$u(p) = p + \frac{2}{3} p(p - 1)(p - 2)N_s.$$

APPLICATIONS AND EXAMPLES

Kerr and Wright (1954) made a three-part study of genetic drift presented in three continuous articles. In the first, a study of genetic drift due to inbreeding, he used the trait "forked". The other two experiments were with "Bar" and "spineless". It is stated that for the forked case, the selection differential is much less than ten per cent so that the results illustrate random drift from inbreeding in an almost pure form. Of 96 lines carried to fixation or to 16 generations, f^t became fixed in 41 lines, f(forked) in 29 lines, and 26 lines were still unfixed. The conclusion was that the amount of selection against forked is slight.

The Bar experiment was more extensive and use was made of the Fokker-Planck equation. One hundred eight small populations were used and little selective mortality was found but severe selection against Bar from low productivity of homozygous Bar females and Bar males. Starting from 50 per cent Bar genes in each case, the distribution soon reached approximate stability of form (about four generations) as type came to be fixed at a rate of 22 per cent per generation and Bar at a rate of 0.7 per cent per generation. After generation 10, type had been fixed in 95 lines, Bar in three, and 10 were still unfixed. The form of the distribution agreed well with that expected from a population of effective size, 72 per cent of actual size, and an empirically determined rate of change of the frequency (q) of Bar, $\delta q = -.35q(1 - q)$.

Crow and Morton (1955) derived a formula for the variance of random drift of gene frequency and for effective population number. If N_e is effective size, then this variance is $q(1 - q)/2N$, where q is the frequency of allele under discussion. The formula derived is

$$V_{\delta q} = \frac{q(1 - q)}{4N} \left[1 - F' + (1 + F') \frac{V_k}{\mu_k} \right]$$

where N is total number of offspring, μ_k and V_k are the mean and variance of the number of surviving offspring per parent, and F' is Wright's coefficient of inbreeding. Also

$$N_e = \frac{2N}{1 - F' + (1 + F') \dfrac{V_k}{\mu_k}} .$$

They also indicate that V_k/μ_k is a measure of the degree of departure from idealized conditions and thus propose that this ratio be used as an index of variability in progeny number. The authors then give an account of an experiment with drosophila in which they applied these methods.

In a small population experiment Merrell (1953) followed gene frequency changes in sex-linked recessive genes of Drosophila Melanogaster. Population sizes were from 10 to 100. The percentage of wild type flies rose rapidly and remained above 90 per cent, while some strains decreased in frequency.

Large fluctuations occurred due to genetic drift, in some cases leading to loss of the recessive gene. The results were interpreted as due to the combined effects of natural selection and genetic drift.

Spencer (1947) analyzed a sample of 110 wild flies showing a frequency of 10 per cent for the gene "stubble bristles". In a sample of identical size, collected at a point almost one-fourth mile distant from the first collection area and two years later, he found the gene frequency seven per cent for the "stubble" bristle. The genes "brick" eye color and "dubonnet" eye color were also recovered more than once in both samples. The concentration of these genes in the population is explained as caused by genetic drift brought about by seasonal fluctuations of population size.

An example of the difference in large and small populations is shown in Fig. 4.

Computer Simulation

For the case of $\overline{s} = 0$ given by Kimura (1955), Barker and Butcher (1966) developed a Monte Carlo computer program to investigate quasi-fixation of genes due to random fluctuation of selection intensities. They start with a gene frequency of 0.9 for the desirable allele and a constant mean selection coefficient equal to .01. They performed 10 simultaneous experiments with variance of selection coefficient V_s ranging from .02 to 0.2. In terms of the probability of quasi-loss of the desirable allele, the results confirm the theoretical expectation

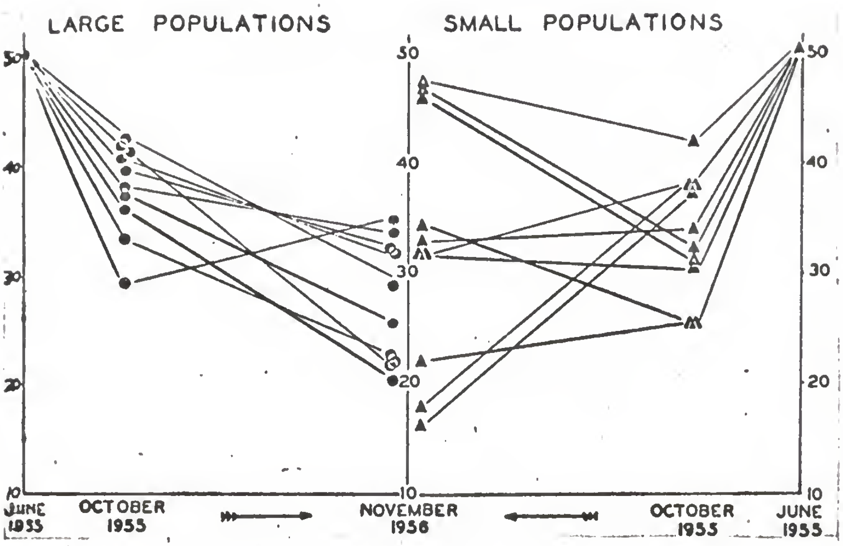

Fig. 4. Difference in gene frequency change when comparing large (4000) and small (20) samples of drosophila, where each sample is divided into 10 groups. (Dobzhansky, 1957.)

of Kimura (1962). The number of generations to final stability of quasi-loss tended to increase as $2\bar{s}/V_s$ increased and could be expected to be at least 1000 for $0.5 \leq 2\bar{s}/V_s < 1.0$.

SUMMARY

Using a differential equation, Fisher (1922) was the first to give a mathematical treatment for the problem of random genetic drift in finite populations due to random sampling of gametes. His result for the rate of decay of unfixed classes was not correct, being only half its true value. Wright (1931), using path coefficients and an integral equation, supplied the first correct solution for the state of steady decay.

In these results, Fisher and Wright both assumed that a steady state of decay had been attained, but nothing was known about how the process leads to the state of steady decay. Kimura (1955), by calculating the moments of the distribution with the help of the Fokker-Planck equation, obtained a solution which assumed an infinite series under the continuous model, showing that the process approaches asymptotically the state of steady decay.

When there is drift due to random fluctuations in selection intensity and random sampling, the process of change in gene frequency in a population can be represented by a stochastic process. Kimura (1954) presented an analysis for this process for the case of no dominance. In the case of random drift in small populations it was found that complete fixation or loss of an allele would be realized. Complete fixation or loss may

not be realized in the case of drift due to fluctuation of selection intensities. It is shown that for large populations, if a sufficient number of generations are allowed, a situation will be realized in which the allele is either almost fixed in the population or almost lost from it. The rate of decay per generation is given as $V_s/8$, where V_s is the variance of the selection coefficient.

Kimura (1954) also made a comparison of drift due to fluctuation intensities with drift due to random sampling. He gives a rather restricted formula by equating the two rates of fixation:

$$\frac{1}{2N} \cong \frac{V_s}{8} .$$

There are several experimental studies on this subject, some of which are listed, dealing with experimental animals. It is noted here that there have been studies of genetic drift in human populations, especially those of Glass (1952, 1954) and Lasker (1952, 1964).

A derivation of the Fokker-Planck equation and its use in deriving the distribution function given by Wright is given in the Appendix.

ACKNOWLEDGMENT

I would like to thank Dr. A. D. Dayton for his help and guidance during the writing of this report. I also would like to thank Dr. Raja Nassar, Dr. Keith Huston, and Dr. J. V. Craig for their stimulating courses on genetics.

BIBLIOGRAPHY

Barker, J. S. F. and Butcher, J. C., 1966. A Simulation Study of Quasi-fixation of Genes Due to Random Fluctuation of Selection Intensities. Genetics 53:261.

Buri, P., 1956. Gene Frequency in Small Populations of Mutant Drosophila. Evolution 10:367.

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

Dobzhansky, T., 1943. Genetics of Natural Populations IX. Genetics 28:162.

Dobzhansky, T., 1957. An Experimental Study of Interaction Between Genetic Drift and Natural Selection. Evolution 11:311.

Falconer, D. S., 1960. Introduction to Quantitative Genetics. New York: Ronald Press.

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

Fisher, R. A., 1921. On the Dominance Ratio. Proceedings Royal Society of Edinburgh 42:321.

Fisher, R. A. and Ford, E. B., 1947. The Spread of a Gene in Natural Conditions in a Colony of the Moth Panaxia Dominula L. Heredity 1:143.

Fisher, R. A. and Ford, E. B., 1950. The Sewall Wright Effect. Heredity 4:117.

Glass, B., etal., 1952. Genetic Drift in a Religious Isolate. American Naturalist 86:145.

Glass, B., 1954. Genetic Changes in Human Populations, Especially Those Due to Gene Flow and Genetic Drift. Advances in Genetics 6:95.

Hagedoorn, A. L. and A. C., 1921. The Relative Value of the Processes Causing Evolution. The Hague: Martinus Nijhoff.

Haldane, J. B. S., 1939. The Equilibrium Between Mutation and Random Extinction. Annals of Eugenics 9:400.

Harris, T. E., 1948. Branching Processes. Annals of Mathematical Statistics 19:474.

House, V. L., 1953. The Use of the Binomial Expansion for a Classroom Demonstration of Drift in Small Populations. Evolution 7:84.

Kerr, W. E. and Wright, S., 1954. Experimental Studies of the Distribution of Gene Frequencies in Very Small Populations of Drosophila Melanogaster. Evolution 8:172, 225, 293.

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, 1952a. Process of Irregular Change of Gene Frequencies Due to the Random Fluctuation of Selection Intensities. National Institute of Genetics Annual Report (Japan) 2:56.

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

Kimura, Motoo, 1953. Stepping-stone Model of Populations. National Institute of Genetics Annual Report (Japan) 3:62.

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 Distribution 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, 1957. Some Problems of Stochastic Processes in Genetics. Annals of Mathematical Statistics 28:882.

Kimura, Motoo, 1962. On the Probability of Fixation of Mutant Genes in a Population. Genetics 47:713.

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.

Kimura, Motoo and Crow, J. F., 1964. The Number of Alleles That Can be Maintained in a Finite Population. Genetics 49:725.

Lasker, G. W. and Kaplan, B. A., 1964. The Coefficient of Breeding Isolation: Population Size, Migration Rates, and the Possibilities for Random Genetic Drift in Six Human Communities in Northern Peru. Human Biology 36:327.

Lasker, G. W., 1952. Mixture and Genetic Drift in Ongoing Human Evolution. American Anthropologist 54:433.

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

Merrell, D. J., 1953. Gene Frequency Changes in Small Laboratory Populations of Drosophila Melanogaster. Evolution 7:95.

Robertson, A., 1960. A Theory of Limits in Artificial Selection. Proceedings Royal Society of London 153:234.

Sheppard, P. M., 1951. Fluctuations in the Selective Value of Certain Phenotypes in the Polymorphic Land Snail Cepaea Nemoralis (L). Heredity 5:125.

Spencer, W. P., 1947. Genetic Drift in a Population of Drosophila Immigrans. Evolution 1:103.

Wright, S., 1921. Systems of Mating. Genetics 6:111.

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., 1938a. The Distribution of Gene Frequencies under Irreversible Mutation. Proceedings National Academy of Science 24:253.

Wright, S., 1938b. The Distribution of Gene Frequencies in Populations of Polyploids. Proceedings National Academy of Science 24:372.

Wright, S., 1942a. Statistical Genetics and Evolution. Bulletin American Mathematical Society 48:223.

Wright, S., etal., 1942b. Genetics of Natural Populations. Genetics 27:363.

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

Wright, S., 1948. On the Roles of Directed and Random Changes in Gene Frequency in the Genetics of Populations. Evolution 2:279.

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

APPENDIX

APPENDIX

Derivation of Fokker-Planck Equation and Its Use in Deriving the Distribution Equation Given by Wright

Using a method given by Kimura (1955c), let $\phi(x, t)$ represent the curve for probability distribution of gene frequencies at time t. The distribution is approximated with histograms, each column having width h, as shown in Fig. 5. The gene frequency of each class is represented by the middle point of the column. Consider the class with gene frequency x. For sufficiently small h, the area of the column $\phi(x, t)h$ gives the probability that the population has gene frequency $x \pm 1/2\ h$.

By considering a small change in time Δt, it is sufficient to consider the movement of the gene frequency to its adjacent classes. This population, with gene frequency x, will move to another class due to systematic as well as random changes.

Let $m(x)\ \Delta t$ be the probability that the population moves to the higher class $(x + h)$ by systematic pressure. Let $v(x)\ \Delta t$ be the probability that it moves outside the class by random fluctuation, half of the time to the left class $(x - h)$ and the other half to the right class $(x + h)$. Movement to other than adjacent classes is neglected due to the very small probability.

Thus the probability that the population will have gene frequency $x \pm 1/2\ h$ after Δt is obtained by considering the exchange of gene frequencies between these adjacent classes.

$$\phi(x,t + \Delta t)h = \phi(x,t)h - \left[v(x) + m(x)\right] \Delta t\ \phi(x,t)h$$

$$+ \left[\frac{v(x - h)}{2}\right] \Delta t\ \phi(x - h,\ t)h$$

$$+ \left[\frac{v(x + h)}{2}\right] \Delta t\ \phi(x + h,\ t)h$$

$$+ m(x - h)\ \Delta t\ \phi(x - h,\ t)h \qquad (45)$$

The second term on the right is the amount of loss due to movement to other classes, the third term is contribution from left class, the fourth by the right class both due to random change, and the last term is the contribution from the left class due to systematic change.

Let $\sigma^2(x,t)\,\Delta t$ be the variance of the change in x per Δt due to random change,

$$\sigma^2(x,t)\ \Delta t = h^2 \left[\frac{v(x)}{2}\right] \Delta t + (-h)^2 \left[\frac{v(x)}{2}\right] \Delta t$$

so

$$\sigma^2(x,t) = h^2\ v(x)\ . \qquad (46)$$

Let $M(x,\ t)\,\Delta t$ be the mean change in x per Δt,

$$M(x,\ t)\ \Delta t = h\ m(x)\ \Delta t$$

so

$$M(x,\ t) = m(x)h. \qquad (47)$$

Now substitute (46) and (47) into (45) and divide both sides by $\Delta t \cdot h$. Then on rearrangement

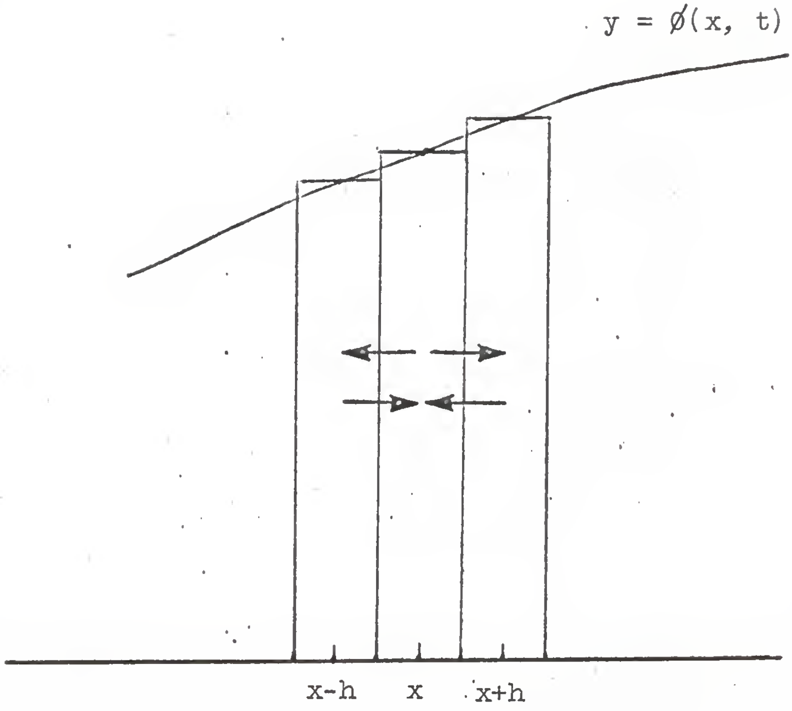

Fig. 5.

$$\frac{\phi(x,t+\Delta t)-\phi(x,t)}{\Delta t}=$$

$$\frac{1}{2}\,\frac{\dfrac{\sigma^2(x+h,t)\phi(x+h,t)-\sigma^2(x,t)\phi(x,t)}{h}-\dfrac{\sigma^2(x,t)\phi(x,t)-\sigma^2(x-h,t)\phi(x-h,t)}{h}}{h}$$

$$-\frac{M(x,t)\phi(x,t)-M(x-h,t)\phi(x-h,t)}{h} \qquad (48)$$

Taking the limit $\Delta t \rightarrow 0$, $h \rightarrow 0$ gives:

$$\frac{\partial\phi(x,t)}{\partial t}=\frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\sigma^2(x,t)\phi(x,t)\right]-\frac{\partial}{\partial x}\left[M(x,t)\phi(x,t)\right] \qquad (49)$$

This is known as the Fokker-Planck equation and also as the Kolmogorov forward solution.

Rewriting (49) where Δq represents the tendency toward a stable equilibrium point due to systematic pressure and δq is tendency to drift away from that point due to random deviation and where the mean change is taken as zero:

$$\frac{\partial}{\partial q}\left[\Delta q\cdot\phi(q)\right]=\frac{1}{2}\frac{\partial^2}{\partial q^2}\left[\sigma_{\delta q}{}^2\,\phi(q)\right] \qquad (50)$$

Then, according to Li (1955), integrating (50) gives

$$\Delta q\cdot\phi(q)=\frac{1}{2}\frac{\partial}{\partial q}\left[\sigma_{\delta q}{}^2\,\phi(q)\right]+\text{constant.} \qquad (51)$$

At this point it is seen that the left-hand member $\Delta q\cdot\phi(q)$ represents the fraction of the distribution that tends to be carried past a given value of q by the systematic pressure Δq in each generation. Since the distribution is stationary, the right-hand side

$$\frac{1}{2}\frac{\partial}{\partial q}\left[\sigma_{\delta q}^{2}\,\phi(q)\right] = \frac{1}{4N}\,\frac{d}{dq}\left[q(1 - q)\,\phi(q)\right]$$

must be the fraction of the distribution which tends to be scattered away in the opposite direction by random deviations in each generation.

Rewriting (51)

$$\frac{\Delta q}{\sigma_{\delta q}^{2}}\left[\sigma_{\delta q}^{2}\,\phi(q)\right] = \frac{1}{2}\,\frac{d}{dq}\left[\sigma_{\delta q}^{2}\,\phi(q)\right]$$

$$\frac{2\;\Delta q}{\sigma_{\delta q}^{2}} = \frac{d/dq\left[\sigma_{\delta q}^{2}\,\phi(q)\right]}{\sigma_{\delta q}^{2}\,\phi(q)}$$

then integrating again,

$$2\int\frac{\Delta q}{\sigma_{\delta q}^{2}}\,dq = \log\left[\sigma_{\delta q}^{2}\,\phi(q)\right] + \text{constant.}$$

Therefore

$$\phi(q) = \frac{C}{\sigma_{\delta q}^{2}}\exp\left[2\int\frac{\Delta q}{\sigma_{\delta q}^{2}}\,dq\right] \tag{52}$$

and where C is a constant such that

$$\int_{0}^{1}\phi(q)\,dq = 1.$$

This is the general formula (52) for the distribution of a gene frequency when a steady state (under the joint actions of Δq and δq) has been reached, as given by Wright (1937, 1938a, 1938b, 1942a).

GENETIC DRIFT--A STOCHASTIC PROCESS

by

CLEMENT JOHN MAURATH

A. B., Fort Hays Kansas State College, 1965

AN ABSTRACT OF A MASTER'S REPORT

submitted in partial fulfillment of the

requirements for the degree

MASTER OF SCIENCE

Department of Statistics

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1967

Genetic drift due to random sampling of gametes and due to fluctuation of selection intensities is presented. Both ideas are considered as stochastic processes and are treated as such. Fisher, using a differential equation, was the first to give a mathematical treatment for the first case in finite populations. His result for the rate of decay of variance was not correct, being only half large enough. Wright, using path coefficients and an integral equation, gave the correct solution as $1/2 N$ per generation. This rate of steady decay was later expanded by Kimura. By using the Fokker-Planck equation and computing the moments of the distribution, he agreed with Wright's results and also obtained a solution which assumed an infinite series under the continuous model, showing that the process approaches asymptotically the state of steady decay. It is found that given enough generations, the gene in question will be either completely lost or completely fixed in the population.

For the case of drift due to fluctuation of selection intensities, it is found that again the gene frequency becomes fixed and reaches this fixation asymptotically, but not necessarily is completely lost or fixed at gene frequency 1.0. It is found that the rate of decay is about $V_s/8$, where V_s is the variance of the selection intensities.

A comparison is made of these two types of genetic drift and examples are given.

A derivation of the Fokker-Planck equation and its use to derive the distribution function given by Wright is given in the Appendix.