VARIANCE OF GENE FREQUENCIES[1]

C. CLARK COCKERHAM
North Carolina State University at Raleigh

Received June 21, 1968

Inbreeding, gene frequency variance, and their corresponding effective population numbers are now commonplace terms in population genetics. The concepts and much of the theory are classical (Wright, 1921, 1931; Fisher, 1930). More recent refinements and extensions of the theory by Crow and associates (Crow, 1954; Crow and Morton, 1955; Kimura and Crow, 1963*a*, *b*) have been primarily concerned with distinguishing between the inbreeding effect on heterozygosity and on the variance of gene frequencies which are so intimately connected in finite populations. The purpose of the present paper is to relate the two in a way which the author thinks is meaningful and easy to grasp. Further, correlational measures are made compatible with probability measures of identity by descent and a simple basis is provided for the analysis of the variance of gene frequencies in experimental or natural populations.

The procedure is to work with the variance of a linear function and to incorporate the role that the inbreeding and coancestry of individuals play in this variance. First, let us develop this role. We let a_{ij} index the jth allele in the ith individual and introduce a measure of frequency x_{ij} defined by

$$x_{ij} = 1 \quad \text{if} \quad a_{ij} = A,$$

$$x_{ij} = 0 \quad \text{if} \quad a_{ij} = \bar{A} \neq A. \tag{1}$$

Letting the population frequency of A be

$$P(a_{ij} = A) = p, \tag{2}$$

then, for random genes,

$$Ex_{ij} = p, \quad Ex^2{}_{ij} = p, \quad \sigma^2{}_{x_{ij}} = p(1-p), \tag{3}$$

where E denotes expectation. For the covariance, $\sigma_{x_{i1}x_{i2}}$, of frequencies of alleles in the ith diploid individual we need the population frequency, $P(a_{i1} = A, a_{i2} = A)$. Making use of the probability of genes being identical by descent (Malécot, 1948), i.e., $P(a_{i1} \equiv a_{i2}) = F_i =$ the inbreeding coefficient in this case,

$$\begin{aligned} &P(a_{i1} = A, a_{i2} = A) \\ &\quad = P(a_{i1} = A)\, P(a_{i2} = A \,|\, a_{i1} = A) \\ &\quad = p[F_i + (1-F_i)p]. \end{aligned} \tag{4}$$

Consequently,

$$\sigma_{x_{i1}x_{i2}} = Ex_{i1}x_{i2} - Ex_{i1}Ex_{i2} = F_i p(1-p), \tag{5}$$

and F_i is the correlation. Next, consider the covariance, $\sigma_{x_{ij}x_{kl}}$, between frequencies of genes from two individuals. Using the same argument as in (4),

$$\begin{aligned} &P(a_{ij} = A, a_{kl} = A) \\ &\quad = P(a_{ij} = A)\, P(a_{kl} = A \,|\, a_{ij} = A) \\ &\quad = p[P(a_{kl} \equiv a_{ij}) + \{1 - P(a_{kl} \equiv a_{ij})\}p], \end{aligned} \tag{6}$$

and

$$\sigma_{x_{ij}x_{kl}} = P(a_{ij} \equiv a_{kl})\, p(1-p). \tag{7}$$

Collecting together the four covariances for the four pairing of alleles between two diploid individuals

$$\begin{aligned} &\sigma_{x_{i1}x_{k1}} + \sigma_{x_{i1}x_{k2}} + \sigma_{x_{i2}x_{k1}} + \sigma_{x_{i2}x_{k2}} \\ &= [P(a_{i1} \equiv a_{k1}) + P(a_{i1} \equiv a_{k2}) \\ &\quad + P(a_{i2} \equiv a_{k1}) + P(a_{i2} \equiv a_{k2})]\, p(1-p) \\ &= 4\theta_{ik} p(1-p), \end{aligned} \tag{8}$$

where θ_{ik} is the coancestry between the two individuals. In this context θ_{ik} is the aver-

[1] Published as Paper No. 2661 of the Journal Series of the North Carolina State University Agricultural Experiment Station, Raleigh, N. C. This investigation was supported in part by Public Health Service Research Grant GM 11546 from the Division of General Medical Sciences.

age of the four correlations or alternatively, the *correlation between the frequency of a random allele from one individual with that of a random allele from another*. Although a side issue, it is interesting to contrast θ with Wright's coefficient of relationship, ρ, which is the correlation between the means of (or sums of) gene frequencies of the individuals, i.e., between $\bar{x}_{i.} = (x_{i1} + x_{i2})/2$ and $\bar{x}_{k.} = (x_{k1} + x_{k2})/2$,

$$\rho_{ik} = \frac{2\theta_{ik}}{\sqrt{1+F_i}\,\sqrt{1+F_k}}. \qquad (9)$$

It probably should be pointed out that the foregoing variances and covariances are for neutral genes and the expectations are for classes of individuals with the same pedigree histories.

Now consider a collection of N diploid individuals. The mean gene frequency is

$$\hat{p} = \bar{x}_{..} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{2} x_{ij}}{2N}. \qquad (10)$$

Expressing the variance of the mean as the variance of a linear function, we have

$$\sigma^2_{\hat{p}} = \Big[\sum_{i=1}^{N}\sum_{j=1}^{2}\sigma^2_{x_{ij}} + 2\sum_{i=1}^{N}\sigma_{x_{i1}x_{i2}} + 2\sum_{i<k}^{N}\sum_{j=1}^{2}\sum_{l=1}^{2}\sigma_{x_{ij}x_{kl}}\Big] \Big/ 4N^2. \qquad (11)$$

Substitution of the variances and covariances into (11) leads to

$$\sigma^2_{\hat{p}} = \left[\frac{1+\bar{F}}{2N} + \frac{N-1}{N}\bar{\theta}\right] p(1-p) = \theta_l p(1-p), \qquad (12)$$

where θ_l is the coancestry of the line or group with itself (Cockerham, 1967). The bar on F will generally be dropped since in most applications no ambiguity will arise, but will be retained on θ to distinguish that it is a group measure. This is the essence of the development; that variance among gene frequencies for neutral genes of groups of individuals depends on the inbreeding and relatedness of individuals in the group as well as the number of individuals.

The group coancestry is dominated by the relatedness of individuals in the group except for very small N. If the members of the group are noninbred and unrelated, there is variance for finite grouping,

$$\sigma^2_{\hat{p}} = p(1-p)/2N. \qquad (13)$$

If the members are unrelated, the largest effect that inbreeding can have is to double the variance

$$\sigma^2_{\hat{p}} = (1+F)p(1-p)/2N, \qquad (14)$$

which in amount is small unless N is small. Even with extremely large groups, $N \to \infty$, there is variance among the groups in proportion to the relatedness of individuals in the groups. As an example, consider unrelated full sib families of infinite size from unrelated parents. In this case $\bar{\theta} = 1/4$, and

$$\sigma^2_{\hat{p}} = p(1-p)/4. \qquad (15)$$

All the simplifications that result from an additive situation may be realized. The total variance may be subdivided exactly into components corresponding to the various subdivisions of the genes. Consider the linear model

$$x_{kij} = p + a_k + b_{ki} + w_{kij}, \qquad (16)$$

where k indexes the groups, and the effects—a for groups, b for individuals, and w within individuals—are all random and uncorrelated, and have variances σ^2_a, σ^2_b and σ^2_w, respectively. The expectations of quadratics, now from a different point of view since pedigree histories are assumed unknown, over classes of genes are

$$\begin{aligned} &Ex_{kij}x_{k'i'j'} \\ &= p^2 + \sigma^2 && \text{if } k=k',\ i=i',\ j=j' \\ &= p^2 + \mathrm{Cov}_{ab} && \text{if } k=k',\ i=i',\ j\neq j' \\ &= p^2 + \mathrm{Cov}_a && \text{if } k=k',\ i\neq i' \\ &= p^2 + \mathrm{Cov}_g && \text{if } k\neq k'. \end{aligned} \qquad (17)$$

For uncorrelated groups, $\mathrm{Cov}_g = 0$, and parametrically we have in terms of correlations,

$$\sigma^2 = p(1-p),\ \mathrm{Cov}_{ab} = \rho_{ab}p(1-p),\ \mathrm{Cov}_a = \rho_a p(1-p). \qquad (18).$$

The correlations, ρ_a and ρ_{ab}, are related to the components of variance as follows,

$$\begin{aligned}(1-\rho_{ab})\,p(1-p) &= \sigma^2{}_w,\\ (\rho_{ab}-\rho_a)\,p(1-p) &= \sigma^2{}_b,\\ \rho_a p(1-p) &= \sigma^2{}_a,\end{aligned} \qquad (19)$$

and

$$\sigma^2 = \sigma^2{}_w + \sigma^2{}_b + \sigma^2{}_a\,. \qquad (20)$$

The intraclass correlations, the covariances and the components of variance, while of different forms, are equivalent parameterizations. They are phrased as general measures to accommodate all causative factors that have led to the variation among gene frequencies. If the variation arises entirely because of drift and mating system, the consequences are generally summarized in terms of F and $\bar{\theta}$. No difficulties will arise, however, by extending the definitions of F and $\bar{\theta}$ to be general measures of correlations, as long as these extensions are kept clearly in mind. The correlation between frequencies of genes of different individuals in the same group is

$$\rho_a = \frac{\sigma^2{}_a}{\sigma^2{}_w + \sigma^2{}_b + \sigma^2{}_a} = \bar{\theta}, \qquad (21)$$

of genes within random individuals from different groups is

$$\rho_{ab} = \frac{\sigma^2{}_b + \sigma^2{}_a}{\sigma^2{}_w + \sigma^2{}_b + \sigma^2{}_a} = F, \qquad (22)$$

of genes within individuals *within* groups is

$$\rho_b = \frac{\sigma^2{}_b}{\sigma^2{}_w + \sigma^2{}_b} = \frac{F-\bar{\theta}}{1-\bar{\theta}}, \qquad (23)$$

and of genes with themselves is always one whether viewed over groups,

$$\rho_{abw} = \frac{\sigma^2{}_w + \sigma^2{}_b + \sigma^2{}_a}{\sigma^2{}_w + \sigma^2{}_b + \sigma^2{}_a} = 1, \qquad (24)$$

within groups,

$$\rho_{bw} = \frac{\sigma^2{}_w + \sigma^2{}_b}{\sigma^2{}_w + \sigma^2{}_b} = 1, \qquad (25)$$

or within individuals,

$$\rho_w = \sigma^2{}_w/\sigma^2{}_w = 1, \qquad (26)$$

although in (23), (25), and (26) the correlations are undefined for groups which are fixed.

Various averages of correlations are found by simply counting the numbers of correlations in each class and averaging them. Of course F and $\bar{\theta}$ are averages, F for the N pairs of genes within individuals $\bar{\theta}$ for the $2N(N-1)$ pairs of genes among individuals. Averaging for all $N(2N-1)$ possible pairs of different genes leads to the group inbreeding coefficient (Cockerham, 1967)

$$F_l = \frac{F-\bar{\theta}}{2N-1} + \bar{\theta}, \qquad (27)$$

while averaging for $4N^2$ pairings including genes with themselves leads to the group coancestry coefficient, θ_l (12). The last measure, θ_l, is also the correlation between the frequencies of genes in two gametes drawn at random from the group of individuals.

The assumptions involving the model and the components of variance were presented in a manner which is most familiar. In a broader view of the subdivision of variation, certain intraclass correlations and components of variance (or their renamed equivalences such as Cov's in (18)) may be negative. Thus, if the system of mating is such that mates are less related than the average within the group, $\bar{\theta} > F$, $\sigma^2{}_b < 0$, and $\rho_b < 0$. Another reason for $\sigma^2{}_b < 0$ would be certain types of selection.

We are now in a position to formulate expectations, variances and so on in terms of the different parametrical equivalences. For random groups,

$$\begin{aligned}\sigma^2{}_{\hat{p}} = E\frac{(\bar{x}_{k..}-\bar{x}_{k'..})^2}{2} &= E(\bar{x}_{k..}-p)^2\\ &= \frac{\sigma^2{}_w}{2N} + \frac{\sigma^2{}_b}{N} + \sigma^2{}_a\\ &= \left(\frac{1-F}{2N} + \frac{F-\bar{\theta}}{N} + \bar{\theta}\right) p(1-p)\\ &= \theta_l p(1-p).\end{aligned} \qquad (28)$$

The variance among individual mean gene frequencies within groups is

$$\sigma^2_{b'} = E\frac{(\bar{x}_{ki.} - \bar{x}_{ki'.})^2}{2} = \frac{\sigma^2_w}{2} + \sigma^2_b$$

$$= \left[\frac{1-F}{2} + (F - \bar{\theta})\right] p(1-p), \quad (29)$$

while the variance of individual mean gene frequencies around the group mean is

$$\sigma^2_{b''} = E(\bar{x}_{ki.} - \bar{x}_{k..})^2 = \frac{N-1}{N}\sigma^2_{b'}. \quad (30)$$

Of interest is the following expectation, often reputed to be the expected frequency of heterozygotes with random mating,

$$E2\bar{x}_{k..}(1-\bar{x}_{k..}) = 2E\bar{x}_{k..} - 2E\bar{x}^2_{k..}$$
$$= 2p - 2p^2 - 2\sigma^2_p$$
$$= 2p(1-p)(1-\theta_l). \quad (31)$$

The expected frequency of heterozygotes is of course

$$Q_{A\bar{A}} = 2\sigma^2_w = 2p(1-p)(1-F). \quad (32)$$

The term

$$1 - \theta_l = \frac{2N-1}{2N}(1-\bar{\theta}) - \frac{F-\bar{\theta}}{2N} \quad (33)$$

can almost never be equivalent to 1–F in finite populations. With random union of gametes $F = \theta$ and only a finite correction is required,

$$\frac{2N}{2N-1}E2\bar{x}_{k..}(1-\bar{x}_{k..}) = Q_{A\bar{A}}. \quad (34)$$

Another small but significant point is suggested: random union of gametes and random mating are synonymous in finite populations *only* in monoecious organisms when self-fertilization is allowed. The avoidance of self-fertilization, separate sexes, or special mating systems lead to other than random union of gametes and $F \neq \bar{\theta}$.

For several variations in gametic sampling in monoecious groups with random mating, θ_l is equivalent to F in the offspring in which case the expectation in (31) correctly gives the expected frequency of heterozygotes in a sample of N offspring. Deviations from this situation will become clearer in a later section on systems of mating. Note, however, that we are concerned with the variation of genes as they are constituted in the individuals and not in their projected progeny.

To identify frequencies of genotypes with the other parameters we note the relationships in (32) for the heterozygotes, and the other frequencies are found to be

$$Q_{AA} = p - (1-F)p(1-p) = p - \sigma^2_w$$
$$Q_{\bar{A}\bar{A}} = 1 - p - (1-F)p(1-p)$$
$$= 1 - p - \sigma^2_w. \quad (35)$$

Genotypic frequencies are quadratic functions of p except in the case of fixation.

Systems of Mating

Included are all of the classical systems (Wright, 1921) and all others for which the breeding recipe and group size remain the same over time. Letting the initial members be unrelated and noninbred, the initial variance of mean gene frequencies is the same for all systems with the same N.

$$\sigma^2_{\hat{p}0} = p(1-p)/2N \quad (36)$$

Then, they diverge in the manner that they diverge in F and $\bar{\theta}$ such that

$$\sigma^2_{\hat{p}t} = \left[\frac{1-F_t}{2N} + \frac{F_t - \bar{\theta}_t}{N} + \bar{\theta}_t\right] p(1-p) \quad (37)$$

for each.

For the maximum avoidance systems, i.e., the avoidance of any inbreeding for as long as possible or the maximum avoidance of the mating of relatives, there is a simple relationship between θ_l and F,

$$\theta_{lt} = F_{t+v+1}, \quad (38)$$

where v is the degree of avoidance and $2^v = N$. For other systems (unpublished) and v' degree, less than maximum, avoidance

$$\theta_{lt} = F_{t+v'+1}. \quad (39)$$

These relationships emphasize that while inbreeding is being avoided initially, the variance proceeds at the same rate as the inbreeding v or v' generations hence.

As a base of reference, one might use the idealized monoecious population (Kimura and Crow, 1963*a*) for which $v = 0$, and

$$F_t = \bar{\theta}_t = F_{lt}\,, \tag{40}$$

and all correlations as well as $\sigma^2_{\hat{p}}$ proceed with time at a constant rate. Deviations from this system lead to nonconstant rates for both F and $\bar{\theta}$ which are antagonistically related. The detailed results are being published elsewhere. Briefly, for populations which have the same gametic variance and do not have permanent sublines, a mating system which increases F more than another must increase $\bar{\theta}$ less, and vice versa. Asymptotically, F and $\bar{\theta}$ must converge, but the asymptotic rate of a system is dominated by $\bar{\theta}$ which maintains the same rank relative to other systems. The closest mating of relatives without subdividing the population is that of the circular half sib design (Kimura and Crow, 1963*b*), treated also by Robertson (1964) and Wright (1965), and for which the rate of increase in $\bar{\theta}$ is minimal.

For systems of avoidance of the mating of relatives, $F < \bar{\theta}$, and the correlation between gene frequencies of individuals within groups, ρ_b, and the component of variance for individuals within groups, σ^2_b, are negative. For monoecious populations, $F = \bar{\theta}$ and $\rho_b = \sigma^2_b = 0$, while for systems for which mates are more related than random $F > \bar{\theta}$ and ρ_b, $\sigma^2_b > 0$.

It is for systems of mating that we can clearly identify the measures used herein with Wright's (1965) F statistics, F_{IT}, F_{IS} and F_{ST}. In this connection it may be noted that the coancestry, θ, for a class of individuals corresponds to his r, i.e., $\theta_0 = r_0$ for gametes from the same individual or the coancestry of an individual with itself. By identification,

$$F_{IT} = F_t\,, \qquad F_{ST} = \theta_{lt-1}\,,$$
$$F_{IS} = \frac{F_{IT} - F_{ST}}{1 - F_{ST}} = \frac{F_t - \theta_{lt-1}}{1 - \theta_{lt-1}}. \tag{41}$$

The parameters, F and $\bar{\theta}$, correctly elucidate the correlational structure of genes within and among individuals and among groups while the F statistics generally do not. Only for some systems of mating, e.g. the idealized monoecious and dioecious populations and full sib mating, are F_{IT} and F_{ST} phrased in terms of a single transitional gametic set such that they reflect the correlations of genes within and among the individuals under *consideration*. For the monoecious population, if the effective number is the census number of parents and if the number of offspring individuals under *consideration* in generation t is also N,

$$F_{IT} = F_t\,, \qquad F_{ST} = \theta_{lt-1} = \bar{\theta}_t = F_t\,,$$
$$F_{IS} = \frac{F_t - \bar{\theta}_t}{1 - \bar{\theta}_t} = 0. \tag{42}$$

For a dioecious population with equal numbers of each sex and other conditions similar to those stipulated for the monoecious population, and for full sib mating,

$$F_{IT} = F_t\,, \qquad F_{ST} = \theta_{lt-1} = \bar{\theta}_t = F_{t+1}\,,$$
$$F_{IS} = \frac{F_t - \bar{\theta}_t}{1 - \bar{\theta}_t}. \tag{43}$$

The two sets of parameters would be the same if $F_{ST} = \bar{\theta}$ always.

Separate Sexes

For separate sexes, one must sometimes distinguish between some of the various measures for the two sexes, N_m and N_f, F_m and F_f, and $\bar{\theta}_m$ and $\bar{\theta}_f$. As they differ so will the other measures, F_{lm} and F_{lf}, and θ_{lmm} and θ_{lff}, and the various intraclass correlations and components of variance. In addition a probability or correlational measure between random genes from different sexes is needed, $\theta_{lmf} = \bar{\theta}_{mf}$. These probabilities have been studied in some detail for variations in sampling of gametic sets and in

gametic variance (Cockerham, 1967). With random mating and sex of offspring random,

$$\bar{\theta}_{mt} = \bar{\theta}_{ft} = \theta_{lmft} = F_{t+1}, \quad F_m = F_f, \quad (44)$$

while for independent sampling of gametic sets, which includes the experimental controlling of the numbers of offspring of each sex from each parent,

$$\frac{\theta_{lmmt} + \theta_{lfft} + 2\theta_{lmft}}{4} = \theta_{lt} = F_{t+2}$$

$$F_m = F_f, \quad \bar{\theta}_{mt} \neq \bar{\theta}_{ft} \text{ (some cases)},$$
$$\theta_{lmft} = F_{t+1} \text{ (all cases)}. \quad (45)$$

The list of components of variance may be extended to include one for sexes within groups

$$\sigma^2{}_s = \left(\frac{\bar{\theta}_m + \bar{\theta}_f}{2} - \bar{\theta}_{mf}\right) p(1-p), \quad (46)$$

and the component for independent groups must be modified to

$$\sigma^2{}_a = \bar{\theta}_{mf} p(1-p). \quad (47)$$

The components for within individuals and between individuals may be kept separate by sexes or averaged, depending on the purpose for which they are to be used. The same applies to various averages of correlations. Sometimes it is advantageous to give each sex equal weight in the average, while in many situations weighting in proportion to the number (if different) in each sex is appropriate. Problems in definitions, as well as in analysis, are the usual statistical ones when dealing with unequal numbers. Some of these features will be illustrated in the following treatment of the variance among gene frequencies of unrelated groups.

Two ways of calculating the mean gene frequency,

$$\hat{p} = \frac{\hat{p}_m + \hat{p}_f}{2}, \qquad \hat{p}' = \frac{N_m\hat{p}_m + N_f\hat{p}_f}{N},$$
$$N = N_m + N_f, \quad (48)$$

are the same only with equal numbers of each sex, $N_m = N_f$. The corresponding variances are

$$\sigma^2_{\hat{p}} = (\theta_{lmm} + \theta_{lff} + 2\theta_{lfm})\frac{p(1-p)}{4},$$

$$\sigma^2_{\hat{p}'} = (N^2{}_m\theta_{lmm} + N^2{}_f\theta_{lff} + 2N_mN_f\theta_{lmf}) \times \frac{p(1-p)}{N^2}. \quad (49)$$

For the situation described for (45),

$$\sigma^2_{\hat{p}t} = \theta_{lt}p(1-p) = F_{t+2}p(1-p), \quad (50)$$

while for the one described for (44),

$$\sigma^2_{\hat{p}t} = \left[\frac{1-F_t}{2N_s} + \frac{F_t - F_{t+1}}{N_s} + F_{t+1}\right] \times p(1-p), \quad (51)$$

$$N_s = \frac{4N_mN_f}{N}$$

for all variations in gametic variance, which is equivalent to (50) only with the gametic variance appropriate for equal chance. It is only for sex of offspring random that the variance of weighted mean frequency is related simply to F,

$$\sigma^2_{\hat{p}'t} = \left[\frac{1-F_t}{2N} + \frac{F_t - F_{t+1}}{N} + F_{t+1}\right] \times p(1-p), \quad (52)$$

but is equal to $F_{t+2}\ p(1-p)$ only when $N_m = N_f$ and sampling of gametes from parents of each sex is random. For most purposes $\hat{p}$ would seem to be preferred to $\hat{p}'$ since the sexes have equal weight in determining the next generation.

Effective Numbers

Effective numbers are used in several contexts, some of which are: to accommodate variations in the variance of gametes per parent, to approximate one recurrence formula with another more simple one, and to denote rates of inbreeding and of increase in variance of gene frequency. As mentioned previously, only for monoecious populations are the rates constant, particularly in early generations. The real distinction between the effective numbers as rates may be seen by comparing rates for F and $\bar{\theta}$ in

the early generations within and between systems, for example, maximum avoidance and circular half sib (Kimura and Crow, 1963*b*).

The rate, δ, for the variance when N is constant,

$$\delta_{\sigma^2\hat{p}t} = \frac{\sigma^2_{\hat{p}t} - \sigma^2_{\hat{p}t-1}}{p(1-p) - \sigma^2_{\hat{p}t-1}} = \frac{W_{Ft}\delta_{Ft} + W^-_{\theta t}\delta^-_{\theta t}}{W_{Ft} + W^-_{\theta t}}, \quad (53)$$

is a weighted average of the other two rates,

$$W_{Ft} = \frac{1-F_{t-1}}{2N}, \quad W^-_{\theta t} = \frac{N-1}{N}(1-\bar{\theta}_{t-1}), \quad (54)$$

$$\delta_{Ft} = \frac{F_t - F_{t-1}}{1-F_{t-1}}, \quad \delta^-_{\theta t} = \frac{\bar{\theta}_t - \bar{\theta}_{t-1}}{1-\bar{\theta}_{t-1}},$$

but almost entirely dominated by $\delta^-_{\theta t}$. In some systems of mating, such as maximum avoidance, one has the following relation between rates,

$$\delta_{\sigma^2\hat{p}t} = \delta_{F_{t+v+1}}. \quad (55)$$

It is only asymptotically, however, that the rates, δ_{Ft} and $\delta^-_{\theta t}$, converge and also approach a constant.

With constant N, changes in $\sigma^2_{\hat{p}}$ are reflected exactly by changes in F and $\bar{\theta}$. With fluctuating numbers there are effects of the numbers themselves in addition to those reflected in F and $\bar{\theta}$. Kimura and Crow (1963*a*) give an example in which each parent contributes to only one offspring in the next generation and thereby reducing the population by one half each generation. Starting with an idealized monoecious population, $F = \bar{\theta}$,

$$\sigma^2_{\hat{p}t} = \left[\frac{1-F_t}{2N_t} + F_t\right] p(1-p), \quad (56)$$

and the gene frequency variance in successive generations is

$$\sigma^2_{\hat{p}t+t'} = \left[\frac{1-F_t}{2N_{t+t'}} + F_t\right] p(1-p), \quad (57)$$

since no additional inbreeding is accrued and the relatedness of individuals remains the same. However, $N_{t+t'} = N_t/2^{t'}$ so that the variance increases considerably, actually doubling each generation if $F_t = 0$. The rate of increase in the first generation is

$$\delta_{\sigma^2\hat{p}t+1} = \frac{1}{2N_t - 1} = \frac{1}{2[2N_{t+1} - \frac{1}{2}]}, \quad (58)$$

and the number, $2N_{t+1} - \frac{1}{2} = N_t - \frac{1}{2}$, of offspring from the same N_t parents but produced as in an idealized population would give the same result. The halving operation rapidly reduces the population to an individual with the limiting variance of $(1 + F_t)\, p(1-p)/2$.

Further Structuring of the Population

Structuring of the population is accommodated by accounting for the relationships of various classes of individuals. Let θ_d, $d = 1, 2, 3, \ldots, u$, correspond to the dth class of relationship, less with ascending order. The covariances are $Fp(1-p)$, $\theta_1 p(1-p)$, $\theta_2 p(1-p)$, $\ldots$, $\theta_u p(1-p)$. While $\sigma^2_{\hat{p}}$ may always be written as a function of the components of variance,

$$\sigma^2_{\hat{p}} = \left[\frac{1-F}{2N} + \frac{F-\theta_1}{q_1} + \frac{\theta_1-\theta_2}{q_2} + \ldots + \frac{\theta_u}{q_u}\right] p(1-p), \quad (59)$$

where the q's are dependent on the numbers in the various classes and the structuring, this form is simple and the most useful only when the structuring is hierarchical. For example, if each group consists of n random subdivisions,

$$\sigma^2_{\hat{p}} = \left(\frac{1-F}{2N} + \frac{F-\theta_1}{N} + \frac{\theta_1-\theta_2}{n} + \theta_2\right) \times p(1-p). \quad (60)$$

Most mating systems are structured and the structures of some were considered in detail by Wright (1965). This amounts to elucidating the θ_d's of which $\bar{\theta}$ is an aver-

age. The advantage of this elucidation is that it provides a basis for a detailed analysis of a suspected nonneutral gene in the experimental results from a system of mating.

Structuring may be explored from the standpoint of minimizing $\sigma^2_{\hat{p}}$. By inspection of (60) $\sigma^2_{\hat{p}}$ is smallest when n is largest since $\theta_1 - \theta_2 \geqslant 0$. At the extreme $n = N$ and the subdivisions are of individuals which requires self-fertilization to maintain a constant size of groups. The variance is then a minimum as was shown by Kimura and Crow (1963*a*), and for unrelated initial individuals is

$$\sigma^2_{\hat{p}t} = \frac{1 + F_t}{2N} p(1-p). \qquad (61)$$

The limiting variance for n unrelated subdivisions, $\theta_2 = 0$,

$$\sigma^2_{\hat{p}\infty} = \frac{p(1-p)}{n}, \qquad (62)$$

shows the effectiveness of subdivisions in reducing the variance.

Base Populations

If an experimental population is initiated from a finite group, the loss in variance due to the constitution of the initial or base group can be summarized simply in terms of θ_l for the base group. The variance among such groups is

$$\sigma^2_{\hat{p}} = \theta_l p(1-p), \qquad (63)$$

leaving a maximum of

$$(1-\theta_l)p(1-p) \qquad (64)$$

on the average within offspring from the base group. This maximum can be attained only within an infinite number of offspring contributed to equally by each parent of the base group. Other losses in variance due to sampling will depend upon the number of individuals and the manner in which they are used to develop the experimental population from the base group. This result illustrates also the effect of bottlenecks on populations.

Estimation and Tests of Hypotheses

For purposes of illustration let the numbers of individuals of each genotype be

	AA	A$\bar{\text{A}}$	$\bar{\text{A}}\bar{\text{A}}$	Total
kth group	N_{k2}	N_{k1}	N_{k0}	$N_{k.}$
all groups	$N_{.2}$	$N_{.1}$	$N_{.0}$	$N_{..}$

Various means and quadratics are related to the numbers as follows (omission of bar on x designates a sum over subscripts replaced by dots):

$$\bar{x}_{k..} = x_{k..}/2N_{k.} = (2N_{k2} + N_{k1})/2N_{k.},$$

$$\bar{x}_{...} = (2N_{.2} + N_{.1})/2N_{..}$$

$$\sum_k \Big(\sum_{i,j} x^2_{kij} \Big) = \sum_k (2N_{k2} + N_{k1}) = 2N_{.2} + N_{.1},$$

$$\sum_i x^2_{ki.} = 4N_{k2} + N_{k1}$$

$$\sum_k x^2_{k..} = \sum_k (2N_{k2} + N_{k1})^2,$$

$$x^2_{...} = (2N_{.2} + N_{.1})^2. \qquad (66)$$

For one group of individuals the analysis of variance is given in Table 1. By equating mean squares to their expectations the estimators of the components of variance are found to be

$$\hat{\sigma}^2_w = N_{k1}/2N_{k.}, \qquad (67)$$

$$\hat{\sigma}^2_b = \frac{4N_{k2} + N_{k1}}{4(N_{k.} - 1)} - \frac{(2N_{k2} + N_{k1})^2}{4N_{k.}(N_{k.} - 1)} - \frac{N_{k1}}{4N_{k.}}.$$

No information is available within a single group about group differences for p, i.e., for θ or for the variance of p_k around p. For simplicity one reparameterizes for single groups so that expectations over sample groups are unbiased.

$$\mathrm{E}\, p_k(1-p_k) = (1-\bar{\theta})p(1-p),$$

$$\mathrm{E}\, F_k p_k(1-p_k) = (F-\bar{\theta})p(1-p),$$

$$\mathrm{E}\, F_k = (F-\bar{\theta})/(1-\bar{\theta}). \qquad (68)$$

With these new parameters the estimators (where $\hat{=}$ denotes 'estimated by' for functions of parameters) are

$$\hat{p}_k = \bar{x}_{k..}, \qquad F_k p_k(1-p_k) \hat{=} \hat{\sigma}^2_b,$$

$$(1-F_k)p_k(1-p_k) \hat{=} \hat{\sigma}^2_w,$$

$$p_k(1-p_k) \hat{=} \hat{\sigma}^2_w + \hat{\sigma}^2_b, \; \hat{F}_k = \frac{\hat{\sigma}^2_b}{\hat{\sigma}^2_w + \hat{\sigma}^2_b},$$

TABLE 1. *Analysis of variance of gene frequencies of a single group.*

Source	df	Mean Squares	Expectations of Mean Squares
Between individuals	$N_{k.}-1$	$S_b = \frac{4N_{k2}+N_{k1}}{2(N_{k.}-1)} - \frac{(2N_{k2}+N_{k1})^2}{2N_{k.}(N_{k.}-1)}$	$\sigma^2{}_w + 2\sigma^2{}_b$
Within individuals	$N_{k.}$	$S_w = N_{k1}/2N_{k.}$	$\sigma_w{}^2$

$$\hat{Q}_{kAA} = \hat{p}_k - \hat{\sigma}^2{}_w = N_{k2}/N_{k.},$$
$$\hat{Q}_{kA\bar{A}} = 2\hat{\sigma}^2{}_w = N_{k1}/N_{k.},$$
$$\hat{Q}_{k\bar{A}\bar{A}} = 1 - \hat{p}_k - \hat{\sigma}^2{}_w = N_{k0}/N_{k.}. \qquad (69)$$

That the observed relative frequencies, $\hat{Q}_k$'s, of the genotypes are unbiased estimators of the expected frequencies, Q_k's, may seem surprising at first, but no assumptions about F_k have been made. If one assumes $F_k = 0$, i.e., the genes united at random, then $\sigma^2{}_b = 0$, and the two mean squares (Table 1) have the same expectation. They are pooled to provide the best unbiased estimator of $\sigma^2{}_w$,

$$p_k(1-p_k) \triangleq \tilde{\sigma}^2{}_w = \frac{N_{k.}S_w + (N_{k.}-1)S_b}{2N_{k.}-1} = \frac{2N_{k.}}{2N_{k.}-1}\hat{p}_k(1-\hat{p}_k), \qquad (70)$$

and the frequencies are now estimated as

$$\hat{Q}'_{kAA} = \hat{p}_k - \tilde{\sigma}^2{}_w, \quad \hat{Q}'_{kA\bar{A}} = 2\tilde{\sigma}^2{}_w,$$
$$\hat{Q}'_{k\bar{A}\bar{A}} = 1 - \hat{p}_k - \tilde{\sigma}^2{}_w. \qquad (71)$$

As for testing hypotheses with data from a single group, only one test, that genes were united at random or $F_k = 0$, is possible. Two approximate tests are suggested. One is to assume the test statistic,

$$T_{bw} = S_b/S_w, \qquad (72)$$

follows the F distribution. Another is to perform the χ^2 goodness of fit test of $\hat{Q}_k$'s in (69) to the Q'_k's in (71). Both of these tests can be evaluated numerically and compared since the exact distribution of N_{k1} given p_k and $N_{k.}$ is easily derived, but is outside the scope of this paper.

In studies of isolate populations, one often obtains data only on a sample of the isolate group. Sometimes of interest is the variance of $\hat{p}_k$ about the mean gene frequency, p_k, of the K members of the isolate group. This variance is

$$\sigma^2{}_{\hat{p}_k, p_k} = \mathrm{E}(\hat{p}_k - p_k)^2 = \frac{1+F_k}{2}p_k(1-p_k)\frac{K-N_{k.}}{KN_{k.}}, \qquad (73)$$

TABLE 2. *Combined analysis of variance of gene frequencies.*

Source	df	Mean Squares	Expectations of Mean Squares
Between groups	$M-1$	$S_a = \frac{\Sigma(2N_{k2}+N_{k1})^2}{2(M-1)N_{.}} - \frac{(2N_{.2}+N_{.1})^2}{2(M-1)N_{..}}$	$\sigma^2{}_w + 2\sigma^2{}_b + 2N_{.}\sigma^2{}_a$
Between individuals	$M(N_{.}-1)$	$S_b = \frac{4N_{.2}+N_{.1}}{2M(N_{.}-1)} - \frac{\Sigma(2N_{k2}+N_{k1})^2}{2N_{..}(N_{.}-1)}$	$\sigma^2{}_w + 2\sigma^2{}_b$
Within individuals	$N_{..}$	$S_w = \frac{N_{.1}}{2N_{..}}$	$\sigma^2{}_w$

and is estimated by

$$\hat{\sigma}^2{}_{\hat{p}_k,\, p_k} = \frac{\hat{\sigma}^2{}_w + 2\hat{\sigma}^2{}_b}{2}\,\frac{K - N_{k.}}{KN_{k.}}. \quad (74)$$

Both the variance and the estimator go to zero with complete sampling, $N_{k.} = K$.

The combined analysis of variance for M groups, each with a constant number of individuals, i.e., $N_{k.} = N_.$ and $MN_. = N_{..}$, is given in Table 2. Without making any assumptions about the parameters, the various estimators are

$$\hat{p} = \bar{x}_{...}, \qquad \hat{\sigma}^2{}_w = S_w,$$
$$\hat{\sigma}^2{}_b = \frac{S_b - S_w}{2} \qquad \hat{\sigma}^2{}_a = \frac{S_a - S_b}{2N_.}.$$

$$(1-F)p(1-p) \triangleq \hat{\sigma}^2{}_w,$$
$$(F-\bar{\theta})p(1-p) \triangleq \hat{\sigma}^2{}_b,$$
$$\bar{\theta}p(1-p) \triangleq \hat{\sigma}^2{}_a,$$
$$(1-\bar{\theta})p(1-p) \triangleq \hat{\sigma}^2{}_w + \hat{\sigma}^2{}_b,$$
$$p(1-p) \triangleq \hat{\sigma}^2{}_w + \hat{\sigma}^2{}_a + \hat{\sigma}^2{}_b, \quad (75)$$
$$\hat{\bar{\theta}} = \frac{\hat{\sigma}^2{}_a}{\hat{\sigma}^2{}_w + \hat{\sigma}^2{}_a + \hat{\sigma}^2{}_b},$$
$$(F-\bar{\theta}) \triangleq \frac{\hat{\sigma}^2{}_b}{\hat{\sigma}^2{}_w + \hat{\sigma}^2{}_b + \hat{\sigma}^2{}_a},$$
$$\hat{F} = \frac{\hat{\sigma}^2{}_b + \hat{\sigma}^2{}_a}{\hat{\sigma}^2{}_w + \hat{\sigma}^2{}_a + \hat{\sigma}^2{}_b},$$
$$\hat{Q}_{A\bar{A}} = 2\hat{\sigma}^2{}_w = N_{.1}/N_{..}.$$

Again, without making any assumptions, the observed frequencies of the genotypes are the unbiased estimators of the expected frequencies. If one assumes genes to unite at random within groups, $F = \bar{\theta}$, but that frequencies have differentiated among groups, $\bar{\theta} \neq 0$, then

$$Q_{A\bar{A}} = 2(1-\bar{\theta})p(1-p), \quad (76)$$

and the best unbiased estimator is

$$\hat{Q}'_{A\bar{A}} = \frac{2[N_{..}S_w + M(N_. - 1)S_b]}{2N_{..} - M}$$
$$= 2\tilde{\sigma}^2{}_w, \quad (77)$$

similar to the case for a single group. If one further assumes that genes are distributed entirely at random within and among groups, i.e., $F = \bar{\theta} = 0$, then

$$Q_{A\bar{A}} = 2p(1-p), \quad (78)$$

and the best estimator is

$$\hat{Q}''_{A\bar{A}} = 2\frac{N_{..}S_w + M(N_. - 1)S_b + (M-1)S_a}{2N_{..} - 1}$$
$$= 2\hat{\tilde{\sigma}}^2{}_w = \frac{4N_{..}}{2N_. - 1}\hat{p}(1-\hat{p}). \quad (79)$$

Various hypotheses may be tested.

H_{01}—genes within groups united at random, $F - \bar{\theta} = 0$ or $\sigma^2{}_b = 0$.

H_{02}—genes randomly distributed among groups, $\bar{\theta} = 0$ or $\sigma^2{}_a = 0$. (80)

H_{03}—genes randomly distributed among and within groups, $F = \bar{\theta} = 0$ or $\sigma^2{}_a = \sigma^2{}_b = 0$.

Test statistics involving functions of mean squares are

H_{01}—T_{bw} similar to the one for a single group.

H_{02}—$T_{ab} = S_a/S_b$. (81)

H_{03}—$T_{(ab)w} = \dfrac{(N_. - 2)S_b + 2S_a}{N_.S_w}$.

The first two may be approximated by the F-distribution and the last one, $T_{(ab)w}$, by an approximate F-distribution appropriate for ratios of combined mean squares. The corresponding χ^2 goodness of fit tests are

H_{01}—fit $\hat{Q}_k$'s to $\hat{Q}'_k$'s over all groups with M d.f. Included is one d.f. for the fit of $\hat{Q}$'s to $\hat{Q}'$'s which may be singled out.

H_{02}—fit all $\hat{Q}'_k$'s to $\hat{Q}''$'s with $M-1$ d.f. which includes one d.f. for the fit of $\hat{Q}'$'s to $\hat{Q}''$'s.

H_{03}—fit all $\hat{Q}_k$'s to $\hat{Q}''$'s with $2M-1$ d.f.

Which of the test systems, T or χ^2, is more robust is not known. Most likely each is more sensitive in some tests and to some alternative hypotheses.

The foregoing analyses, estimators and tests of hypotheses are not confined strictly

to inbreeding concepts as has been pointed out previously. One might sample experimental or natural populations over space and/or time, the samples corresponding to groups. The various estimators or tests of hypotheses may be more related to suspected selection than inbreeding. All contributory causes of association and differentiation in gene frequencies are included in alternatives to the test. One sometimes has sufficient history—pedigrees or group sizes, mating system, and so on—of the populations to project the consequences of inbreeding, and thus to test whether the results are commensurate with inbreeding only. Also, when mating is random large negative or positive values of $F-\bar{\theta}$ cannot arise without other causes. Different results from the analysis of different genes for the same groups provide evidence that inbreeding is not the sole cause.

The analyses may be expanded to include other hierarchical or factorial classifications, but the foregoing suffices for the illustration. In monogamous species, one may perform an analysis of mate pairs and test for a correlation between genes of mates.

In sampling natural populations, particularly humans, it is not practical to maintain samples or groups of equal size. Also the structuring is not exactly hierarchical or into balanced sets of groups. The problems of analysis are the statistical ones of dealing with these types of data. One procedure, not well advertised, in the quadratic treatment of such data is to utilize linear functions of means of products. This may be illustrated for the balanced case. Note that

$$\overline{x^2_{kij}} = \frac{\sum_{k,\,i,\,j} x^2_{ijk}}{2N_{..}} \tag{83}$$

is the mean of the squared values of the frequencies of all genes, and

$$\overline{x_{kij}x_{kij'}} = \frac{\sum_{k,\,i} x_{ki1}x_{ki2}}{N_{..}} \tag{84}$$

is the mean of the products of the frequencies of different genes for the same individual. In these forms

$$[(1-F)p(1-p)] \mathrel{\hat{=}} \hat{\sigma}^2_w = \overline{x^2_{kij}} - \overline{x_{kij}x_{kij'}}. \tag{85}$$

Proceeding to the next component,

$$\overline{x_{kij}x_{ki'j'}} = \frac{\sum_{k,\,i<i'} x_{ki.}x_{ki'.}}{2\sum_{k} N_{k.}(N_{k.}-1)} \tag{86}$$

is the mean of the products of gene frequencies of different individuals in the same group, and

$$[(F-\theta)p(1-p)] \mathrel{\hat{=}} \hat{\sigma}^2_b = \overline{x_{ki1}x_{ki2}} - \overline{x_{kij}x_{ki'j'}}. \tag{87}$$

The last type of term needed is the mean of the products of the frequencies of genes in different groups,

$$\overline{x_{kij}x_{k'i'j'}} = \frac{\sum_{k<k'} x_{k..}x_{k'..}}{\sum_{k<k'} N_{k.}N_{k'.}}, \tag{88}$$

and

$$[\bar{\theta}p(1-p)] \mathrel{\hat{=}} \hat{\sigma}^2_a = \overline{x_{kij}x_{ki'j'}} - \overline{x_{kij}x_{k'i'j'}}. \tag{89}$$

The means of products were expressed in forms that will accommodate variations in the size of groups. For the balanced situation

$$\overline{x_{kij}x_{ki'j'}} = \frac{2\sum_{k,\,i<i'} \bar{x}_{ki.}\bar{x}_{ki'.}}{N_{..}(N_{.}-1)},$$

$$\overline{x_{kij}x_{k'i'j'}} = \frac{2\sum_{k<k'} \bar{x}_{k..}\bar{x}_{k'..}}{M(M-1)}. \tag{90}$$

There are other ways for the unbalanced situation of computing the means of products, e.g., $\overline{x_{kij}x_{k'i'j'}}$ as in (90) and

$$\overline{x_{kij}x_{ki'j'}} = \sum_{k}\left[\sum_{i<i'} \frac{x_{ki.}x_{ki'.}}{2N_{k.}(N_{k.}-1)}\right] \Big/ M, \tag{91}$$

but the ones, (86) and (88), weighting all pairs of genes of a class equally would

often be preferred. We now have in principle a basis for accommodating any structuring of the groups and imbalances involving sizes of groups as far as estimation is concerned. If the structuring is based on distances among the groups, for example, with some reasonable lumping of similar distances into classes, then the mean of the products of gene frequencies of genes separated by each distance class can be computed with its own correlational parameter, θ_d, the parameter for the greatest distance being set equal to zero since no information beyond this distance is available from the data. Models may be fitted relating the parameters to distance, estimates of the parameters may be compared for different genes, and so on.

It should be emphasized that the elucidation of the variance of gene frequencies, the correlations, and components of variance depends in no way on the dominance or detection system for the genes, but the analysis of data does. Therefore, the analyses as described hold only for genes which show codominance, i.e., are detected in both homozygotes and heterozygotes. The observational variable for a recessive gene has characteristics very different from that of a codominant gene.

Quantitative Characteristics

In the foregoing the gene was the basic unit. For individuals (diploids) as the basic unit one must work with the sum of two alleles which has variance of

$$\sigma^2_{(x_{i1}+x_{i2})} = (1 + F_i)2p(1-p). \quad (92)$$

For a quantitatively varying characteristic let y_{ki} be the measure for the ith individual in the kth group. If all genes contribute additively, the total variance among unrelated individuals, excluding environmental variance, is $(1 + F)\ \sigma^2_A$, where σ^2_A is the additive variance for non-inbred individuals. The variance among group means,

$$\bar{y}_{k.} = \frac{\sum_i y_{ki}}{N}, \quad (93)$$

is

$$\sigma^2_{\bar{y}_{k.}} = \left[\frac{1 + F - 2\bar{\theta}}{N} + 2\bar{\theta}\right]\sigma^2_A = 2\theta_l\sigma^2_A, \quad (94)$$

which involves just two components of variance: one for groups,

$$\sigma^2_a = 2\theta\sigma^2_A, \quad (95)$$

and one for individuals within groups,

$$\sigma^2_b = (1 + F - 2\bar{\theta})\sigma^2_A \quad (96)$$

which sum to the total variance, $(1 + F)\ \sigma^2_A$.

These results were obtained by Lush (1948). They do not agree with those of Wright (1965) primarily because of the difference between F_{ST} and $\bar{\theta}$ pointed out previously (41).

Summary

The role that inbreeding and coancestry play in the distribution of neutral genes is incorporated into the variance of a linear function to provide a simple cumulative expression of the variance among mean gene frequencies of groups of individuals. The total variance of gene frequencies is subdivided into components corresponding to genes within individuals, among individuals within groups, and among groups. Various intraclass correlations, some of which may be negative, of gene frequencies are formulated. The various types of parameters are considered for and extended to include further structuring of the population, separate sexes, systems of mating, and effective numbers. Estimators and tests of hypotheses for the parameters are developed.

Literature Cited

Cockerham, C. Clark. 1967. Group inbreeding and coancestry. Genetics 56:89–104.

Crow, J. F. 1954. Breeding structure of populations. II. Effective population number, p. 543–556. *In* Statistics and mathematics in biology. Iowa State College Press.

Crow, J. F., and N. E. Morton. 1955. Measurement of gene frequency drift in small populations. Evolution 9:202–214.

FISHER, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford, Rev. ed. 1958.

KIMURA, M., AND J. F. CROW. 1963*a*. The measurement of effective population number. Evolution 17:279–288.

——. 1963*b*. On the maximum avoidance of inbreeding. Genet. Res., Cambridge 4:399–415.

LUSH, J. L. 1948. The genetics of populations. Ames, Iowa. Mimeo.

MALÉCOT, G. 1948. Les mathématiques de l'hérédité. Masson, Paris.

ROBERTSON, A. 1964. The effect of non-random mating within inbred lines on the rate of inbreeding. Genet. Res. 5:164–167.

WRIGHT, S. 1921. Systems of mating. I–V. Genetics 6:111–178.

——. 1931. Evolution in Mendelian populations. Genetics 16:97–159.

——. 1938. Size of population and breeding structure in relation to evolution. Science 87: 430–431.

——. 1965. The interpretation of population structure by F-Statistics with special regard to systems of mating. Evolution 19:395–420.