ANALYSES OF GENE FREQUENCIES*

C. CLARK COCKERHAM

Department of Statistics, North Carolina State University, Raleigh, North Carolina 27607

Manuscript received July 31, 1972
Revised copy received April 12, 1973
Transmitted by TIMOTHY PROUT

ABSTRACT

Models of variance components and their intraclass correlational equivalences are developed for genes falling into various categories of subdivisions within a population. Estimable functions are elaborated demonstrating that intraclass correlations can be estimated only relative to that for the least related genes in the informational system. The effects of different types of subdivisions —and of ignoring them—on the parameters are demonstrated. Small sample estimators are formulated for all of the parameters by three different methods, including both a weighted and an unweighted method of analysis of the variation among subpopulations. How estimators change with assumptions about the parameters is illustrated. Various tests of hypotheses are outlined in χ^2 and F-test terminology. Discussed are factors which may affect the correlations and the manner in which their effects are manifest, hopefully in clarification of some of the misconceptions that have arisen in this connection.

IN a previous paper (COCKERHAM 1969) I pointed out the role that coefficients of inbreeding and coancestry play in the variance of frequencies of neutral genes. In a parallel development a linear model of effects was introduced with a corresponding decomposition of total variance into components of variance. Components of variance were manipulated to produce intraclass correlations, which in turn were shown to correspond to functions of the coefficients of inbreeding and coancestry. It was emphasized that when applied in analyses of data on gene frequencies, the coefficients must be given correlational definitions to include all causative factors affecting them, not just those due to drift and mating system.

General relationships between components of variance and intraclass correlations were demonstrated, and formulated for an analysis of a sample of observations from a single subdivision and of equal-sized samples from several subdivisions of a population. A formulation not well known was given for unequal sizes of samples for the subdivisions, the thought being that no one would have difficulty in adapting analyses to unequal sample sizes by some of the usual methods. Various tests of hypotheses were formulated in χ^2 terminology, and some tests of hypotheses, it was pointed out, could be phrased in F-test terminology.

I have had sufficient feedback to indicate that to some readers some of my notions appear strange, that the models and analyses may be alright for balanced experiments but not for unequal-sized samples, and that such large sample theory is not applicable in practice where sample sizes are often small. On review, part of the confusion may have stemmed from my lack of emphasizing and elaborating certain points, and lack of relating procedures directly to other published ones.

* Paper No. 3843 of the Journal Series of the North Carolina State University Agricultural Experiment Station, Raleigh, North Carolina. This investigation was supported in part by Public Health Service Grant GM 11546 from the Division of General Medical Sciences.

The purposes of this paper are to correct some of these omissions, to formulate alternative analyses for unequal sample sizes, to emphasize differences between hypothesis testing and estimation, and to discuss the correlational parameters as they relate to causative factors.

Since I shall be referring to COCKERHAM (1969) repeatedly I shall use a shorthand notation (1969). The main, but actually trivial, development in that paper was the designation of the gene—each and every gene—not the frequency of a class of genes or of genotypes, as the observational unit. Then, by providing a measure of frequency, 1 or 0, for each gene there obtained the metrical situation subject to treatment by all appropriate modern statistical theory and methods. The rest followed automatically, including the small sample estimation and testing of hypotheses.

BACKGROUND CONSIDERATIONS

Two of the most noted and widely used developments in population genetics theory are WRIGHT's F-statistics and "WAHLUND's Principle." In WRIGHT's F-statistics I include not just F_{IT}, F_{IS} and F_{ST} (WRIGHT 1951) but all those (WRIGHT 1943, 1946, 1965 and others) defined for specific classes of gametes. PROFESSOR WRIGHT has repeatedly pointed out their correspondence with all types of structuring of individuals within populations, such as in systems of mating or in various kinds of subdivisions as may happen in natural populations. He has repeatedly stressed their general correlational nature whenever applied to populations, regardless of the factor(s) leading to the correlations.

Then why ever introduce MALÉCOT's (1948) probabilities of identity by descent in this vein? One reason was to rebut statements to the effect that these probabilities could not lead to negative correlations, and it was demonstrated that they could do so even for neutral genes in a randomly-mating dioecious population where the probabilities apply exactly. However, it was pointed out that when applied to real populations these parameters, too, must be given general correlational definitions for realism. Then why distinguish them from WRIGHT's correlations? For purposes of clarity, most distinctions being small and some negligible. WRIGHT defines correlations for genes in terms of those in gametes. For example, for individuals A and B it is for genes in the gametes from A with genes in the gametes from B. I use the average of the correlations of genes of A with genes of B. The two correlations are identical (with regular meiosis and this is not a point at issue). The distinction in this case only emphasizes that the present treatment is of genes as constituted in individuals. A real difference can arise when applied to a group of individuals. The average of the correlations of genes among distinct individuals is often different from that of the average of all correlations of genes in a sample of gametes from them.

I found in the treatment of a single hierarchy of subpopulations (groups, 1969) that it was only necessary to define two correlations—simply, that between genes within individuals, F, and that between genes of different individuals in the same subpopulation, $\bar{\Theta}$. Another correlation, entirely a function of these two, is that between genes within individuals *within* subpopulations

$$f = (F - \bar{\Theta})/(1 - \bar{\Theta}) = (\rho_b,\ 1969)\ .$$

These three correlations manipulate just as the F-statistics, and obviously $F = F_{IT}$, $f = F_{IS}$ and $\bar{\Theta} = F_{ST}$ for all intents and purposes. Another distinction is that F and $\bar{\Theta}$ have not been defined relative to some standard, although they become so in practice and F_{IT} and F_{ST} are defined relative to a total. This matter is best discussed later.

Turning now to "WAHLUND's Principle," it is most often presented as a form of

$$\mathcal{E}\, p_k^2 = (\mathcal{E} p_k)^2 + \sigma_{p_k}^2$$

where $\mathcal{E}$ is the expected or average value (E in 1969), p_k is the gene frequency for the k*th* subpopulation and $\sigma_{p_k}^2$ is the variance among the frequencies. This rather old principle involving moments was not new at that time in application to many areas, including genetics. WAHLUND (1928) showed, among other things, the role that variance due to differences in gene frequencies among subpopulations played in the total genotypic frequencies from amalgamating the subpopulations. More important, however, he extended the subdivision concept to isolates within subpopulations with the identification of an additional source of variance, and, by analogy, on to a third subdivision and another source of variance. This development can be construed to be the identification and definition of components of variance as we know them today for populations with various levels of subdivisions.

I simply extended the variance component concept to include the component, σ_w^2, due to variation of genes within individuals and the component, σ_b^2, due to variation of genes among individuals within subpopulations (1969). Then, designating the component of variance due to differences among subpopulations as σ_a^2, these three components which sum to the total variance, σ^2, were related to the correlations as follows:

	Components of variance		Correlational equivalences		Covariance equivalences
	σ_a^2	=	$(\bar{\Theta}-\bar{\Theta}_g)pq$	=	$\mathrm{Cov}_a - \mathrm{Cov}_g$
	σ_b^2	=	$(F-\bar{\Theta})pq$	=	$\mathrm{Cov}_{ab} - \mathrm{Cov}_a$
	σ_w^2	=	$(1-F)pq$	=	$\sigma_T^2 - \mathrm{Cov}_{ab}$
Total	σ^2	=	$(1-\bar{\Theta}_g)pq$	=	$\sigma_T^2 - \mathrm{Cov}_g$

(1)

(p is gene frequency and $q = 1-p$)

Covariances, *Cov*'s, of genes were also introduced (1969) since they provide a rationale for the correlations. The covariance representation was found very useful (COCKERHAM 1963) for the analyses of relatives for quantitative characteristics. The covariances are σ_T^2 for genes with themselves, Cov_{ab} for distinct genes of the same individual, Cov_a for genes of different individuals in the same subpopulation, and Cov_g for genes in different subpopulations.

Another correlation, $\bar{\Theta}_g$, among genes in different subpopulations is introduced in the model (1) corresponding to Cov_g. This is for the least related genes, least related in the sense that they are farthest apart in the hierarchy. If the model properly represents our sphere of information, then for practical purposes for

representation and estimation, since $\bar{\Theta}_g$ is not estimable, we set $p'q' = (1-\bar{\Theta}_g)pq = \sigma^2$, $\bar{\Theta}' = (\bar{\Theta}-\bar{\Theta}_g)/(1-\bar{\Theta}_g)$ and $F' = (F-\bar{\Theta}_g)/(1-\bar{\Theta}_g)$ so that $\sigma_a^2 = \bar{\Theta}'p'q'$, $\sigma_b^2 = (F'-\bar{\Theta}')p'q'$ and $\sigma_w^2 = (1-F')p'q'$, and it is the primed parameters that are estimable. Note that if $\bar{\Theta}_g = 0$ then the primed and unprimed parameters are the same. In either case $f' = (F'-\bar{\Theta}')/(1-\bar{\Theta}') = (F-\bar{\Theta})/(1-\bar{\Theta}) = f$ is the same. Thus in practice the model is reduced to accommodate the parameters estimable from the information or data available, and the estimable correlations are always relative to the correlation of genes farthest apart (generally least related and least correlated) in the informational system. While in practice one drops the primes, introduction of these primed parameters points out the result of enforcing the constraint that $\bar{\Theta}_g = 0$ when it actually is not zero.

If $\bar{\Theta}_g$ is not zero, to estimate it requires information on unrelated subdivisions among which the genes are not correlated. Then, there is available a component of variance, σ_g^2, due to differences among unrelated subdivisions with $\sigma_g^2 = \bar{\Theta}_g pq$. The estimable correlations are now $\bar{\Theta}_g$, $\bar{\Theta}$ and F as well as those formed by them, f, F' and $\bar{\Theta}'$, while with information on only one subdivision of related subpopulations it is only the latter three that are estimable.

In theory involving inbreeding and drift one generally determines the parameters or correlations relative to unrelated genes—that is genes among completely unrelated subdivisions tracing separately back to the noninbred and nonsubdivided founder population. I think this is what PROFESSOR WRIGHT means when he says relative to the total for F_{IT} and F_{ST}. From a practical standpoint in the analysis of observations one need only define those correlations necessary to characterize the observed structure and relative to the correlation of the most distantly related genes. Only with ancillary information can one relate back to some founder population or to more distantly related populations.

Now suppose there were actually subdivisions, isolates, within subpopulations but otherwise model (1) with $\bar{\Theta}_g = 0$ is correct. New components of variance are introduced, σ_{b1}^2 due to differences among individuals within isolates and σ_{b2}^2 due to isolates within subpopulations. Correspondingly, $\bar{\Theta}_1$ is the correlation of genes between individuals within the same isolate and $\bar{\Theta}_2$ is the correlation of genes between isolates in the same subpopulation. The new model equivalences are

	Components of variance		Correlational equivalences
	σ_{a1}^2	=	$\bar{\Theta}_2 pq$
	σ_{b2}^2	=	$(\bar{\Theta}_1-\bar{\Theta}_2)pq$
	σ_{b1}^2	=	$(F-\bar{\Theta}_1)pq$
	σ_w^2	=	$(1-F)pq$
Total	σ^2	=	pq

(2)

Since $\bar{\Theta}$ is the average of the correlations among all genes of different individuals in the same subpopulation it is an average of $\bar{\Theta}_1$ and $\bar{\Theta}_2$. Let the proportion of the number of pairs of genes from different individuals in the same isolate to the total number of pairs of genes from different individuals in the same subpopulation be α. Then, $\bar{\Theta} = \alpha\,\bar{\Theta}_1 + (1-\alpha)\bar{\Theta}_2$. Using this relationship we can relate the components of variance for the two models

$$\sigma_a^2 = \alpha\,\sigma_{b2}^2 + \sigma_{a1}^2 \qquad = [\alpha(\bar{\Theta}_1 - \bar{\Theta}_2) + \bar{\Theta}_2]pq = \bar{\Theta}pq$$

$$\sigma_b^2 = \sigma_{b1}^2 + (1-\alpha)\sigma_{b2}^2 = [F - \bar{\Theta}_1 + (1-\alpha)(\bar{\Theta}_1 - \bar{\Theta}_2)]pq = (F - \bar{\Theta})pq$$

$$\sigma_w^2 = \sigma_w^2 \qquad = (1-F)pq$$

These formulations do not imply that if one just ignores isolates in theory or estimation that the same results are obtained as when there are no isolates. Rather, they show the parametrical relationships which allow the evaluation of ignoring isolates or subdivisions within a lower category. We have available using model (1) $\bar{\Theta}$, F, f and the total variance as before but only F and the total variance are appropriate for the new situation.

For the new model (2), in addition to F, the appropriate correlations are

$$\bar{\Theta}_2 = \sigma_{a1}^2/\sigma^2 \qquad \bar{\Theta}_1 = (\sigma_{a1}^2 + \sigma_{b2}^2)/\sigma^2\,,$$

which determine other correlations: between genes within individuals within isolates

$$f_1 = (F - \bar{\Theta}_1)/(1 - \bar{\Theta}_1) = \sigma_{b1}^2/(\sigma_{b1}^2 + \sigma_w^2)\,,$$

between genes of different individuals within isolates within subpopulations

$$f_2 = (\bar{\Theta}_1 - \bar{\Theta}_2)/(1 - \bar{\Theta}_2) = \sigma_{b2}^2/(\sigma_{b1}^2 + \sigma_{b2}^2 + \sigma_w^2)$$

and between genes within individuals relative to genes between isolates within subpopulations

$$f_3 = f_1 + f_2 - f_1 f_2 = (F - \bar{\Theta}_2)/(1 - \bar{\Theta}_2) = (\sigma_{b1}^2 + \sigma_{b2}^2)/(\sigma_{b1}^2 + \sigma_{b2}^2 + \sigma_w^2)\,.$$

These examples should be sufficient to indicate the various correlations that can be formulated.

Next, consider the situation where model (1) is correct for variation within subpopulations, but the subpopulations fall into related groups within areas which are ignored in model (1). The model is now

$$\begin{array}{lcc} \text{Components of variance} & & \text{Correlational equivalences} \\ \sigma^2_{a4} & = & (\bar{\Theta}_3 - \bar{\Theta}_4)pq \\ \sigma^2_{a3} & = & (\bar{\Theta} - \bar{\Theta}_3)pq \\ \sigma^2_b & = & (F - \bar{\Theta})pq \\ \underline{\sigma^2_w} & = & \underline{(1 - F)pq} \\ \text{Total} \quad \sigma^2 + \beta\,\sigma^2_{a4} & = & (1 - \bar{\Theta}_4)pq \end{array} \tag{3}$$

The new components of variance are σ^2_{a3} due to subpopulations within areas and σ^2_{a4} due to areas. The corresponding correlations are $\bar{\Theta}_3$ for genes of different subpopulations in the same area and $\bar{\Theta}_4$ for genes in different areas. Obviously, if we apply model (1) to this situation, $\bar{\Theta}_g \neq 0$. Let β be the proportion of the number of pairs of genes among different subpopulations in the same area to the total number of pairs of genes among different subpopulations. Then

$$\bar{\Theta}_g = \beta\,\bar{\Theta}_3 + (1-\beta)\bar{\Theta}_4 .$$

Since $\bar{\Theta}_4$ is for the least related genes in the system it is set equal to zero, or we consider all the correlations modified to satisfy this constraint and $\bar{\Theta}_g = \beta\,\bar{\Theta}_3$. This leads to

$$\begin{array}{lll} \sigma^2_a = (\bar{\Theta} - \beta\bar{\Theta}_3)pq & = \sigma^2_{a3} + (1-\beta)\sigma^2_{a4} \\ \sigma^2_b = (F - \bar{\Theta})\,pq & = \sigma^2_b \\ \underline{\sigma^2_w = (1-F)pq} & = \underline{\sigma^2_w} \\ \text{Total} \quad \sigma^2 = (1-\beta\bar{\Theta}_3)pq & = \sigma^2_T - \beta\,\sigma^2_{a4} \end{array}$$

If we apply model (1) to this situation only f is available, and everything else is measured relative to less than the total variance

$$F_1 = \frac{\sigma^2_a + \sigma^2_b}{\sigma^2} = \frac{F - \beta\bar{\Theta}_3}{1 - \beta\bar{\Theta}_3} \neq F , \qquad \bar{\Theta}_5 = \frac{\sigma^2_a}{\sigma^2} = \frac{\bar{\Theta} - \beta\bar{\Theta}_3}{1 - \beta\bar{\Theta}_3} \neq \bar{\Theta} .$$

From these f can be obtained directly or by components of variance.

$$f = (F_1 - \bar{\Theta}_5)/(1 - \bar{\Theta}_5) = (F - \bar{\Theta})/(1 - \bar{\Theta}) .$$

All of the other correlations and the total variance are available only by application of model (3).

The concern so far has been the development of a model for fitting to populations. There may be many reasons why the model does not fit even though it appropriately takes into account the logical structuring of the genes. Interpretations of estimates depend first on the goodness of fit, and of course one looks for reasons or interpretations for lack of fit. These considerations are left until during

and after a treatment of estimation. Before that a few comments will be made about components of variance and intraclass correlations and their estimation.

Components of variance, although not necessarily known by that name, have been used extensively in population genetics. For example, FISHER (1918) and WRIGHT (1935) examined the decomposition of the genetic variance into components or parts ascribable to average, dominance and interaction effects of genes. Variance components were finding use in other fields also, and the earliest papers on their estimation, to my knowledge, are TIPPETT (1931), YATES and ZACOPANAY (1935), DANIELS (1939), and WINSOR and CLARKE (1940). All the ingredients of variance components in a one-way classification are given by AIRY (1861). A recent and comprehensive treatment of variance components and their estimation is given by SEARLE (1971a; also see SEARLE 1971b). The method of average products (1969) for estimation is treated by KOCH (1967). He later suggested an improvement in terms of averages of squares of symmetrical differences (KOCH 1968). An exhaustive survey of the one-way classification was made by HARVILLE (1969). Several contributions have been made by RAO (1971a, 1971b, 1972).

Intraclass correlations in terms of F-statistics and similar adaptations have played an extremely important role in population genetics, more so than components of variance. Also, the estimation of an intraclass correlation was introduced prior to that for variance components (FISHER 1925). Yet intraclass correlations have not been looked upon with much favor by statisticians and they are mentioned only casually, if at all, in most statistical texts. LUSH (LUSH and MOLLN 1942; LUSH and STRAUS 1942; LUSH 1947, 1948) develops the equivalences between intraclass correlations and variance components by application, including correspondence in estimation. Through the influence of the LUSH school both intraclass correlations and variance components have found much use in animal breeding theory and experimental research. The most extensive general treatment of intraclass correlations and their estimation appears to be by KEMPTHORNE (1957).

ESTIMATION AND HYPOTHESES TESTING

I shall in general use the same notation as previously (1969), the exceptions being noted. For the k^{th} subpopulation the numbers of individuals of each genotype in the sample are

	AA	$A\bar{A}$	$\bar{A}\bar{A}$	Total
Observed	N_{k2}	N_{k1}	N_{k0}	N_k

For summation over samples, $\sum_k N_{ki} = N_i$, $\sum_k N_k = N$ [blanks replace the (.)'s (1969) in summation]. Note that N's here are for the samples. I previously (1969) omitted pointing out that in estimation and testing of hypotheses the various N's were for samples, and had used the same notation for size of subpopulation.

I shall illustrate the application of three methods of estimation: analysis of variance, $\mathcal{S}$; average symmetrical products, $\mathcal{P}$ (KOCH 1967); and average sym-

metrical squared differences, $\mathcal{D}$ (KOCH 1968). Analysis of variance methods are of course the most familiar. Methods $\mathcal{P}$ and $\mathcal{D}$ are similar in approach, easy to apply in any situation, and adaptable to fitting to any population structuring. The estimators for the three methods differ only when sample sizes are unequal.

The procedure in each method is to arrive at quadratic forms of the observations. Then by equating these quadratic forms to their expectations in terms of components of variance, unbiased estimators of the components of variance are readily found as linear functions of the quadratic forms. In applying the methods we use x_{ki1} and x_{ki2} for the two alleles of the i^{th} individual in the k^{th} subpopulation, where x takes the value 1 if the gene is A, and the value 0 otherwise. For

TABLE 1

Average symmetrical products

Category of gene pairs	Numbers of pairs	Average product*
		a. One subpopulation
Between individuals	$W_{bk}=2N_k(N_k-1)$	$\mathcal{P}_{bk}=\dfrac{\sum_{i<i'}\sum_{j,j'} x_{kij}x_{ki'j'}}{2N_k(N_k-1)}=\dfrac{4N^2{}_k\hat{p}^2{}_k-4N_k\hat{p}_k+N_{k1}}{4N_k(N_k-1)}$
Within individuals	$W_{wk}=N_k$	$\mathcal{P}_{wk}=\dfrac{\sum_i x_{ki1}x_{ki2}}{N_k}=\dfrac{N_{k2}}{N_k}=\hat{p}_k-\dfrac{N_{k1}}{2N_k}$
Themselves	$W_{ck}=2N_k$	$\mathcal{P}_{ck}=\dfrac{\sum_i\sum_j x^2_{kij}}{2N_k}=\dfrac{2N_{k2}+N_{k1}}{2N_k}=\hat{p}_k$
		b. Several subpopulations
Between subpopulations	$2(N^2-\sum_k N^2{}_k)$	$\mathcal{P}_a=\dfrac{\sum_{k<k'}\sum_{i,i'}\sum_{j,j'} x_{kij}x_{k'i'j'}}{4\sum_{k<k'} N_kN_{k'}}=\dfrac{N^2\hat{p}^2-\sum_k N^2{}_k\hat{p}^2{}_k}{N^2-\sum_k N^2{}_k}$
Between individuals	$2(\sum_k N^2{}_k-N)$	$\mathcal{P}_b=\dfrac{\sum_k W_{bk}\mathcal{P}_{bk}}{\sum_k W_{bk}}=\dfrac{4\sum_k N^2{}_k\hat{p}^2{}_k-4N\hat{p}+N_1}{4(\sum_k N^2{}_k-N)}$
Within individuals	N	$\mathcal{P}_w=\dfrac{\sum_k W_{wk}\mathcal{P}_{wk}}{\sum_k W_{wk}}=\dfrac{N_2}{N}=\hat{p}-\dfrac{N_1}{2N}$
Themselves	$2N$	$\mathcal{P}_c=\dfrac{\sum_k W_{ck}\mathcal{P}_{ck}}{\sum_k W_{ck}}=\dfrac{2N_2+N_1}{2N}=\hat{p}$

$\mathcal{E}\,\mathcal{P}_{ck}=p_k=p^2{}_k+\sigma^2{}_{wk}+\sigma^2{}_{bk}=p^2{}_k+\sigma^2{}_k$, $\mathcal{E}\,\mathcal{P}_{wk}=p^2{}_k+\sigma^2{}_{bk}$, $\mathcal{E}\,\mathcal{P}_{bk}=p^2{}_k$
$\mathcal{E}\,\mathcal{P}_c=p=p^2+\sigma^2{}_w+\sigma^2{}_b+\sigma^2{}_a=p^2+\sigma^2$, $\mathcal{E}\,\mathcal{P}_w=p^2+\sigma^2{}_b+\sigma^2{}_a$, $\mathcal{E}\,\mathcal{P}_b=p^2+\sigma^2{}_a$
$\mathcal{E}\,\mathcal{P}_a=p^2$

* $\hat{p}_k$ is the sample gene frequency from the k^{th} subpopulation and $\hat{p}=\sum_k N_k\hat{p}_k/N$ is the weighted average of the sample frequencies.

method $\mathcal{P}$ we simply take the average of the products of the x's for all pairs of genes in the same category including the category of genes with themselves. These average products ($\mathcal{P} = \overline{xx}$, 1969) are given in Table 1 for observations from a single subpopulation sample. With samples from several subpopulations, the weighted average of the sample average products is obtained, where each weight is the number of products in that category for the sample. This amounts to adding the products in a category over samples and dividing by the total number of products. These pooled products and the new average product, $\mathcal{P}_a$, for genes in different subpopulations are also given in Table 1. Then by equating the $\mathcal{P}$'s to their expectations (bottom of Table 1) the estimators of the components of variance are found as linear functions of the $\mathcal{P}$'s, e.g., $\hat{\sigma}^2_w = \mathcal{P}_c - \mathcal{P}_w$.

For method $\mathcal{D}$ the same pairings of genes are utilized except one half of the square of the difference between the two x's replaces the product used in $\mathcal{P}$. Of course $\mathcal{D}_{ck}$ and $\mathcal{D}_c$ for genes with themselves are zero. The other average squared differences are displayed in Table 2 for a single sample and for samples from several subpopulations. Note that the manner of pooling over samples is the same as for method $\mathcal{P}$, and that the estimators of the variance components, utilizing the expectations at the bottom of Table 2, are linear functions of the $\mathcal{D}$'s.

TABLE 2

Average symmetrical squared differences

Category of gene pairs	Numbers of pairs	(1/2) average of squared differences
		a. One subpopulation
Between individuals	$W_{bk} = 2N_k(N_k-1)$	$\mathcal{D}_{bk} = \dfrac{\sum_{i<i'}\sum_{j,j'}(x_{kij}-x_{ki'j'})^2}{4N_k(N_k-1)} = \dfrac{4N^2{}_k\hat{p}_k\hat{q}_k - N_{k1}}{4N_k(N_k-1)}$
Within individuals	$W_{wk} = N_k$	$\mathcal{D}_{wk} = \dfrac{\sum_i (x_{ki1}-x_{ki2})^2}{2N_k} = \dfrac{N_{k1}}{N_k}$
		b. Several subpopulations
Between subpopulations	$2(N^2 - \sum_k N^2{}_k)$	$\mathcal{D}_a = \dfrac{\sum_{k<k'}\sum_{i,i'}\sum_{j,j'}(x_{kij}-x_{k'i'j'})^2}{4(N^2-\sum_k N^2{}_k)} = \dfrac{N^2\hat{p}\hat{q} - \sum_k N^2{}_k\hat{p}_k\hat{q}_k}{N^2-\sum_k N^2{}_k}$
Between individuals	$2(\sum_k N^2{}_k - N)$	$\mathcal{D}_b = \dfrac{\sum_k W_{bk}\mathcal{D}_{bk}}{\sum_k W_{bk}} = \dfrac{4\sum_k N^2{}_k\hat{p}_k\hat{q}_k - N_1}{4(\sum_k N^2{}_k - N)}$
Within individuals	N	$\mathcal{D}_w = \dfrac{\sum_k W_{wk}\mathcal{D}_{wk}}{\sum_k W_{wk}} = \dfrac{N_1}{2N}$

$\mathcal{E}\,\mathcal{D}_{wk} = \sigma^2{}_{wk}$, $\mathcal{E}\,\mathcal{D}_{bk} = \sigma^2{}_{wk} + \sigma^2{}_{bk} = \sigma^2{}_k$, $\mathcal{E}\,\mathcal{D}_w = \sigma^2{}_w$, $\mathcal{E}\,\mathcal{D}_b = \sigma^2{}_w + \sigma^2{}_b$, $\mathcal{E}\,\mathcal{D}_a = \sigma^2{}_w + \sigma^2{}_b + \sigma^2{}_a = \sigma^2$

Analysis of variance estimators of variance components are phrased in terms of mean squares and their expectations given in Table 3. In pooling mean squares over samples the weighting factors are the degrees of freedom. For the variation among subpopulations and among individuals in subpopulations there is no best method of analysis with unequal-sized samples. Two commonly used ones are a weighted and an unweighted one, both given in Table 3. The unweighted term, $\mathcal{S}'_a$, is simply the unweighted variance among sample means, and its expectation involves the harmonic mean, N_h, of sample sizes. In contrast the expectation of the mean square $\mathcal{S}_a$ involves N_c which is SNEDECOR's (1946) k_0 average for a simple hierarchy.

Each of the three methods, including both the weighted and unweighted analysis for method $\mathcal{S}$, provides unbiased estimators of the components of variance and of any linear functions or translations of them,

TABLE 3

Analysis of variance

Source	df	Mean squares*
		a. One subpopulation
Individuals	$w_{bk} = N_k - 1$	$\mathcal{S}_{bk} = \frac{\sum_i (\sum_j x_{kij})^2}{2(N_k-1)} - \frac{(\sum_i \sum_j x_{kij})^2}{2N_k(N_k-1)} = \frac{4N_k\hat{p}_k\hat{q}_k - N_{k1}}{2(N_k-1)}$
Within individuals	$w_{wk} = N_k$	$\mathcal{S}_{wk} = \frac{\sum_i \sum_j x^2_{kij}}{N_k} - \frac{\sum_i (\sum_j x_{kij})^2}{2N_k} = \frac{N_{k1}}{2N_k}$
		b. Several subpopulations
Subpopulations	$M - 1$	Weighted $\mathcal{S}_a = \sum_k N_k(\hat{p}_k - \hat{p})^2/(M-1)$; Unweighted $\mathcal{S}'_a = \sum_k (\hat{p}_k - \bar{p})^2/(M-1)$
Individuals	$N - M$	$\mathcal{S}_b = \frac{\sum_k w_{bk}\mathcal{S}_{bk}}{\sum_k w_{bk}} = \frac{4\sum_k N_k\hat{p}_k\hat{q}_k - N_1}{2(N-M)}$
Within individuals	N	$\mathcal{S}_w = \frac{\sum_k w_{wk}\mathcal{S}_{wk}}{\sum_k w_{wk}} = \frac{N_1}{2N}$

$\mathcal{E}\,\mathcal{S}_{wk} = \sigma^2_{wk}$, $\mathcal{E}\,\mathcal{S}_{bk} = \sigma^2_{wk} + 2\sigma^2_{bk}$, $\mathcal{E}\,\mathcal{S}_w = \sigma^2_w$, $\mathcal{E}\,\mathcal{S}_b = \sigma^2_w + 2\sigma^2_b$,

$\mathcal{E}\,\mathcal{S}'_a = (\sigma^2_w + 2\sigma^2_b + 2N_h\sigma^2_a)/2N_h$ where $N_h = M/\sum_k \frac{1}{N_k}$,

$\mathcal{E}\,\mathcal{S}_a = \sigma^2_w + 2\sigma^2_b + 2N_c\sigma^2_a$ where $N_c = (N - \sum_k N^2_k/N)/(M-1)$

* $\bar{p} = \sum_k \hat{p}_k/M$ is the unweighted average of the sample gene frequencies.
M is the number of subpopulations sampled.

$$\hat{\sigma}^2_{wk} = [(1-\hat{f}_k)\hat{p}_k\hat{q}_k] , \quad \hat{\sigma}^2_{bk} = (\hat{f}_k\hat{p}_k\hat{q}_k) , \quad \hat{\sigma}^2_{wk} + \hat{\sigma}^2_{bk} = \hat{\sigma}^2_k = (\hat{p}_k\hat{q}_k)$$

$$\hat{\sigma}^2_w = [(1-\hat{F})\hat{p}\hat{q}] , \quad \hat{\sigma}^2_b = [(\hat{F}-\hat{\Theta})\hat{p}\hat{q}] , \quad \hat{\sigma}^2_a = (\hat{\Theta}\hat{p}\hat{q})$$

$$\hat{\sigma}^2_a + \hat{\sigma}^2_b = (\hat{F}\hat{p}\hat{q}) , \quad \hat{\sigma}^2_w + \hat{\sigma}^2_b + \hat{\sigma}^2_a = \hat{\sigma}^2 = (\hat{p}\hat{q}) ,$$

where it is the entire term involving the correlation(s) and gene frequencies that is being estimated. The correlations are estimated as ratios of the estimators of components of variance and are not necessarily unbiased,

$$\hat{f}_k = \frac{\hat{\sigma}^2_{bk}}{\hat{\sigma}^2_k} , \quad \hat{F} = \frac{\hat{\sigma}^2_a + \hat{\sigma}^2_b}{\hat{\sigma}^2} , \quad \hat{\Theta} = \frac{\hat{\sigma}^2_a}{\hat{\sigma}^2} , \quad \hat{f} = \frac{\hat{\sigma}^2_b}{\hat{\sigma}^2_w + \hat{\sigma}^2_b} .$$

Whichever method of estimation is used it should be noted that

$$\hat{f} = \frac{\hat{F}-\hat{\Theta}}{1-\hat{\Theta}}$$

so that the estimators operate just as the parameters or the F-statistics.

It may be verified that each method produces the same estimators for a single sample and for samples of equal sizes from several subpopulations. With unequal sample sizes only $\hat{\sigma}^2_w$ does not vary among the three methods, and $\hat{\sigma}^2_a$ may differ for the weighted and unweighted analyses. Some aspects of these differences will be considered after a single sample has been considered in some detail.

One Subpopulation

The relativity of correlations is exemplified by a single subpopulation. We have genes within and between individuals, and the correlation, f_k (=F_k, 1969), is relative to this comparison. In the context of what happens on the average over subpopulations, $\mathcal{E} f_k = (F-\bar{\Theta})/(1-\bar{\Theta})$, but other subpopulations are required to estimate the other correlations.

We may estimate f_k in a reasonably straightforward manner without making any assumptions. Each $(x_{kij} - x_{ki'j'})^2/2$, i and i' distinct, in $\mathcal{D}_{bk}$ provides an estimate of the variance among the least related genes in the hierarchy, i.e., among genes of different individuals, and $(\hat{p}_k\hat{q}_k) = \mathcal{D}_{bk}$ is the estimator of the variance. The covariance of genes within individuals is found relative to that between individuals. Except for corrections for the mean these sample covariances are provided by $\mathcal{P}_{wk}$ and $\mathcal{P}_{bk}$, respectively, so that $(\hat{f}_k\hat{p}_k\hat{q}_k) = \mathcal{P}_{wk} - \mathcal{P}_{bk}$, and

$$\hat{f}_k = \frac{(\hat{f}_k\hat{p}_k\hat{q}_k)}{(\hat{p}_k\hat{q}_k)} = 1 - \frac{2N_{k1}(N_k-1)}{4N^2_k\,\hat{p}_k\hat{q}_k - N_{k1}}$$

Since we can compute correlations and variances without resorting to the analysis of variance table or components of variance then why do so? The main

reason, of course, is that it is in the context of analysis of variance and variance components that methods of estimation and testing of hypotheses are best documented, taught, and most readily available. One difficulty with components of variance as they are usually presented is that they can never be truly negative—only estimates can. FISHER (1925) noted that the lower limit on the intraclass correlation in a simple hierarchy was $-1/(k-1)$ where k is the number of items in each class. For f the number of genes within individuals is 2 so that the lower limit is -1. This limit can be attained only when $p = (1/2)$ and in general is limited to $-q/p$ where $q \leq p$. This dependence on frequency need cause no concern, however. I shall argue later that most likely f should be slightly negative on the average. Thus, it is necessary in the adaptation of variance component methods to recognize the acceptability of negative values for σ_b^2.

Turning now to tests of hypotheses, there is only that concerning f_k or the degree of randomness with which alleles occur together. The hypothesis of most interest is $f_k = 0$ or complete randomness. While no assumptions have been made so far, a hypothesis involves a statement of assumptions to be tested and of course any test statistic involves assumptions about distributions. It was noted (1969) that the analysis of variance in Table 3 provided an approximate F-test of the hypothesis $f_k = 0$. This test is dependent on the assumption of normality, which of course is not true. One problem with the F-test in this particular situation is that we wish to test $f_k = 0$ against the alternatives $f_k < 0$ and $f_k > 0$, and would take the ratio of the larger of $\mathcal{S}_{bk}$ and $\mathcal{S}_{wk}$ to the smaller; but the F-test is tabulated as a one-tailed test and modification of the probability levels of rejection is required. On the other hand the χ^2 test is two-tailed, testing deviations in both directions.

Expected values are required for χ^2 tests. Let those for the model be

AA	$A\bar{A}$	$\bar{A}\bar{A}$	Total
η_{k2}	η_{k1}	η_{k0}	N_k ,

and with f_k general, $\eta_{k1} = 2N_k p_k q_k (1-f_k)$, $\eta_{k2} = N_k p_k - \eta_{k1}/2$ and $\eta_{k0} = N_k q_k - \eta_{k1}/2$. One must construct expected values since gene frequencies are not known. In doing so it is noted that

$$\mathcal{E}\hat{p}_k\hat{q}_k = p_k q_k - \sigma^2_{\hat{p}_k} = p_k q_k - p_k q_k (1+f_k)/2N_k .$$

Under the hypothesis that $f_k = 0$, $\mathcal{E}\hat{p}_k\hat{q}_k = p_k q_k (2N_k-1)/2N_k$, so that $\hat{\eta}_{k1} = \hat{p}_k\hat{q}_k (2N_k)^2/(2N_k-1)$, and consequently, $\hat{\eta}_{k2} = N_k\hat{p}_k - \hat{\eta}_{k1}/2$, $\hat{\eta}_{k0} = N_k\hat{q}_k - \eta_{k1}/2$ are unbiased estimators of the η's ($f_k = 0$), and are minimum variance estimators. This is the same result (1969) but formulated differently. The result is the same as that provided by HOGBEN (1946) and by LEVENE (1949) based on random pairings of genes subject to fixed gene frequencies. While the argument presented here is philosophically different, it is not surprising that the two arguments lead to the same expected values since it is only the nature of pairing of the genes that can be tested. A test of random pairings of genes subject to fixed sample gene frequencies can be made using the sample distribution provided by LEVENE (1949) or HALDANE (1954).

Absence of bias in the constructed expected values depends on the assumption $f_k = 0$, which seems appropriate for testing that hypothesis. If $f_k \neq 0$

$$\mathcal{E}\,\hat{\eta}_{k1} = 2N_k p_k q_k - f_k p_k q_k\, 2N_k/(2N_k - 1)\ .$$

From some points of view it may be desirable to have an expected value constructed such that it is unbiased for all f_k. For example, if p_k were known, then $\eta_{k1} = 2N_k p_k q_k$ would be the expected value used for N_{k1}. Under test now would be deviations due to sample gene frequencies as well as random pairing. In any case one can construct expected values with f_k general which are unbiased estimators of η's for $f_k = 0$

$$\tilde{\eta}_{k1} = 2N_k(\hat{pq}) = 2N_k[\hat{p}_k\hat{q}_k\, N_k/(N_k - 1) - N_{k1}/4N_k(N_k - 1)]$$

$$\tilde{\eta}_{k2} = N_k\hat{p}_k - \tilde{\eta}_{k1}/2, \qquad \tilde{\eta}_{k0} = N_k\hat{q}_k - \tilde{\eta}_{k1}/2$$

I have difficulty in rationalizing that these are appropriate expected values for testing $f_k = 0$. For one thing there is a one-to-one transform between the $\tilde{\eta}$'s and N's; i.e., given the $\tilde{\eta}$'s one can produce the observed values. Also, they are not the best estimators when $f_k = 0$.

SMITH (1970) arrived at a value, H, which he proposed to compare to its standard error. By identification,

$$H \equiv (\hat{f}_k\hat{p}_k\hat{q}_k) = \hat{\sigma}_b^2.$$

Actually, a test of $H = 0$ is a test of $f_k = 0$ and essentially this is what is involved in the analysis of variance F-test. SMITH noted when considering combining estimates from several samples that "for a more accurate study of the situation, some kind of variance component analysis would be desirable."

One of the problems in making corrections to data is that they may invalidate estimation procedures applied to the corrected data. For example, one may note that even with $f_k = 0$ that $\mathcal{E}\, 2N_k\hat{p}_k\hat{q}_k < \mathcal{E}\, N_{k1}$, and make some correction in the observed values to account for this difference. Now, to turn around and ignore these corrections and to estimate f_k, for example, is to compound biases generally, depending of course on the nature of the corrections.

CANNINGS and EDWARDS (1969) made a curious development. They considered the sample genotypic frequencies to be fixed, and within the limits of the sample frequencies let the genes be randomly distributed between maternal and paternal gametic arrays. The expectations of these *constructed* but *unobservable* variables were evaluated; for the heterozygotes, for example, $\mathring{\eta}_{k1} = 2N_k\hat{p}_k\hat{q}_k + N_{k1}/2N_k$. Everything is fixed in this case and there is a one-to-one transform between $\mathring{\eta}$'s and N's just as there is for the $\tilde{\eta}$'s. Even if we ignore the fixed nature of the terms, they are biased if $f_k \neq 0$, $\mathcal{E}\,\mathring{\eta}_{k1} = 2N_k p_k q_k - 2f_k p_k q_k$.

With separate sexes the estimated gene frequency is, of course, an average of the maternal, $\hat{p}'_k$, and paternal, $\hat{p}''_k$, gene frequencies, $\hat{p}_k = (\hat{p}'_k + \hat{p}''_k)/2$. If $\mathcal{E}\,\hat{p}'_k = \mathcal{E}\,\hat{p}''_k = p_k$, i.e., if only sampling differences are involved in the differences between the two gene frequencies, then all of the results presented follow through in the same manner. Some difficulties do arise when the gene frequencies are

truly different between the sexes, $p'_k \neq p''_k$, but the way to investigate this matter is by comparing the gene frequencies of the two sexes.

Several Subpopulations

There is some balance in the data, two genes per individual, so that all three methods provide the same estimator of $\hat{\sigma}^2_w$ which is the minimum variance unbiased estimator. With unequal sample sizes the estimators of σ^2_b and σ^2_a and any functions of them vary among the three methods, and of σ^2_a differ between the weighted and unweighted analysis $\mathcal{S}$. No documentation of the best procedure is available, the main problem being that the best procedure is dependent upon the unknown parameters to be estimated. For example, consider the weighted and unweighted analysis for method $\mathcal{S}$. The unweighted analysis is best if $\bar{\Theta} = 1$ and the weighted analysis is best if $\bar{\Theta} = 0$, and the best is usually somewhere in between. Method $\mathcal{P}$ is somewhere in between. So is method $\mathcal{D}$, and KOCH (1968) points out that the variances of estimators of variance components by method $\mathcal{D}$ do not involve the mean for normally distributed x's in contrast to estimators by method $\mathcal{P}$. While this is a desirable property its relevance to the situation at hand where variances of estimators always involve p is not clear. One argument in favor of the unweighted analysis over the weighted one is that if the variation within subpopulations is heterogeneous the estimator of σ^2_a by the unweighted analysis is still unbiased. All of these considerations are for estimators of variance components and do not necessarily carry over to estimators of the intraclass correlations about which little is known.

Tests of significance by χ^2 or F-tests are all on much less sound grounds than estimation. The F-test for H_{02} or $\bar{\Theta} = 0$ (1969) appears to have, in practice, the one-tailed F-test orientation against the alternative of $\bar{\Theta} > 0$. For the weighted analysis it is $T_{ab} = \mathcal{S}_a/\mathcal{S}_b$ and for the unweighted analysis, $T_{ab} = 2N_h\mathcal{S}'_a/\mathcal{S}_b$. The other main hypothesis, H_{01} or $F-\bar{\Theta} = 0$, has, in practice, the alternatives $F-\bar{\Theta} < 0$ and $F-\bar{\Theta} > 0$. This may appear clearer in terms of f. Since $f = (F-\bar{\Theta})/(1-\bar{\Theta})$, then $1-F = (1-f)(1-\bar{\Theta})$ and $1-F+2(F-\bar{\Theta}) = (1+f)(1-\bar{\Theta})$ so that $\mathcal{E}\,\mathcal{S}_b = (1+f)(1-\bar{\Theta})pq$, $\mathcal{E}\,\mathcal{S}_w = (1-f)(1-\bar{\Theta})pq$. Obviously, the argument on the mean squares reverses as f is positive or negative just as for f_k in a single sample. Possibly χ^2 tests are better for hypotheses involving variation within subpopulations. However, there are several ways of making χ^2 tests, with the best method generally unknown.

How estimators vary with assumptions about the parameters of the model is illustrated in conjunction with arriving at expected frequencies for χ^2 tests of significance. For example for a χ^2 test of the composite hypothesis $\bar{\Theta} = f_k = 0$ (all k), i.e. all genes associated at random, we need only an estimator of pq. For these assumptions the three mean squares for the weighted analysis are pooled to provide the best estimator

$$(\overline{pq}) = \frac{(M-1)\mathcal{S}_a + (N-M)\mathcal{S}_b + N\mathcal{S}_w}{2N-1} = \frac{2N\hat{p}\hat{q}}{2N-1}$$

which can also be obtained by appropriate pooling of the quadratic forms for methods $\mathcal{D}$ and $\mathcal{P}$. Expected values are constructed as

$$\bar{\eta}_{k1} = 2N_k(\overline{pq}), \quad \bar{\eta}_{k2} = N_k\hat{p} - \bar{\eta}_{k1}/2, \quad \bar{\eta}_{k0} = N_k\hat{q} - \bar{\eta}_{k1}/2$$

and

$$\chi^2_1 = \sum_{k,i} (N_{ki} - \bar{\eta}_{ki})^2/\bar{\eta}_{ki}$$

with $2M-1$ df. If this test is significant it may be for one or both of two reasons, all $f_k \neq 0$ or $\bar{\Theta} \neq 0$. For all $f_k = 0$ the expected values are those given previously for each subpopulation,

$$\hat{\eta}_{k1} = 4N_k^2\hat{p}_k\hat{q}_k/(2N_k-1), \quad \hat{\eta}_{k2} = N_k\hat{p}_k - \hat{\eta}_{k1}/2, \quad \hat{\eta}_{k0} = N_k\hat{q}_k - \hat{\eta}_{k1}/2$$

$$\chi^2_2 = \sum_{k,i} (N_{ki} - \hat{\eta}_{ki})^2/\hat{\eta}_{ki}$$

with M df is just the sum of the M individual χ^2's for $f_k = 0$. Before proceeding to $\bar{\Theta} = 0$ consider the fitting of the mean f in testing the hypothesis all $f_k = f$. The expected values are

$$\check{\eta}_{k1} = 2N_k(1-\hat{f})\,(\hat{p_k q_k}), \quad \check{\eta}_{k2} = N_k\hat{p}_k - \check{\eta}_{k1}/2, \quad \check{\eta}_{k0} = N_k\hat{q}_k - \check{\eta}_{k1}/2,$$

in which is utilized the unbiased estimator $(\hat{p_k q_k})$ in each sample for $p_k q_k$ whatever f_k is, and $\hat{f} = \hat{\sigma}_b^2/(\hat{\sigma}_w^2 + \hat{\sigma}_b^2)$ estimated by one of the three methods. Then,

$$\chi^2_3 = \sum_{k,i} (N_{ki} - \check{\eta}_{ki})^2/\check{\eta}_{ki}$$

with $M-1$ df is a test of homogeneity of the f_k's about their mean. It corresponds in some ways to BARTLETT's test of homogeneity of variances, but in this case, of homogeneity of variances within subpopulations after adjusting for differences in gene frequencies and overall f.

The usual method for testing $\bar{\Theta} = 0$ is to compute the interaction χ^2 for the $2 \times M$ table of gene frequencies

$$\chi^2_4 = \frac{\sum_k 2N_k(\hat{p}_k - \hat{p})^2}{\hat{p}\hat{q}} = \frac{(M-1)\mathcal{S}_a}{\hat{p}\hat{q}}$$

with $M-1$ df, where $\mathcal{S}_a$ is the mean square for the weighted analysis. There is also the difference $\chi^2_{1-2} = \chi^2_1 - \chi^2_2$ with $M-1$ df which provides a test of $\bar{\Theta} = 0$. Alternatively, there is the F-test, $T_{ab} = \mathcal{S}_a/\mathcal{S}_b$. Which of these tests is better is a matter for further investigation.

An overall test of $f = 0$ or $F = \bar{\Theta}$ may be obtained in the following manner. Under the hypothesis $f = 0$, $\mathcal{E}\,\mathcal{S}_b = \mathcal{E}\,\mathcal{S}_w$, which implies that

$$\mathcal{E}\,N_1 = \mathcal{E}\,4N \sum_k N_k\hat{p}_k\hat{q}_k/(2N-M), \text{ and}$$

$$\hat{\eta}_1 = 4N \sum_k \hat{p}_k\hat{q}_k/(2N-M), \quad \hat{\eta}_2 = N\hat{p} - \hat{\eta}_1/2, \quad \hat{\eta}_0 = N\hat{q} - \hat{\eta}_1/2$$

are unbiased expected values for $\chi^2_5 = \sum_i (N_i - \hat{\eta}_i)^2/\hat{\eta}_i$ with 1 df. The difference $\chi^2_{2-3} = \chi^2_2 - \chi^2_3$ with 1 df is also a test of $f = 0$. Alternative unbiased expected values are found for χ^2_5 by similar applications of methods $\mathcal{P}$ and $\mathcal{D}$.

A rough measure of the importance of the sources of variation can be obtained by apportioning the total χ^2_1 as follows:

Source	df	χ^2 Variation
f	1	χ^2_5
$(f_k - f)$'s	$M-1$	$\chi^2_2 - \chi^2_5$
$\bar{\Theta}$	$M-1$	$\chi^2_1 - \chi^2_2$

One divides through by df to obtain the relative importance of each source. However, these mean χ^2 variations suffer just as do mean squares in reflecting the portion of the total variation due to each source.

DISCUSSION

It is interesting that WRIGHT's F-statistics and WAHLUND's components of variance for subdividing are in principle two sides of the same coin, although the latter parameterization has not been documented like the F-statistics. It is common (e.g., BARRAI 1971) to work with both f and variances, f for the variation within subpopulations and variance for variation among subpopulations. Often in theory subpopulations are considered infinite (WAHLUND 1928). There is nothing wrong with mixtures of parameterizations or of treating infinite subpopulations as long as matters are kept straight. It is actually f_k's that are being treated in this case and f does not reflect the forces under the general label of inbreeding as does F. As an example, consider an array of subpopulations each of infinite size and characterized by f_k and p_k. Let

$$p = \mathcal{E}p_k, \qquad f = \mathcal{E}f_k p_k q_k / \mathcal{E}p_k q_k, \qquad \mathcal{E}(p_k - p)^2 = \sigma_a^2$$

($\mathcal{E}$ in this case just replaces summation notation divided by the number of subpopulations if it is finite). The variance among means of subpopulations of infinite size and around a known mean, p, is the component of variance, $\sigma_a^2 = \bar{\Theta}pq$. Then, letting all $N_k \to \infty$ for $\mathcal{S}_w$ and $\mathcal{S}_b$,

$$\sigma_w^2 = \mathcal{E}(1-f_k)p_k q_k = (1-f)(1-\bar{\Theta})pq = (1-F)pq$$

$$\sigma_w^2 + 2\sigma_b^2 = \mathcal{E}[2p_k q_k - (1-f_k)p_k q_k] = (1+f)(1-\bar{\Theta})pq = [1-F+2(F-\bar{\Theta})]pq,$$

and $\sigma_b^2 = (F-\bar{\Theta})pq$. By turning the relationship around

$$F = f + \bar{\Theta}(1-f),$$

and even if all $f_k = 0$ or if they average to $f = 0$, $F = \bar{\Theta}$. This demonstrates again the composite nature of F which takes into account all contributions of genes being alike within individuals in contrast to unrelated genes in the population. When LI (1969) amalgamates two isolates and shows that although each has

no correlation there is a correlation in the amalgamated population, it is the f's that are zero, and $F = \bar{\Theta}$ that is produced in the total population. This point is stressed because f is often discussed in the context of F and they are usually very different in practice.

The methods of estimation presented are small sample methods and are simply an adaptation of those developed in the field of statistics for variance components. When it was recognized that components of variance could be logically defined for genes within individuals and between individuals in the same subpopulation, the rest followed.

I have treated the components of variance and the correlational parameterizations in parallel. While they are equivalent parameterizations, each appears to have advantages of clarity in some contexts and the two together give a better documentation of the situation, both in theory and in estimation. It is the components of variance for which unbiased estimators are available from small samples. The small sample estimators of the correlations are ratios of unbiased estimators; but no unbiased estimators of the correlations appear to be available.

One gets the impression that some authors are trying to describe finite subpopulations or populations as they are exactly currently constituted, as would forecasters or econometricians in evaluating the production of a crop in a particular area either before or after harvest. Sampling procedures and methods of treatment, generally found in textbooks under sampling from finite populations, are for so-called fixed populations and are different from those outlined. The methods outlined provide unbiased estimators (from random samples) of those quantities produced by applying precisely the same methods to the entire population or subpopulations although they be finite. In practice, and in the context of the information available by these procedures, it seems appropriate to view natural populations as samples in time and space.

Turning now to factors affecting the correlations, those mentioned often are inbreeding, selection, migration, differential fertility, assortative mating, differences in frequencies of genes in male and female gametes, and mixtures of subdivisions. In discussing these factors it is necessary to distinguish those effects which lead to differences among subpopulations from those effects which arise from what goes on within subpopulations. If genes are randomly distributed within subpopulations then differences among them contribute equally to F and $\bar{\Theta}$ and nothing to f. Thus f is almost entirely a consequence of effects within subpopulations.

First consider the effects of inbreeding for neutral genes. It is necessary to distinguish between drift inbreeding, Θ, among subpopulations due almost entirely to finite sizes of subpopulations over time, and inbreeding, f, within subpopulations due entirely to the system of mating within subpopulations. With random mating and separate sexes $\bar{\Theta}_t = F_{t+1}$ (t indexes generations) so that

$$f_t = (F_t - F_{t+1})/(1 - F_{t+1}) \cong -1/(2n_e - 1)$$

where n_e is the effective size of subpopulations. The result is a slightly negative f. Random mating includes brother-sister matings and other matings of relatives. Any avoidance of mating of these relatives will make f more negative. Only if

mates are more related than a random pair of individuals (ROBERTSON 1964), including parents and offspring and all the other relatives generally found in data, will f be positive. Inbreeding subdivisions or isolates must be continued largely separated over time to make f positive. The extreme form of this is obligatory self-fertilization. In essence the effect on f of inbreeding isolates, with interchange among them, corresponds to the effect on $\bar{\Theta}$ of subpopulations with migration. Other types of isolates will be considered under mixtures of subdivisions.

While the system of mating within subpopulations determines f it also influences the rate of increase in $\bar{\Theta}$ if one includes factors such as the variation in gametes per parent (or family sizes) and inbreeding isolates. These factors are discussed in detail by COCKERHAM (1970). However, $\bar{\Theta}$ is a result of reasonably long-term effects and without historical information one cannot know the rate with which a particular $\bar{\Theta}$ was attained. On the other hand f reflects the more immediate mating system.

Selection should also be considered in two ways. One is the long-term consequences with primary effects equally on F and $\bar{\Theta}$ and not f and the other is the consequences of the immediate effects of selection reflected primarily in f. The long-term effects of selection of any sort operating consistently in the subpopulations would tend to make them more alike, and thus the effect is to reduce $\bar{\Theta}$ since it can only be measured as $\bar{\Theta} - \bar{\Theta}_g$. The change, $\bar{\Theta}_g$, in the entire population from some founder population may be large, but information other than status quo is required to know or estimate it. This conclusion must be modified if selection is different among the localities occupied by the subpopulations. If there is an interaction between selection and environmental niches, different genes being favored, the result is to lead to differentiation among the subpopulations and to increase $\bar{\Theta}$. Also, drift due to finite size of subpopulations may occasionally lead to fixation of different alleles. With interaction of nonalleles this may cause differentiation in selection among subpopulations and an increase in $\bar{\Theta}$. Migration (to be discussed) would tend to avert differential fixation, however.

The immediate consequences on f of selection for alleles at a locus were given by WALLACE (1958), by LEWONTIN and COCKERHAM (1959) and were discussed in detail by WORKMAN (1969). In essence, unless the viability of the heterozygote is less or equal to the geometric mean of the viabilities of the two homozygotes, the effect is to make f negative.

I shall group the effects of differential fertility and unequal male and female gametic gene frequencies together, since the former (PURSER 1966) has its effect through the latter. Of course any long-term differential fertility effects would come under selection. ROBERTSON (1965) discusses the consequences of the unequal frequencies which have a negative effect on f. Apparently BRUCE (1910) was the first to notice that differential frequencies in uniting male and female gametes produced a decrease in homozygotes and an increase in heterozygotes. Interestingly enough if there is random union of male and female genes they are independent, but there is an excess of heterozygotes over Hardy-Weinberg frequencies (also independent) when the two frequencies differ. If only sampling differences are involved then there is no problem, as was pointed out previously,

when small sample methods are used. If a consistent difference in gene frequencies between the sexes is suspected, as might result from differential selection, then a comparison of the gene frequencies of males with that of females provides the most pertinent information. Differential fertility is more difficult to study and requires an analysis of parents and children.

Assortative mating will make a positive contribution to f and disassortative mating will make a negative contribution to f. The analyses considered herein really provide no direct information on the mating system. An analysis of mates is required to produce direct information on this point. It is a simple matter to adapt these small sample methods to the analysis of mates.

Mixtures of subdivisions do not always have the effects imputed to them. I shall consider the two types, isolates within subpopulations (model 2) and areas of subpopulations (model 3). There can, of course, be many types of isolates within subpopulations. First, consider the isolates to be families, i.e. parents and children and possibly some other close relatives such as grandparents. The analysis can be partitioned accordingly to etimate σ^2_{b1} and σ^2_{b2} and the correlations (model 2). Now, if there is, in general, random mating in the subpopulations, the members of the families will be much more related than they are inbred, and $f_1 = (F-\bar{\Theta}_1)/(1-\bar{\Theta}_1)$ could be reasonably negative, say $-.15$. On the other hand immediate family members will also be much more related than members of different families, and $f_2 = (\bar{\Theta}_1-\bar{\Theta}_2)/(1-\bar{\Theta}_2)$ will be reasonably positive. The net effect of ignoring family structure is to average these two f's, $f = (F-\bar{\Theta})/(1-\bar{\Theta})$ where $\bar{\Theta}$ is an average of $\bar{\Theta}_1$ and $\bar{\Theta}_2$, but f will still be zero or slightly negative unless parents are more related than random pairs of individuals in the subpopulations.

Next consider the isolates to be inbreeding units with no exchange among them. With random mating within isolates $f_1 = 0$, but because of drift or other effects among isolates f_2 can be reasonably positive. In this case to ignore the isolates leads to a positive f as an average. It does not take a great deal of interchange (migration) among the isolates to undo this positive effect, as we shall see, and the question of f positive or negative still reduces to whether mates are more- or less-related than random members in the subpopulations. One would want to take account of inbreeding isolates in any case.

It is at the subpopulation level that subdivisions probably cause the most trouble in practice. Some subpopulations may have been from a relatively recent common origin, these being more related than more distantly separated subpopulations in time. Of course historical information is required to take account of these matters in an analysis. Migration tends to make populations in close proximity more alike than are those more distantly separated in space. It is much more important to take account of these differences than those within subpopulations. As was shown with model (3) if certain groups of subpopulations are more closely related than others then to ignore this is to work with less than the total variance available, with the correlations being correspondingly reduced.

The extreme of subdividing is to consider the population a continuum with relationships indexed according to distances, as did MORTON, MIKI and YEE (1968) in fitting distance models. For codominant genes considered herein one

can apply methods $\mathcal{P}$ or $\mathcal{D}$ for genes separated by distance d, with some reasonable grouping of genes into classes of distances (1969). Here we consider pairs of genes and not genotypes as did MORTON, MIKI and YEE (1968). Let $d = 0, 1, 2, \ldots l$ where 0 is for genes with themselves, 1 is for genes in the same individual, and $d > 1$ for genes separated according to the distance structure with l corresponding to the greatest distance. Then for method $\mathcal{D}$,

$$\mathcal{D}_0 \equiv 0, \quad \mathcal{E}\,\mathcal{D}_1 = (1-F)pq, \quad \mathcal{E}\,\mathcal{D}_l = (pq)$$

and

$$\hat{F} = 1 - \mathcal{D}_1/\mathcal{D}_l, \quad \hat{\bar{\Theta}}_d = 1 - \mathcal{D}_d/\mathcal{D}_l, \quad 1 < d < l.$$

A comparable adaptation of method $\mathcal{P}$ is

$$\hat{F} = (\mathcal{P}_1 - \mathcal{P}_l)/(\mathcal{P}_0 - \mathcal{P}_l), \quad \hat{\bar{\Theta}}_d = (\mathcal{P}_d - \mathcal{P}_l)/(\mathcal{P}_0 - \mathcal{P}_l), \quad 1 < d < l.$$

Having estimated the correlations, then, one may fit distance models. Alternatively, one may use different groupings for which to fit migration models. Note that the correlation of genes separated by the greatest distance is set to zero and that the other correlations are relative to this correlation if it is not zero. There is some difficulty, in practice, in obtaining the local, f_0, and overall mean, $\bar{f}$, correlations—utilized in some detail by CROW and MARUYAMA (1971)—since quadratic functions of gene frequencies must also be estimated. While these authors were considering multiple alleles, the same principle applies. Implicit in their formulations is that the mean gene frequency, p_i, over replicates of their global finite system is zero and consequently $\sum_i p^2_i = 0$, which is inherent in the assumption of an infinite number of neutral alleles.

The obvious long-term effects of migration are to reduce F and $\bar{\Theta}$ from what they would be without migration. The immediate effect of a migrant gamete on f is nil. The immediate effect of a migrant individual on f is to increase f very slightly and this effect is lost with mating and death of the migrant. The theory of isolation with migration has been considered in detail by WRIGHT (1943, 1946 and 1951), by MALÉCOT (1948, 1967) and more recently by MARUYAMA (1970).

With relation to MARUYAMA's (1970) island model, using his f_0 and f_1,

$$F = \bar{\Theta} = f_0, \quad \bar{\Theta}_g = f_1$$

and what we would estimate as $\bar{\Theta}$ is $(f_0-f_1)/(1-f_1)$. For his other models $\bar{\Theta}_g$ is the average of all his f's save f_0.

The effect of isolation at equilibrium, using MARUYAMA's island population measures, is

$$\bar{\Theta} = (f_0-f_1)/(1-f_1) = 1/[2N/(1-u)^2(1-m)(1-m-2m/n)-2N+1]$$

where N is size of subpopulation (not sample), nN is the size of the total population, m is the migration rate, and u is the mutation rate. (MARUYAMA's formulations for the island model need a slight modification due to the finiteness of n but are accurate enough for present purposes.) For $u \ll m$ and both small

$$\bar{\Theta} \cong 1/[1+4Nm(n+1)/n] \cong 1/(1+4Nm) \text{ if } n \text{ large.}$$

The last expression was given by WRIGHT (1943). While for purposes of fixation

one migrant gamete per subpopulation per generation is sufficient to make the population behave as one panmictic unit (or slightly larger), this rate of migration does allow considerable differentiation among subpopulations, $\bar{\Theta}$ slightly less than 1/3. A migrant individual per subpopulation per generation reduces $\bar{\Theta}$ to something less than 1/5 and so on. For other models, (MARUYAMA 1970), the average differentiation among subpopulations is even larger. However, it certainly does not take a lot of migration to make $\bar{\Theta}$ very small. The same principle is involved with inbreeding isolates within subpopulations with interchange or migration of individuals among isolates. It does not take much interchange among them to make f very small. It would appear that the mating system within many human subpopulations would tend to make f very slightly negative.

Many helpful suggestions and comments were made by DRS. PETER M. BURROWS and BRUCE S. WEIR.

LITERATURE CITED

AIRY, G. B., 1861 *On the algebraical and numerical theory of errors of observations and the combination of observations*. Macmillan & Co., London.

BARRAI, I., 1971 Subdivision and inbreeding. Am. J. Human Genet. **23**: 95–96.

BRUCE, A. B., 1910 The Mendelian theory of heredity and the augmentation of vigor. Science **32**: 627–628.

CANNINGS, C. and A. W. F. EDWARDS, 1969 Expected genotypic frequencies in a small sample: Deviation from Hardy-Weinberg equilibrium. Am. J. Human Genet. **21**: 245– 247.

COCKERHAM, C. C., 1963 Estimation of genetic variances. Statistical Genetics and Plant Breeding, NAS-NRC **982**: 53–94. ——, 1969 Variance of gene frequencies. Evolution **23**: 72–84. ——, 1970 Avoidance and rate of inbreeding. pp. 104–127. In: *Mathematical Topics in Population Genetics*. Edited by K. Kojima. Springer-Verlag, New York.

CROW, J. F. and T. MARUYAMA, 1971 The number of neutral alleles maintained in a finite, geographically structured population. Theoret. Pop. Biol. **2**: 437–453.

DANIELS, H. E., 1939 The estimation of components of variance. J. Roy. Stat. Soc. Supp. **6**: 186–197.

FISHER, R. A., 1918 The correlation between relatives on the supposition of Mendelian inheritance. Trans. Roy. Soc. Edinburgh **52**: 399–433. ——, 1925 *Statistical methods for research workers* (and later editions). Oliver and Boyd, Edinburgh.

HALDANE, J. B. S., 1954 An exact test for randomness of mating. J. Genet. **52**: 631–635.

HARVILLE, D. A., 1969 Variance component estimation for the unbalanced one-way random classification—a critique. *Aeronautical Research Laboratories Report 69–0180*. Department of Commerce, Washington, D. C.

HOGBEN, L., 1946 *An introduction to mathematical genetics*. W. W. Norton and Company, Inc., New York.

KEMPTHORNE, O., 1957 *An introduction to genetic statistics*. John Wiley and Sons, Inc., New York.

KOCH, G. G., 1967 A general approach to the estimation of variance components. Technometrics **9**: 93–118. ——, 1968 Some further remarks concerning "A general approach to the estimation of variance components." Technometrics **10**: 551–558.

LEVENE, H., 1949 On a matching problem arising in genetics. Ann. Math. Stat. **20**: 91–94.

LEWONTIN, R. C. and C. C. COCKERHAM, 1959 The goodness-of-fit test for detecting natural selection in random mating populations. Evolution **13**: 561–564.

LI, C. C., 1969 Population subdivision with respect to multiple alleles. Ann. Human Genet. **33**: 23–29.

LUSH, J. L., 1947 Family merit and individual merit as bases for selection. Parts I, II. The American Naturalist **81**: 241–261, 362–379. ——, 1948 The genetics of populations. Ames, Iowa. (Mimeo.)

LUSH, J. L. and A. E. MOLLN, 1942 Litter size and weight as permanent characteristics of sows. U. S. Department of Agriculture Technical Bulletin No. 836, 40. pp.

LUSH, J. L. and F. S. STRAUS, 1942 The heritability of butterfat production in dairy cattle. Jour. Dairy Sci. **25**: 975–982.

MALÉCOT, G., 1948 *Les mathématiques de l'hérédité*. Masson et Cie, Paris. ——, 1967 Identical loci and relationship. In: "Proceedings Fifth Berkeley Symposium Mathematical Statistics and Probability" **4**: 317–332. Univ. Calif. Press, Berkeley.

MARUYAMA, T., 1970 Effective number of alleles in a subdivided population. Theoret. Pop. Biol. **1**: 273–306.

MORTON, N. E., C. MIKI, and S. YEE, 1968 Bioassay of population structure under isolation by distance. Am. J. Human Genet. **20**: 411–419.

PURSER, A. F., 1966 Increase in heterozygote frequency with differential fertility. Heredity **21**: 322–327.

RAO, C. R., 1971a Estimation of variance and covariance components—MINQUE theory. J. Multivariate Anal. **1**: 257–275. ——, 1971b Minimum variance quadratic unbiased estimation of variance components. J. Multivariate Anal. **1**: 445–456. ——, 1972 Estimation of variance and covariance components in linear models. J. Amer. Stat. Assoc. **67**: 112–115.

ROBERTSON, A., 1964 The effect of non-random mating within inbred lines on the rate of inbreeding. Genet. Res. **5**: 164–167. ——, 1965 The interpretation of genotypic ratios in domestic animal populations. Animal Prod. **7**: 319–324.

SEARLE, S. R., 1971a Topics in variance component estimation. Biometrics **27**: 1–76. ——, 1971b *Linear models*. John Wiley and Sons, Inc. New York.

SMITH, C. A. B., 1970 A note on testing the Hardy-Weinberg Law. Ann. Human Genet. **33**: 377–383.

SNEDECOR, G. W., 1946 *Statistical methods*. The Iowa State College Press, Ames, Iowa.

TIPPETT, L. H. C., 1931 *The methods of statistics*. Williams and Norgate, London.

WAHLUND, S., 1928 Zusammensetzung von populationen und korrelationserscheinungen vom standpunkt der vererbungslehre aus betrachtet. Hereditas **11**: 65–106.

WALLACE, B., 1958 The comparison of observed and calculated zygotic distributions. Evolution **12**: 113–115.

WINSOR, C. P. and G. L. CLARKE, 1940 A statistical study of variation in the catch of plankton nets. Journal of Marine Research **3**: 1–34.

WORKMAN, P. L., 1969 The analysis of simple genetic polymorphisms. Human Biol. **41**: 97–114.

WRIGHT, S., 1935 The analysis of variance and the correlations between relatives with respect to deviations from an optimum. J. Genet. **30**: 243–256. ——, 1943 Isolation by distance. Genetics **28**: 114–138. ——, 1946 Isolation by distance under diverse systems of mating. Genetics **31**: 39–59. ——, 1951 The genetical structure of populations. Annals of Eugenics **15**: 323–354. ——, 1965 The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution **19**: 395–420.

YATES, F. and I. ZACOPANAY, 1935 The estimation of the efficiency of sampling, with special reference to sampling for yield in cereal experiments. J. Agri. Sci. **25**: 545–577.