Journal paper no. J-2549 of the
IOWA AGRICULTURAL EXPERIMENT STATION, AMES, IOWA. Project no. 890
Reprinted without change of pagination from the
Proceedings of the Royal Society, B, *volume* 143, pp. 103–113,

The correlation between relatives in a random mating population*

By O. Kempthorne
Iowa State College

(*Communicated by Sir Ronald Fisher, F.R.S.—Received* 6 *April* 1954)

The variance of a characteristic in a random mating diploid population is partitioned into two components, the genotypic variance and the environmental variance. The genotypic variance is then separated into components which can be attributed to the additively genetic effects, the dominance deviations and the epistatic effects. The total ep [illegible] subdivided into components arising from interactions among sets of 2, 3, [illegible] interaction variance from a particular number of loci is shown to be compos [illegible] attributable in a certain sense to the interaction of additive and/or dominance [illegible] genes. For example, the epistatic variance arising from two loci consists of three [illegible] parts: additive-by-additive, additive-by-dominance and dominance-by-dominance variance.

It is then shown that the covariances, and hence the correlations, between relatives which are not inbred, can be expressed in terms of the components of genotypic variance. The basic result is that if σ^2_X is a component of the genotypic variance, where X contains r additive and s dominance effects, the covariance between two relatives with respect to σ^2_X is equal to

$$[(\tfrac{1}{2})\,(\phi+\phi^1)]^r\,[\phi\phi^1]^s,$$

where ϕ and ϕ^1 are easily calculated from the pedigree of the relatives.

Introduction

Numerous workers have examined the theoretical values for the correlations between relatives since 1900. Pearson (1903) examined the very particular case when there are two alleles with equal frequencies. Weinberg (1908, 1909) in 1908 examined in some detail the more general case when the number of alleles is arbitrary. Most of his work on alternative inheritance was concerned with the case when the alleles were additive in their effects. He also appears to have made an attack on the case when there is dominance in that he talks about 'Prevalenz'. His solution is not at all general, and it does not appear to the present author that any claims can be made that Weinberg obtained an exact solution.

The first exact solution with respect to an arbitrary number of alleles appears to be that given by Fisher (1918). In this paper Fisher made an exact analysis with no assumptions at all except those included generally under the term 'random mating', i.e. no selection, no differential viability of genotypes, no maternal effects etc.

It is also worth noting that Fisher was the first to tackle in an exact manner the question of epistacy, although he restricted consideration to what he termed 'dual epistacy', that is, interaction between pairs of loci. Fisher also dealt with homogamy; his treatment on this matter was approximate, first, in the mathematical sense in that quantities of second and higher order were ignored and secondly, in that epistacy was ignored. This work appears to be the only treatment in the literature

---

* Journal Paper No. J-2549 of the Iowa Agricultural Experiment Station, Ames, Iowa. Project 890.

on the subject which allows for dominance and does not appear to have been extended. Fisher's application of his results to human data is questionable because the correlations, parent-offspring and full-sib, which he took from Pearson & Lee (1903), are in reality based on a mixture of at least two bodies of data and contain both environmental and genotypic components. Also Fisher made an attack on the influence of linkage showing that under his particular assumptions a pair of linked loci behaved essentially as one locus with an appropriate number of alleles.

Wright (1921) in a series of papers examined the matter of correlations between relatives by his method of path coefficients and obtained a large number of results for the simple case when there is no dominance and no interaction between loci. The method of path coefficients gave the correct result for this situation, but cannot from its nature give the correct result in general. Thus the correlation between full sibs with regard to dominance deviations was not discovered by Wright using the method of path coefficients.

Malécot (1948) gave an exact meaning to Wright's coefficient F, which Wright had labelled as the 'correlation between uniting gametes' by distinguishing between genes which are alike in state, as, for instance, two genes A and A, and two genes which are alike by descent because one of the two genes is an exact copy (excepting mutation) of the other or the two genes are both exact copies of some one gene which occurred prior to both in the ancestry. By synthesizing Wright's notion of correlation, modified to a probability and Fisher's notions in his 1918 paper, Malécot showed that the covariance of two individuals is equal to

$$\tfrac{1}{2}(\phi+\phi')\,\sigma_A^2+\phi\phi'\sigma_D^2,$$

where σ_A^2 is the additively genetic variance, σ_D^2 is the dominance variance and ϕ and ϕ' are the coefficients of relation between the two individuals via their two loci. Malécot's treatment then provides a general answer for the case of random mating when there is one locus with arbitrary dominances amongst the alleles, of which there is an arbitrary number. The only restriction in Malécot's work is that there is no inbreeding, that is, there is zero probability of any individual possessing two copies of the same gene in the original random mating population.

The purpose of the present paper is to demonstrate the completely general partition of variance in a random mating population and to exhibit the covariances of any two individuals with no inbreeding in the sense given above. The practical utility of the results will not be discussed. The aim here is merely to record the general complete mathematical solution for a mathematical population which is the usual one considered. It would be nice to generalize the results to the case when linkage occurs, but this will not be done herein.

BASIC DEFINITIONS

We have an arbitrary random mating population with n loci which will be denoted by superfixes running from 1 to n. At the ith locus it will be supposed that there are m_i alleles which will be denoted by A_j^i, where j runs from 1 to m_i. (In fact the numbers m_i will never appear in the formula.) The gene frequency of the jth allele at the ith locus will be denoted by p_j^i. The array of genotypes in the population, because of random mating, is equal to

$$\prod_{i=1}^{n}\{\Sigma p_j^i A_j^i\}^2.$$

This will be referred to as the genotypic array.*

In the idealized situation usually considered the observation on an individual with a certain genotype, which is usually called the phenotypic value of the individual, is made up additively of two parts:

(1) The average of the possible phenotypic values for a genotype over a population of environments; this is usually called the genotypic value of the genotype, following Fisher (1918).†

(2) An environmental deviation, which will have an average value of zero over the population of possible environments.

As a rough and quick shorthand one may write

$$P = G + E,$$

where the letters have an obvious connotation and G and E are assumed to be uncorrelated. Thus in this work, as indeed in all work in the area that the author knows, genotype-environment interactions are assumed to be zero. We shall need terms for the variances of P, G and E, and the symbols σ_P^2, σ_G^2 and σ_E^2‡ will be used.

THE PARTITION OF VARIANCE

One method by which one may obtain the various components of variance is a stepwise one. The additively genetic variance which will be denoted by σ_A^2‡ may be defined as that portion of the total variance in the phenotypic values which can be attributed, in a least-squares sense, to the effects of the genes. If we have one locus, for example, with genes A_1, A_2, A_3, then σ_A^2 is equal to

$$\sigma_G^2 - \min_{(\alpha_1,\alpha_2,\alpha_3)} \sum_{j,j'} p_j p_{j'}(y_{jj'} - \alpha_j - \alpha_{j'})^2.$$

In other words, σ_A^2 is the sum of squares attributable to a regression on a known linear function of unknown gene effects (see any statistical text (e.g. Kempthorne 1952, chap. 5)).

We may then define the variance attributed to loci by finding the sum of squares attributable to a regression of genotypic values on a linear function of unknown loci effects, one for each combination of alleles of each locus. This sum of squares we might denote by σ_L^2 say, and we would then obtain the dominance variance σ_D^2 as

$$\sigma_D^2 = \sigma_L^2 - \sigma_A^2.$$

---

* Some people prefer the term zygotic array for this quantity.

† The author has heard criticisms of the use of this term for its present purpose. No one appears, however, to have suggested any substitute.

‡ The symbol σ_G^2 is frequently used to denote the variance of additively genetic values. For obvious reasons it is desirable to have a symbol for the variance of genotypic values and for the variance of additively genetic values. The use of the symbol σ_G^2 for additively genetic variance can give the impression that this variance may be attributed to genotypes. Fisher used the term genetic variance to characterize what we call here the additively genetic variance.

Finally, we would obtain the epistatic variance, σ_I^2, by the equation

$$\sigma_I^2 = \sigma_G^2 - \sigma_L^2.$$

In total then we would have, since G and E are uncorrelated,

$$\sigma_P^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2 + \sigma_F^2.$$

The above process is not only tedious but also obscures the basic fact that the variance σ_I^2 is partitionable and should be partitioned. Fisher's (1918) work on dual epistacy mentioned in the introduction was sufficient to show that certain components of interaction or epistacy enter into covariances between relatives while others do not.

The procedure which will be followed is based very much on the derived models in factorial experiments (see Kempthorne 1952, §17·3) and was used by Anderson (1953) and Anderson & Kempthorne (1954) in a particular case when parents are completely inbred. The procedure is very simple, and to illustrate to the non-mathematically minded we shall first demonstrate its application to the case when there is one locus and an arbitrary number of alleles, m say. We use the symbol $A_j A_k$ to denote *both* the genotype of an individual and the genotypic value of that individual. It will be clear which meaning is to be attached to the symbol, or the meaning will be specified. The gene frequencies in this case will be denoted by p_j. Now we may write as an algebraic identity

$$\begin{aligned} A_j A_k = {} & [\sum_m p_m A_m][\sum_n p_n A_n] + [A_j - \sum_m p_m A_m][\sum_n p_n A_n] + [\sum_m p_m A_m][A_k - \sum_n p_n A_n] \\ & + [A_j - \sum_m p_m A_m][A_k - \sum_n p_n A_n]. \qquad (1) \end{aligned}$$

This is an algebraic identity in the symbols whatever interpretation we may care to make of them. Now let us expand each of the four distinct terms on the right-hand side of equation (1), regard all terms like $A_p A_q$ as genotypes and insert the genotypic values in place of the genotypic symbols. Then we may write

$$A_j A_k = \mu + \alpha_j + \alpha_k + d_{jk}, \qquad (2)$$

where $A_j A_k$ is now the genotypic value of the $A_j A_k$ genotype.

It is clear that μ is the population mean and referring to Fisher's work (1918, 1930, 1941) that α_j and α_k are the additive genetic values of the jth and kth alleles and that d_{jk} is the dominance deviation. In fact

$$\alpha_j = \sum_n p_n A_j A_n - \mu,$$

where $A_j A_n$ is the genotypic value, and this is exactly the quantity Fisher derived. The standard relationships among the components on the right-hand side hold. For instance, it is obvious that

$$\begin{aligned} \sum_j p_j \alpha_j &= \sum_j p_j [A_j - \sum_m p_m A_m][\sum_n p_n A_n] \\ &= \sum_{j,n} p_j p_n A_j A_n - \sum_{m,n} p_m p_n A_m A_n \\ &= \mu - \mu = 0. \end{aligned}$$

If we consider a random gene in the population, the probability of its being A_j is p_j, so that expectation of the additively genetic value of the individual is

$$\sum_{j,k} p_j p_k(\alpha_j+\alpha_k) = 2\sum_j p_j\alpha_j,$$

which is zero. Also $\sum\limits_{j,k} p_j p_k \alpha_j \alpha_k = (\sum\limits_j p_j\alpha_j)^2 = 0$. Similarly, the expectation of d_{ij} is zero since

$$\sum_{jk} p_j p_k d_{jk} = 0.$$

Also the additively genetic value of an individual is uncorrelated with its dominance deviation because

$$\sum_{jk} p_j p_k(\alpha_j+\alpha_k)\,d_{jk} = 2\sum_{j,k} p_j p_k \alpha_j d_{jk} = 2\sum_j p_j\alpha_j\{\sum_k p_k d_{jk}\}$$

and

$$\begin{aligned}\sum_k p_k d_{jk} &= \sum_k p_k(A_jA_k-\mu-\alpha_j-\alpha_k)\\ &= \mu+\alpha_j-\mu-\alpha_j-0\\ &= 0.\end{aligned}$$

We shall now use the standard symbol $\mathscr{E}(x)$ to denote the expectation or average value of X. We have then

$$\begin{aligned}&\mathscr{E}(A_jA_k) = \mu,\\ &\mathscr{E}(\alpha_j) = 0, \quad \mathscr{E}(\alpha_k) = 0, \quad \mathscr{E}(d_{jk}) = 0,\\ &\mathscr{E}(\alpha_j\alpha_k) = 0, \quad \mathscr{E}(\alpha_j d_{jk}) = 0,\end{aligned}$$

and if we denote variances as follows

$$\begin{aligned}&\mathscr{E}(A_jA_k-\mu)^2 = \sigma_G^2,\\ &\mathscr{E}(\alpha_j^2) = \tfrac{1}{2}\sigma_A^2, \quad \text{so that} \quad \sigma_A^2 = 2\Sigma p_j\alpha_j^2 \quad \text{(cf. Fisher (1918))},\\ &\mathscr{E}(d_{jk})^2 = \sigma_D^2,\end{aligned}$$

then

$$\sigma_G^2 = \sigma_A^2+\sigma_D^2.$$

Now that the particular case has been exhibited the general development is merely a matter of writing things down. We have the general identity

$$\begin{aligned}\prod_{i=1}^n A^i_{j_i}A^i_{k_i} = \prod_{i=1}^n \{&[\sum_m p^i_m A^i_m][\sum_n p^i_n A^i_n]+[A^i_{j_i}-\sum_m p^i_m A^i_m][\sum_n p^i_n A^i_n]\\ &+[\sum_m p^i_m A^i_m][A^i_{k_i}-\sum_n p^i_n A^i_n]+[A^i_{j_i}-\sum_m p^i_m A^i_m][A^i_{k_i}-\sum_n p^i_n A^i_n]\}.\end{aligned}$$

We now expand this product, substitute the genotypic values in place of the genotypes, make some identifications of terms to be described, and we get

$$\begin{aligned}\prod_{i=1}^n A^i_{j_i}A^i_{k_i} &= \mu+\sum_{i=1}^n(\alpha^i_{j_i}+\alpha^i_{k_i})+\sum_{i=1}^n d^i_{j_ik_i}+\sum_{\substack{i,i'\\ i<i'}}\\ &\times\{(\alpha^i\alpha^{i'})_{j_ij_{i'}}+(\alpha^i\alpha^{i'})_{j_ik_{i'}}+(\alpha^i\alpha^{i'})_{k_ij_{i'}}+(\alpha^i\alpha^{i'})_{k_ik_{i'}}\}+\sum_{\substack{i,i'\\ i<i'}}\{(\alpha^id^{i'})_{j_ij_{i'}k_{i'}}+(\alpha^id^{i'})_{k_ij_{i'}k_{i'}}\}\\ &+\sum_{\substack{i,i'\\ i<i'}}\{(\alpha^{i'}d^i)_{j_{i'}j_ik_i}+(\alpha^{i'}d^i)_{k_{i'}j_ik_i}\}+\sum_{\substack{i,i'\\ i<i'}}(d^id^{i'})_{j_ik_ij_{i'}k_{i'}}+S\{\sum_{\substack{i,i',i''\\ i<i'<i''}}(\alpha^i\alpha^{i'}\alpha^{i''})_{j_ij_{i'}j_{i''}}\}\\ &+S\{\sum_{\substack{i,i',i''\\ i\neq i'\neq i''\\ i<i'}}(\alpha^i\alpha^{i'}d^{i''})_{j_ij_{i'}j_{i''}k_{i''}}\}+S\{\sum_{\substack{i,i',i''\\ i\neq i'\neq i''\\ i'<i''}}(\alpha^id^{i'}d^{i''})_{j_ij_{i'}k_{i'}j_{i''}k_{i''}}\}\\ &+\sum_{\substack{i,i',i''\\ i<i'<i''}}(d^id^{i'}d^{i''})_{j_ik_ij_{i'}k_{i'}j_{i''}k_{i''}}+\text{etc.,}\end{aligned} \tag{3}$$

where $\prod_{i=1}^{n} A^{i}_{j_i} A^{i}_{k_i}$ is now the genotypic value,

$$\mu = \prod_{i=1}^{n} [\sum_{m} p^{i}_{m} A^{i}_{m}] [\sum_{n} p^{i}_{n} A^{i}_{n}] = \text{mean of population},$$

$$\alpha^{i}_{j_i} = [A^{i}_{j_i} - \sum_{m} p^{i}_{m} A^{i}_{m}] [\sum_{n} p^{i}_{n} A^{i}_{n}] \prod_{i' \neq i} [\sum_{m} p^{i'}_{m} A^{i'}_{m}] [\sum_{n} p^{i'}_{n} A^{i'}_{n}]$$

= additive effect of gene $A^{i}_{j_i}$ in the population,

$$d^{i}_{j_i k_i} = [A^{i}_{j_i} - \sum_{m} p^{i}_{m} A^{i}_{m}] [A^{i}_{k_i} - \sum_{n} p^{i}_{n} A^{i}_{n}] \prod_{i' \neq i} [\sum_{m} p^{i'}_{m} A^{i'}_{m}] [\sum_{n} p^{i'}_{n} A^{i'}_{n}]$$

= dominance deviation due to genes $A^{i}_{j_i}$, $A^{i}_{k_i}$ in the population,

$$(\alpha^{i}\alpha^{i'})_{j_i j_{i'}} = [A^{i}_{j_i} - \sum_{m} p^{i}_{m} A^{i}_{m}] [\sum_{n} p^{i}_{n} A^{i}_{n}] [A^{i'}_{j_{i'}} - \sum_{m} p^{i'}_{m} A^{i'}_{m}], \\ [\sum_{n} p^{i'}_{n} A^{i'}_{n}] \prod_{i'' \neq i \neq i'} [\sum_{m} p^{i''}_{m} A^{i''}_{m}] [\sum_{n} p^{i''}_{n} A^{i''}_{n}]$$

= additive × additive deviation arising from gene $A^{i}_{j_i}$ at locus i and gene $A^{i'}_{j_{i'}}$ at locus i',

$$(\alpha^{i}d^{i'})_{j_i j_{i'} k_{i'}} = [A^{i}_{j_i} - \sum_{m} p^{i}_{m} A^{i}_{m}] [\sum_{n} p^{i}_{n} A^{i}_{n}] [A^{i'}_{j_{i'}} - \sum_{m} p^{i'}_{m} A^{i'}_{m}] [A^{i'}_{k_{i'}} - \sum_{n} p^{i'}_{n} A^{i'}_{n}] \\ \times \prod_{i'' \neq i \neq i'} [\sum_{m} p^{i''}_{m} A^{i''}_{m}] [\sum_{n} p^{i''}_{n} A^{i''}_{n}]$$

= additive × dominance deviation arising from gene $A^{i}_{j_i}$ at locus i, and genes $A^{i'}_{j_{i'}}$ $A^{i'}_{k_i}$ at locus i',

$$(d^{i}d^{i'})_{j_i k_i j_{i'} k_{i'}} = [A^{i}_{j_i} - \sum_{m} p^{i}_{m} A^{i}_{m}] [A^{i}_{k_i} - \sum_{n} p^{i}_{n} A^{i}_{n}] [A^{i'}_{j_{i'}} - \sum_{m} p^{i'}_{m} A^{i'}_{m}] [A^{i'}_{k_{i'}} - \sum_{n} p^{i'}_{n} A^{i'}_{n}] \\ \times \prod_{i'' \neq i \neq i'} [\sum_{m} p^{i''}_{m} A^{i''}_{m}] [\sum_{n} p^{i''}_{n} A^{i''}_{n}]$$

= dominance × dominance deviation arising from genes $A^{i}_{j_i}, A^{i}_{k_i}$ at locus i and genes $A^{i'}_{j_{i'}}, A^{i'}_{k_{i'}}$ at locus i'.

$(\alpha^{i}\alpha^{i'}\alpha^{i''})_{j_i j_{i'} j_{i''}}$ = additive × additive × additive deviation arising from genes $A^{i}_{j_i}$ at locus i, $A^{i'}_{j_{i'}}$ at locus i', and $A^{i''}_{j_{i''}}$ at locus i'', etc.

and S denotes summation of all possible terms of the particular type, obtained by replacing any j subscript which does not occur with its corresponding k subscript by the corresponding k subscript.

Thus the genotypic value of an individual is made up of the following parts:

(1) the general mean,

(2) the sum of $2n$ additive gene effects,

(3) the sum of n dominance deviations,

(4) the sum of $2n(n-1)$ additive × additive interaction deviations,

(5) the sum of $2n(n-1)$ additive × dominance interaction deviations,

(6) the sum of $\frac{n(n-1)}{2}$ dominance × dominance interaction deviations,

(7) the sum of $\frac{8n(n-1)(n-2)}{6}$ additive × additive × additive interaction deviations,

(8) the sum of $\frac{12n(n-1)(n-2)}{6}$ additive × additive × dominance interaction deviations,

(9) the sum of $\frac{6n(n-1)(n-2)}{6}$ additive × dominance × dominance interaction deviations,

(10) the sum of $\frac{n(n-1)(n-2)}{6}$ dominance × dominance × dominance interaction deviations,

and so on.

So far we have labelled terms as being additive effects or dominance deviations etc. without justifying the terminology. This we now proceed to do. The basic point is that there exists an orthogonality like that which arises in the design of experiments. First it is easily seen that the expectation of any term apart from μ is equal to zero; for example

$$\begin{aligned}\mathscr{E}(\alpha_j^i) &= \sum_j p_j^i \alpha_j^i \\ &= \sum_j p_j^i [A_j^i - \sum_m p_m^i A_m^i] [\sum_n p_n^i A_n^i] \prod_{i' \neq i} (\Sigma p_m^{i'} A_m^{i'})^2 \\ &= [(\Sigma p_j^i A_j^i) - (\sum_m p_m^i A_m^i)] [\sum_n p_n^i A_n^i] \prod_{i' \neq i} (\Sigma p_m^{i'} A_m^{i'})^2 \\ &= 0.\end{aligned}$$

Every term is of the form of a product of quantities like

$$[A_j^i - \sum_m p_m^i A_m^i]$$

and

$$[\sum_n p_n^{i'} A_n^{i'}],$$

where i' and i take values from 1 to n. The probability of a term containing a quantity like $[A_j^i - \sum_m p_m^i A_m^i]$ is just p_j^i, so that when we multiply terms by their probabilities and sum we get quantities occurring like

$$\sum_j p_j^i A_j^i - \sum_m p_m^i A_m^i.$$

Whatever such a term is multiplied by and genotypic values inserted we shall get zero as a result. Hence we have all the possible relationships of the form

$$\mathscr{E}_j(\alpha_j^i) = 0, \quad \mathscr{E}_{j,k}(\alpha_{jk}^i) = 0, \quad \mathscr{E}_j(\alpha^i \alpha^{i'})_{jj'} = 0,$$

$$\mathscr{E}_{j \text{ or } j' \text{ or } k} (\alpha^i d^{i'})_{j,j',k'} = 0 \quad \text{and so on.}$$

Exactly the same reasoning may be followed to see that the terms in the equation or model (3) are uncorrelated. For example:

$$\begin{aligned}\mathscr{E}_{j_i, k_i, j_{i'}, k_{i'}} [(\alpha^i \alpha^{i'})_{j_i j_{i'}}] [(d^i d^{i'})_{j_i k_i j_{i'} k_{i'}}] &= \mathscr{E}_{j_i, j_{i'}} [(\alpha^i \alpha^{i'})_{j_i j_{i'}} \mathscr{E}_{k_i, k_{i'}} (d^i d^{i'})_{j_i k_i j_{i'} k_{i'}}] \\ &= 0.\end{aligned}$$

Finally, to obtain the partition of genotypic variance we make the following symbolism:

$$\mathscr{E}\{\sum_{i}[\alpha^{i}_{j_i}]^2\} = \tfrac{1}{2}\sigma^2_A,$$

$$\mathscr{E}\{\sum_{i}[d^{i}_{j_ik_i}]^2\} = \sigma^2_D,$$

$$\mathscr{E}\{\sum_{\substack{i,i'\\ i<i'}}[(\alpha^{i}\alpha^{i'})_{j_ij_{i'}}]^2\} = \tfrac{1}{4}\sigma^2_{AA},$$

$$\mathscr{E}\{\sum_{\substack{i,i'\\ i\neq i'}}[(\alpha^{i}d^{i'})_{j_ij_{i'}k_{i'}}]^2\} = \tfrac{1}{2}\sigma^2_{AD},$$

$$\mathscr{E}\{\sum_{\substack{i,i'\\ i<i'}}[(d^{i}d^{i'})_{j_ik_ij_{i'}k_{i'}}]^2\} = \sigma^2_{DD},$$

$$\mathscr{E}\{\sum_{\substack{i,i',i''\\ i\neq i'\neq i''\\ i<i'<i''}}[(\alpha^{i}\alpha^{i'}\alpha^{i''})_{j_ij_{i'}j_{i''}}]^2\} = \tfrac{1}{8}\sigma^2_{AAA},$$

and so on. Because of the orthogonality the variance of the sum of the items in the model (3) is equal to the sum of the variances of the items. A step-by-step procedure of evaluating the additive variance, the dominance variance, the additive-by-additive variance, the additive-by-dominance variance and so on, will merely pick off in order the variances of the successive items.

We then have

$$\sigma^2_G = \sigma^2_A + \sigma^2_D + \sigma^2_{AA} + \sigma^2_{AD} + \sigma^2_{DD} + \sigma^2_{AAA} + \sigma^2_{AAD} + \sigma^2_{ADD} + \sigma^2_{DDD} + \text{etc.}, \qquad (4)$$

where a variance like σ^2_{AAADD}, for instance, is the variance due to interactions between 5 loci, the interactions being of the type

additive × additive × additive × dominance × dominance.

THE CORRELATIONS BETWEEN RELATIVES

We shall now develop a relationship for the covariance between two relatives with an arbitrary degree of relationship. There will be two restrictions on the development: (1) we shall suppose that there is no linkage and (2) we shall suppose that the relatives under consideration are not inbred, that is, there is zero probability that the two alleles of an individual at a particular locus are direct copies (apart from mutation) of any gene at the head of the joint ancestry of the two relatives.

Actually each of the two relatives is a random member from particular populations, and the first step is to develop a representation of the two populations, say I and II. For definiteness we may suppose that the two populations are related through both their sires and their dams. The gametic arrays of the sires will be supposed to be related and, apart from random population gametes, for both populations will be derived for each locus from a set of say 'a' original genes, say A^i_{sj} $(j = 1, 2, \dots, a)$, where A^i_{sj} are a random set of the genes in the whole population. Likewise the gametic arrays for the dams will be derived from a set of 'b' original genes say A^i_{dk} $(k = 1, 2, \dots, b)$, where again A^i_{dk} are a random set of genes from the population. We use P^{I}_{sj} for the probability that gene A^i_{sj} is on the gamete contributed by the sire to population I and similarly we define $P^{\mathrm{II}}_{sj}, P^{\mathrm{I}}_{dk}$ and P^{II}_{dk}. Hence the genotypic arrays of the two populations may be expressed as follows:

Population I:

$$\prod_{i=1}^{n}\left[\sum_{j=1}^{a} P^{\mathrm{I}}_{sj} A^{i}_{sj} + \left(1 - \sum_{j=1}^{a} P^{\mathrm{I}}_{sj}\right)\left(\sum_{m} P^{i}_{m} A^{i}_{m}\right)\right]\left[\sum_{k=1}^{b} P^{\mathrm{I}}_{dk} A^{i}_{dk} + \left(1 - \sum_{k=1}^{b} P^{\mathrm{I}}_{dk}\right)\left(\sum_{n} P^{i}_{n} A^{i}_{n}\right)\right];$$

Population II:

$$\prod_{i=1}^{n}\left[\sum_{j=1}^{a} P^{\mathrm{II}}_{sj} A^{i}_{sj} + \left(1 \sum_{j=1}^{a} P^{\mathrm{II}}_{sj}\right)\left(\sum_{m} P^{i}_{m} A^{i}_{m}\right)\right]\left[\sum_{k=1}^{b} P^{\mathrm{II}}_{dk} A^{i}_{dk} + \left(1 - \sum_{k=1}^{b} P^{\mathrm{II}}_{dk}\right)\left(\sum_{n} P^{i}_{n} A^{i}_{n}\right)\right].$$

To obtain the covariance of the two populations we have to expand each genotypic array, substitute for each genotype its genotypic value, take the product of the resulting expressions, find its expectation and finally subtract μ^2, where μ is the population mean.

This process is aided by noting that the genotypic value of

$$\prod_{i=1}^{n} A^{i}_{ji} A^{i}_{ki}$$

can be written formally as $$\prod_{i=1}^{n} (\mathrm{I} + a^{i}_{ji})(\mathrm{I} + \alpha^{i}_{ki}),$$

in which after expansion treating the 'I' as unity, 'I' is replaced by μ, terms like α^{i}_{ji}, α^{i}_{ki} are left alone, a term like $\alpha^{i}_{ji}\alpha^{i}_{ki}$ is replaced by $d^{i}_{j_i k_i}$, a term like $\alpha^{i}_{j_i}\alpha^{i'}_{k_{i'}}$ is replaced by $(\alpha^{i}\alpha^{i'})_{j_i k_{i'}}$, a term like $\alpha^{i}_{ji}\alpha^{i}_{ki}\alpha^{i'}_{ji}\alpha^{i''}_{ji''}\alpha^{i''}_{ki''}$ is replaced by $(d^{i}\alpha^{i'}d^{i''})_{j_i k_i j_{i'} j_{i''} k_{i''}}$, and so on.

Hence the array of genotypic values of population I may be written formally as

$$\prod_{i=1}^{n}\left[\sum_{j=1}^{a} P^{\mathrm{I}}_{sj}(\mathrm{I} + \alpha^{i}_{sj}) + \left(1 - \sum_{j=1}^{a} P^{\mathrm{I}}_{sj}\right)\sum_{m} p^{i}_{m}(\mathrm{I} + \alpha^{i}_{m})\right]$$
$$\times\left[\sum_{k=1}^{b} P^{\mathrm{I}}_{dk}(\mathrm{I} + \alpha^{i}_{dk}) + \left(1 - \sum_{k=1}^{b} P^{\mathrm{I}}_{dk}\right)\sum_{n} p^{i}_{n}(\mathrm{I} + \alpha^{i}_{n})\right],$$

with a similar expression for population II. This may be simplified further by noting that we may replace $\sum_{m} p^{i}_{m}(\mathrm{I} + \alpha^{i}_{m})$ by unity wherever it occurs because $\sum_{m} p^{i}_{m} = 1$ and $p^{i}_{m}\alpha^{i}_{m} = 0$. Hence in its simplest formal expression the array of genotypic values in population I is

$$\prod_{i=1}^{n}\left[\sum_{j=1}^{a} P^{\mathrm{I}}_{sj}(\mathrm{I} + \alpha^{i}_{sj}) + \left(1 - \sum_{j=1}^{a} P^{\mathrm{I}}_{s}\right)\mathrm{I}\right]\left[\sum_{k=1}^{b} P^{\mathrm{I}}_{dk}(\mathrm{I} + \alpha^{\mathrm{I}}_{dk}) + \left(1 - \sum_{k=1}^{b} P^{\mathrm{I}}_{dk}\right)\mathrm{I}\right], \qquad (5)$$

with a similar expression for population II. We now have to form the product of these two arrays and take expectations over the α^{i}_{sj} and α^{i}_{dj}. The terms involving I only clearly give μ^2. The terms involving α^{i}_{sj} alone and α^{i}_{dj} alone from each population are

$$\sum_{i=1}^{n}\left(\sum_{j=1}^{a} P^{\mathrm{I}}_{sj}\alpha^{i}_{sj} + \sum_{k=1}^{b} P^{\mathrm{I}}_{dk}\alpha^{i}_{dk}\right)\left(\sum_{j=1}^{a} P^{\mathrm{II}}_{sj}\alpha^{i}_{sj} + \sum_{k=1}^{b} P^{\mathrm{II}}_{dk}\alpha^{i}_{dk}\right),$$

and since the A^i_{sj} and A^i_{dk} are random members we have

$$\mathscr{E}\left[\left(\sum_{i=1}^{n} \alpha^i_{sj}\right)^2\right] = \mathscr{E}\left[\left(\sum_{i=1}^{n} \alpha^i_{dk}\right)^2\right] = \tfrac{1}{2}\sigma^2_A,$$

and all different terms are uncorrelated. Hence the covariance of the two populations with respect to σ^2_A is equal to

$$\left[\left(\sum_{j=1}^{a} P^{\mathrm{I}}_{sj} P^{\mathrm{II}}_{sj}\right) + \left(\sum_{k=1}^{b} P^{\mathrm{I}}_{dk} P^{\mathrm{II}}_{dk}\right)\right] \frac{\sigma^2_A}{2}.$$

With respect to dominance deviations we get

$$\sum_i \left[\sum_{j=1}^{a} \sum_{k'=1}^{b} P^{\mathrm{I}}_{sj} P^{\mathrm{I}}_{dk} d^i_{(sj)(dk')}\right] \left[\sum_{j=1}^{a} \sum_{k'=1}^{b} P^{\mathrm{II}}_{sj} P^{\mathrm{II}}_{dk} d^i_{(sj)(dk')}\right],$$

which since $\mathscr{E}[\sum_i (d^i_{j_i k_i})^2] = \sigma^2_D$, and otherwise the expectation of a product of two d's is zero, gives

$$\left(\sum_{j=1}^{a} \sum_{k=1}^{b} P^{\mathrm{I}}_{sj} P^{\mathrm{I}}_{dk'} P^{\mathrm{II}}_{sj} P^{\mathrm{II}}_{dk'}\right) \sigma^2_D$$

or

$$\left(\sum_{j=1}^{a} P^{\mathrm{I}}_{sj} P^{\mathrm{II}}_{sj}\right) \left(\sum_{k=1}^{b} P^{\mathrm{I}}_{dk} P^{\mathrm{II}}_{dk}\right) \sigma^2_D.$$

It should be noted that the results so far are exactly those given by Malécot (1948, p. 24), although his derivation applies to the case of one locus. Malécot's notation is as follows:

$$\phi = \sum_{j=1}^{a} P^{\mathrm{I}}_{sj} P^{\mathrm{II}}_{sj},$$

$$\phi' = \sum_{k=1}^{b} P^{\mathrm{I}}_{dk} P^{\mathrm{II}}_{dk},$$

so that with respect to within-locus effects the covariance is equal to

$$(\phi + \phi')\frac{\sigma^2_A}{2} + \phi\phi'\sigma^2_D.$$

We could expand equation (5) to evaluate the contribution of the other variances, but it is simpler to use the following argument. The quantities ϕ and ϕ' determine the whole of the covariance between the populations because the covariance between two individuals with respect to any term in the model, i.e. involving an arbitrary number of loci, is zero unless the two individuals have exactly the same genes from the original set leading to the relationship at each of the loci. Now ϕ is the probability that the individuals will receive the same gene by way of the sire and ϕ' is the probability that they will receive the same gene by way of the dam at any one locus. Consider an interaction which involves the additive deviations at r loci and the dominance deviations at s loci. In the case of the dominance deviations, the contribution will be zero unless the two individuals have received the same gene by way of the sires and the same gene by way of the dams at each of the loci, and the probability of this is $(\phi\phi')^s$. In the case of additive deviations, the additive deviations can be the same by way of the sire and/or by way of the dam at each locus, so that we obtain as the probability $(\phi+\phi')^r$. Finally, the expectation of such a term added over all possible distinct sets of r and s loci is $\frac{1}{2^r}\sigma_X^2$, where X contains r A's and s D's.

Hence, the covariance with regard to this interaction variance is

$$(\phi+\phi')^r(\phi\phi')^s\left(\frac{1}{2^r}\right)\sigma_X^2.$$

Thus, for example, the covariance with regard to the simplest interaction terms are as follows:

term involving	coefficient
AA	$(\frac{1}{4})(\phi+\phi')^2$
AD	$(\frac{1}{2})(\phi+\phi')\phi\phi'$
DD	$\phi^2\phi'^2$
AAA	$(\frac{1}{8})(\phi+\phi')^3$
AAD	$(\frac{1}{4})(\phi+\phi')^2(\phi\phi')$
ADD	$(\frac{1}{2})(\phi+\phi')(\phi\phi')^2$
DDD	$(\phi'\phi')^3$

It may be noted that the results with respect to two-loci interactions for simple degrees of relationship were obtained by Fisher (1918), in that the covariance between parent and offspring contains $\frac{1}{4}$ of a part of the variance from dual epistacy, and it is easily verified that this part of the variance is what is here called the additive × additive variance. This result was verified by Cockerham (1952).

REFERENCES

Anderson, V. L. 1953 Ph.D. Thesis, Iowa State College Library.
Anderson, V. L. & Kempthorne, O. 1954 (In the press).
Cockerham, C. C. 1952 Ph.D. Thesis, Iowa State College Library.
Fisher, R. A. 1918 *Trans. Roy. Soc. Edinb.* **52**, 399–433.
Fisher, R. A. 1930 *The genetical theory of natural selection.* Oxford University Press.
Fisher, R. A. 1941 *Ann. Eugen., Lond.*, **11**, 53–63.
Kempthorne, O. 1952 *The design and analysis of experiments.* New York: John Wiley and Sons.
Malécot, G. 1948 *Les mathématiques de l'hérédité.* Paris: Masson et Cie.
Pearson, K. 1903 *Phil. Trans.* A, **203**, 53–87.
Pearson, K. & Lee, A. 1903 *Biometrika*, **2**, 357–462.
Weinberg, W. 1908 *Z. indukt. Abstamm.- u. VererbLehre*, **1**, 337–392, 440–460.
Weinberg, W. 1909 *Z. indukt. Abstamm.- u. VererbLehre*, **2**, 276–330.
Wright, S. 1921 *Genetics*, **6**, 111–178.

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, CAMBRIDGE
(BROOKE CRUTCHLEY, UNIVERSITY PRINTER)