STATISTICAL CONCEPTS IN GENETICS

Mimeograph No. 1

I. Introduction - Rates of genetic improvement of plants and animals are affected by numerous variables, e.g., system of breeding, method of selection, mode of inheritance of important characteristics, the extent to which phenotype responds to non-genetic variables, reproductive rate, etc. While we recognize the qualitative effects of many such factors, we do not always know the magnitude of these effects as precisely as we would like. In this course we shall be concerned with the quantitative effects of factors which influence the means and variances of genetic populations.

The material to be presented will fall far short of what might be considered. Limiting factors will be time, the fact that the quantitative effects of certain known variations in genetic mechanism have not been investigated, and the competency of the instructor. The general procedure will be to consider simple genetic situations first, then generalize to more complex situations where possible.

The subject matter will be mathematical in nature but for the most part nothing more complicated than rather simple algebra will be called for. In the few exceptional instances unfamiliarity with the mathematical tools should not interfere with understanding what is involved.

Perhaps the most important thing to remember in considering a subject such as this is that mathematics like all forms of inductive logic leads to correct answers only when basic assumptions are correct. Consequently each member of the groups should be on guard against violation of genetic principles and against application of formulae derived in situations where assumptions involved in the derivations do not hold. On the other hand it will frequently prove useful to analyze an artificial situation as a guide to what occurs in a situation which is related but cannot be completely specified.

II. Statistical formulae

A. Variance (σ^2)

1. Consider a population the individuals of which are

$X_1, X_2, X_3, \ldots\ldots$

The mean ($\bar{X}$) is

$$\frac{X_1 + X_2 + X_3 \ldots\ldots}{N}$$

$$\text{The variance, } \sigma^2 = \frac{S(X - \bar{X})^2}{N} = \frac{S(X^2) - \frac{(SX)^2}{N}}{N} \qquad (1)$$

Let $x_1 = X_1 - \bar{X}$, $x_2 = X_2 - \bar{X}$, etc.

$$\text{Then } \sigma^2 = \frac{Sx^2}{N} \qquad (2)$$

2. Form a new population by dividing each X by c.

$$\frac{X_1}{c}, \frac{X_2}{c}, \frac{X_3}{c} \ldots .$$

$$\text{Its mean equals } \frac{\frac{X_1}{c} + \frac{X_2}{c} + \frac{X_3}{c} \ldots .}{N} = \frac{\bar{X}}{c}$$

$$\frac{X_1}{c} - \frac{\bar{X}}{c} = \frac{x_1}{c}, \quad \frac{X_2}{c} - \frac{\bar{X}}{c} = \frac{x_2}{c}, \text{ etc.}$$

$$\text{The variance} = \frac{S(\frac{x}{c})^2}{N} = \frac{Sx^2}{Nc^2} = \frac{\sigma_x^2}{c^2} \qquad (3)$$

B. The correlation coefficient (r) and the regression coefficient (b)

Consider two populations

$X_1, X_2, X_3, \ldots .$ with mean $\bar{X}$, and

$Y_1, Y_2, Y_3, \ldots .$ with mean $\bar{Y}$.

$$r = \frac{S(X - \bar{X})(Y - \bar{Y})}{\sqrt{S(X - \bar{X})^2 . S(Y - \bar{Y})^2}} \qquad (4)$$

or if $x_1 = X_1 - \bar{X}$, $x_2 = X_2 - \bar{X}$, etc.

and $y_1 = Y_1 - \bar{Y}$, $y_2 = Y_2 - \bar{Y}$, etc.

$$r = \frac{Sxy}{\sqrt{Sx^2 . Sy^2}} = \frac{Sxy/N}{\sqrt{\frac{Sx^2}{N} \cdot \frac{Sy^2}{N}}} = \frac{Sxy}{N\sigma_X\sigma_Y} \qquad (5)$$

$$b_{Y.X} = \frac{S(X - \bar{X})(Y - \bar{Y})}{S(X - \bar{X})^2} = \frac{Sxy}{Sx^2} \qquad (6)$$

$$b_{X.Y} = \frac{S(X - \bar{X})(Y - \bar{Y})}{S(Y - \bar{Y})^2} = \frac{Sxy}{Sy^2} \qquad (6a)$$

$$r = \frac{Sxy}{\sqrt{Sx^2 . Sy^2}} = \sqrt{\frac{Sxy . Sxy}{Sx^2 . Sy^2}} = \sqrt{b_{Y.X} \cdot b_{X.Y}} \qquad (7)$$

The regression equation for estimating Y from X is

$$Y_e = \bar{Y} + bx$$

$Y - Y_e$ is the difference between an observed Y and the value that would have been estimated from the regression equation on the basis of the associated value of X, hence is called a deviation from regression.

$$Y - Y_e = Y - \bar{Y} - bx = y - bx$$

$$S(Y - Y_e)^2 = S(y - bx)^2 = S(y^2 - 2bxy + b^2x^2)$$

$$= Sy^2 - 2bSxy + b^2Sx^2 = Sy^2 - 2\frac{Sxy . Sxy}{Sx^2} + \frac{Sxy . Sxy . Sx^2}{Sx^2 . Sx^2}$$

$$= Sy^2 - (Sxy)^2/Sx^2 \qquad (8)$$

This value is called the sum of squares for deviations from regression.

(Note: The method of computing b is such that $S(Y - Y_e)^2$ is a minimum, i.e. it is smaller than would result if any other value were substituted in the regression equation for the b obtained.)

$Y_e - \bar{Y}$ is the deviation of an estimated value of Y from the mean of Y. $S(Y_e - \bar{Y})^2$ is accordingly called the regression sum of squares.

$$Y_e - \bar{Y} = \bar{Y} + bx - \bar{Y} = bx$$

$$S(Y_e - \bar{Y})^2 = Sb^2x^2 = b^2Sx^2 = \frac{Sxy . Sxy . Sx^2}{Sx^2 . Sx^2} = \frac{(Sxy)^2}{Sx^2} \qquad (9)$$

$$S(Y - Y_e)^2 + S(Y_e + \bar{Y})^2 = Sy^2 - \frac{(Sxy)^2}{Sx^2} + \frac{(Sxy)^2}{Sx^2} = Sy^2 \qquad (10)$$

We can therefore state that the sum of squares for Y is the sum of two parts: (1) The regression sum of squares, and (2) the sum of squares of the deviations from regression. Or we can say that the total variance $\frac{Sy^2}{N}$, in Y can be divided into two parts:

$\frac{S(Y_e - \bar{Y})^2}{N}$, the variance in Y associated with variance in X, and

$\frac{S(Y - Y_e)^2}{N}$ (commonly called the standard error of estimate), the variance in Y independent of variance in X.

C. The Variance of sums and means

Consider two populations

$A_1, A_2, A_3 \ldots\ldots$ with mean $\bar{A}$, and

$B_1, B_2, B_3 \ldots\ldots$ with mean $\bar{B}$.

$$\sigma_A^2 = \frac{Sa^2}{N}, \qquad \sigma_B^2 = \frac{Sb^2}{N}$$

Now let a third population be formed as follows:

$C_1 = A_1 + B_1, \quad C_2 = A_2 + B_2,$ etc.

$$\bar{C} = \frac{A_1 + B_1 + A_2 + B_2 \ldots\ldots}{N} = \bar{A} + \bar{B}$$

$$c_1 = C_1 - \bar{C} = A_1 + B_1 - \bar{A} - \bar{B} = a_1 + b_1$$

$$c_2 = C_2 - \bar{C} = A_2 + B_2 - \bar{A} - \bar{B} = a_2 + b_2$$

etc.

$$\sigma_C^2 = \frac{Sc^2}{N} = \frac{S(a + b)^2}{N} = \frac{S(a^2 + 2ab + b^2)}{N}$$

$$= \frac{Sa^2}{N} + \frac{Sb^2}{N} + \frac{2Sab}{N} = \sigma_A^2 + \sigma_B^2 + 2r_{AB}\,\sigma_A\,\sigma_B \qquad (11)$$

(see equation 5)

If A and B are uncorrelated ($r_{AB} = 0$) as they would be if they were designated A_1, A_2, A_3, etc., and B_1, B_2, B_3, etc., at random, the variance of $C(\sigma_C^2)$ is found to be the sum of the variances of A and B.

This can be extended for sums of any number of variables. Thus

$$\sigma^2_{(A + B + C + D)} = \sigma^2_A + \sigma^2_B + \sigma^2_C + \sigma^2_D \qquad (12)$$

provided none of the variables are correlated.

Or

$$\sigma^2_{(X_1 + X_2 + \ldots X_N)} = \sigma^2_{X_1} + \sigma^2_{X_2} \ldots + \sigma^2_{X_N}$$

If $X_1, X_2 \ldots X_N$ are all drawn from the same population

$$\sigma^2_{X_1} = \sigma^2_{X_2} = \ldots = \sigma^2_{X_N} = \sigma^2$$

And

$$\sigma^2_{(X_1 + X_2 \ldots + X_N)} = N\sigma^2$$

Now if this sum, $(X_1 + X_2 \ldots + X_N)$ is divided by N to obtain a mean, $\bar{X}$, applying (3) we find

$$\sigma^2_{\bar{X}} = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \qquad (13)$$

the formula for the variance of a mean.

III. The composition of phenotypic variance.

Consider a population of organisms (they may be any kind of plant or animal) and a single characteristic of those organisms, e.g., height at a specified age. The height of an individual is the resultant of its genotype and the various environmental factors which effect height. A uniform environment for all individuals of a group is an abstraction never an actuality. We must recognize that the environments (defined to encompass the effects of all factors other than genotype) of plants vary even though the plants are growing adjacent to each other and that the environments of animals vary even though all are handled as nearly alike as is humanly possible. A little thought about soil variation, competition, incidence of parasites, accidents of various and subtle sorts, etc., will suggest many uncontrollable sources of environmental variation.

Let the individuals in the population be numbered

1, 2, 3 • • •

P_1, P_2, P_3 • • • be the measured heights of those individuals,

G_1, G_2, G_3 • • • be the genotypes of those individuals,

and E_1, E_2, E_3 • • • be the environments under which they develop.

$\bar{P}$ will be the mean height of the population.

Now suppose individuals all of genotype G_1 could be developed one under each of the environments E_1, E_2, E_3 • • •, and that the same could be done for individuals of each of the other genotypes G_2, G_3, G_4 • • •. For an individual with genotype G_1 and raised in environment E_1, let the measured height be P_{11}; and for an individual with genotype G_2 and raised in environment E_3, let the measured height be P_{23}. Now let

$$P_{11} - \bar{P} = \varepsilon_{11} = o_{11} \qquad (14)$$

$$P_{23} - \bar{P} = \varepsilon_{23} = o_{23}$$

Note: In all cases the first subscript number attached to P, ε, or o, indicates genotype and the second environment.

Then $\dfrac{\varepsilon_{11} + \varepsilon_{12} + \varepsilon_{13} + \cdots \varepsilon_{1N}}{N} = \varepsilon_1$, the effect of the genotype G_1 averaged over all environments. (15)

$$\frac{g_{21} + g_{22} + g_{23} + \cdots g_{2N}}{N} = g_2,$$ the effect of genotype G_2 averaged over all environments.

$e_{11} + e_{21} + e_{31} + \cdots e_{N1} = e_1$, the effect of environment E, averaged over all genotypes. (16)

etc.

Note: When only one subscript number is used it refers to genotype if used with g, environment if used with e.

Now $P_{11} - \bar{P}$ is not necessarily equal to $g_1 + e_1$. A specific genotype will not have the same effect in all environments; a specific environment will not have the same effect on the development of individuals of different genotypes. Is it the effect of the genotype or the effect of the environment that changes? We cannot distinguish which is true and resolve the situation by saying that genotype and environment interact.

Let $P_{11} - \bar{P} = g_1 + e_1 + i_{11}$ or $i_{11} = P_{11} - \bar{P} - g_1 - e_1$

$P_{23} - \bar{P} = g_2 + e_3 + i_{23}$ or $i_{23} = P_{23} - \bar{P} - g_2 - e_3$

You will note that the i's (interaction terms) are the amount by which the deviation of the phenotype from the mean fails to be the sum of the average deviations for the genotype and environment involved. In general,

$$P_{ij} - \bar{P} = g_i + e_j + i_{ij} \qquad (17)$$

The phenotypic variance is

$$\frac{S\, p^2}{N} \qquad p = P - \bar{P}$$

and from (12), II, remembering (17)

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + \sigma_i^2$$

provided there are no correlations among g, e, and i.

It is obvious that q and e will not ordinarily be correlated in plants and animals (except man) since an individual organism does not find itself in as specific environment as a consequence of its genotype; the environmental and genotypic variations are random with respect to each other. In humans this is much less likely to be true since, to an extent, the genotypes of a child's parents will be a source of variation in the same direction in both the genotype and environment of the child. The effect of this will vary of course depending on the trait being considered.

As a consequence of the way in which it is defined i will not be correlated with either c or g. Hence we can state that

$$\sigma_p^2 = \sigma_g^2 + \sigma_c^2 + \sigma_i^2 \qquad (18)$$

or that phenotypic variance is the sum of three parts (1) variance arising from the differences among the average effects of genotypes, (2) variance arising from the differences among the average effects of environments, and (3) variance arising from the interaction of genotype and environment.

For most purposes in this course σ_c^2 and σ_i^2 will not be distinguished from each other but will be referred to together as environmental variance. However, it should be remembered where variance arising from the interaction of genotype and environment falls and that it has not been ignored. There are problems in which σ_i^2 would need to be considered separately.

IV. Preliminary consideration of mass selection for a single trait.

The effectiveness of mass selection depends on four primary factors:

1. The proportion of individuals available that must be selected.

2. The variance of the population from which selections are to be made.

3. The proportion of the phenotypic variance which arises from differences in genotype.

4. Whether the average genotypic value of the progeny of selected parents is as great as the average genotypic value of the selected parents.

The selection differential (hereafter to be denoted by the letter $\underline{s}$) is the mean phenotypic difference between selected individuals and the population (including the selected individuals) from which they were selected. Obviously $\underline{s}$ cannot be as large if a high proportion of the population must be used in producing the next generation as if only small number of individuals need be selected. It is equally clear that $\underline{s}$ cannot be large if there is little variation in the population from which selections are to be made.

With regard to the third factor listed above it is clear that if phenotypic variance is largely of non-genetic origin a large selection differential may mean very little in terms of genotypic superiority of the selected groups. The relationship of phenotypic superiority (s) of selected individuals to their genotypic superiority can easily be put into quantitative form. It is statistically a regression problem. We are dealing with two variables (phenotype and genotype) and wish to predict one (genotype) knowing the other (phenotype). We need to know the regression of genotype on phenotype (b_{gp}) in order to set up a prediction equation.

Note: Genotypic value is defined as performance expected of a genotype under average environment not as value for breeding. Thus when dominance is complete Aa and AA have the same genotypic value.

Let genotypes be measured in terms of their effects (g's) and phenotypes in terms of deviations from the mean phenotype ($p = P - \bar{P}$).

$$p = g + e + i \qquad (17), \text{ III}$$

Since $\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + \sigma_i^2$, (18), III

$$Sp^2 = Sg^2 + Se^2 + Si^2.$$

$$Sgp = Sg(g + e + i) = Sg^2 + Sge + Sgi,$$

but since g is not correlated with e or i,

Sge and Sgi are each zero.

Hence $Sgp = Sg^2$

and $$bgp = \frac{Sg^2}{Sg^2 + Se^2 + Si^2} = \frac{Sg^2}{Sp^2} = \frac{\sigma_g^2}{\sigma_p^2}$$

Note: There is no correction term to be subtracted from Sg^2, Sge, or Sgi, since Sg, Se, and Si, like Sp, are all equal to zero. This can be shown from (14), (15), (16), and (17); III.

Now we can set up the prediction equation for genotypic superiority of selected individuals.

Letting g_s = average genotypic value of selected group

$g_s - \bar{g}$ = genotypic superiority of selected group

and in regular regression equation form

$$g_s - \bar{g} \; (=) \; b_{gp} (P_s - \bar{P})$$

Since $\bar{g} = 0$ and $P_s - \bar{P} = s$, the selection differential

(P_s = average phenotypic value of selected group)

we may write,

$$g_s \ (=) \ sb_{gp} \quad \text{where } g_s = \text{genotypic superiority}$$

Note: (=) is used to indicate estimation rather than strict equality. Sampling error is present.

Example: Suppose one has records on butterfat production in one lactation for each of 100 dairy cows for which the mean production was 300 lbs. of butterfat and the standard deviation 40 lbs. Suppose further that the 20 cows with the best records are to be used as the foundation of another herd. How much will these 20 be expected to be superior genotypically to the entire group of 100?

Assuming that the distribution of fat production was approximately "normal" s can be estimated making use of characteristics of the "normal" curve. With "normal" distribution $s = z/p$ standard deviations where p is the proportion selected and z is the height of the ordinate which divides the area under the curve into portions relative in magnitude to the proportion selected and the proportion rejected. The value of z can be obtained using tables I and II of the Statistical Tables published by Fisher and Yates (1). Table I is entered with P equal to either 2p or 2(1 - p), whichever is 1.0 or less. (The factor, 2, is introduced since table I gives the relative deviate, x, beyond which a given proportion of the population, P, is found when both tails of the curve are considered; we are interested in only one tail of the curve.) Table II is then entered with the x obtained from table I to find z. In our case $2p = .4$, x for $P = .4$ is .8416, and z for $x = .8416$ is .2799.

$$s = \frac{z}{p} = \frac{.2799}{.2} = 1.4 \text{ standard deviations}$$

$$= 1.4 \times 40 = 56 \text{ lbs.}$$

$$g_s = 56 \, b_{gp}$$

b_{gp} is probably about .3 - .35 for cows in the same herd (2,3)

Using $b_{gp} = .3$, we get

$g_s = .3 \times 56 = 16.8$ lbs. as an estimate of the genotypic superiority of the top 20 cows.

Note: Values of s for a given series of p are listed by Lush (4).

It should be noted that while g_s is directly proportioned to b_{gp} it does not necessarily vary linearly with b_{gp} because when b_{gp} is reduced as a result of an increase in $So^2 + Si^2$, $\bar{s}$ will be larger, other things being equal. For example, in the above problem

$$\sigma_p^2 = 40 \times 40 = 1600$$

If b_{gp} were .3 as assumed

$$b_{gp} = \frac{\sigma_g^2}{\sigma_p^2} = .3 \quad \text{and} \quad \sigma_g^2 = .3 \times 1600 = 480$$

$$\sigma_o^2 + \sigma_i^2 = 1600 - 480 = 1120$$

Now suppose b_{gp} had been only .15 as a consequence of $\sigma_o^2 + \sigma_i^2$ being 2720 instead of 1120.

In that case the standard deviation would have been

$$\sqrt{2720 + 480} = \sqrt{3200} = 56.6$$

and $\bar{s}$ would have been $1.4 \times 56.6 = 79.2$ instead of 56. Then we would have found

$$g_s = .15 \times 79.2 = 11.88$$

which is 70% of the g_s expected when b_{gp} was .3 and the standard deviation 40 instead of 50% as might have been expected.

Whether the average difference between genotypic value of the progeny of selected and unselected parents will be as large as the difference in genotypic value of the selected and unselected parents (or half that large if selection is practiced only among females and both selected and unselected female mated to the same male) will depend on whether gene action is strictly additive. It will not be as large if either dominance or gene interactions are involved. This matter will be given detailed attention in a later section.

References:

1. Fisher, R. A., and F. Yates (1938)Statistical Tables for Biological, Agricultural, and Medical Research. Oliver and Boyd. London and Edinburgh.

2. Lush, Jay L., H. W. Norton III, and Floyd Arnold (1941) Effects which Selection of Dams May Have on Sire Indexes. J. Dairy Sci. 24:695-721.

3. Dickerson, G. E. Estimates of Producing Ability in Dairy Cattle. J. Agric. Res. 61:561-585.

4. Lush, Jay L. Animal Breeding Plans. The Iowa State College Press. Ames. 2nd ed.

V. Gene frequency and the distribution of genotypes when mating is random.

The frequency of a particular gene is defined as the proportion between the number of that gene in a population and the total number of loci at which it might have been present. Thus, in a population of 100 dipolid organisms there are 200 loci at which a specific gene might be present. Suppose that among 100 organisms, 30 are AA, 60 Aa, and 10 aa. Then there are 120 A genes and the frequency of that gene is 120/200 = .6.

We will use the letter q to designate the frequency of desirable genes. The frequency of their allelomorphs will consequently be 1 - q.

In a random mating population the ratio of genotypes homozygous for the desirable gene, heterozygous, and homozygous for the less desirable gene will tend toward $q^2 : 2q(1 - q) : (1 - q)^2$. This can easily be demonstrated as follows:

The probability of a gamete containing A is obviously q and the probability of 2 specific gametes, i.e. two combining to form a zygote, both containing A is q^2. Thus the proportion of AA zygotes will be q^2.

The probability of a gamete containing a is 1 - q and the probability of 2 specific gametes both containing a is $(1 - q)^2$. Thus the proportion of aa zygotes will be $(1 - q)^2$.

The remainder of the zygotes must be of the Aa type and the proportion in which they appear will consequently be $1 - q^2 - (1 - q)^2 = 1 - q^2 - 1 + 2q - q^2 = 2q - 2q^2 = 2q(1 - q)$.

When two gene pairs, say Aa and Bb, segregate independently, the various possible genotypes will tend to have the following relative frequencies:

Genotype	Frequency	Frequency of AA, Aa, and aa
AABB	$q_a^2 q_b^2$	
AABb	$q_a^2 2q_b (1 - q_b)$	q_a^2
AAbb	$q_a^2 (1 - q_b)^2$	
AaBB	$2q_a (1 - q_a) q_b^2$	
AaBb	$2q_a (1 - q_a) 2q_b (1 - q_b)$	$2q_a (1 - q_a)$
Aabb	$2q_a (1 - q_a) (1 - q_b)^2$	
aaBB	$(1 - q_a)^2 q_b^2$	
aaBb	$(1 - q_a)^2 2q_b (1 - q_b)$	$(1 - q_a)^2$
aabb	$(1 - q_a)^2 (1 - q_b)^2$	

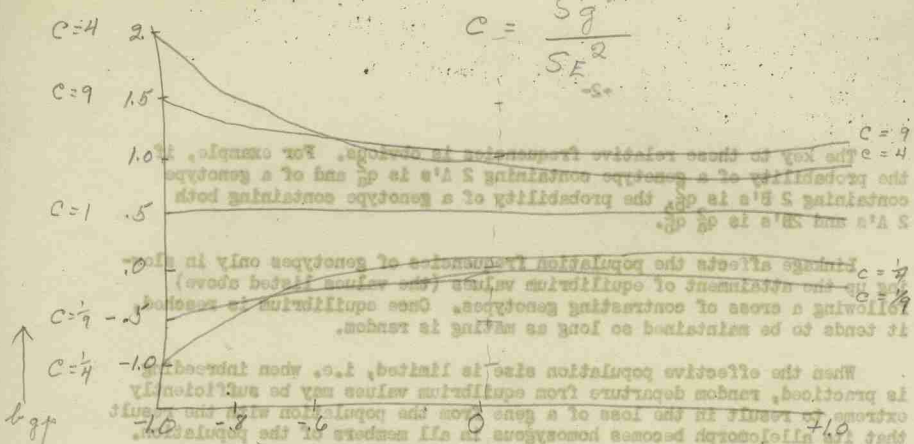
The key to these relative frequencies is obvious. For example, if the probability of a genetype containing 2 A's is q_a^2 and of a genotype containing 2 B's is q_b^2, the probability of a genotype containing both 2 A's and 2B's is $q_a^2 q_b^2$.

Linkage affects the population frequencies of genotypes only in slowing up the attainment of equilibrium values (the values listed above) following a cross of contrasting genotypes. Once equilibrium is reached, it tends to be maintained so long as mating is random.

When the effective population size is limited, i.e. when inbreeding is practiced, random departure from equilbrium values may be sufficiently extreme to result in the loss of a gene from the population with the result that its allelomorph becomes homozygous in all members of the population. That is why inbreeding results in increased homozygosis. We will return to this subject later.

	R					
N.	.1	.2	.4	.6	.8	1.0
1	1	1	1	1	1	1
2	1.35					1.0
3	1.58					
4						
6						
∞	3.16					

$$C = \frac{S_g^2}{S_E^2}$$



Herd ave. 400#

H	600#	1 record	$\frac{1}{3} \times 200 = 67 = 467^{\#}$
B	565	2 "	$\frac{1}{2} \times 165 = 82\frac{1}{2} = 482.5$
C	560	4 "	$\frac{2}{3} \times 160 = 104 = 507^{\#}$

e_1 permanent envir.
e_2 random "

$$\sigma_g^2 + \sigma_{e_1}^2 + \sigma_{e_2}^2 + \sigma_i^2$$

$$\text{repeatability} = \frac{\sigma_g^2 + \sigma_{e_1}^2}{\sigma_g^2 + \sigma_{e_1}^2 + \sigma_{e_2}^2 + \sigma_i^2} = "r" \quad \leftarrow \text{(future record)} \quad \leftarrow \text{(1st record)}$$

$$\downarrow$$

$$\frac{\sigma_{e_2}^2 + \sigma_i^2}{2} \quad \text{(if 2 record are available)}$$

Regress of one record on average of n records

$$\frac{\sigma_g^2 + \sigma_{e_1}^2}{\sigma_g^2 + \sigma_{e_1}^2 + \frac{\sigma_{e_2}^2 + \sigma_i^2}{n}} = \frac{n\,r}{1 + (n-1)\,r}$$

$e_1 = c \qquad e_2 = v$

$$b_{(g+c)} \cdot \bar{p}_{\text{(past record)}} = \frac{nR}{1 + (n-1)R}$$

VI "Repeatability" and related concepts

Animals and perennial plants give expression to some of their traits recurrently. The yield of a plot of strawberries, an apple tree, or a plot of alfalfa; litter size in swine; annual milk production in dairy cows are examples. By measuring a trait more than once the genotype of an individual is more accurately estimated. How many observations is it worth making on the same individual or plot (group of individuals)? Three factors are involved

1. The proportion of the phenotypic variance which results from genotypic variation.

2. The extent to which environmental effects remain constant for an individual from one expression of the trait to another.

3. The occurrence of interactions of genotype with age of the organism.

The third factor will be omitted from consideration at first since its relation to selection is distinct in nature from that of the first two. The situation will be simplified further by treating variance arising from the interaction of genotype and environment as though it arose from variation in environment.(This has no effect on conclusions to be reached).

Then $p = g + e$

where p, g, and e are defined as before.

Now let $e = c + v$

where c is the average effect of the portion of an individual's environment which remains constant, and
v is the effect of the portion of an individual's environment which is variable from one expression of the trait to another.

Then for any single expression of the trait

$$p = g + c + v$$

Successive expressions by the same individual may be symbolized as

$$p_1 = g + c + v_1$$

$$p_2 = g + c + v_2$$

$$\underline{p_n = g + c + v_n}$$

$$S(p) = ng + nc + v_1 + v_2 + \cdots v_n$$

$$\bar{p} = g + c + \bar{v}$$

When g and c are uncorrelated, as will usually be the case in normally

interbreeding populations of plants and animals (other than humans),

$$V_{\bar{p}} = V_g + V_c + V_{\bar{v}} = V_g + V_c + \frac{V_v}{n} \qquad (19)$$

Notes: (1) V will be used in place of σ^2 from this point on - less work in typing.

(2) $\bar{v}$ will be uncorrelated with c since otherwise it would be in part a constant effect of environment. Correlation of $\bar{v}$ with g would also require that $\bar{v}$ be in part an effect of environment constant for an individual since g is constant for individuals.

The covariance of two variables is their sum of products divided by N. Therefore

$$r = \frac{Sxy}{\sqrt{Sx^2 \cdot Sy^2}} = \frac{Cov\ XY}{\sqrt{V_X \cdot V_Y}} \qquad \text{and}$$

$$b = \frac{Sxy}{Sx^2} = \frac{Cov\ XY}{V_X}$$

The covariance of $\bar{p}$ and $(g + c)$ is

$$\frac{S(g + c + \bar{v})(g + c)}{N} = \frac{Sg^2 + Sc^2}{N} = V_g + V_c \qquad (20)$$

and the regression of $(g + c)$ on $\bar{p}$ is

$$b_{(g + c) \cdot \bar{p}} = \frac{V_g + V_c}{V_g + V_c + \frac{V_v}{n}} \qquad (21)$$

Note: The regression of $(g + c)$ on $\bar{p}$ is equivalent to the regression of future expressions of the trait on $\bar{p}$ since $(g + c)$ is the value around which future expression will vary.

When n is one, this is the regression of $(g + e)$ on one observation of the trait; it then becomes

$$\frac{V_g + V_c}{V_g + V_c + V_v}$$

Lush defines "repeatability" as the correlation between single observations on the same individual and uses r to denote it. We will use R to avoid confusion with the use of r as a symbol for correlation in general.

$$R = \frac{Cov\ P_1P_2}{\sqrt{V_{P_1} \cdot V_{P_2}}} = \frac{V_g + V_c}{\sqrt{(V_g + V_c + V_{v_1})(V_g + V_c + V_{v_2})}}$$

If $V_{v_1} = V_{v_2}$ (Variance of effects due to temporary differences in environment is of the same size for all expressions of the trait)

$$R = \frac{V_g + V_c}{V_g + V_c + V_v} \tag{22}$$

We shall consider "repeatability" as the regression of (g + c) on a single observation of the trait. R can be substituted in eq.(21) to put it into more compact form.

$$b_{(g+c)\cdot\bar{p}} = \frac{V_g + V_c}{V_g + V_c + \frac{V_v}{n}} = \frac{n(V_g + V_c)}{n(V_g + V_c)+V_v}$$

$$= \frac{n(V_g + V_c)}{V_g + V_c + V_v + (n-1)(V_g + V_c)}$$

Dividing numerator and denominator by $(V_g + V_c + V_v)$ and substituting R according to eq.(22).

$$b_{(g+c)\cdot\bar{p}} = \frac{nR}{1 + (n-1)R} \tag{23}$$

Now $$\frac{V_g + V_c}{V_g + V_c + \frac{V_v}{n}} \cdot \frac{V_g}{V_g + V_c} = \frac{V_g}{V_g + V_c + V_v/n}$$

which is easily shown to be the regression of genotype on the mean of n observations of the trait, hence we have

$$b_{g\bar{p}} = \frac{nR}{1 + (n-1)R} \cdot \frac{V_g}{V_g + V_c} \tag{24}$$

It is apparent that increasing n causes the same percentage change in $b_{g\bar{p}}$ and $b_{(g+c)\cdot\bar{p}}$. It is also obvious that $b_{g\bar{p}} \leq b_{(g+c)\cdot\bar{p}}$ and hence that future performance is more accurately estimated from past performance than is the genotypic value, unless there are no permanent differences (between individuals) in environment.

The effect of increased numbers of observations on progress made through selection can now be demonstrated. Genotypic superiority of selected individuals as shown in Section IV can be estimated as

$$g_s (=) s b_{gp}$$

where p was the phenotypic expression of the trait

in terms of deviation from the phenotypic mean.
In this case we must substitute $b_{g\bar{p}}$, and

$$g_s \ (=) \ s b_{g\bar{p}}$$

When n (the number of observations per individual is increased $V_{\bar{p}}$ is decreased - see eq. (19). Since $V_{\bar{p}}$ is the denominator of $b_{g\bar{p}}$, the latter is increased accordingly. However, the size of s (the selection differential) is proportional to $\sqrt{V_{\bar{p}}}$ so s for any given proportion to be selected is decreased when n is increased.

Let s, $b_{g\bar{p}}$, and V_p represent the selection differential, the regression coefficient, and the phenotypic variance when n = 1 and s^1, $b^1_{g\bar{p}}$, and V_{p^-} represent their values when n is a number other than 1.0.
Now suppose that for a given n

$$V_{\bar{p}} = \frac{V_p}{a}$$

Then

$$b^1_{g\bar{p}} = \frac{V_g + V_c}{V_{p/a}} = a b_{g\bar{p}} ,$$

$$s^1 = s/\sqrt{a}, \text{ and}$$

$$g^1_s \ (=) \ a b_{g\bar{p}} \ s/\sqrt{a} = \sqrt{a} \ s b_{g\bar{p}} = \sqrt{a} g_s \qquad (25)$$

For example, if n is increased enough to make a equal to 2, i.e. to halve $V_{\bar{p}}$, genotypic superiority obtained through selection is $\sqrt{2}$ or 1.414 times as great if the same proportion is selected in each case.

Values of $\sqrt{a}$ are tabulated below for a small set of values of n and R.

	R					
n	.1	.2	.4	.6	.8	1.0
2	1.35	1.29	1.20	1.12	1.05	1.0
3	1.58	1.46	1.29	1.17	1.07	1.0
4	1.75	1.58	1.35	1.20	1.08	1.0
5	1.89	1.67	1.39	1.21	1.09	1.0
6	2.00	1.73	1.41	1.22	1.20	1.0
∞	3.16	2.24	1.58	1.29	1.12	1.0

It is quite obvious that unless V_v is large relative to $V_g + V_c$ added records on the same individuals is an inefficient means of speeding progress by selection.

In the case of plants that can be propagated vegetatively the

situation is modified. The question then has two parts. How many plants of the same genotype and how many observations per plant will be worthwhile? By an extension of reasoning used in the simple case

$$V_{\bar{p}} = V_g + \frac{V_c}{m} + \frac{V_v}{nm}$$ where $\bar{p}$ is a mean for n observations on each of m plants.

$$b_{g\bar{p}} = \frac{V_g}{V_g + \frac{V_c}{m} + \frac{V_v}{nm}}$$

As before when n and/or m are changed from one let

$$V_{\bar{p}} = \frac{V_p}{a}$$

Then as before

$$g^1{}_s = \sqrt{a}\, g_s$$

Given V_g/V_p, the proportion of phenotypic variance of a single observation which is of genotypic origin, and $V_c/(V_c + V_v)$, the proportion of environmental variance arising from environment constant for individual plants, $\sqrt{a}$ can be computed for varying values of n and m. A table of such values is given below.

			V_g/V_p =	.1			.4			.8		
nm	m	n	$V_c/(V_c+V_v)$ =	.1	.4	.8	.1	.4	.8	.1	.4	.8
1	1	1		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1	2		1.30	1.17	1.05	1.17	1.10	1.03	1.05	1.03	1.01
	2	1		1.35	1.35	1.35	1.19	1.19	1.19	1.05	1.05	1.05
4	1	4		1.60	1.30	1.08	1.30	1.17	1.05	1.08	1.05	1.02
	2	2		1.70	1.55	1.41	1.33	1.28	1.22	1.08	1.07	1.06
	4	1		1.75	1.75	1.75	1.35	1.35	1.35	1.08	1.08	1.08
8	1	8		1.85	1.38	1.09	1.38	1.21	1.06	1.09	1.06	1.02
	2	4		2.01	1.70	1.44	1.42	1.33	1.24	1.10	1.08	1.06
	4	2		2.11	1.97	1.82	1.44	1.41	1.37	1.10	1.09	1.09
	8	1		2.17	2.17	2.17	1.45	1.45	1.45	1.10	1.10	1.10
16	1	16		2.01	1.42	1.10	1.42	1.23	1.06	1.10	1.06	1.02
	2	8		2.23	1.79	1.46	1.47	1.36	1.24	1.10	1.09	1.06
	4	4		2.36	2.11	1.85	1.49	1.44	1.38	1.11	1.11	1.09
	8	2		2.44	2.36	2.23	1.51	1.49	1.46	1.11	1.11	1.10
	16	1		2.49	2.49	2.49	1.51	1.51	1.51	1.11	1.11	1.11
∞	∞			3.16	3.16	3.16	1.58	1.58	1.58	1.12	1.12	1.12

certain facts stand out.

1. When V_g/V_p is high increasing either n or m is of little value.

2. When $V_c/(V_c + V_v)$ is even moderately high increasing n is of little value, e.g. with m = 2, V_g/V_p = .1 and $V_c/(V_c + V_v)$ = .4 increasing n from 1 to 8 changes $\sqrt{a}$ from 1.35 to 1.79. Put differently making 8 observations per plant instead of 1 increases the effectiveness of selection only 33% as a return for carrying the material 8 times as long.

3. $\sqrt{a}$ is a maximum for any value of nm when n = 1. However, we must not construe this to mean that it is always inefficient to increase n. The matter of genotype-age interaction (not considered previously) now comes into the picture. Obviously, if such an interaction has any possibility of occurrence observation at any one age might be misleading. What can be concluded is that if more than one observation per plant is to be made because of the possibility of an age-genotype interaction, the observations should be spaced widely in years. A few observations spaced over a wide period of time will catch the interaction and for reducing environmental variance of the mean it is more efficient to increase m.

Suppose we were working with a plant for which V_g/V_p = .1 and $V_c/(V_c + V_v)$ = .4 and that we wanted to make two observations (spaced several years apart) on each plant as a check on change in performance with age. If we tested 8 plants per genotype our progress would be (2.36 - 1.97)/ 1.97 or 20% greater than if 4 plants were tested from each of the same number of genotypes, However, if the limitation on our work is in terms of total numbers of plants that can be tested instead of the number of genotypes that can be tested, we could test twice as many genotypes if only 4 plants each were raised. And if the number selected is to be constant regardless of number tested the proportion selected would only be half as great if twice as many are tested. Suppose 10% are to be selected if 8 per genotype are tested and 5% if 4 per genotype are tested. In the latter case the selection differential will be 2.06 standard deviation as against 1.76 standard deviations in the former. Consequently, the ratio of progress made using 4 plants per genotype to that made using 8 per genotype would actually be (1.97 x 2.06): (2.36 x 1.76) or 4.06:4.15, and advantage of only 2% for 8 plants per genotype. If the proportion to be saved were higher it would actually prove advantageous to use only 4 plants per genotype. Thus if the proportion to be saved were 40% and 20% instead of 10% and 5%, 4 plants per genotype would have a 22% advantage in expected progress.

The estimation of V_g, V_c, + V_v

V_g and V_c cannot be estimated separately (except indirectly) except in plants that can be propogated vegetatively. However, only in such cases is it worth much to know anything but their sum, which can be estimated.

Suppose we have records for n years on each of m cows. An analysis of variance of the data would be as follows:

	d.f.	M.S.	Quantity estimated by M.S.
Years	n - 1		
Cows	m - 1	A	$V_v + n(V_g + V_c)$
Y x C	(n -1)(m -1)	B	V_v
Total	mn - 1		

Hence B estimates V_v and (A - B)/n estimates $V_g + V_c$

Suppose in the case of a plant in which vegetative propagation can be practiced we made observations in n years on each of m plants of each of g genotypes. The analyses of variance of the data would be as follows:

	d.f.	M.S.	Quantity estimated by M.S.
Years	n - 1		
Genotypes	g - 1	A	$V_v + mV_{ag} + nmV_g$
Plants in Genotypes	g(m - 1)	B	$V_v + nV_c$
Y x G	(n - 1)(g - 1)	C	$V_v + mV_{ag}$
Y x P in Genotypes	g(n - 1)(m-1)	D	V_v
Total	gnm - 1		

Hence D estimates V_v, (B-D)/n estimates V_c, and (A-C)/nm estimates V_g. V_{ag} is variance due to age-genotype interaction.

For examples of "repeatability" estimates for characteristics of animals see references (2) and (3) cited in Section IV. See also references (5) and (6) cited below. Stewart estimated "repeatability" directly from the regression of 2nd record on 1st since the animals involved had been selected on the basis of their first records and hence the technique described above would have given a biased estimate.

References:

5. Stewart, H. A. The Inheritance of Prolificacy in Swine. Jour. An. Sci. 4:359 - 366.

6. Jour. Dairy Sci. 25:45 - 56.

VII. Genotypic variance arising from a single pair of genes.

If the gene pair (A,a) is present in a population three genotypes (with respect to the A locus) will appear; AA, Aa, and aa. Their frequencies will be in the ratio $q^2 : 2q(1-q) : (1-q)^2$. Let the average effect of aa on the organism be $\bar{x}$, that of AA be $(\bar{x} + 2u)$, and that of Aa be $(\bar{x} + u + d)$. Clearly, if A is completely dominant, d = u and if dominance is entirely absent, d = o. d/u can be taken as a measure of dominance (this scheme for symbolizing the effects of the three genotypes was taken from Fisher, Immer and Tedin (7)). The situation is summarized in tabular form below.

Genotype	Frequency	Y^1	Y	X
AA	q^2	$z + 2u$	u	2
Aa	$2q(1-q)$	$z + u + d$	d	1
aa	$(1-q)^2$	z	-u	o

Y is obtained by coding Y^1 ; $Y = Y^1 - (z + u)$
X is the number of A's in the genotype

$$V_Y = \frac{S(Y^2) - (SY)^2/N}{N}$$

Note: N(total frequency) = 1, hence $V_Y = SY^2 - (SY)^2$

$$SY = q^2u + 2q(1-q)d - (1-q)^2 u$$

$$= u(q^2 - 1 + 2q - q^2) + 2q(1-q)d$$

$$= (2q-1)u + 2q(1-q)d \qquad (27)$$

$$SY^2 = q^2u^2 + 2q(1-q)d^2 + (1-q)^2 u^2$$

$$V_Y = u^2\left[q^2 + (1-q)^2\right] + 2q(1-q)d^2 - \left[(2q-1)u + 2q(1-q)d\right]^2$$

$$= 2q(1-q)u^2 + 2q(1-q)\left[1 - 2q(1-q)\right]d^2 + 4q(1-q)(1-2q)ud \qquad (28)$$

Note:

(1) When d = o; dominance absent

$$V_Y = 2q(1-q)u^2$$

V_Y is a maximum when q = .5

(2) When d = u; dominance complete

$$V_Y = 4q(2-q)(1-q)^2 u^2$$

V_Y is a maximum when q = .293

(3) Assuming the same value for u,

V_Y is $2(2 - q)(1 - q)$ times as great when

$d = u$ as when $d = 0$.

As $q \longrightarrow 0$, $2(2 - q)(1 - q) \longrightarrow 4.0$

As $q \longrightarrow 1$, $2(2 - q)(1 - q) \longrightarrow 0.0$

(4) Since Y will be essentially uncorrelated with z (which is a function of the remainder of the organisms genotype and of environment) the total phenotypic variance will be $V_z + V_Y$.

Let the additive effect of A be defined as the regression of Y^1 (or of Y, which will be equivalent) on X, the number of A's in the genotype. Remembering (see Section 2) that the regression coefficient is computed in a manner that minimizes the unexplained portion of the variance, you will see that by defining the additive effect of A in this way we are giving it the value which allows the greatest possible portion of V_Y to be explained on the basis of additive gene action. It happens that in population where mating is random, the additive effect computed according to the above definition is also the response that would be obtained from substituting A for a, averaged for all loci in the population where a is present (and hence the substitution is theoretically possible). Thus, the concept of additive effects applied in cases where gene action is not of the simple additive type is not so abstract as it might at first appear.

$$\text{Cov}_{XY} = SXY - (SX)(SY), \text{ and}$$

$$V_X = SX^2 - (SX)^2 \qquad \text{(remembering that } N = 1.0\text{)}$$

$$S(X) = 2q^2 + 2q(1 - q) = 2q$$

$$V_X = 4q^2 + 2q(1 - q) - 4q^2 = 2q(1 - q)$$

$$\text{Cov}_{XY} = 2q^2u + 2q(1 - q)d - 2q\left[(2q - 1)u + 2q(1 - q)d\right]$$

$$= \left[2q^2 - 4q^2 + 2q\right]u + \left[2q(1 - q) - 4q^2(1 - q)\right]d$$

$$= 2q(1 - q)u + 2q(1 - q)(1 - 2q)d$$

$$b_{YX} = \frac{2q(1 - q)u + 2q(1 - q)(1 - 2q)d}{2q(1 - q)} = u + (1 - 2q)d \qquad (29)$$

Note:

(1) When $d = o$, b_{XY} (the additive effect of A) $= u$
This is obvious since considering average phenotypic values $(AA - Aa) = (Aa - aa) = u$

(2) When $d = u$, $b_{XY} = 2u(1 - q)$

(3) When $q = .5$, $b_{XY} = u$ regardless of the value of d. Thus, when $q = .5$ the additive effect of A is always half the difference between the average phenotypic values of AA and aa.

The additively genetic variance is the variance associated with the additive effect of A which is, from the definition of additive effects, the portion of the variance in Y due to regression on X(the number of A genes in a genotype).

Additively genetic variance

$$= V_g = \frac{(\text{Cov } XY)^2}{V_X} = \frac{[2q(1-q)]^2 [u + (1-2q)d]^2}{2q(1-q)} \qquad (30)$$

$$= 2q(1-q)[u + (1-2q)d]^2$$

As a fraction of V_Y it is

$$\frac{[u + (1-2q)d]^2}{u^2 + [1 - 2q(1-q)]d^2 + 2(1-2q)ud}$$

Let $a = d/u$ as a measure of degree of dominance. Then $d = au$ and substituting au for d in the above expression, we obtain

$$\frac{1 + 2(1-2q)a + (1-2q)^2 a^2}{1 + 2(1-2q)a + [1 - 2q(1-q)]a^2} = V_g / V_Y \qquad (31)$$

When some degree of dominance is present ($a \neq o$) V_g will always be smaller than V_Y. The difference is variance caused by dominance deviations from the additive scheme. (In the case of genotypic variance arising from two or more gene pairs there may also be variance caused by gene interaction deviations).

Variance due to dominance deviations

$$= V_d = V_Y - V_g$$

As a fraction of V_Y it is

$$V_d / V_Y = 1 - V_g / V_Y$$

Note: From Eq. (31)

(1) When $a = o$ (no dominance), $V_g / V_Y = 1.0$.

(2) When $d = u$ (complete dominance), $V_g / V_Y = \frac{2(1-q)}{(2-q)}$

The ratio V_g / V_Y is listed below for various values of a and q It will be noted that

	q				
a	.1	.3	.5	.7	.9
0.0	1.0	1.0	1.0	1.0	1.0
.2	.99	.99	.91	.98	.99
.4	.98	.95	.83	.91	.94
.6	.97	.91	.77	.79	.81
.8	.96	.87	.71	.63	.53
1.0	.95	.82	.67	.46	.18

dominance deviations are a comparatively minor source of variance unless a, the degree of dominance, is above 0.6 and q is 0.5 or over. With complete dominance and q above 0.7 the proportion of variance due to dominance deviations becomes large.

Reference:

Fisher, R.A., Immer, F.R., Tedin, Olaf
1932: Genetical Interpretation of Statistics
of the third degree in the study of
Quantitative Inheritance.
Genetics 17: 107: 121.

Supplement to VII

A matter of some significance that should have been clarified in Mimeograph No. 5 is the magnitude of V_g (additively genetic variance) when a degree of dominance is present relative to its magnitude when dominance is absent.

$$V_g = 2q(1-q)\left[u + (1-2q)d\right]^2 \qquad \text{from (30)}$$

Substituting au for d as before $\quad a = \frac{d}{u}$

$$V_g = 2q(1-q)\left[1 + 2(1-2q)a + (1-2q)^2 a^2\right]u^2 \qquad (32)$$

When a = o

$$V_g = 2q(1-q)u^2 = V_Y \qquad (33)$$

Let V'_g signify the value of V_g when a = o

Then (assuming u to have the same value in both cases)

$$V_g/V'_g = 1 + 2(1-2q)a + (1-2q)^2a^2$$

$$= 1 + (1-2q)a\left[2 + (1-2q)a\right] \qquad (34)$$

Note that this quantity will be greater than 1.0 when $q < .5$ but less than 1.0 when $q > .5$. Thus when dominance is present but $q < .5$ the additively genetic portion of the variance will be greater than the total variance (which is all additively genetic) when dominance is absent; however, when $q > .5$ the additively genetic variance in the presence of dominance will be less than if dominance were absent.

Another matter that deserved attention is the formula for V_d.

$$V_d = V_Y - V_g$$

which from eqs. (28) and (32), substituting au for d

$$= 2q(1-q)\left[1 + 2(1-2q)a + \left[1 - 2q(1-q)\right]a^2\right]u^2$$

$$- 2q(1-q)\left[1 + 2(1-2q)a + (1-2q)^2a^2\right]u^2$$

$$= 2q(1-q)\left[1 - 2q(1-q) - (1-2q)^2\right]a^2u^2$$

$$= 4q^2(1-q)^2a^2u^2 \qquad (35)$$

Consider the case where q = .5 and a = 1 (dominance complete). By eq. (32), $V_g = \frac{1}{2} u^2$; and by eq. (35), $V_d = \frac{1}{4} u^2$. $V_Y = V_g + V_d = 3/4\ u^2$ and $V_g/V_Y = 2/3$, the value tabulated on page 4 of Mimeograph No. 5.

VIII Selection considered relative to a single gene pair.

In Section IV the relation between the genotypic superiority of selected individuals and their phenotypic superiority (s) was defined. The relationship of their genotypic value to that of their progeny remains to be clarified.

In a population within which mating among selected individuals is random the genotypic mean of the population is a function of gene frequency. Consider again the situation set up in Section VII (using au = d).

Genotype	Frequency	Y'	Y	X
AA	q^2	z + 2u	u	2
Aa	2q(1 - q)	z + u + au	au	1
aa	$(1 - q)^2$	z	-u	0

$$SY = (2q - 1)u + 2q(1 - q)au = u\left[(2q - 1) + 2q(1 - q)a\right] \qquad (36)$$

and since N(total frequency) = 1.0 this is also the value of $\overline{Y}$. Consequently the difference in gene frequency between selected individuals and the population from which they were selected is the key to the genotypic superiority of their offspring. This difference can be estimated through use of the regression of X and Y.

$$b_{XY} = \frac{\text{Cov } XY}{V_Y}$$

As X goes from zero to 2.0, q goes from zero to 1.0, hence

$$b_{qY} = b_{XY}/2 = \frac{\text{Cov } XY}{2V_Y} \qquad (37)$$

The change in genotypic mean per unit change in q is (for an infinitesimal change in q) the derivative of $\overline{Y}$ with respect to q.

$$\overline{Y} = u\left[(2q - 1) + 2q(1 - q)a\right]$$

$$\frac{d\overline{Y}}{d_q} = 2u\left[1 + (1 - 2q)a\right] \qquad (38)$$

Note: The expression $\frac{d\overline{Y}}{d_q}$ is read, the change in $\overline{Y}$ per unit change in q. It is called the $q\overline{Y}$ derivative of $q\overline{Y}$ with respect to q.

Now consider the effect of selection on genotypic value of offspring in three steps.

(1) For every unit of phenotypic superiority of selected parents they are superior genotypically by b_{gp} units.

(2) For every unit of genotypic superiority of selected parents the average gene frequency is b_{qg} units above the average frequency for the population from which the parents were selected.

(3) The genotypic mean of the offspring of selected parents increases at the rate of $\frac{d\bar{Y}}{dq}$ per unit change in q as gene frequency increases.

Therefore the regression of genotypic mean of offspring on phenotype of parents is

$$b_{gp}\ b_{qg}\ \frac{d\bar{Y}}{dq}$$

$$= \frac{V_Y}{V_p} \cdot \frac{Cov\ XY}{2V_Y} \cdot 2u\left[1 + (1 - 2q)a\right]$$

$$= \frac{Cov\ XY}{V_p} \cdot u\left[1 + (1 - 2q)a\right]$$

Taking value of Cov XY from Section VII and substituting au for d, this becomes

$$\frac{2q(1 - q)\ u^2\left[1 + (1 - 2q)a\right]^2}{V_p}$$

Referring to eq. (32) we see that this is equal to V_g/V_p. Therefore, genetic advance (resulting from change in frequency of the A gene) of the progeny of selected parents over the average genotypic value expected in the absence of selection is

$$s\ V_g/V_p \qquad (39)$$

where s is the selection differential (the average phenotypic superiority of selected parents, V_g is the additively genetic portion of the genotypic variance arising from segregation at the A locus, and V_p is the total phenotypic variance.

Since s for a given proportion of individuals selected is proportional to $\sqrt{V_p}$

$$s = k\sqrt{V_p}$$

and (39) becomes

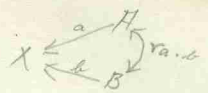
$$k\sqrt{V_p}\ \ V_g/V_p = k\,V_g/\sqrt{V_p}$$

Attention to the effects of dominance on the size of V_g and V_p will bring out that when q is less than .5 genetic advance will frequently be greater if dominance is present than if it were not (this assumes u to have the same value in either case). How much q must be below .5 to make this true depends on the relative magnitudes of genotypic and environmental variance. When q is greater than .5 genetic advance will always be less when dominance is present than it would be if dominance were not present.

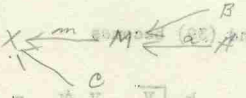
Wright. Amer. Nat. 1922 Inbreeding Coef.

Path Coef. 1921

$$P_{coef\ X \cdot A} = \frac{\sigma_{X \cdot A}}{\sigma_X}$$



$$\frac{\sqrt{V_A}}{\sqrt{V_X}} = P_{X \cdot A} = a$$



$$P_{X \cdot A} = P_{m \cdot A} P_{X \cdot m}$$

r_{xy} is sum of all paths back of X+Y

$$r_{xy} = bb' + a r_{ab} b' + a r_{ac} c + b r_{bc} c$$

IX. Deviations from the additive scheme caused by gene interactions.

Deviations caused by gene interactions are referred to by Wright and Lush as epistatic deviations from the additive scheme or simply as epistatic deviations. The computation of generalized formulae for variance resulting from epistatic deviations and its magnitude relative to additively genetic variance is very tedious. As an alternative we will study two or three specific cases.

Case 1 - Consider the type of interaction referred to ordinarily as complimentary gene action. Genes A and B have no effect by themselves but when both are present either in simplex or duplex an "effect" is observed. Let the "effect" be quantitized as 2 y. Then the situation will be as below.

Genotype	Frequency	Y'	Y	X_a	X_b
AABB	$q_a^2 q^2_b$	z + 2y	2y	2	2
AABb	$q_a^2 2q_b(1-q_b)$	z + 2y	2y	2	1
AAbb	$q_a^2(1-q_b)^2$	z	0	2	0
AaBB	$2q_a(1-q_a)q_b^2$	z + 2y	2y	1	2
AaBb	$2q_a(1-q_a)2q_b(1-q_b)$	z + 2y	2y	1	1
Aabb	$2q_a(1-q_a)(1-q_b)^2$	z	0	1	0
aaBB	$(1-q_a)^2 q_b^2$	z	0	0	2
aaBb	$(1-q_a)^2 2q_b(1-q_b)$	z	0	0	1
aabb	$(1-q_a)^2(1-q_b)^2$	z	0	0	0

$$\begin{aligned} SY &= 2q_a^2 q_b^2 y + 4q_a^2 q_b(1-q_b)y + 4q_a(1-q_a)q_b^2 y + 8q_a q_b(1-q_a)(1-q_b)y \\ &= 2q_a q_b\left[q_a q_b + 2q_a(1-q_b) + 2q_b(1-q_a) + 4(1-q_a)(1-q_b)\right] y \\ &= 2q_a q_b(2-q_a)(2-q_b)y \end{aligned} \tag{40}$$

$$\begin{aligned} SY^2 &= 4q_a^2 q_b^2 y^2 + 8q_a^2 q_b(1-q_b)y^2 + 8q_a(1-q_a)q_b^2 y^2 + 16q_a(1-q_a)q_b(1-q_b)y^2 \\ &= 4q_a q_b\left[q_a q_b + 2q_a(1-q_b) + 2q_b(1-q_a) + 4(1-q_a)(1-q_b)\right] y^2 \\ &= 4q_a q_b(2-q_a)(2-q_b)y^2 \end{aligned}$$

$$\begin{aligned} V_Y &= 4q_a q_b(2-q_a)(2-q_b)y^2 - 4q_a^2 q_b^2(2-q_a)^2(2-q_b)^2 y^2 \\ &= 4q_a q_b(2-q_a)(2-q_b)\left[1-q_a q_b(2-q_a)(2-q_b)\right] y^2 \end{aligned} \tag{41}$$

Note that $q_a q_b(2-q_a)(2-q_b)$ = the frequency of genotypes containing at least one A and one B gene. The variance will be a maximum when half the genotypes have the value, 2y, and half the value, zero. Hence V_Y will be a maximum when $q_a q_b(2-q_a)(2-q_b) = 1/2$.

The average values for genotypes AA, Aa, and aa will be as follows:

$$Y_{AA} = 2y\left[q_b^2 + 2q_b(1-q_b)\right] = 2q_b(2-q_b)y$$

$$Y_{Aa} = 2y\left[q_b^2 + 2q_b(1-q_b)\right] = 2q_b(2-q_b)y$$

$$Y_{aa} = 0$$

In corresponding fashion

$$Y_{BB} = 2q_a(2-q_a)y$$

$$Y_{Bb} = 2q_a(2-q_a)y$$

$$Y_{bb} = 0$$

Note that in the case of A the dominance deviation, $d = u = q_b(2-q_b)y$ and in the case of the B gene, $d = u = q_a(2-q_a)y$. When $d = u$, V_g for a single gene pair is from eq. (30)

$$8q(1-q)^3u^2$$

Hence V_g resulting from the A, a pair will be

$$8q_a(1-q_a)^3q_b^2(2-q_b)^2y^2$$

and V_g resulting from the B,b pair will be

$$8q_b(1-q_b)^3q_a^2(2-q_a)^2y^2$$

Again when $d = u$, V_d for a single gene pair is from eq. (35)

$$4q^2(1-q)^2u^2$$

and V_d resulting from the A, a pair will be

$$4q_a^2(1-q_a)^2q_b^2(2-q_b)^2y^2$$

and from the B, b pair will be

$$4q_b^2(1-q_b)^2q_a^2(2-q_a)^2y^2$$

The variance from epistatic deviations is the total V_Y minus the sum of the four portions given above. This becomes

$$4q_aq_b(2-q_a)(2-q_b)\left[1-q_b(2-q_b)(1-q_a)^2-q_a(2-q_a)\right]y^2$$

Note that this expression becomes zero when either q_a or q_b becomes either zero or 1.0. Values of V_g, V_d, and V_e are listed on the next page for various values of q_a and q_b.

q_b		q_a = .1	.3	.5	.7	.9
.1	V_y	.139	.350	.489	.572	.611
	V_g	.042	.181	.347	.488	.572
	V_d	.002	.015	.027	.033	.033
	V_i	.095	.154	.115	.050	.006
.3	V_Y	.350	.770	.945	.995	1.000
	V_g	.182	.428	.593	.721	.809
	V_d	.014	.092	.164	.192	.181
	V_i	.154	.250	.187	.082	.010
.5	V_Y	.489	.945	.98	.867	.765
	V_g	.346	.593	.56	.499	.494
	V_d	.027	.164	.28	.306	.263
	V_i	.115	.187	.14	.061	.007
.7	V_Y	.572	.995	.867	.569	.357
	V_g	.488	.721	.499	.250	.154
	V_d	.033	.192	.306	.292	.200
	V_i	.050	.082	.061	.027	.003
.9	V_Y	.611	1.000	.765	.357	.0780
	V_g	.572	.809	.494	.154	.0142
	V_d	.033	.181	.263	.200	.0636
	V_i	.006	.010	.007	.003	.0004

From Section VII note that when d = u, Cov XY = $4q(1-q)^2u$ (42)

Hence

$$\text{Cov } X_aY = 4q_a(1-q_a)^2q_b(2-q_b)y \tag{43}$$

since $u = q_b(2-q_b)y$ and since the variation in Y arising from segregation at the B locus and from the joint effect of A and B is uncorrelated with X_a and therefore does not contribute to the covariance of X_a and Y.

$$b_{X_aY} = \frac{\text{Cov } X_aY}{V_Y} \quad \text{and} \quad b_{q_aY} = \frac{\text{Cov } X_aY}{2V_Y}$$

Taking the partial derivative of $\bar{Y}$ with respect to q_a

$$\frac{\partial \bar{Y}}{\partial q_a} = 2y\left[2q_b(2-q_b)(1-q_a)\right]$$

The change in $\bar{Y}$ that would result if selection acted on q_a and not q_b would be expressed by the regression of $\bar{Y}$ on phenotypic superiority of selected parents.

$$b_{Yp} \cdot b_{q_aY} \cdot \frac{\partial \bar{Y}}{\partial q_a} = \frac{V_Y}{V_p} \cdot \frac{\text{Cov } X_aY}{2V_Y} \cdot 2y\left[2q_b(2-q_b)(1-q_a)\right]$$

$$= \frac{8q_a(1-q_a)^3q_b^2(2-q_b)^2y^2}{V_p}$$

Note that the numerator is the additively genetic variance arising from segregation at the A locus.

In an analogous fashion it can be shown that if selection acted only on q_b the regression of offsprings genotype on phenotypic superiority of parents would be

$$\frac{8q_b(1-q_b)^3q_a^2(2-q_a)^2y^2}{V_p}$$

or the ratio of additively genetic variance associated with the B,b pair to the total phenotypic varinnce. If there were no gene interaction the regression of genotype of offspring on phenotype of parent would be the sum of the regressions for the separate gene pairs. However, examination of the expression for $\bar{Y}$, eq. (40), reveals that when q_a and q_b are increased simultaneously $\bar{Y}$ increases somewhat more than the sum of the increases that would occur (1) if q_a were increased while q_b remained constant, and (2) if q_b increased while q_a remained constant. This means that the regression of offsprings genotype on parents phenotype is a little greater than the proportion of the total phenotypic variance which is additively genetic. Wright (8) has shown that it will be larger by one-half the fraction of total phenotypic variance which arises from epistatic deviations.

In an interaction system of the type under discussion the gene frequency will tend to remain equal for the two pairs of genes. This is apparent from the fact that selection causes the greatest increase in q for the gene pair for which q is lowest at the time. The regression of q_a on

phenotype is $b_{Yp} b_{q_a Y}$ and the change in q_a with a given amount of selection is

$$\triangle q_a = s b_{Yp} b_{q_a Y} = k\sqrt{V_p} \cdot \frac{V_Y}{V_p} \cdot \frac{\text{Cov } X_a Y}{2V_Y} = \frac{k \text{ Cov } X_a Y}{2\sqrt{V_p}}$$

where k = selection differential in standard deviations. In like manner

$$\triangle q_b = \frac{k \text{ Cov } X_b Y}{2\sqrt{V_p}}$$

Then
$$\frac{\triangle q_a}{\triangle q_b} = \frac{\text{Cov } X_a Y}{\text{Cov } X_b Y} \qquad (44)$$

$$= \frac{4q_a(1-q_a)^2 q_b(2-q_b)y}{4q_b(1-q_b)^2 q_a(2-q_a)y} = \frac{(1-q_a)^2(2-q_b)}{(1-q_b)^2(2-q_b)}$$

(Cov $X_a Y$ is given in eq. (43) and Cov $X_b Y$ is easily obtained from eq. (42) and the value of u for the B, b gene pair).

$\frac{\triangle q_a}{\triangle q_b} = \frac{(1-q_a)^2(2-q_b)}{(1-q_b)^2(2-q_a)}$ is always greater than one when $q_a < q_b$ and less than one when $q_b < q_a$ from which it is apparent that selection will tend to keep q_a and q_b equal.

Case 2. - Consider two gene pairs with equal and additive effects on a character that is at its optimum when two plus genes are present regardless of the locus at which they are present. Either one or three plus genes would be less desirable than two and none or four still less desirable. The situation might then be as below.

Genotype	Frequency	Y	X_a	X_b
AABB	$q_a^2 q_b^2$	0	2	2
AABb	$2q_a^2 q_b(1-q_b)$	y	2	1
AAbb	$q_a^2(1-q_b)^2$	2y	2	0
AaBB	$2q_a(1-q_a)q_b^2$	y	1	2
AaBb	$4q_a q_b(1-q_a)(1-q_b)$	2y	1	1
Aabb	$2q_a(1-q_a)(1-q_b)^2$	y	1	0
aaBB	$(1-q_a)^2 q_b^2$	2y	0	2
aaBb	$2(1-q_a)^2 q_b(1-q_b)$	y	0	1
aabb	$(1-q_a)^2(1-q_b)^2$	0	0	0

For the A, a pair

$$Y_{AA} = \frac{2q_a^2q_b(1-q_b)y + 2q_a^2(1-q_b)^2y}{q_a^2} = 2(1-q_b)y$$

$$Y_{Aa} = (1+2q_b-2q_b^2)y$$

$$Y_{aa} = 2q_by$$

$$u = \frac{2(1-q_b)y - 2q_by}{2} = (1-2q_b)y$$

$$d = (1+2q_b-2q_b^2)y - \frac{2(1-q_b)y + 2q_by}{2}$$

$$= 2q_b(1-q_b)y$$

For the B,b pair

$$u = (1-2q_a)y$$

$$d = 2q_a(1-q_a)y$$

As before (see eq. 44)

$$\frac{\Delta q_a}{\Delta q_b} = \frac{\text{Cov } X_aY}{\text{Cov } X_bY}$$

which in this case is

$$\frac{2q_a(1-q_a)(1-2q_b) + 2q_aq_b(1-q_a)(1-q_b)(1-2q_a)}{2q_b(1-q_b)(1-2q_a) + 2q_aq_b(1-q_a)(1-q_b)(1-2q_b)}$$

(from the general formula for Cov XY, section VII and the values of u and d listed above)

Two things should be noted from the expression for the ratio $\Delta q_a/\Delta q_b$.

1. When $q_a = q_b = .5$ both numerator and denominator are zero. This means that selection has no tendency to change either q_a or q_b, i.e., the system is in equilibrium. However, it is an unstable equilibrium as will be shown.

2. Whenever q_a is larger than q_b, Δq_a is larger than Δq_b. Both may be negative but if $q_a > q_b$, Δq_a will be the smaller negative value, i.e., larger in the algebraic sense. This means that except when $q_a = q_b$ selection will tend to increase the difference in gene frequency between the two gene pairs. Ultimately this would tend to bring one q value to 1.0, the other to zero.

When $q_a = q_b$, both will tend toward .5 at the same rate. When $q_a = q_b = .5$ no further permanent change in the population mean will result from selection, though genotypic variance will be present, unless the equilibrium (which is unstable) is disturbed. This can best be accomplished by a generation of close inbreeding during which there is a good chance of throwing q_a and q_b out of equality. If this is done selection in succeeding outbreeding will increase the larger of the two, decrease the smaller and thereby increase the population mean and decrease genotypic variance by bringing the genotypic value of the whole population to the level of the best obtained when q_a and q_b were equal to .5.

While the situation studied above is empirical and of extreme simplicity it brings out some important things about characters for which selection is for an optimum rather than an extreme. A typical example is body conformation which in certain meat animals is of considerable economic importance. Type has received more attention in selection than any other trait yet variation around optimum type in herds where strict selection is practiced is extreme. It seems likely that type variation could be reduced without permanent increase in inbreeding by alternating inbreeding with outbreeding between inbred animals.

Summary

Variance due to epistatic deviations may be a very small portion of total genotypic variance even though definite gene interactions are present. This is exemplified in Case 1. In such cases mass selection may be highly effective in outbred populations.

At the other extreme are types of interaction such as Case 2 where the additively genetic portion of the variance may be close to zero and consequently mass selection will be ineffective. In such cases a generation of close inbreeding may upset the equilibrium and lead to a situation where mass selection will again be effective.

In complex characters such as yield the effects of primary characters may combine by multiplication to produce the trait in question. In this type of non-additive action of non-alleles the portion of the genotypic variance due to epistatic deviations will be of the order observed in Case 1.

X. Selection and gene frequency.

Formulae worth fixing in mind and which will be used in this section are the following:

1. Additively genetic variance

$$V_g = 2q(1-q)\left[1 + (1-2q)a\right]^2 u^2$$

2. Variance due to dominance deviations

$$V_d = 4q^2(1-q)^2a^2u^2$$

3. Total genotypic variance

$$V_Y = 2q(1-q)\left[1 + 2(1-2q)a + (1-2q+2q^2)a^2\right]u^2 = V_g + V_d$$

4. $\text{Cov } XY = 2q(1-q)\left[1 + (1-2q)a\right]u$

The change in gene frequency resulting from selection based on phenotype will be the product of the selection differential and the regression coefficient of gene frequency on phenotype.

$$\Delta q = sb_{qp} \tag{45}$$

In accordance with the argument of Section 8

$$b_{qp} = b_{Yp}\, b_{qY}$$

But $b_{Yp} = V_Y / V_p$ (see Section IV)

and $b_{qY} = \dfrac{\text{Cov } XY}{2V_Y}$ (eq. 37)

Hence
$$b_{qp} = \frac{V_Y}{V_p} \cdot \frac{\text{Cov } XY}{2V_Y} = \frac{\text{Cov } XY}{2V_p} \tag{46}$$

The selection differential, s, is understood to be measured in the units by which phenotype is measured. Obviously we can write

$$s = k\sqrt{V_p} \tag{47}$$

where k is the selection differential in standard units.

Now substituting in eq. (45)

$$\Delta q = \frac{k \sqrt{V_p}\ \text{Cov } XY}{2V_p} \quad \frac{k}{2} \cdot \frac{\text{Cov } XY}{\sqrt{V_p}} \qquad (48)$$

In order to make use of eq. (48) for prediction of Δ q, change in q resulting from selection, assumptions must be made (1) about the magnitude of additively genetic variance relative to total phenotypic variance, and (2) about the number of segregating genes affecting the trait for which selection is to be practiced. Assume n gene pairs each responsible for an equal amount of additively genetic variance, V_g. The total additively genetic variance arising from the n pairs will then be n V_g.

Let $nV_g = \frac{1}{t} V_p$

Then $V_p = ntV_g$

Substituting in eq. (48) we have

$$\Delta q = \frac{k}{2} \cdot \frac{\text{Cov } XY}{\sqrt{ntV_g}}$$

$$= \frac{k}{2\sqrt{nt}} \cdot \frac{2q(1-q)\left[1 + (1-2q)a\right] u}{\sqrt{2q(1-q)\left[1+(1-2q)a\right]^2 u^2}}$$

$$= \frac{k}{2\sqrt{nt}} \sqrt{2q(1-q)} \qquad (49)$$

Values of Δ q for t = 4, ie. $V_g = V_p/4$, and varying values of q and n are given below. All values listed for Δq are in k units. Thus for q = .7, t= 4, and n = 20;

	q				
n	.1	.3	.5	.7	.9
1	.105	.160	.175	.160	.105
2	.075	.115	.125	.115	.075
3	.060	.095	.010	.095	.060
5	.045	.070	.080	.070	.045
10	.035	.050	.055	.050	.035
20	.025	.035	.040	.035	.025
50	.015	.020	.025	.020	.015
100	.010	.015	.020	.015	.010

$\triangle q$ will be .035 on the average if k = 1.0, i.e. if the selection differential amounts to one phenotypic standard deviation.

The assumption that the n genes contribute equal amounts of additively genetic variance is an artificial one only if the n values of q are also assumed equal as they would be, for example, in the F_1 and F_2 of a cross of pure lines. Starting with such material the frequency of genes contributing the greatest amount of additively genetic variance (those for which $[1+(1-2q)a]u$ was largest) would respond more rapidly to selection pressure. As q changed from one-half, the term $2q(1-q)$ in the formula for V_g would decrease more rapidly in the variance of genes for which q changed most rapidly. The result would be an approach to a moving equilibrium (moving so long as selection pressure continued and was effective) at values of q for the various genes which would make V_g equal for all.

It is of some interest to examine eq.(49) for certain special cases. Consider a trait conditioned by the action of one gene pair, not affected by environment, and for which dominance is not involved. The mean will be

$$(2q-1)u + u + z = 2qu + z$$

(Obtained from eq. (27) adding u + z, the amounts subtracted when original phenotypic values were coded.)

The genotypic variance will be $2q(1-q)u^2$ and this is also the phenotypic variance since there is no effect of environment. The best individuals will have the phenotype

$$2u + z$$

and consequently the greatest possible selection differential will be

$$2u + z - 2qu - z = 2u - 2qu = 2(1-q)u$$

and the value of k will be

$$k = \frac{s}{\sqrt{V_p}} = \frac{2(1-q)u}{\sqrt{2q(1-q)u^2}} = \frac{2(1-q)}{\sqrt{2q(1-q)}}$$

Substituting in eq. (49)

$$\Delta q = \frac{2(1-q)}{2\sqrt{nt}\sqrt{2q(1-q)}} \cdot \sqrt{2q(1-q)} = \frac{1-q}{\sqrt{nt}}$$

But $n = 1$ since there is only one gene pair and $t = 1$ since $V_g = V_p$ (no environmental variance); hence $\sqrt{nt} = 1$, and

$$\Delta q = 1-q$$

Thus q changes from q to

$$q + \Delta q = q + 1-q = 1.0$$

The result is obvious since choice of only AA individuals automatically makes q equal to one; however, such simple checks on formulae help one to become acquainted with them.

As another example consider selection for a trait controlled by one gene pair for which dominance is complete (a = 1, d = u) and unaffected by environment. Assume q = .5. From eq. (27) the mean will be

$$\begin{aligned} &(2q-1)u + 2q(1-q)d + u + z \\ &= \left[2q-1 + 2q(1-q) + 1\right] u + z \qquad (d = u) \\ &= (4q-2q^2)u + z \\ &= 3/2\, u + z \qquad (q = .5) \end{aligned}$$

Three-fourths of the individuals will have the genotypic value $2u + z$ and of these 2/3 will be Aa and 1/3 AA. Hence selection of the best individuals will result in q becoming 2/3 and Δq will have been 1/6. The selection differential, s, will have been 1/2 u. Since there is no effect of environment.

$$V_p = V_Y = \frac{3u^2}{4}$$

$$V_g = \tfrac{1}{2} u^2 \quad \text{and}$$

$$t = V_p/V_g = 1.5$$

$$k = s/\sqrt{V_p} = \frac{u}{2} \sqrt{\frac{4}{3u^2}} \quad \sqrt{\frac{1}{3}}$$

$$\Delta q = \frac{k}{2\sqrt{nt}} \cdot \sqrt{2q(1-q)} \quad \frac{\sqrt{1/3}}{2\sqrt{1.5}} \quad \sqrt{1/2} = 1/6$$

as observed above.

It will be noted that Δq depends on the proportion of phenotypic variance which is additively genetic, the number of genes affecting the trait, and the size of k. Large numbers of genes involved, response of phenotype to environment, and small selection differentials (measured in standard deviations, k values) all reduce the change in gene frequency effected by selection. It is easy to see that in many instances selection may result in only very small changes in gene frequency, but if gene number is large, may at the same time result in considerable increase in the population mean.

Attention should be drawn to the fact that when selection is being practiced for more than one trait, the k values for any one trait must necessarily be reduced. If the traits are considered of equal importance the reduction will be between $1/N$ and $1/\sqrt{N}$ depending on how individuals to be selected are decided upon. This point will be returned to in connection with selection indices.

<u>The composition of genotypic variance when a > 1 (super-dominance) and the effect on Δq under mass selection.</u>

Values of V_g, V_d, and V_Y are listed below for q = .5, .6, and .7 and for values of a in the interval 0 - 3.0. All variances are in terms of the unit u^2.

q	a	V_g	V_d	V_Y
.5	0	.5	0	.5
	.5	.5	.063	.563
	1.0	.5	.250	.750
	1.5	.5	.563	1.063
	2.0	.5	1.000	1.500
	2.5	.5	1.563	2.063
	3.0	.5	2.250	2.750
.6	0	.48	0	.48
	.5	.39	.06	.45
	1.0	.31	.23	.54
	1.5	.24	.52	.76
	2.0	.17	.92	1.09
	2.5	.12	1.44	1.56
	3.0	.08	2.07	2.15
.7	0	.42	.00	.42
	.5	.27	.04	.31
	1.0	.15	.17	.32
	1.5	.07	.40	.47
	2.0	.02	.70	.72
	2.5	.00	1.10	1.10
	3.0	.02	1.58	1.60

Note that strong super-dominance results in additively genetic variance becoming a relatively small portion of total genotypic variance. This obviously means mass selection in random bred populations is relatively ineffective in the presence of super-dominance.

Now note that where q is .7 and a = 2.5 there is no additively genetic variance (V_g =0). You can easily show that for a = 2.5 and q = .8, V_g = $.08u^2$. However, at that value of q, Cov XY is a minus quantity, - .16u. This means that when a = 2.5, mass selection will reduce q when it is over .7 and increase q when it is less than .7. In general, in the presence of super-dominance (a > 1.0) mass selection will bring q to an equilibrium value rather than to one. This equilibrium value will be the value of q for which the population mean, $\bar{Y}$, is a maximum.

Now if the equation for $\bar{Y}$ is differentiated with respect to q, we obtain the rate at which $\bar{Y}$ changes as q changes. Obviously, at the point where $\bar{Y}$ ceases to increase with increasing q and starts to decrease, $\bar{Y}$ is at a maximum. Hence if we form the derivative (by differentiation) of $\bar{Y}$, set it equal to zero, and solve for q, we obtain the value of q for which $\bar{Y}$ is a maximum which is the equilibrium value of q approached under mass selection.

$$\bar{Y} = (2q-1)u + 2q(1-q)d \qquad \text{(eq.27)}$$

$$= (2q-1)u + 2q(1-q)au$$

$$= u(2q-1+2qa-2q^2a)$$

$$\frac{dY}{dq} = u(2+2a-4qa)$$

$$u(2+2a-4qa) = 0$$

$$q = \frac{2+2a}{4a}$$

Equilibrium values of q are listed below for various values of a.

a	q
1.0	1.0
1.5	.833
2.0	.750
2.5	.700
3.0	.667
3.5	.643
4.0	.625

As a increases without limit the equilibrium value of q approaches .5 as a limit.

This matter will be returned to in connection with methods for detecting super-dominance.

XI. The Path Coefficient

In 1921 Sewall Wright (8) described a method for studying the interrelations of correlated variables. The important feature of the method was the use of a priori knowledge concerning cause and effect relationships among the variables concerned. In the article referred to he introduced a statistic which he termed the path coefficient. Actually the path coefficient is a standard partial regression coefficient but specifically the standard partial regression coefficient of a dependent variable on an independent variable and only to be used in that sense since certain of its useful attributes are thereon dependent. A valuable aspect of the path coefficient method as presented by Wright is the graphic presentation of systems of interrelated variables. While it is not a fundamental addition, it aids in clear visualization of what can be very complex situations.

The path coefficient has been used very little in fields other than genetics and even in genetics is not an indispensable tool. However, the conditions necessary for its application are very often met in genetical problems since in many cases there is no uncertainty regarding what is cause and what is effect.

Wright used the path coefficient to obtain a general solution to the once perplexing problem of the effect of inbreeding on homozygosis. It is worth understanding path coefficients if for no other reason than to be able to read with understanding his writings (9,10) on that subject.

The path coefficient is defined as the ratio of the standard deviation of a dependent variable when all independent variables (which exert effects on it) are held constant except the one in question, the variability of which is kept unchanged, to the total standard deviation of the dependent variable. It is understood that the designation of a variable as "dependent" must not be empirical but must rest on known cause and effect relationships of the variables in question. Symbolically,

$$p_{X.A} = \frac{\sqrt{V_{X.A}}}{\sqrt{V_X}} \qquad (50)$$

where $p_{X.A}$ is the path coefficient of X on A, $\sqrt{V_X}$ the standard deviation of X and $\sqrt{V_{X.A}}$ the standard deviation of X resulting solely from variation in A. The symbol $d_{X.A}$ specifies the proportion of the variance (squared standard deviation) of X due solely to variance in A.

$$d_{X.A} = p^2_{X.A} = \frac{V_{X.A}}{V_X} \qquad (51)$$

Wright calls $d_{X.A}$ the coefficient of determination.

The important attributes of the path coefficient and the coefficient of determination are best made clear by consideration of a series of situations graded with respect to complexity. Consider first a variable X which is the

simple sum of two other variables A and B, e.g. the number of dominant genes in a zygote as a function of the number in the two gametes that combined to form it. The situation is presented graphically in Figure 1. Cause and effect paths of relationship are indicated by a single headed arrow pointing from the causative factor to the factor affected. Thus an arrow points from A to X. Correlations not stemming from a cause and effect relationship are indicated by a double-headed (non-directional) arrow as in the case of the correlation between A and B. The notation is lightened by use of small case letters to designate path coefficients. From the position of the letter, "a" in the figure it is understood that "a" will be used instead of $p_{X.a}$.

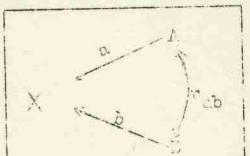


Figure 1

Since $X = A + B$

$$V_X = V_A + V_B + 2\, r_{ab}\sqrt{V_A\, V_B} \qquad \text{(see eq. 11) (52)}$$

It is obvious that if B were held constant without altering variation in A, V_X would then be equal to V_A, hence by the definition of the path coefficient

$$a = p_{X.A} = \sqrt{V_A}/\sqrt{V_X}$$

$$b = p_{X.B} = \sqrt{V_B}/\sqrt{V_X}$$

$$a^2 = d_{X.A} = V_A/V_X \quad \text{and}$$

$$b^2 = d_{X.B} = V_B/V_X$$

Dividing eq. 52 by V_X we obtain

$$\frac{V_X}{V_X} = 1 = \frac{V_A}{V_X} + \frac{V_B}{V_X} + 2r_{ab}\frac{\sqrt{V_A\, V_B}}{V_X}$$

The last term of the above equation is called the coefficient of joint determination and is symbolized by $d_{X.\overline{AB}}$. It is twice the product of the correlation between the two independent variables involved and the two path coefficients.

$$2r_{ab}\frac{\sqrt{V_A\, V_B}}{V_X} = 2r_{ab}\, p_{X.A}\, p_{X.B} \qquad (53)$$

Note that the sum of the two coefficients of direct determination and the coefficient of joint determination is one. Note also that when $r_{ab} = 0$, $p_{X.A} = r_{XA}$ and $p_{X.B} = r_{X.B}$. When $r_{ab} = 0$, Cov $AX = V_A$ and

$$r_{AX} = \frac{V_A}{\sqrt{V_A\, V_X}} = \sqrt{V_A}/\sqrt{V_X}$$

$$= p_{X.A} \qquad (54)$$

Case 2 - Let X be the following function of A,B, and C.

$$X = uA + rB = wC$$

where u, v, and w are constants.

Figure 2 depicts the relationships among the variables.

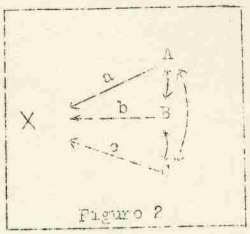


Figure 2

$$V_X = S(uA + vB + wC)^2/N$$

$$= u^2V_A + v^2V_B + w^2V_C + 2r_{ab}uv\sqrt{V_AV_B}$$

$$+ 2r_{ac}uw\sqrt{V_AV_C} + 2r_{bc}vw\sqrt{V_BV_C}.$$

Again if B and C were held constant without altering variation in A, V_X would be equal to u^2V_A. Therefore

$$a = u\sqrt{V_A}/\sqrt{V_X} \quad \text{and} \quad a^2 = d_{X.A} = u^2V_A/V_X$$

In like manner

$$b = v\sqrt{V_B}/\sqrt{V_X}, \quad b^2 = d_{X.B} = v^2V_B/V_X$$

$$c = w\sqrt{V_C}/\sqrt{V_X}, \quad \text{and } c^2 = d_{X.C} = w^2V_C/V_X$$

The coefficient of joint determination by A and B,

$$d_{X.\overline{AB}} = 2r_{ab}\frac{uv\sqrt{V_AV_B}}{V_X} = 2r_{ab}\,p_{X.A}p_{X.B}$$

In like fashion

$$d_{X.\overline{AC}} = 2r_{ac}p_{X.A}p_{X.C}, \text{ and}$$

$$d_{X.\overline{BC}} = 2r_{bc}p_{X.B}p_{X.C}.$$

Note again that dividing the equation for V_X through by V_X we have the sum of the coefficients of direct determination plus the sum of the coefficients of joint determination equal to one. This can be shown true for any system in which all courses of variation in a dependent variable are taken into account. Written in general form

$$\Sigma\, d_{X.i} + \Sigma\, d_{X.\overline{ij}} = 1 \qquad (55)$$

$$\text{or } \Sigma\, p^2_{X.i} + 2\Sigma\, r_{ij}p_ip_j = 1 \qquad (56)$$

Case 3 - Consider a situation such as that depicted in figure 3 in which there is a chain of cause and effect relationships.

Let $X = uA + nM$
and $M = vB + wC$

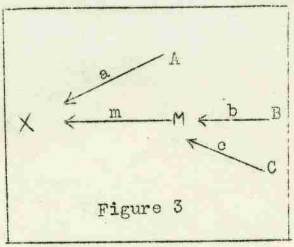


Figure 3

If A remained constant without changing the variance of M, V_X would be equal to n^2V_M. Hence,

$$m = n\sqrt{V_M}/\sqrt{V_X}$$

In like fashion,

$$b = v\sqrt{V_B}/\sqrt{V_M}$$

Now if A and C were both constant with no change in variance of B, V_M would equal v^2V_B and V_X would equal $n^2v^2V_B$. Hence

$$P_{X.B} = nv\sqrt{V_B}/\sqrt{V_X}$$

$$= \frac{n\sqrt{V_M}}{\sqrt{V_X}} \cdot v\frac{\sqrt{V_B}}{\sqrt{V_M}} = mb$$

This reasoning can easily be extended to show that for chains of events (regardless of the number of events involved) the path coefficient from the first to the last event is the product of all the path coefficients between adjacent events. Thus for the situation depicted in figure 4.

$P_{X.D} = abcd$ and $P_{A.C} = bc$.

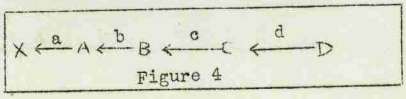


Figure 4

One of the most useful attributes of path coefficients is that the correlation between two variables can be expressed as a function of the path coefficients connecting causative factors to the variables in question and the correlation coefficients among those causative factors. Consider the situation depicted in figure 5. Suppose

$X = uA + vB$, and

$Y = zB + wC$.

$$r_{XY} = \frac{S(uA + vB)(zB + wC)}{\sqrt{SX^2 \cdot SY^2}}$$

Expanding the numerator we have

$$(uzSAB + uwSAC + vwSBC + vzSB^2)/\sqrt{SX^2 \cdot SY^2}$$

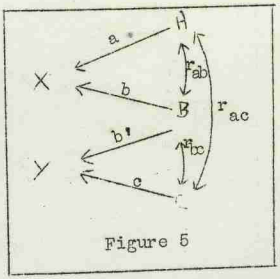


Figure 5

Dividing numerator and denominator by N this becomes

$$\frac{uz\ Cov\ AB}{\sqrt{V_X V_Y}} + \frac{uw\ Cov\ AC}{\sqrt{V_X V_Y}} + \frac{vw\ Cov\ BC}{\sqrt{V_X V_Y}} + \frac{vz\ V_B}{\sqrt{V_X V_Y}}$$

Examination of figure 5 will show there are four connecting paths from X to Y which do not involve more than one correlation coefficient each. It will be shown that if the coefficients along each of these paths are multiplied together the sum of the four products obtained is the correlation coefficient between X and Y. The four paths are as follows:

1. X, A, B, Y
2. X, A, C, Y
3. X, B, C, Y
4. X, B, Y

For the first the product of coefficients is $ab'r_{ab}$

$$a = u\sqrt{V_A} / \sqrt{V_X}$$

$$b' = z\sqrt{V_B} / \sqrt{V_Y}$$

$$r_{ab} = Cov\ AB / \sqrt{V_A V_B}$$

and $ab'r_{ab} = \dfrac{uz\ Cov\ AB}{\sqrt{V_X V_Y}}$ which is the first term in the equation for the correlation of X and Y. In like manner it is easily shown that the products of coefficients for paths 2, 3, and 4 are equal to the other three terms of the equation for r_{XY} as follows:

$$acr_{ac} = \frac{uw\ Cov\ AC}{\sqrt{V_X V_Y}}$$

$$bcr_{bc} = \frac{vw\ Cov\ BC}{\sqrt{V_X V_Y}}$$

$$bb' = \frac{vz\ V_B}{\sqrt{V_X V_Y}}$$

The proof can be extended to show that the correlation of two variables is always the sum of the products of coefficients for the various paths connecting the variables in question. There are two simple rules to remember in application of this fact.

1. No path may include more than one correlation coefficient. Thus, X,A,B,C,Y is not a path contributing to r_{XY} in the above example.

2. Avoid duplicating a path. Suppose in the example a variable D had been a factor influencing both A and B. The paths X, A,B, Y and X,A,D,B.Y would then have been duplicates. This will be avoided if no two paths are allowed to include a common pair of variables (those being correlated excluded).

The applications of path coefficients in genetics revolve around their use in synthesizing correlation coefficients. It should be noted that in all proofs given above all interrelations were linear and additive. These assumptions will be seen to hold in the instances where the method is to be applied. For more extensive proofs see Wright's publication cited below.

References

8. Wright, Sewall (1921) Correlation and Causation. Jour. Agr. Res. 20:557
9. Wright, Sewall (1921) Systems of Mating, I-V Gen. 6:111-178.
10. Wright, Sewall (1922) Coefficients of Inbreeding and Relationship. Amer. Nat. 56:330-338.

XII. The coefficient of inbreeding.

Mating between related individuals is called inbreeding. It is well known that inbreeding reduces heterozygosity. This obviously means that when inbreeding is practiced uniting gametes tend to be more similar in composition than would gametes of the entire population paired at random, i.e. that there is a degree of correlation between uniting gametes. The first step in the derivation of the correlation coefficient (originally derived by Wright (10)) will be to demonstrate that when inbreeding is practiced in the absence of selection the correlation between the numbers of one of a pair of allemorphic genes in uniting gametes is equal to the proportion by which heterozygosity has been decreased by the system of breeding that has been followed. Consider the following situation.

Genotype of gametes

Female	Male	Frequency	X	Y
A	A	$q^2 + L$	1	1
A	a	$q(1-q)-L$	1	0
a	A	$q(1-q)-L$	0	1
a	a	$(1-q)^2 + L$	0	0

X is the number of A genes in the female gamete and Y the number in the male gamete. 2L is the decrease in heterozygous fertilizations from that which occurs in random mating.

2q of the A genes are involved, q from each sex. In random mating $2q^2$ go into homozygous and 2q(1-q) into heterozygous fertilizations

$$2q^2 + 2q(1-q) = 2q^2 + 2q - 2q^2 = 2q$$

When inbreeding is practiced, the number involved in heterozygous fertilizations is reduced. Since all A genes not involved in heterozygous fertilizations must enter into homozygous unions the frequency of the latter type is easily computed to be

$$\frac{2q - 2q(1-q) + 2L}{2} = q^2 + L$$

In like manner, it is easily shown that the frequency of aa unions is $(1-q)^2 + L$ which makes the total frequency one as it must be.

$$SX = \left[q^2 + L + q(1-q) - L\right] = q$$

$$V_X = \left[q^2 + L + q(1-q) - L\right] - q^2$$

$$= (q-q^2) = q(1-q)$$

Obviously

$$SY = q \text{ , and}$$

$$V_Y = q(1-q)$$

since the X and Y distributions are identical.

$$\text{Cov } XY = (q^2+L) - q^2 = L$$

$$r_{XY} = \frac{L}{\sqrt{q(1-q) \cdot q(1-q)}} = \frac{L}{q(1-q)}$$

The percent heterozyhosity under random mating was

$$P = 2q(1-q)$$

Hence

$$r_{XY} = \frac{L}{P/2} = \frac{2L}{P}$$

and the correlation of uniting gametes is obviously the percent decrease in heterozygosis since 2L is the decrease in heterozygosis and P is the heterozygosis originally present.

The next step is to show that the correlation can be put into a form from which it can be conviently computed from pedigree information. Figure 1 depicts the relationships among

1. A zygote
2. The gametes that combined to form it
3. The genotypes of sire and dam.

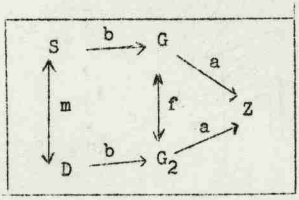


Figure 1.

a is the path coefficient from gamete to zygote.

b is the path coefficient from genotype to gamete formed from it.

$f = r_{XY}$, the correlation between uniting gametes (the coefficient of inbreeding),

and m is the correlation between genotypes of sire and dam. Computing f according to the principal developed in the foregoing section

$$f = b^2 m \tag{57}$$

From eq. 56 of the foregoing section

$$1 = 2a^2 + 2a^2 f$$

$$a = \sqrt{\frac{1}{2(1+f)}} \tag{58}$$

Since the path coefficient, b, represents the only path connecting the gamete and the genotype of the individual producing it, b is equal to the correlation between the gamete and the genotype from which it came. This correlation can be computed using the following information.

Genotype	Gametes produced	Frequency	X	Y
AA	A	$q^2 + f_1 q(1-q)$	2	1
Aa	A	$q(1-q)(1-f_1)$	1	1
Aa	a	$q(1-q)(1-f_1)$	1	0
aa	a	$(1-q)^2 + f_1 q(1-q)$	0	0

X is the number of A genes in the genotype.
Y is the number of A genes in the gamete.
f_1 is the correlation between gametes which gave use to the parents genotype.

$$SX = 2q^2 + 2f_1 q(1-q) + 2q(1-q)(1-f_1) = 2q$$

$$SY = q^2 + f_1 q(1-q) + q(1-q)(1-f_1) = q$$

$$V_X = 4q^2 + 4f_1q(1-q) + 2q(1-q)(1-f_1) - 4q^2$$

$$= 2q(1-q) + 2f_1q(1-q) = 2q(1-q)(1+f_1)$$

$$V_Y = q^2 + f_1q(1-q) + q(1-q)(1-f_1) - q^2$$

$$= q(1-q)$$

$$\text{Cov } XY = 2q^2 + 2f_1q(1-q) + q(1-q)(1-f_1) - 2q^2$$

$$= q(1-q) + f_1q(1-q) = q(1-q)(1+f_1)$$

$$b = r_{XY} = \frac{q(1-q)(1+f_1)}{\sqrt{2q(1-q)(1+f_1)q(1-q)}} \sqrt{\frac{1+f_1}{2}} \qquad (59)$$

If f is to be greater than zero, the genotypes of sire and dam must be correlated. In the absence of selection this will result only when they are related, i.e. when they have one or more common ancestors or when one is an ancestor of the other. This will result in a path connecting G_1 and G_2 which passes through the individual common to both sides of the pedigree. For example, consider the situation depicted in figure 2.

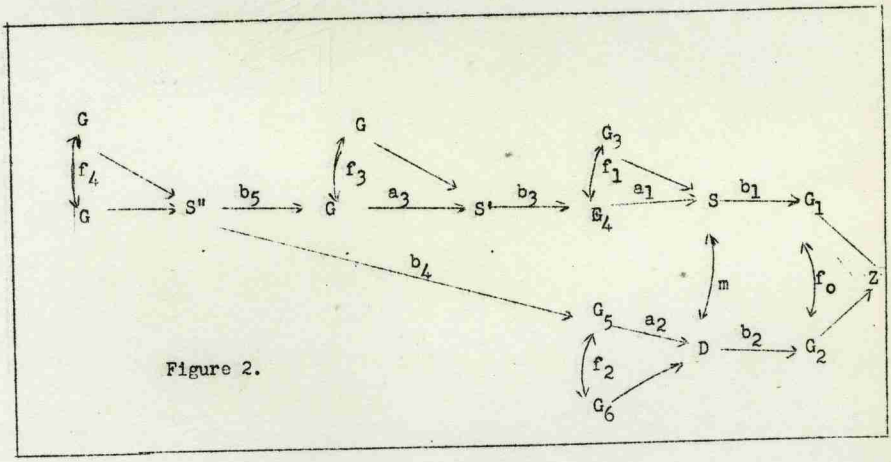


Figure 2.

$$b_1 = \sqrt{\frac{1+f_1}{2}} \qquad a_1 = \sqrt{\frac{1}{2(1+f_1)}}$$

$$b_2 = \sqrt{\frac{1+f_2}{2}} \qquad a_2 = \sqrt{\frac{1}{2(1+f_2)}}$$

$$b_3 = \sqrt{\frac{1+f_3}{2}} \qquad a_3 = \sqrt{\frac{1}{2(1+f_3)}}$$

$$b_4 = b_5 = \sqrt{\frac{1+f_4}{2}}$$

$$f_0 = a_1 b_1 a_3 b_3 b_5 b_4 a_2 b_2$$

$$a_1 b_1 = \sqrt{\frac{1+f_1}{2}} \sqrt{\frac{1}{2(1+f_1)}} = \frac{1}{2}$$

$$a_3 b_3 = \sqrt{\frac{1+f_3}{2}} \sqrt{\frac{1}{2(1+f_3)}} = \frac{1}{2}$$

$$a_2 b_2 = \sqrt{\frac{1+f_2}{2}} \sqrt{\frac{1}{2(1+f_2)}} = \frac{1}{2}$$

$$b_4 b_5 = \frac{1+f_4}{2}$$

$$f_0 = \left(\frac{1}{2}\right)^4 (1+f_4) \qquad (60)$$

If f_o is the inbreeding coefficient of the individual z, f_4 must be the inbreeding coefficient of S". Consideration of the source of the exponent of $\frac{1}{2}$ in eq. (60) will show that in general this exponent will have the value $n+n_1+1$ where n is the number of generations by which the common ancestor is removed from the sire and n_1 is the number of generations by which the common ancestor is removed from the dam.

If there is more than one common ancestor or one common ancestor appears more than once in the ancestry of either the sire or dam of the individual for which the inbreeding coefficient is being computed, there will be more than one path connecting G_1 and G_2. In accordance

with the principle developed in the preceding section, F_0 (the correlation between G_1 and G_2) will be the sum of the contributions of all paths, and

$$f_0 = \Sigma \left[\left(\frac{1}{2}\right)^{n+n_1+1} (1+f_A) \right] \tag{61}$$

where f_A is the inbreeding of the common ancestor in question and Σ indicates summation for all paths connecting G_1 and G_2 .

It should be noted that

$$\begin{aligned} m &= f_0 \,/\, b_1 b_2 \\ &= \Sigma \left[\left(\frac{1}{2}\right)^{n+n_1+1} (1+f_A) \right] \Big/ \sqrt{\frac{1+f_s}{2}} \sqrt{\frac{1+f_d}{2}} \\ &= 2\Sigma \left[\left(\frac{1}{2}\right)^{n+n_1+1} (1+f_A) \right] \Big/ \sqrt{(1+f_s)(1+f_d)} \\ &= \Sigma \left[\left(\frac{1}{2}\right)^{n+n_1} (1+f_A) \right] \Big/ \sqrt{(1+f_s)(1+f_d)} \end{aligned} \tag{62}$$

where f_s = inbreeding of the individual S and f_d = inbreeding of the individual D. This correlation is the so-called Coefficient of Relationship of Wright (10). Note there is nothing about it that restricts its use to individuals that have been or are to be mated together or to individuals of opposite sex.

> Note: The correlations between uniting gametes and between parents genotype and gametes produced were computed with reference to only a single gene pair. This means the inbreeding coefficient refers to the decrease in proportion of heterozygotes for a single gene pair in the entire population. But if it gives the decrease to be expected with respect to one gene pair, it gives it for any gene pair and hence for all.

XIII. Special formulae for the coefficient of inbreeding where a systematic scheme of inbreeding is followed.

Wright (9) gives a number of special formulae for use in computing the inbreeding coefficient resulting from the systematic use of certain inbreeding systems. The derivation of one will be considered. Figure 3 depicts the situation involved in continuous sibbing.

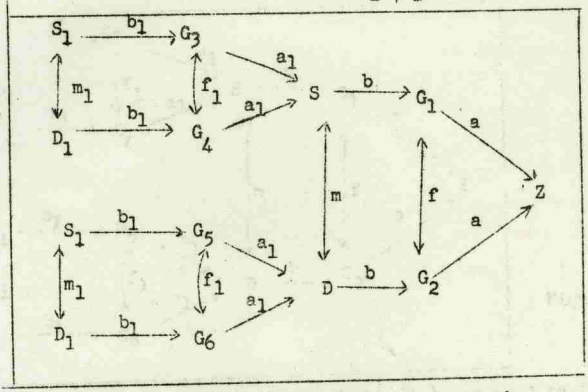


Figure 3.

Subscripts (except in the case of the G's) are used to indicate the generation for which a value applies, thus f is the inbreeding of the most recent generation, f_1 of the one just preceding it, f_2 of the generation before that, etc. Remember that while S_1 is shown twice in the figure, the same individual is indicated in each case. Thus, for example, $a_1^2 b_1^2$ is the contribution of the path S - S_1 - D to the correlation m. The same applies for D.

$$m = a_1^2 b_1^2 + a_1^2 b_1^2 + a_1^2 b_1^2 m_1 + a_1^2 b_1^2 m_1$$

$$= 2a_1^2 b_1^2 (1+m_1) = 2a_1^2 (b_1^2 + b_1^2 m_1) \qquad (63)$$

$$f_1 = b_1^2 m_1 \qquad (64)$$

$$a_1^2 = \frac{1}{2(1+f_1)} \qquad \text{(from eq. 58)} \qquad (65)$$

$$b_1^2 = \frac{1+f_2}{2} \qquad \text{(from eq. 59)} \qquad (66)$$

Substituting in eq. (63) using eqs. (64,65 and 66) we obtain

$$m = 2 \cdot \frac{1}{2(1+f_1)} \left[\frac{1+f_2}{2} + f_1\right]$$

$$= \frac{1 + 2f_1 + f_2}{2(1+f_1)}$$

Now

$$f = mb^2 \tag{57}$$

and $$b^2 = \frac{1 + f_1}{2} \tag{59}$$

Hence $$f = \frac{1+2f_1 + f_2}{2(1+f_1)} \cdot \frac{1 + f_1}{2} = (1 + 2f_1 + f_2)/4 \tag{67}$$

Applying this formula we find the inbreeding for successive generations of sib mating to be as follows:

Generation	f
1	.25
2	.375
3	.500
4	.594
5	.672
6	.734
7	.785
8	.826

Other formula of this type are as follows:

System	f	Limit of f
1. Random sire with daughters, granddaughters, etc.	$(1+2f_1)/4$	.50
2. Contimuous backcrossing to homozygous parent.	$(1+f_1)/2$	1.00
3. Offspring with younger parent.	$(1+2f_1+f_2)/4$	1.00
4. Double first cousins.	$(4f_1+2f_2+f_3+1)/8$	1.00
5. One male with large number of half sisters (half sisters of each other).	$(6f_1+f_2+1)/8$	1.00
6. Half first cousins.	$(1+4f_2+f_3)/32$	.037

Limit of f is the highest level of inbreeding obtainable by following the system of mating indefinitely. It is found by substituting f for f_1, f_2, etc. and solving for f. Thus, in the first formula, this gives

$$f = (1+2f)/4$$

$$4f = 1+2f$$

$$2f = 1, \quad f = .50 \quad \text{(the limit of f)}$$

Formulae for still other systems are given by Wright (9). System 5 is of special interest in animal breeding because it represents the most intense type of inbreeding that does not break a herd up into non-interbreeding lines.

A useful approximate formula (Wright 11) for the percent of remaining heterozygosis which is eliminated in each generation when mating is at random within a population of limited size is

$\frac{1}{8M} + \frac{1}{8F}$ where M is the number of males and F the number of females in the breeding population. For a breeding population of 2 males and 7 females, this would equal 1/16 + 1/56 = .08 and inbreeding would result approximately as follows:

Generation of Inbreeding	f
1	.08
2	.15
3	.22
4	.28
5	.34

References:

11. Wright, Sewall (1931) Evolution in Mendelian Populations Genetics 16: 97-159.

STATISTICAL CONCEPTS IN GENETICS

II. Statistical formulae

A small group of population parameters (the mean, variance, linear regression coefficient, and correlation coefficient) will be used repeatedly. The student should have at his command the basic formulae involving them. The most important of such formulae are given below together with some attention to derivation. The derivations given should receive the students' attention because they exemplify sorts of algebraic manipulation to be used extensively in the material that follows. Some of the formulae are given in several algebraic forms; the student will find it helpful to have an easy familiarity with all of them.

A. Variance

Consider a population of which the individual members are designated as

$$1, 2, 3, \cdots, N$$

and their magnitudes in some measured character are

$$X_1, X_2, X_3, \cdots, X_N$$

The arithmetic mean ($\bar{X}$) is

$$\frac{X_1 + X_2 + \cdots X_N}{N}$$

The variance (σ^2) is, by definition,

$$\sigma_X^2 = \frac{S(X - \bar{X})^2}{N} = \frac{S(X^2) - \frac{(SX)^2}{N}}{N} \qquad (1)$$

Let $x_1 = X_1 - \bar{X}$, $x_2 = X_2 - \bar{X}$, etc.

Then, $\sigma^2 = \frac{Sx^2}{N}$ (2)

Small case letters will be used throughout to symbolize deviations of a variable from its mean. Other uses will, of course, be made of the lower case letters but when a variable is symbolized b a specified letter the corresponding lower case letter will be reserved for the deviation of that variable from its mean.

B. The effects of coding on mean and variance.

1. Let the measured values of the above population be coded by subtraction of a constant value, a, from each. This yields a new population of values,

$$X_1 - a,\ X_2 - a,\ X_3 - a,\ \ldots$$

The mean is

$$\overline{X - a} = \frac{X_1 - a + X_2 - a \cdots X_N - a}{N} = \frac{\sum X - Na}{N} = \bar{X} - a \qquad (3)$$

The variance is

$$\sigma^2_{(X-a)} = \frac{S(X - a - \bar{X} + a)^2}{N} = \frac{S(X - \bar{X})^2}{N} = \sigma^2_X \qquad (4)$$

Note that the coding changes the mean in the same way that it changes each individual but does not affect the variance. Obviously coding by addition would have similar consequences.

2. Next, consider coding by multiplication. Let each value of the original population be multiplied by a constant, c. We then have the values

$$cX_1,\ cX_2,\ cX_3\ \ldots$$

Their mean is

$$\overline{cX} = \frac{cX_1 + cX_2 + \ldots cX_N}{N} = c\bar{X} \qquad (5)$$

Their variance is

$$\sigma^2_{cX} = \frac{S(cX - c\bar{X})^2}{N} = \frac{c^2 S(X - \bar{X})^2}{N} = c^2 \sigma^2_X \qquad (6)$$

Again the mean is affected in the same manner as the individuals. Clearly the above covers coding by division since c may take a fractional value such as 1/10 and multiplication by 1/10 is the equivalent of division by 10.

C. The correlation coefficient (ρ) and the regression coefficient (β).

Consider two populations

$X_1, X_2, X_3, \cdots$ with mean $\bar{X}$, and

$Y_1, Y_2, Y_3, \cdots$ with mean $\bar{Y}$.

By definition the correlation coefficient is

$$\rho = \frac{S(X - \bar{X})(Y - \bar{Y})}{\sqrt{S(X - \bar{X})^2 \cdot S(Y - \bar{Y})^2}} = \frac{Sxy}{\sqrt{Sx^2 \cdot Sy^2}} = \frac{Sxy/N}{\sqrt{\frac{Sx^2}{N} \cdot \frac{Sy^2}{N}}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad (7)$$

where $\sigma_{xy} = \sigma_{XY}$ = Covariance of X and Y.

$$\beta_{Y \cdot X} = \frac{S(X - \bar{X})(Y - \bar{Y})}{S(X - \bar{X})^2} = \frac{S_{xy}}{Sx^2} = \frac{\sigma_{xy}}{\sigma_x^2} \quad (8)$$

$$\beta_{X \cdot Y} = \frac{S(X - \bar{X})(Y - \bar{Y})}{S(Y - \bar{Y})^2} = \frac{S_{xy}}{Sy^2} = \frac{\sigma_{xy}}{\sigma_y^2} \quad (8a)$$

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \sqrt{\frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2}} = \sqrt{\beta_{Y \cdot X} \cdot \beta_{X \cdot Y}} \quad (9)$$

The equation of the regression (Y on X) model is

$$Y = \bar{Y} + \beta x + e$$

or subtracting $\bar{Y}$ from each side of the equation

$$y = \beta x + e$$

where e is a random error in Y not correlated with variation in X.

Rearranging

$$e = y - \beta x$$

$$Se^2 = S(y - \beta x)^2 = S(y^2 - 2\beta xy + \beta^2 x^2)$$

$$= Sy^2 - 2\beta Sxy + \beta^2 Sx^2 = Sy^2 - \frac{2Sxy \cdot Sxy}{Sx^2} + \frac{Sxy \cdot Sxy \cdot Sx^2}{Sx^2 \cdot Sx^2}$$

$$= Sy^2 - (Sxy)^2/Sx^2 \qquad (10)$$

Se^2 is, of course, the sum of squares of the deviations of Y from regression on X. It should be noted that if any value other than Sxy/Sx^2 were assigned β, Se^2 would be increased, i.e. β satisfied the least squares criterion.

The equation for the regression line is

$$\hat{Y} = \bar{Y} + \beta x$$

where $\hat{Y}$ is the predicted value of Y for any value of x inserted.

$$\beta x = Y - \bar{Y}$$

symbolizes the deviations of predicted values from the mean of Y and the sum of squares of such deviations is what is called the regression sum of squares

$$S(\beta x)^2 = \beta^2 Sx^2 = \frac{(Sxy)^2}{(Sx^2)^2} \cdot Sx^2 = \frac{(Sxy)^2}{Sx^2} \qquad (11)$$

From (10) and (11)

$$Se^2 + S(\beta x)^2 = Sy^2 - \frac{(Sxy)^2}{Sx^2} + \frac{(Sxy)^2}{Sx^2} = Sy^2 \qquad (12)$$

or dividing through by N

$$\sigma_e^2 + \sigma_{\beta x}^2 = \sigma_y^2 \qquad (13)$$

σ_e^2 is frequently symbolized as $\sigma_{y \cdot x}^2$. From (12) and (13) we see that either the sum of squares or the variance of Y is divisible into two parts (1) that due to regression, and (2) that due to deviation from regression.

D. Variance of sums, differences and means.

Consider two populations

$A_1, A_2, A_3 \cdots$ with mean $\bar{A}$, and

$B_1, B_2, B_3 \cdots$ with mean $\bar{B}$.

$$\sigma_A^2 = \frac{Sa^2}{N}, \quad \sigma_B^2 = \frac{Sb^2}{N}$$

Now let a third population be formed as follows

$$C_1 = A_1 + B_1, \; C_2 = A_2 + B_2, \text{ etc.}$$

$$\bar{C} = \frac{A_1 + B_1 + A_2 + B_2 \cdots}{N} = \bar{A} + \bar{B}$$

$$c_1 = C_1 - \bar{C} = A_1 + B_1 - \bar{A} - \bar{B} = a_1 + b_1$$

$$c_2 = C_2 - \bar{C} = A_2 + B_2 - \bar{A} - \bar{B} = a_2 + b_2$$

etc.

$$\sigma_C^2 = \frac{S_c^2}{N} = \frac{S(a+b)^2}{N} = \frac{Sa^2 + 2Sab + Sb^2}{N}$$

$$= \frac{Sa^2}{N} + \frac{Sb^2}{N} + \frac{2Sab}{N} = \sigma_A^2 + \sigma_B^2 + 2\rho_{AB}\sigma_A\sigma_B \qquad (14)$$

since from equation (7)

$$\frac{Sab}{N} = 2\rho_{AB}\sigma_A\sigma_B .$$

The reader can easily verify that if the C's are taken as the differences between A's and B's rather than as the sums,

$$\sigma_C^2 = \sigma_A^2 + \sigma_B^2 - 2\rho_{AB}\sigma_A\sigma_B$$

In either case if $\rho_{AB} = 0$, i.e. A and B are uncorrelated

$$\sigma_C^2 = \sigma_A^2 + \sigma_B^2$$

When the members of two such populations are paired at random ρ will always be zero. There will be many instances important to us in which this will be the case. Equation (13) is true because there is no correlation between X and e.

The above can easily be extended to sums (or differences) involving any number of variables. Thus

$$\sigma^2_{(A \pm B \pm C \pm D \cdots)} = \sigma_A^2 + \sigma_B^2 + \sigma_C^2 + \sigma_D^2 \cdots \qquad (15)$$

provided none of the variabiles are correlated. Or

$$\sigma^2_{(X_1 + X_2 + X_3 \cdots X_N)} = \sigma^2_{X_1} + \sigma^2_{X_2} + \sigma^2_{X_3} + \cdots \sigma^2_{X_N}$$

If $X_1, X_2, \cdots X_N$ are all drawn from the same population

$$\sigma^2_{X_1} = \sigma^2_{X_2} = \cdots = \sigma^2_{X_N} = \sigma^2$$

and

$$\sigma^2_{(X_1 + X_2 + X_3 \cdots X_N)} = N\sigma^2$$

Now if this sum, $(X_1 + X_2 + X_3 \cdots X_N)$, is divided by N to obtain a mean, $\bar{X}$, applying (6) we find

$$\sigma^2_{\bar{X}} = \frac{1}{N^2} N\sigma^2 = \frac{\sigma^2}{N} \qquad (16)$$

the formula for the variance of a mean.

Note that all of the foregoing has dealt with population parameters rather than sample statistics. This is the reason, for example, why the denominator in the variance formula is N rather than N - 1. It also explains why equations such as (15) involving variances of sums and differences can be given as strict equalities rather than as approximations. The reason for approaching the subject in this way is that the parameters (the population values of the mean, variance, etc.) are the "expected" values of the analagous statistics in sample data. Much of our time will be spent in the derivation of "expected" values and for that purpose the above forms will be appropriate. Obviously, when expectations are to be checked against observations or parameters are to be estimated using sample data the appropriate sample statistics must be used as "estimators" of the parameters. For example, if s^2 is to be an unbiased estimate of σ^2 it must be computed from the sample data as $\frac{S(X - \bar{X})^2}{N - 1}$ rather than as $\frac{S(X - \bar{X})^2}{N}$.

Problems:

1. Give the variance of $aX + \frac{Y}{b} - Z$ in terms of the variances of X, Y, and Z assuming that a and b are constants and that X, Y, and Z are uncorrelated.
2. Give the variance of $(\bar{X} + \bar{Y})$ assuming the correlation (ρ_{xy}) to be .5, σ_x^2 to be 10, σ_y^2 to be 20, and N to be 12.
3. If $X = A + b_1 + b_2 + \cdots b_r + e_1 + e_2 + \cdots e_n$ and there are no correlations among the summed values, what is the variance of X? Consider the b's as randomly drawn from one population and the e's as randomly drawn from another. What is the variance of X/n?
4. Given $\sigma_C^2 = 64$, $\sigma_D^2 = 100$, and $\sigma_{CD} = 32$. What is the variance of $rC + \frac{D}{t}$ where r and t are constants?

Genotypic value — phenotypic

100 heifers, dairy records
10 best selected
Selection always on phenotype

Model

	E_1	E_2	$\cdots$	E_m ← environments	mean
G_1	$p_{11} = g_1 + e_{11}$	p_{12}		p_{1m}	g_1
G_2	p_{21}	p_{22}		p_{2m}	g_2
⋮					
G_n	p_{n1}			p_{nm}	g_n
mean	e_1	e_2		e_n	0

P = phenotype

$P_{ij} - \bar{P} = p_{ij}$ j environ. i genotype

$$g_1 = \frac{p_{11} + p_{12} + \cdots p_{1m}}{m} = \frac{\sum_{j=1}^{m} p_{1j}}{m} = \text{genotypic effect of } G_1$$

$$g_i = \frac{p_{i1} + p_{i2} + \cdots p_{im}}{m} = \frac{\sum_{j=1}^{m} p_{ij}}{m} = \text{effect of the } i \text{ th genot.}$$

$$\frac{\sum_{i=1}^{n} p_{ij}}{n} = \text{effect of the } E_j$$

$p_{ij} = g_i + e_j + \ell_{ij}$ interaction effect.

$p_{ijk} = \cdots + z_{ijk}$ measurement errors

show that there are no r between these four.

$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + \sigma_i^2 + \sigma_z^2$ variance

$p = g + e$

regression problem

$(P_s - \bar{P}) B_{gp} (=) g_s - \bar{g}$

$$B_{gf} = \frac{\sigma_{gp}}{\sigma_p^2}$$

$$\sigma_{gp} = \frac{\sum gp}{N} = \sum \frac{g(g+e)}{N} = \frac{\sum g^2}{N} + \frac{\sum ge}{N}$$

III. The genotypic improvement resulting from selection.

The value of a genotype may be measured either in terms of (1) its effect on the phenotype of the individual which possesses it, or (2) the mean genotype of progeny of that individual. For illustration, consider a pair of genes (A,a) of which the gene A is completely dominant to its allele. Measured in terms of their effects on the phenotype of their possessors, the genotypes AA and Aa have equal value while by the second criterion of measurement (the progeny) their values are obviously not equal. The value of a genotype with respect to the phenotype of its possessor will be referred to henceforward as the genotypic value or the value of the genotype. The deviations of such values from their population mean will be termed the effect of the genotype. One individual will be said to be genotypically superior to another if the value or effect of its genotype is greater (this assumes that the method of measurement is such that high values are more desirable than low ones).

We shall be concerned first with the effect of selection upon the difference in genotypic value (as defined above) between selected individuals (or lines, varieties, hybrids, etc.) and the mean of the population or sample from which selections were made, i.e., the genotypic superiority of selected individuals. The effect of selection upon the mean genotypic value in later generations is a more complex problem, consideration of which must be postponed until further groundwork has been laid. However, as will be shown later, the gain reflected in later generations is proportional to the genotypic superiority of the selected individuals.

Selection is based on the phenotype. The phenotypic expression of any characteristic of an individual, e.g., height at a specified age, is the resultant of the genotype of the individual and the environment in which it develops. The relative

importance of heredity and environment as sources of phenotypic variation is known to vary greatly for different characteristics. Traits such as color are in general almost completely under genetic control though there are classical exceptions. The so-called quantitative characters are for the most part much more responsive to environmental variation.

We recognize intuitively that given a certain amount of genetic variation selection will increase in effectiveness as variation in phenotype from environmental sources decreases. It would appear that control of environment should offer a means for making selection more effective and some effort in this direction is productive. However, it is essential to recognize that a completely uniform environment for all individuals of a group is an abstraction, never an actuality. The environments (defined to encompass all non-genetic variables that affect phenotype) of plants vary even though the plants are growing adjacent to each other and the environments of animals vary even though all are handled as nearly alike as is humanly possible. A little consideration of soil variation, competition, the random distribution of parasitic and pathogenic organisms, accidents of various and subtle sorts, etc. will suggest many uncontrollable sources of environmental variation.

As a preliminary exercise to gain familiarity with the meaning of certain terms and notations consider a population of genotypes

$$G_1, G_2, \cdots, G_N$$

of any plant or animal, and a population of environments

$$E_1, E_2, \cdots, E_M .$$

Assume that one individual of each genotype is raised in each of the M environments. Symbolize by P the measure of any characteristic such as height and assume that the measurement is not subject to error. Let P_{ij} be height of the individual with the

ith genotype raised in the jth environment. Thus, for example, P_{23} will be the height of the individual with the genotype G_2 raised in the environment E_3. Let $\bar{P}$ be the mean height of the NM individuals, and

$$P_{ij} - \bar{P} = p_{ij} \tag{17}$$

Thus

$$P_{11} - \bar{P} = p_{11}$$

$$P_{27} - \bar{P} = p_{27}$$

etc.

Then

$$\frac{P_{11} + P_{12} + \cdots P_{1M}}{M} = g_1,$$

the effect of the genotype G_1 for the population of environments involved. In general

$$\frac{P_{i1} + P_{i2} + \cdots P_{iM}}{M} = g_i , \tag{18}$$

the effect of the ith genotype, and

$$\frac{P_{1j} + P_{2j} + \cdots P_{Nj}}{N} = e_j, \tag{19}$$

the effect of the jth environment for the population of genotypes involved.

Note that the effect of a genotype is defined in terms of the average, for some specified population of environments, of the associated phenotypes. In any practical selection problem it is important that the population of environments be delineated. For example, in selection among corn hybrids it might be composed of those environments in which corn is raised in the Coastal Plain area of North Carolina. In like manner the effect of an environment is defined in terms of the average phenotype, for a population of genotypes, of individuals raised in that environment.

Now $P_{11} - \bar{P}$ is not necessarily equal to the sum of g_1 and e_1. The phenotypic response to a given variation in genotype may not be the same in all environments;

individuals differing in genotype may respond differently to a specific variation in environment. Is it the response to genotype or to environment which varies? We cannot distinguish and resolve the situation by saying that there are genotype-environment interactions. Let

$$P_{11} - \bar{P} = g_1 + e_1 + i_{11} \qquad \text{or} \qquad i_{11} = P_{11} - g_1 - e_1$$

$$P_{23} - \bar{P} = g_2 + e_3 + i_{23} \qquad \text{or} \qquad i_{23} = P_{23} - g_2 - e_3$$

etc.

You will note that the i (the interaction term) is the amount by which the deviation of the phenotype from the general mean fails to be the sum of the average deviations for the genotype and environment involved. In general

$$P_{ij} = P_{ij} - \bar{P} = g_i + e_j + i_{ij} \qquad (20)$$

which is our mathematical model for phenotype. For practical purposes it is not quite complete since it does not recognize errors of measurement which are always involved in practical situations. Such errors are recognized by addition of another term to the model.

$$P_{ijk} = g_i + e_j + i_{ij} + z_{ijk} \qquad (21)$$

The third subscript is necessary because more than one individual of a given genotype might theoretically be raised in any given environment. Thus P_{ijk} specifies the kth individual of the ith genotype raised in the jth environment, and z_{ijk} is a random error in measurement of that individual's phenotype. The phenotypic variance (σ_p^2) for the population specified is

$$\sigma_p^2 = \sigma_g^2 + \sigma_e^2 + \sigma_i^2 + \sigma_z^2 \qquad \text{[see (12), II]} \qquad (22)$$

since the definitions of g, e, i, and z were such that there can be no correlations among them. The reasons for considering this hypothetical population were (1) to

establish definitions of genotypic effect, environmental effect, and genotype-environment interaction, and (2) the population described is of the type from which the material with which we deal in practical problems may frequently be considered a sample.

In certain instances, however, the parent population cannot be assumed to be of the type considered above. The most important thing to be on guard against is lack of independence in the distribution of genotypes and environments. (The specification that each genotype was to be raised in each environment was one way of describing the situation that would obtain in the population if genotypes and environments were independently distributed.) It is probable that in the case of certain human traits genotypes and environments are not independently distributed. For example, both the IQ of children and the environment in which they are raised are probably correlated to some extent with the IQ of their parents. Again in natural populations which extend over areas which differ in average environment, the distributions are likely to be non-independent since in each area there would be a tendency for the best adapted genotypes to reproduce most rapidly and those best adapted might well differ from area to area. Our primary concern, however, will be with applications to controlled breeding programs with farm animals or agronomic crops with which random distribution can be affected.

For some of our purposes there will be no point in distinguishing environmental effects, genotype environment interactions, and errors of measurement. In those instances the following simplified model will be used.

$$p = g + e$$

in which $\underline{e}$ is the sum of the three non-genetic effects. In other cases, $\underline{e}$ will be defined to include the measurement error, the effect of environment, and a fraction of the genotype-environment interaction with $\underline{i}$ being defined to include a specified component of the interaction. However, the complete model should always be kept in mind as a check on the applicability of any variate version.

The estimation of the genotypic superiority to be expected on the average in selected individuals is a regression problem. We can measure phenotype so if we know the regression of genotype on phenotype we can predict the amount by which phenotypically superior individuals are superior in genotype. Let

$$p = g + e$$

where g is genotypic effect and e is the deviation of g from p due to environmental variation, genotype environment interaction, and measurement error. Then

$$\sigma_{pg} = \frac{\sum pg}{N} = \frac{\sum g(g + e)}{N} = \frac{\sum g^2}{N} + \frac{\sum ge}{N} \qquad (23a)$$

Note the absence of a correction term in the equation. It would be zero since the means of p, g, and e are all zero. If genotypes are distributed at random relative to variations in environment and the measurement error is random, there will be no correlation between g and e, i.e. $\sum$ ge will equal zero. Hence

$$\sigma_{pg} = \frac{\sum g^2}{N} = \sigma_g^2 \qquad (23b)$$

The regression of genotype on phenotype will be

$$\beta_{gp} = \frac{\sigma_{pg}}{\sigma_p^2} = \frac{\sigma_g^2}{\sigma_p^2} \qquad (24)$$

We can now set up the prediction equation for genotypic superiority of selected individuals. Let

P_s be the selection differential (the mean difference in phenotype between selected individuals and the group from which they are selected), and

g_s be the mean genotypic superiority of the selected individuals.

Then

$$g_s \ (=) \quad \rho_{gp} p_s \tag{25}$$

where (=) is used to indicate estimation rather than strict equality in view of sampling error present. If p_s is measured in terms of the phenotypic standard deviation we set

$$p_s = k \sigma_p \tag{26}$$

where k is the number of standard deviations by which selected individuals are phenotypically superior to the entire group. Then

$$g_s \ (=) \quad \rho_{gp} k \sigma_p = k \frac{\sigma_g^2}{\sigma_p^2} \sigma_p = \frac{k \sigma_g^2}{\sigma_p} \tag{25a}$$

This form of the equation has special utility for measuring the advantage gained by reduction of the environmental component of phenotypic variance through control of environmental variation. k is on the average constant when the proportion of total individuals which are selected is constant. Since σ_g^2 is not a function of environmental variation the only term in the equation which will be affected by change in environmental variance is σ_p. Let us consider the effect of variation in σ_p on g_s assuming k constant, i.e. that the selected proportion of the population remains constant. Let $\Delta \sigma_p$ symbolize the amount by which σ_p is changed and Δg_s be the corresponding change in genotypic gain resulting from selection. Then

$$g_s \ (=) \ \frac{k \sigma_g^2}{\sigma_p} \tag{25a}$$

$$g_s + \Delta g_s \ (=) \ \frac{k \sigma_g^2}{\sigma_p + \Delta \sigma_p} \tag{27}$$

and

$$\Delta g_s \; (=) \; \frac{k\sigma_g^2}{\sigma_p + \Delta\sigma_p} - \frac{k\sigma_g^2}{\sigma_p}$$

The expected increment in g_s as a fraction of g_s will be

$$\frac{\Delta g_s}{g_s} = \frac{\Delta g_s \sigma_p}{k\sigma_g^2} = \frac{\sigma_p}{\sigma_p + \Delta\sigma_p} - 1 \qquad (28)$$

In cases where σ_p is reduced (as presumably it would be if environment is made more uniform) $\Delta\sigma_p$ is a negative quantity. Let the change in σ_e^2 resulting from greater control on environment be symbolized by $(a - 1)\sigma_e^2$ where $0 < a < 1.0$. This is the amount of change in σ_p^2. Then

$$\frac{\Delta g_s}{g_s} = \frac{\sigma_p}{\sigma_p + \Delta\sigma_p} - 1 = \sqrt{\frac{\sigma_p^2}{\sigma_p^2 + (a - 1)\sigma_e^2}} - 1 \qquad (29)$$

Substituting $\sigma_p^2 - \sigma_g^2$ for σ_e^2 (since $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$)

$$\frac{\Delta g_s}{g_s} = \sqrt{\frac{\sigma_p^2}{\sigma_p^2 + (a - 1)(\sigma_p^2 - \sigma_g^2)}} - 1 \qquad (29a)$$

If numerator and denominator of the term under the radical are both divided by σ_p^2, we obtain

$$\frac{\Delta g_s}{g_s} = \sqrt{\frac{1}{a - (a - 1)\sigma_g^2/\sigma_p^2}} - 1 \qquad (29b)$$

which is the fraction by which genotypic gain from selection is expected to be increased as a consequence of changing environmental variance by $(a - 1)\sigma_e^2$, i.e. from σ_e^2 to $\sigma_e^2 + (a - 1)\sigma_e^2 = a\sigma_e^2$. Values of this fraction are listed in Table 1. They indicate that when σ_g^2 amounts to as much as one-half of σ_p^2, control on environment is relatively inefficient in increasing the effect.

of selection. Even when σ^2_g is rather low relative to σ^2_p, environmental variance

Table 1. Values of Δg_s as a fraction of g_s for varying values of a and σ^2_g / σ^2_p (from equation 29).

σ^2_g / σ^2_p

a	.1	.3	.5	.7	.9
.9	.048	.037	.026	.015	.005
.7	.170	.125	.085	.048	.015
.5	.348	.240	.154	.085	.026
.3	.644	.400	.240	.125	.037
.1	1.294	.644	.348	.170	.048

must be reduced sharply if the effect of selection is to be increased very much. For example when σ^2_g is one-tenth of σ^2_p reduction of σ^2_e to .7 σ^2_e increases the expected genotype gain from selection by only 17%.

The expected value of k.

When the data on which selection is based are available the selection differential, p_s, can always be obtained directly. However, in theoretical problems the expected magnitude of k, the selection differential in units of the population phenotypic variance, is often required.

If the sample from which selections are to be made is of size 50 or less, expected k may be obtained from Table XX of Fisher and Yates (1). This table gives, for samples from N = 2 to N = 50, the number of standard deviations by which the 1st, 2nd, • • • Nth individuals may on the average be expected to deviate from the mean if the individuals of the sample are placed in rank order. For example, if we are going to take the best 4 of a sample of 40, we find from the table that on the average they will deviate 2.16, 1.75, 1.52, and 1.34

standard deviations from the mean, respectively. The average of these figures, 1.69, is taken as the expected value of k. Assumptions involved are that the sample can be considered as randomly drawn from a parent population of the "normal" form.

If selection is to be from samples of more than 50, the expected value of k can be obtained from the attributes of the normal curve as z/w where w is the proportion selected and z is the height of the ordinate which divides the area under the "normal" curve into portions relative in magnitude to the proportions selected and rejected. The value of z can be obtained using Tables I and II of Fisher and Yates (1) or from the table in the Handbook of Chemistry and Physics which relates to the "normal" curve. If the tables of Fisher and Yates are used, Table I is entered with P equal to either 2w or 2(1 - w) whichever is 1.0 or less. (The factor, 2, is introduced since Table I gives the relative deivate, x, beyond which a given proportion of the population, P, is found when both tails of the curve are considered; we are interested in only the positive tail of the curve.) Table II is then entered with the x obtained from Table I to find z.

Example: 20% of the sample is to be selected, i.e. $w = .2$, $P = 2w = .4$, x for $P = .4$ is .8416 (from Table I), z for $x = .8416$ is .2799 (from Table II).

$$k = z/w = .2799/.2 = 1.4$$

If the table from the Handbook of Chemistry and Physics is used, find the value $.5 - w$ (or if $w > .5$, the value $w - .5$) in the column headed "Area" and read the value of z from the next column to the right which is headed "Ordinate."

There will be a slight upward bias in k taken as z/w since this assumes the sample from which selection is made to be of size approaching infinity. However, the magnitude of the bias will be unimportant unless the proportion to be selected is very small.

Problems:

5. Obtain the expected values of $\underline{k}$ for selection of .05 and .3 of samples of size 20, 30, 40, and 50 using Table XX of Fisher and Yates.

6. Obtain the expected values of $\underline{k}$ for selection of .05 and .3 using Tables I and II of Fisher and Yates or the table in the Handbook of Chemistry and Physics.

7. Assume σ_p^2 for annual egg production of pullets is 625. What is the expected selection differential for the top 15 from a random sample of 100? for the top 3 from a sample of 20?

8. Given $\sigma_g^2 / \sigma_p^2 = .4$ and $\sigma_p^2 = 100$, what genotypic gain will be expected from selection of the best half of samples of size 6, 12, 24, and 48? Obtain these same values assuming σ_e^2 is reduced to $.8\,\sigma_e^2$, $.6\,\sigma_e^2$, and $.4\,\sigma_e^2$.

References:

1. Fisher, R. A. and F. Yates (1938) Statistical Tables for Biological, Agricultural, and Medical Research. Oliver and Boyd. London and Edinburgh.

2. Handbook of Chemistry and Physics, Chemical Rubber Publishing Co., Cleveland, Ohio.

IV. (con'd)

C. Cases where each genotype may be represented by more than one individual but where the individual expresses the character only once.

Individuals of the same genotype are possible with pure lines or with material that can be asexually propagated. Thus these sorts of material belong in this category unless the characters are expressed more than once as in perennials. Specific examples of traits in this group are the characters of inbred lines of corn or single cross corn hybrids, and of varieties of such normally self-fertilized crops as oats and tobacco. Sweet and irish potatoes are examples from among the horticultural crops.

Precision in the comparison of genotypes can be much greater with material of this type. Because large numbers of individuals of each genotype can be produced, the error of comparison can be reduced by replication. However, important questions arise concerning the optimum amount of material to use in genotype comparisons and the optimum distribution of the material relative to the population of environments for which we wish to evaluate the genotypes.

As a concrete example of the practical problem consider the evaluation of strains of oats for use in the Piedmont area of North Carolina. The population of environments involved are those which occur in that area. Specific sources of variation in environment that are recognizable are soil variability within the area and variation in the climate complex from year to year or from one part of the area to another in the same year. It is quite obvious that selection among varieties based on data collected in one year and at one location could be quite ineffective with respect to genotypic value for the population of Piedmont environments since only one of the soils and one of the "climates" would be involved. We recognize that the evaluation of the genotypes should be made on more soils and in more "climates". How can we decide such practical questions as how many locations and years should be involved in a variety trial and how many replications should be

used at each location in each year.

The mathematical model is more involved in this case since we wish to subdivide the total effect of environment into portions arising from several specific sources. Let

$$P_{ijklm} = g_i + a_j + b_k + c_{ij} + d_{ik} + f_{jk} + h_{ijk} + u_{jkl} + v_{ijk\ell} + z_{ijk\ell m} \qquad (48)$$

where

P_{ijklm} is the deviation from mean phenotype for the mth individual of a plot of the ith genotype in the lth replication at the jth location in the kth year.

g_i = the effect of the ith genotype

a_j = the effect of the jth location

b_k = the effect of the kth year

c_{ij} = the interaction of the ith genotype and the jth location

d_{ik} = the interaction of the ith genotype and the kth year

f_{jk} = the interaction of the jth location and the kth year

h_{ijk} = the second order interaction between the ith genotype, the jth location and the kth year.

u_{jkl} = the effect of the lth replication in the jth location and the kth year

v_{ijkl} = the effect of the plot on which the ith genotype is raised in the jth location, the kth year, and the ℓth replication.

z_{ijklm} = measurement error plus the effect of intra-plot environmental variation for the individual in question.

Of course, some of these effects cancel out in the comparison of genotypes if each genotype is represented by the same number of individuals in each replication, location, and year. Suppose each genotype is represented by a plot of n individuals in each of r replications at each of s locations in each of t years. Then

$$Sp_1 = nrstg_1 + nrt\sum_{j=1}^{s} a_j + nrs\sum_{k=1}^{t} b_k + nrt\sum_{j=1}^{s} c_{1j} + nrs\sum_{k=1}^{t} d_{1k}$$

$$+ nr\sum_{j=1}^{s}\sum_{k=1}^{t} f_{jk} + nr\sum_{j=1}^{s}\sum_{k=1}^{t} h_{1jk} + n\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r} u_{jk\ell}$$

$$+ n\sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r} v_{1jk\ell} + \sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r}\sum_{m=1}^{n} z_{1jk\ell m} \qquad (49)$$

The sum for any other genotype will be the same except for substitution of the subscript for that genotype wherever the numeral one appears as a subscript in the above expression. This leaves unchanged the terms involving the year effects, the location effects, the interaction of year and location, and the replication effects; that is, those terms are constants in the sums for the various genotypes. Being constants they will not contribute to the variance of the phenotypic means and will be omitted from the equation for the mean, $\bar{P}_1$.

$$\bar{P}_1 = \frac{Sp_1}{nrst} = g_1 + \sum_{j=1}^{s}(c_{1j})/s + \sum_{k=1}^{t}(d_{1k})/t + \sum_{j=1}^{s}\sum_{k=1}^{t}(h_{1jk})/st$$

$$+ \sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r}(v_{1jk\ell})/rst + \sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r}\sum_{m=1}^{n}(z_{1jk\ell m})/nrst \qquad (49a)$$

The variance of the mean phenotype for a genotype will be (see equations 6 and 15, section II).

$$\sigma^2_{\bar{p}} = \sigma^2_g + \frac{s\sigma^2_c}{s^2} + \frac{t\sigma^2_d}{t^2} + \frac{st\sigma^2_h}{s^2t^2} + \frac{str\sigma^2_v}{s^2t^2r^2} + \frac{strn\sigma^2_z}{s^2t^2r^2n^2}$$

$$= \sigma^2_g + \frac{\sigma^2_c}{s} + \frac{\sigma^2_d}{t} + \frac{\sigma^2_h}{st} + \frac{\sigma^2_v}{str} + \frac{\sigma^2_z}{strn} \qquad (50)$$

where

σ^2_g = genotypic variance

σ^2_c = variance due to interaction of genotype and location

σ^2_d = variance due to interaction of genotype and year

σ^2_h = variance due to interaction of genotype, location, and year

σ^2_v = variance due to environmental variation between plots in the same replication, and

σ^2_z = intra-plot variance between individuals of the same genotype.

If means are computed on the basis of plots as is common where the number of plants per plot is large, we have $n\bar{p}$, the variance of which is

$$\sigma^2_{n\bar{p}} = n^2\sigma^2_g + \frac{n^2\sigma^2_a}{s} + \frac{n^2\sigma^2_d}{t} + \frac{n^2\sigma^2_h}{st} + \frac{n^2\sigma^2_v}{str} + \frac{n^2\sigma^2_z}{strn} \qquad (50a)$$

what is ordinarily called the experimental error or strictly speaking the within replication plot variance is in our notation $n^2\sigma^2_v + n\sigma^2_z$.

If as before we write

$$\bar{p} = g + e$$

then

$$\sigma^2_{\bar{p}} = \sigma^2_g + \sigma^2_e$$

and

$$\sigma^2_{n\bar{p}} = n^2\sigma^2_g + n^2\sigma^2_e$$

where

$$n^2\sigma^2_c = \frac{n^2\sigma^2_c}{s} + \frac{n^2\sigma^2_d}{t} + \frac{n^2\sigma^2_h}{st} + \frac{n^2\sigma^2_v}{str} + \frac{n^2\sigma^2_z}{strn} \qquad (50b)$$

is the non-genotypic, and $n^2\sigma^2_g$ the genotypic portion of σ^2_{np}. Clearly increased replication will not by itself be a very good method for decreasing $n^2\sigma^2_c$ if interactions of genotypes with locations and years are very great. What is needed in that case is to increase s and t, the numbers of locations and years at which the genotypes are compared. As a basis for deciding the optimum ratios between s, t, and r, estimates of the several variance components are required. They can be obtained from the data of variety comparisons conducted in several locations in more than one year. Assume as above that each genotype (variety) has been replicated r times at each of s locations in each of t years and that there were n individuals of the variety per plot. The analysis of variance would be as follows.

Source of Variation	d.f.	M.S.	Expectation of M.S.
Locations	s - 1		
Years	t - 1		
Locations x Years	(s - 1)(t - 1)		
Reps in locations and years	st(r - 1)		
Varieties	G - 1	M_1	$n^2(\sigma^2 + r\sigma^2_h + rt\sigma^2_c + rs\sigma^2_d + rts\sigma^2_g)$
Var. x locations	(s - 1)(G - 1)	M_2	$n^2(\sigma^2 + r\sigma^2_h + rt\sigma^2_c)$
Var x years	(t - 1)(G - 1)	M_3	$n^2(\sigma^2 + r\sigma^2_h + rs\sigma^2_d)$
Var x loc x years	(s - 1)(t - 1)(G -1)	M_4	$n^2(\sigma^2 + r\sigma^2_h)$
Var x reps in locations and years	st(r - 1)(G - 1)	M_5	$n^2\sigma^2$
Total	strG - 1		

As used in the expectations of mean squares $n^2\sigma^2$ is equal to $n^2\sigma_v^2 + n\sigma_z^2$ of (50a).

The mean square expectations are appropriate functions of the population variances derived on the basis of the arithmetic operations involved in computation of the mean squares. Let us first consider the variety mean square. The population variance of variety means is given in (50a). Now remember that the mean square is computed as the sample variance of the variety sums divided by str, the number of plot totals that are summed for each variety. The variance of variety sums is $r^2s^2t^2$ times the variance of variety means and dividing it by rst we have from (50a)

$$n^2(rst\sigma_g^2 + rt\sigma_c^2 + rs\sigma_d^2 + r\sigma_h^2 + \sigma_v^2 + \frac{\sigma_z^2}{n})$$

Substituting σ^2 for $\sigma_v^2 + \frac{\sigma_z^2}{n}$ we have the expectation of the variety mean square as given in the analysis of variance.

The variety x location mean square is computed from the variety sums for single locations. As a working procedure for obtaining its expectation, we will (1) write the equation for such sums, (2) subtract all terms involving effects not specific for a given genotype and location, (3) write the expression for the population variance of the remainder, and (4) divide that variance by the number of plots totaled in these sums as is done in computation of the mean squares.

Step (1)

$$Sp_{ij} = rtng_i + rtna_j + rtnc_{ij} + rn\sum_{k=1}^{t} b_k + rn\sum_{k=1}^{t} d_{ik} + rn\sum_{k=1}^{t} f_{jk}$$

$$+ rn\sum_{k=1}^{t} h_{ijk} + n\sum_{k=1}^{t}\sum_{\ell=1}^{r} u_{jk\ell} + n\sum_{k=1}^{t}\sum_{\ell=1}^{r} v_{ijk\ell}$$

$$+ \sum_{k=1}^{t}\sum_{\ell=1}^{r}\sum_{m=1}^{n} z_{ijk\ell m}$$

Step (2)

$$S'p_{ij} = rtnc_{ij} + rn\sum_{k=1}^{t} h_{ijk} + n\sum_{k=1}^{t}\sum_{\ell=1}^{r} v_{ijk\ell} + \sum_{k=1}^{t}\sum_{\ell=1}^{r}\sum_{m=1}^{n} z_{ijk\ell m}$$

Step (3)

$$\text{Variance} = r^2t^2n^2\sigma_c^2 + r^2n^2t\,\sigma_h^2 + n^2tr\,\sigma_v^2 + rtn\,\sigma_z^2$$

Step (4)

$$\frac{\text{Variance}}{rt} = n^2\left(rt\,\sigma_c^2 + r\,\sigma_h^2 + \sigma_v^2 + \frac{\sigma_z^2}{n}\right)$$

The final expression is the mean square expectation as given in the analysis of variance.

The steps outlined above may now be put in a general form that will be applicable for determining the expectation of the mean square for any source of variance in the analysis of variance table. The first two steps listed above can be combined into one. The steps involved are then as follows: (1) Write the equation for the sums used in computing the mean square, omitting all effects not specific for a single one of these sums. Note that effects specific for single sums will include among its subscripts those necessary to specify the sum. (2) Write the variance of the expression [remembering (6) and (15) of section II]. (3) Divide the variance by the number of plots totaled in each of these sums. For a slightly different rule for writing mean square expectations see Crump (8).

Given the mean square expectations the procedure for estimating the variance components from the mean squares is easily seen.

$$M_5 \longrightarrow n^2\sigma^2$$

$$(M_4 - M_5)/r \rightarrow n^2\sigma_h^2$$

$$(M_3 - M_4)/rs \rightarrow n^2\sigma_d^2$$

$$(M_2 - M_4)/rt \rightarrow n^2\sigma_c^2$$

$$(M_1 - M_2 - M_3 + M_4)/rst \longrightarrow n^2\sigma_g^2$$

where $\longrightarrow$ is read "is an estimate of".

As a numerical example an analysis of variance is presented below of data on yield of 10 varieties of soybeans at each of 5 stations in each of two years. The comparison was replicated 4 times in each location and year. We have then: G = 10, s = 5, t = 2, r = 4.

Source of Variation	d.f.	M.S.	M.S. Expectation
Years	1	424,453	
Locations	4	1,044,298	
Y x L	4	1,632,833	
Varieties	9	422,236	$n^2(\sigma^2 + 4\sigma_h^2 + 20\sigma_d^2 + 8\sigma_c^2 + 40\sigma_g^2)$
V x Y	9	42,950	$n^2(\sigma^2 + 4\sigma_h^2 + 20\sigma_d^2)$
V x L	36	46,359	$n^2(\sigma^2 + 4\sigma_h^2 + 8\sigma_c^2)$
V x Y x L	36	60,744	$n^2(\sigma^2 + 4\sigma_h^2)$
Reps in year and location	270	12,716	$n^2\sigma^2$

A somewhat unexpected result is noted. The V x Y and V x L means squares are smaller than the V x Y x L mean square instead of larger as was to be expected assuming σ_c^2 and σ_d^2 greater than zero. Since variances cannot be negative quantities the most logical procedure probably is to conclude that σ_d^2 and σ_c^2 are zero (or at least of insignificant magnitude) and to use all three of these interaction mean squares as estimates of $n^2(\sigma^2 + 4\sigma_h^2)$. On this premise our estimates of the variance components are as follows:

Component	Estimate
$n^2\sigma^2$	12,716
$n^2\sigma_h^2$	$\left[\frac{9(42,950) + 36(46,359) + 36(60,744)}{81} - 12,716\right]/4$
	$= 9,914$

(con'd)

Component	Estimate
$n^2\sigma^2_c$	zero
$n^2\sigma^2_d$	zero
$n^2\sigma^2_g$	$\left[422{,}236 - \frac{9(42{,}950) + 36(46{,}359) + 36(60{,}744)}{81}\right]$
	$= 9{,}247$

$n^2\sigma^2_e$ for the experiment as conducted is estimated as

$$\frac{9{,}914}{10} + \frac{12{,}716}{40} = 1309 \qquad \left[\text{from (50b)}\right]$$

and as an estimate of $n^2\sigma^2_g/\sigma^2_{\bar{np}}$ we have

$$\frac{9247}{9247 + 1309} = .88$$

The above estimates are, of course, subject to sampling variance. However, if the data were sufficient to assure that the estimates were substantially correct, they would furnish a basis for deciding the manner in which varieties should be tested in the future. Referring to Table 1, Section III we see that when σ^2_g/σ^2_p is as high as .7 to .9 there is little to gain from attempting to decrease σ^2_e, i.e. in the case under discussion there would have been no point in increasing either replication or the number of years and locations at which the comparison was made. If the estimates of the variances were substantially correct, reduction of replication to two in future comparisons would result in reduction of the expected value of $n^2\sigma^2_g/\sigma^2_{\bar{np}}$ only to

$$\frac{9247}{9247 + \frac{9914}{10} + \frac{12716}{20}} = .85$$

The data indicate that very little was gained from using 4 replications instead of two.

A brief summary of implications of the relative magnitudes of the variance components with respect to the design of selection programs is pertinent at this point.

1. Unless σ^2 is very large relative to the other components of non-genetic variance, more than two replications at a single location in a single year will rarely be worthwhile in terms of Δg_s, the increase in progress expected from selection.

2. When σ^2_c, variance due to interaction of location and genotype, is very large, selection for average performance over the entire area from which the locations were drawn should be abandoned if a criterion can be found for establishing sub-areas within which σ^2_c will be much reduced. The reason is that when this interaction is large there is probably no single genotype that will be superior under the conditions of all locations in the area. By subdivision on some meaningful basis such as soil-type, drainage, etc., it may be possible to find genotypes with superior adaptation for all locations of a given sub-area. (Note that the criterion may not divide areas along geographical lines.) The location-genotype interaction that remains after further area subdividion becomes impractical must be dealt with as error variance in the genotype comparisons and controlled by making variety comparisons over a sufficient number of locations.

3. When σ^2_d is very large the only useful action is to increase $\underline{t}$, the number of years over which tests are run, unless measures of climate or effect of climate are available before planting time which will serve to divide the year population into sub-populations within which year-variety interaction would be greatly reduced. For example, it is possible that genotypes respond differentially to soil moisture at planting time and that certain varieties might be recommended for use when soil moisture was high and others when it was low. (The intent is not to say that this is or is not a reasonable possibility. The example is advanced only to indicate the sort of thing that would be required to eliminate year-genotype interaction as a source of error in recommendations of superior genotypes.)

4. Variance due to second order interaction of genotype, year, and location. In general, this variance must be looked on as error variance to be controlled by increasing locations, years, or both. However, to the extent that it is a consequence of variation in weather between locations in the same year there is a possibility of dealing with it as suggested in the case of σ^2_d.

D. Cases where each genotype may be represented by more than one individual and phenotype is expressed more than once by each individual.

The most common examples are the perennial crops capable of asexual propagation. The annually expressed characters of fruit trees, strawberries, blueberries, sugar cane, etc. belong in this group. Traits of perennial forage crops in which seed is produced entirely by apomixis also belong here.

The situation for these traits can be resolved into one very similar to that for category C if we take the point of view that the trait should be measured in terms of its average expression during the lifetime (defined in terms of commercial practice) of an individual. Unless we assume that the expression of the trait is uncorrelated with age (usually an untenable assumption) lifetime performance is expressed but once though, as will be discussed below, sample portions of lifetime performance may be considered as separate expressions of it. Our model will be as follows:

$$P_{ijk\ell} = g_i + a_j + b_k + c_{ij} + d_{ik} + f_{jk} + h_{ijk} + u_{jk\ell} + v_{ijk\ell}$$

$$\sum_{o=1}^{y} v_{ijk\ell o} + \sum_{n=1}^{n} z_{ijk\ell n} + \sum_{n=1}^{n}\sum_{o=1}^{y} z'_{ijk\ell no} \qquad (51)$$

where $P_{ijk\ell}$ is the lifetime sum (of annual measures of the trait) for a plot of the ith genotype in the ℓ th replication at the jth location in the kth testing period. (A testing period is taken to mean a set of years over which a single planting of a

genotype comparison is observed. o = 1, 2, • • • y is used to identify single years of the testing period.)

g_i, a_j, and c_{ij} have the same meaning as is equation (48)

b_k is the effect of the kth testing period

d_{ik} is the interaction of the ith genotype and the kth testing period.

f_{jk} is the interaction of the jth location and the kth testing period.

h_{ijk} is the second order interaction of the ith genotype, the jth location and the kth testing period.

$u_{jk\ell}$ is the effect of the ℓ th replication in the jth location and and kth testing period.

$v_{ijk\ell}$ is the effect of the plot in which the ith genotype is raised in the jth location and kth testing period that is constant for each year of the kth testing period.

$v'_{ijk\ell o}$ is the portion of the plot effect in the oth year that is temporary. (These effects are assumed to be randomly distributed with respect to plots and years.)

$z_{ijk\ell m}$ is the effect of intra-plot environmental variation for the mth individual of the plot that is constant for all years of the kth testing period.

$z'_{ijk\ell mo}$ is the portion of the effect of intra-plot environmental variation that is temporary. (These are assumed to be randomly distributed with respect to plants and years.)

This model differs from that of the preceding section (equation 48) in three ways. (1) The unit of time involved in expression of the phenotype is recognized as a set or series of years instead of only one. (2) Random plot and intra-plot environmental effects are subdivided into a portion that is constant through the testing period and another that varies from year to year. (3) The phenotype is measured on the plot basis instead of the plant basis.

The mean for a genotype (omitting terms that are constant for all genotypes, as in equation 49a) is

$$\bar{P}_i = g_i + \sum_{j=1}^{s} (c_{ij})/s + \sum_{k=1}^{t} (d_{ik})/t + \sum_{j=1}^{s}\sum_{k=1}^{t} (h_{ijk})/st$$

$$+ \sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r} (v_{ijk\ell})/rst + \sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r}\sum_{o=1}^{y} (v'_{ijk\ell o})/rst$$

$$+ \sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r}\sum_{m=1}^{n} (z_{ijk\ell m})/rst$$

$$+ \sum_{j=1}^{s}\sum_{k=1}^{t}\sum_{\ell=1}^{r}\sum_{m=1}^{n}\sum_{o=1}^{y} (z_{ijk\ell mo})/rst \qquad (52)$$

where

s = number of locations
t = number of testing periods
r = number of replications at each location in each testing period
n = number of individuals per plot, and
o = number of years per testing period.

From (52) the variance of a genotype mean is seen to be

$$\sigma^2_{\bar{p}} = \sigma^2_g + \sigma^2_e$$

where

$$\sigma^2_e = \frac{\sigma^2_c}{s} + \frac{\sigma^2_d}{t} + \frac{\sigma^2_h}{st} + \frac{\sigma^2_v}{rst} + \frac{y\sigma^2_{v'}}{rst} + \frac{n\sigma^2_z}{rst} + \frac{ny\sigma^2_{z'}}{rst} \qquad (53)$$

An analysis of variance using plot sums (for the entire testing period) as the unit variable and data collected in r replications at each of s locations in each of t testing periods would take the following form.

Source of Variation	d.f.	Expectation of M.S.
Location	$s - 1$	
Testing period	$t - 1$	
L x T_p	$(s - 1)(t - 1)$	
Reps in L and T_p	$st(r - 1)$	
Varieties	$G - 1$	$\sigma^2 + r\sigma^2_h + rs\sigma^2_d + rt\sigma^2_s + rts\sigma^2_g$
V x L	$(s - 1)(G - 1)$	$\sigma^2 + r\sigma^2_h + rt\sigma^2_c$
V x T_p	$(t - 1)(G - 1)$	$\sigma^2 + r\sigma^2_h + rs\sigma^2_d$
V x L x T_p	$(s - 1)(t - 1)(G - 1)$	$\sigma^2 + r\sigma^2_h$
V x reps in L and T_p	$st(r - 1)(G - 1)$	$\sigma^2 = \sigma^2_v + y\sigma^2_{v'} + n\sigma^2_z + ny\sigma^2_{z'}$
Total	$strG - 1$	

In addition plot values by years may be used to compute a mean square for "varieties x years in replications in locations and testing periods" which will have as its expected value, $\sigma^2_{v'} + n\sigma^2_{z'}$. If large plants such as fruit trees are involved and therefore a minimum number of plants per plots would be desirable estimates of σ^2_z and $\sigma^2_{z'}$ will be useful. These can be obtained using data for individual plants by years (if available) to compute mean squares for "plants in plots" and "years x plants in plots". The expected value of the former will be $\sigma^2_{z'} + y\sigma^2_z$ and of the latter will be $\sigma^2_{z'}$.

From the foregoing it will be noted that all variance components on which information is needed as a basis for deciding how selection experiments may best be conducted can be estimated if appropriate data are available. However, data involving more than one testing period is not likely to be available for long-lived plants. Indeed, extension of trials over two non-over-lapping testing periods

would most likely be impractical for such plants. Alternatives that suggest themselves are (1) spacing locations widely enough so that weather correlation among them is not likely to be high, and (2) over-lapping testing periods. The latter would mean that the planting year and perhaps one or two more would be different for each testing period and the initial years may have the most profound effect on lifetime performance.

If the cost of taking data from producing plants is very great the experimenter should consider collecting data in only a portion of years of the testing period in the case of long-lived plants unless $\sigma^2_{v'}$ or $\sigma^2_{z'}$ are so large that so doing would appreciably increase σ^2_c. The effect can be seen if (53) is rewritten in terms of $\sigma^2_{\bar{p}/y}$, the variance of a genotype mean on a plot per year basis.

$$\left.\begin{aligned} \sigma^2_{\bar{p}/y} &= \frac{\sigma^2_g}{y^2} + \frac{\sigma^2_c}{y^2} \\ \frac{\sigma^2_c}{y^2} &= \frac{\sigma^2_c}{sy^2} + \frac{\sigma^2_d}{ty^2} + \frac{\sigma^2_h}{sty^2} + \frac{\sigma^2_v}{rsty^2} + \frac{\sigma^2_{v'}}{rsty} + \frac{n\sigma^2_z}{rsty^2} + \frac{n\sigma^2_{z'}}{rsty} \end{aligned}\right\} \quad (54)$$

If the form of the performance curve over the lifetime of the organism has economic importance, it will perhaps be taken into account best if considered a separate attribute to be measured in terms of one or more regression coefficients.

Problems:

11. (a). Assume that a comparison of 12 oats varieties had been conducted at two locations with 6 replications in each of 3 years. Write out the form of the analysis of variance including the mean square expectations required for estimation of σ^2_g and the components of σ^2_e.

(b). Suppose the comparison had been conducted at only one station. Write the analysis with mean square expectations and show that a clean estimate of σ^2_g would not be available.

12. Assume that you have made a cross of two varieties of irish potatoes, that F_2's have been tested in a disease nursery, and that adequate vegetative planting material of each of a large number of disease-resistant selections is at hand. Assume further that from previous data you have good variance component estimates which are as follows.

$\sigma^2_g = 8$

$\sigma^2_c = 6$

$\sigma^2_d = 2$

$\sigma^2_h = 4$

$\sigma^2 = 16$

Now suppose you can plant a trial involving 400 plots that can all be at one location or can be divided equally among anywhere from 2 to 10 locations. However, if the trial is extended over 2 years only 200 plots can be used per year. How many varieties would you put in the trial at how many locations in how many years and in how many replications per location-year? The object is to make g_s as large as possible.

13. Below is an analysis of variance of data collected by P. H. Harvey in a comparison of 25 corn hybrids.

	d.f.	M.S.
Locations	4	2772.0
Reps in Loc.	10	72.3
Genotypes	24	32.19
G x L	96	6.78
Reps x genotypes in locations		4.26

Compute estimates of variance components that can be made from these data. Compute g_s for the best hybrid of the 25 on the basis of this trial. Its average yield for the test was 3.8 above the average for the 25.

References:

(8) Crump, S. Lee (1946) The Estimation of Variance Components in Analysis of Variance. Biometrics 2:7-11.

1. $X_1, X_2, X_3 \ldots X_n$

$$I = \beta_1 X_1 + \beta_2 X_2 + \beta_n X_n$$

$$\beta_1 \sigma_{P1} + \beta_2 \sigma_{P12} + \ldots = \sigma_{g1W}$$

$$\beta_1 \sigma_{P12} + \beta_2 \sigma_{P2} + \ldots = \sigma_{g2W}$$

Regression of G_W on $I = 1.0$

$$G_i = k \sqrt{\beta_1 \sigma_{g1W} + \beta_2 \sigma_{g2W} + \ldots}$$

W — real worth of organism

W = Yield

W = Y Quality

$\sigma^2_{g \cdot x}$

$\sigma^2_{g \cdot \bar{x}} = \sigma^2_g \left(1 - \frac{\sigma^2_g}{\sigma^2_x}\right)$

Mimeo. No. 6
R. E. Comstock:5:49

V. Simultaneous selection for several traits.

When the economic value of an organism is a function of more than one characteristic, it is logical that each trait affecting merit should receive attention in selection directed at genetic improvement. Hazel and Lush (9) showed that it is more efficient to consider each such trait in every generation of selection, provided each trait is given its proper weight relative to the others, than to follow the plan of improving the individual traits one at a time. It should be noted that in certain instances the optimum weight to be given a specific trait will be so high that the optimum basis for selection is essentially selection for that trait alone. In such cases the slight gain from attention to other traits may not compensate for the cost of collecting data on them. This situation might be expected, for example, in the first stages of selection following a cross of two strains one of which was resistant and the other susceptible to an important disease to which resistance was an absolute prerequisite for practical purposes.

Giving a specific weight in selection to each of several traits amounts to basing selection on an index of the form

$$I = b_1X_1 + b_2X_2 + \cdots + b_nX_n$$

where $X_1, X_2, \cdots, X_n$ are the phenotypic values of the traits considered and $b_1, b_2, \cdots, b_n$ are the relative weights accorded the n traits. The term, selection index, will be used specifically in reference to such an index in which the b's are given the best estimates of their optimum values.

If phenotype were unaffected by variation in environment every individual would express its genotypic worth phenotypically and selection would be straightforward and efficient. The weights to be given different traits would then depend only on their contributions to the economic value of the organism. However, not only do environmental variations affect phenotype but their effect on phenotypic expression is greater for some traits than others. It is obvious that proper weighting must take

cognizance of these facts. For example, suppose economic worth of an organism were a function of two traits and that phenotypic variation in one of these traits was entirely non-genetic but in the other was largely genetic. Attention to the trait for which all variation was non-genetic would accomplish no genetic improvement. On the other hand, it would mean a lower selection differential and hence less gain from selection in the trait which was varying genetically.

4. Equations for the estimation of optimum weights.

The estimation procedure was first given by Fairfield Smith (10). He demonstrated its application in a consideration of selection among varieties of wheat. Details of the estimation procedure were also given by Hazel (11) who reported on the construction of a selection index for use with swine. Applications to selection in poultry have been considered by Panse (12) and Lerner et al (13).

The following derivations differ in details but not in fundamentals from the presentations by Fairfield Smith and Hazel.

Let W symbolize economic worth and g_W the genotype for economic worth. Genetic improvement implies increase in g_W. To be effective in increasing g_W selection must be on a basis which will result in choice of genotypes for which g_W is above its mean for the population. Using an index of the form

$$I = b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (55)$$

selection will be most effective when $b_1, b_2, \cdots, b_n$ are given values that make the correlation of I with g_W as large as possible. These are the optimum values of the b's referred to earlier. These optimum values will be symbolized as $\beta_1, \beta_2, \cdots, \beta_n$.

The problem is of the multiple regression type. The X's are the independent variables, and g_W the dependent variable to be predicted or estimated from knowledge of the X's. The regression model for relation of g_W to the X's is

$$g'_W = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n \qquad (56)$$

where g_w' is the average value of g_w associated with a given set of X's. The magnitude of β_0 does not concern us since it is a constant and will have no effect on the differences between estimates of g_w. (Only the differences are important in the selection process.) By reference to Fisher (14) or Snedecor (15) we find that appropriate b's (estimates of the β's in multiple regression) are obtained from simultaneous solution of the following set of equations.

$$\left.\begin{aligned} b_1Sx_1^2 + b_2Sx_1x_2 + \cdots + b_nSx_1x_n &= Sx_1g_w \\ b_1Sx_1x_2 + b_2Sx_2^2 + \cdots + b_nSx_2x_n &= Sx_2g_w \\ \vdots \\ b_1Sx_1x_n + b_2Sx_2x_n + \cdots + b_nSx_n^2 &= Sx_ng_w \end{aligned}\right\} \qquad (56)$$

Remember that $x_1 = X_1 - \bar{X}_1$, $x_2 = X_2 - \bar{X}_2$, etc.

Now consider the quantity, Sx_1g_w. Since $x_1 = g_1 + e_1$,

$$Sx_1g_w = S(g_1 + e_1)g_w = Sg_1g_w + Se_1g_w$$

However e_1 is a random environmental effect uncorrelated with genotype, and so the expected value of Se_1g_w is zero. Hence Sx_1g_w is on the average equal to Sg_1g_w. In like manner Sx_2g_w, Sx_3g_w, etc. are on the average equal to Sg_2g_w, Sg_3g_w, etc. Making the substitutions, Sg_1g_w for Sx_1g_w, Sg_2g_w for Sx_2g_w, etc. and dividing each of the equations by (N - 1), we have

$$\left.\begin{aligned} b_1s_{p11} + b_2s_{p12} + \cdots + b_ns_{p1n} &= s_{g1w} \\ b_1s_{p12} + b_2s_{p22} + \cdots + b_ns_{p2n} &= s_{g2w} \\ \vdots \\ b_1s_{p1n} + b_2s_{p2n} + \cdots + b_ns_{pnn} &= s_{gnw} \end{aligned}\right\} \qquad (57)$$

where s_{p11} = an estimate of the phenotypic variance of X_1,

s_{p12} = an estimate of the phenotypic covariance of X_1 and X_2,

s_{g1w} = an estimate of the genotypic covariance of X_1 and w,

etc.

When only two traits are to be considered in selection there will be two equations

$$b_1 s_{p11} + b_2 s_{p12} = s_{g1w}$$

$$b_1 s_{p12} + b_2 s_{p22} = s_{g2w}$$

When three traits are to be considered the set will involve three equations

$$b_1 s_{p11} + b_2 s_{p12} + b_3 s_{p13} = s_{g1w}$$

$$b_1 s_{p12} + b_2 s_{p22} + b_3 s_{p23} = s_{g2w}$$

$$b_1 s_{p13} + b_3 s_{p23} + b_3 s_{p33} = s_{g3w}$$

and so forth.

B. The expected effect of selection.

The efficiency of selection can be measured in terms of the expected genotypic superiority of selected individuals, strains, or varieties over the mean of the group from which they were selected. This will be the product of the selection differential (mean difference in selection index between selected group and the entire group) and the regression of g_w on the selection index.

The selection differential in accordance with previous notation is $k\sigma_I$ and the regression of g_w on I is σ_{Ig_w}/σ_I^2. Thus the expected genotypic advance is

$$k\sigma_I \frac{\sigma_{Ig_w}}{\sigma_I^2} = \frac{k\sigma_{Ig_w}}{\sigma_I} \qquad (58)$$

Since $x_1 = X_1 - \bar{X}$, $x_2 = X_2 - \bar{X}$, etc.,

$$I - \bar{I} = b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

Then

$$\sigma_{Ig_w} = \frac{\sum (I - \bar{I})g_w}{N}$$

$$= \frac{\sum (b_1x_1 + b_2x_2 + \cdots + b_nx_n)g_w}{N}$$

$$= \frac{b_1\sum x_1g_w}{N} + \frac{b_2\sum x_2g_w}{N} + \cdots + \frac{b_n\sum x_ng_w}{N} \qquad (59)$$

and

$$\sigma_I^2 = \frac{\sum (I - \bar{I})^2}{N} = \sum (b_1x_1 + b_2x_2 + \cdots + b_nx_n)^2/N$$

Expanding and collecting terms appropriately this becomes

$$\sigma_I^2 = b_1\left[b_1\sum x_1^2 + b_2\sum x_1x_2 + \cdots + b_n\sum x_1x_n\right]/N$$
$$+ b_2\left[b_1\sum x_1x_2 + b_2\sum x_2^2 + \cdots + b_n\sum x_2x_n\right]/N$$
$$+ \cdots b_n\left[b_1\sum x_1x_n + b_2\sum x_2x_n + \cdots + b_n\sum x_n^2\right]/N$$

Substituting in terms of equations (56) this can be written

$$\sigma_I^2 = \frac{b_1\sum x_1g_w}{N} + \frac{b_2\sum x_2g_w}{N} + \cdots + \frac{b_n\sum x_ng_w}{N} \qquad (60)$$

From (59) and (60) we see that $\sigma_{Ig_w} = \sigma_I^2$ or $\sigma_I = \sqrt{\sigma_{Ig_w}}$.

Hence (58), the expected advance from selection can be written

$$\frac{k\,\sigma_{Ig_w}}{\sqrt{\sigma_{Ig_w}}} = k\sqrt{\sigma_{Ig_w}}$$

$$= k\sqrt{\frac{b_1\sum x_1g_w}{N} + \frac{b_2\sum x_2g_w}{N} + \cdots + \frac{b_n\sum x_ng_w}{N}}$$

Finally, remembering that, in ariving at equations (57), it was shown that $\sum x_1g_w = \sum g_1g_w$, $\sum x_2g_w = \sum g_2g_w$, etc., we can write the expected genetic

advance from selection based on the index as

$$k \sqrt{b_1 \sigma_{g1w} + b_2 \sigma_{g2w} + \cdots + b_n \sigma_{gnw}} \qquad (61)$$

where σ_{g1w} is the genetic covariance of X_1 with g_w,

σ_{g2w} is the genetic covariance of X_2 with g_w,

etc.

C. Estimation of variances and covariances required for solution of equations (57).

The estimates of phenotypic variances and covariances of the X's (s_{p11}, s_{p12}, etc.) must be appropriate to the sort of values to be used in the index. For example, if selection among strains is to be based on means for r replications at each of s locations in each of t years then the estimates should be for such values. They can be made directly from data involving such means or can be built up from estimates of the variance or covariance components involved in accordance with equations (50) and (53).

The estimation of covariance components from the mean products of an analysis of covariance is exactly analogous to the estimation of variance components from the mean squares of an analysis of variance. This was pointed out by both Fairfield Smith (10) and Hazel (11) though the procedure has not been used extensively.

An analysis of covariance involving data on two traits (say X_1 and X_2) from a field comparison of genotypes at different locations and in different years would take the following form.

Source of Variance	d.f.	Expectation of mean product
Locations	$s - 1$	
Years	$t - 1$	
L x Y	$(s - 1)(t - 1)$	
Reps in years and locations	$st(r - 1)$	
Varieties	$G - 1$	$\sigma_{12} + r\sigma_{hl2} + sr\sigma_{dl2} + tr\sigma_{cl2} + rst\sigma_{gl2}$
V x L	$(s - 1)(G - 1)$	$\sigma_{12} + r\sigma_{hl2} + tr\sigma_{cl2}$
V x Y	$(t - 1)(G - 1)$	$\sigma_{12} + r\sigma_{hl2} + sr\sigma_{dl2}$
V x Y x L	$(s - 1)(t - 1)(G - 1)$	$\sigma_{12} + r\sigma_{hl2}$
Var x reps in year and location	$st(G - 1)(r - 1)$	σ_{12}
Total	$strG - 1$	

G, s, t, and r have the same significance as in Section IV, C and D. The numeral subscripts of the σ's indicate the traits involved and the letter subscripts have the same significance relative to the source of the covariance components as they had in Section IV relative to sources of variance components. For example σ_{cl2} is the covariance between the location-genotype interaction effects for X_1 and X_2.

The covariance of genotype means for two traits can be written in terms of the covariance components in a form analagous to that for variance. For example the covariance between genotype means of X_1 and X_2 is

$$\sigma_{\bar{p}12} = \sigma_{g12} + \frac{\sigma_{cl2}}{s} + \frac{\sigma_{dl2}}{t} + \frac{\sigma_{hl2}}{st} + \frac{\sigma_{12}}{rst} \qquad (62)$$

The genetic covariances of the X's with g_w can be estimated from analyses of covariance (like the above) of the X's with economic worth. An obvious prerequisite

is definition of economic worth and all data required to compute the values of it to be used in the analyses. In agronomic or horticultural crops, worth may in some instances be defined simply as yield whereas in others a satisfactory definition will need to involve both yield and some measure of quality. Once w has been defined and tabulated, estimation of its genetic covariances with the X's is analagous to estimation of genetic variances except that, as indicated above, it involves analysis of covariance instead of analysis of variance.

In case selection is to be between individuals (instead of between strains, lines, or varieties) as is often the case with animals, the variances and covarainces of equations (57) must be for individuals, not for group means. The estimation of the phenotypic variances and covariances of the X's is then completely straightforward. However, the estimation of genetic covariances will need to be made somewhat indirectly from data involving related animals. This will be taken up later since the basis for the possible procedures has not yet been laid.

D. Indirect estimation of the genetic covariances with economic worth.

In some instances the available data will not include all items necessary for computation of w. When this is true it is sometimes possible to estimate the genetic covariances indirectly.

Let $a_1, a_2, \cdots, a_n$ be the increases in economic worth that result from unit phenotypic increases in traits 1, 2, $\cdots$, n when each of the other traits remains unchanged. Then the genotype for economic worth can be written as a function of the genotype for the n traits.

$$g_w = a_1 g_1 + a_2 g_2 + \cdots + a_n g_n$$

and the genotypic covariance of w with the ith trait will be

$$\sigma_{giw} = \Sigma g_i(a_1 g_1 + a_2 g_2 + \cdots + a_n g_n)/N$$

$$= \Sigma(a_1 g_1 g_i + a_2 g_2 g_i + \cdots + a_n g_n g_i)/N$$

$$= a_1 \sigma_{gli} + a_2 \sigma_{g2i} + \cdots + a_i \sigma_{gii} + \cdots + a_n \sigma_{gni} \qquad (62)$$

If one of the n traits does not affect economic worth directly, i.e. variation in it causes no change in economic worth when all other traits remain constant, the corresponding a-value is zero and the term involving that a-value drops out of the equation.

Existing data on animals is frequently not complete enough so that direct estimates of the genetic covariances between economic worth and individual traits can be made. On the other hand estimates of the genetic covariances among the individual traits may be possible using different bodies of data for the estimation of different covariances. It is in such cases that equation (62) becomes useful. [For example, see Hazel (11).] In the case of plants the collection of data for the specific purpose of constructing a selection index is feasible and can be accomplished in a reasonable period of time. Thus with plants the worker is not so dependent on existing data collected for other purposes as in the case with animals, especially those which reproduce slowly such as sheep and cattle.

The a-values.

The a's stem directly from definition of economic worth in the organism in question. For example, in work with Sea Island cotton, H. L. Manning (unpublished) defined economic worth in terms of yield. It was not necessary to bring in quality since the entire range of quality in the population with which he was working was within the range suitable for the purposes for which the cotton was to be used. The individual traits on which selection was being based were bolls per plant (X_1),

seeds per ball (X_2), and lint per seed (X_3). Since the cultural practice involved a constant number of plants per unit area of land economic worth (yield) was a function of these traits as follows:

$$W = X_1X_2X_3$$

Obviously, if X_1 were increased by unity, and there were no change in X_2 or X_3, the increase in W would be X_2X_3. If X_1 were increased by unity for the entire population, the increase in mean W would be $S(X_2X_3)/N$. Thus reasonable a-values in this case were

$$a_1 = S(X_2X_3)/N$$

$$a_2 = S(X_1X_3)/N$$

$$a_3 = S(X_1X_2)/N$$

An alternative procedure (used by Fairfield Smith) would have been to define economic worth as the logarithm of yield. Then the expression for W would have been

$$W = \log \text{yield} = \log X_1 + \log X_2 + \log X_3$$

Then if the index were to be based on logarithms of the X's, the a's would all be 1.0 since change in W in response to change in the log of any of the X's is exactly the amount of the change in the log in question. This approach has the advantage that the a's are independent of the population mean values of the X's. What may be a disadvantage is that the use of logs results in giving more weight to variations among low values of the X's than among the high values.

As an example from livestock the following is presented as a possible definition of economic worth in dairy cattle.

$$W = cX_1 - \frac{X_2}{X_3} - X_4$$

X_1 = annual production of 4% Fat Corrected Milk in lbs.

X_2 = cost of rearing minus disposal value

X_3 = length of productive lifetime in years

X_4 = maintainence cost per year

c = value per lb. of Fat Corrected Milk in excess of cost of production exclusive of maintainence cost for the animal.

X_3 would not ordinarily be a useful measure for use in selection because so much time is required to obtain it that its use would seriously retard a breeding program. The alternative and probably more practical procedure would be to treat X_3 as a constant equal in value to the population mean for productive lifetime. The a-values for X_1, X_2, and X_4 would then be

$$a_1 = c$$

$$a_2 = -1/\bar{X}_3$$

$$a_4 = -1$$

The a-values for any traits which do not enter the function for economic worth must be zero since variation in them will not cause variation in W so long as other traits remain constant. Thus a-values for disease resistence or plant habit in plants or for aspects of conformation in dairy cattle would be zero. This does not mean that they should not receive weight in an index. Indeed their genetic correlations with worth or their phenotypic covariances with other traits might be such that they should be weighted heavily.

E. General remarks.

The foregoing serves only to outline what is involved in the problem of efficient use in selection of data on the various important traits of the organism in question. To have given sufficient detail to cover procedure in the variety of situations in which this problem is confronted is far beyond the scope of this course. The essence of the problem is always the same but in details it differs so much from one situation to another that the construction of an index for any single organism constitutes a research project. Such a project will always involve two phases.

1. Definition of economic worth. Sometimes this will call for considerable effort in itself. For example, in the worth function proposed for dairy cattle the appropriate value for c would require careful consideration of the relationships over time between milk prices, feed prices, labor costs, etc.
2. Estimation of variances and covariances required for estimation of the b's.

It should be noted that a great deal of investigation remains to be done relative to statistical aspects of the selection index problem. Among the most pressing questions is that of the amount of data required to furnish satisfactory approximations to the optimum index. The most common procedure in practice probably is to weight traits in proportion to their economic importance. From a practical point of view it is important that any basis for selection adopted as an alternative be at least as efficient. To insure this there is obviously a minimum amount of data on which such an alternative index must be based. Data being accumulated at this station will serve to throw some light on this question within the next few years.

Problems.

14. Following are estimates of phenotypic variances and covariances and genetic covariances in Sea Island cotton (from unpublished data of H. L. Manning).

$s_{p11} = 9.49$, $s_{p12} = 6.85$, $s_{p22} = 11.94$

$s_{g1w} = 3.46$, $s_{g2w} = 7.45$, $s_{gww} = 7.20$

(a) Estimate the b's to be used in the selection index, $b_1X_1 + b_2X_2 = I$.

(b) Estimate the progress expected from selection based on

1. the selection index
2. X_1, bolls per plant
3. X_2, lint per seed
4. W, yield

(It will not be necessary to substitute a numerical value of k, since the gain in units of k will suffice for comparison of the four bases for selection.)

References:

(9) Hazel, L. N. and J. L. Lush (1942). The Efficiency of Three Methods of Selection. Jour. Hered., 33:393-399.

(10) Smith, H. Fairfield. (1936) A Discriminant Function for Plant Selection. Annals of Eugenics, 7:240-250.

(11) Hazel, L. N. (1943) Genetic Basis for Selection Indices. Genetics, 28: 476-490.

(12) Panse, V. G. (1946) An Application of the Discriminant Function for Selection in Poultry. Jour. Genetics, 47:242-248.

(13) Lerner, I. Michael, V. S. Asmundson, and Dorothy M. Cruden (1947). The Improvement of New Hampshire Fryers. Poult. Sci., 26:515-524.

(14) Fisher, R. A. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh and London.

(15) Snedecor, G. W. Statistical Methods. Iowa State College Press, Ames, Iowa.

Mimeo. No. 7.
R. E. Comstock:6:49

STATISTICAL CONCEPTS IN GENETICS

VI. Gene Frequency and the Distribution of Genotypes.

The frequency of a gene is its number in the population expressed as a fraction of the total number of loci at which it or an allele of it is present. For example, suppose that of 100 diploid organisms, 30 are of the genotype BB, 60 are Bb, and 10 are bb. There are 120 B genes and 200 loci occupied either by B or its allele, b. Hence the frequency of B with respect to this population of 100 individuals is 120/200 = .6.

When using letters to symbolize genes, the large case will always be used to designate the most favorable of a set of allelic genes. Thus it will always be understood that B is a more favorable gene than b, C a more favorable gene than an allele c, etc. The letter q will be used to designate the frequency of the most favorable of a set of allelic genes. Then if such a favorable gene has only one allele, the frequency of its allele will be (1-q).

The Distribution of Genotypes in a Random Mating Population.

Segregation of a pair of allelic genes (B and b, for example) yields three genotypes; BB, Bb, and bb. Assuming random mating and equal viability of the different gametes and genotypes, the expected ratio of genotypes homozygous for the favorable gene, heterozygous, and homozygous for the less favorable gene is q^2: $2q(1-q)$: $(1-q)^2$. This can be demonstrated as follows:

The probability of a gamete containing B is obviously q and the probability of 2 specific gametes, i.e. two combining to form a zygote, both containing B is q^2. Thus the expected proportion of BB zygotes will be q^2.

The probability of a gamete containing b is (1-q) and the probability of 2 specific gametes both containing b is $(1-q)^2$. Thus the probability of bb zygotes will be $(1-q)^2$.

The remainder of the zygotes must be of the Bb type. Since the total of the frequencies of the three types must be 1.0, the expected frequency of

heterozygotes is

$$1-q^2 - (1-q)^2 = 1-q^2 - 1 + 2q - q^2 = 2q-2q^2 = 2q(1-q).$$

When two gene pairs (say B, b and C, c) segregate independently, the expected frequencies of the various possible genotypes will be as follows:

Genotype	Frequency	Frequency of BB, Bb, and bb
BBCC	$q_b^2\ q_c^2$	
BBCc	$q_b^2\ 2q_c(1-q_c)$	q_b^2
BBcc	$q_b^2(1-q_c)^2$	
BbCC	$2q_b(1-q_b)q_c^2$	
BbCc	$2q_b(1-q_b)2q_c(1-q_c)$	$2q_b(1-q_b)$
Bbcc	$2q_b(1-q_b)(1-q_c)^2$	
bbCC	$(1-q_b)^2q_c^2$	
bbCc	$(1-q_b)^2 2q_c(1-q_c)$	$(1-q_b)^2$
bbcc	$(1-q_b)^2(1-q_c)^2$	

The key to these frequencies is obvious. For example, if the probability of a genotype containing 2B's is q_b^2 and of a genotype containing 2C's is q_c^2, the probability of a genotype containing both 2B's and 2C's is $q_b^2q_c^2$.

In general, if n gene pairs are segregating independently in a random breeding population and viability is equal among gametes and zygotes, the expected frequencies of the various genotypes can be obtained from expansion of the expression

$$\left[q_1A_1 + (1-q_1)a_1\right]^2 \left[q_2A_2 + (1-q_2)A_2\right]^2 \cdots \left[q_nA_n + (1-q_n)a_n\right]^2 .$$

In each term of the expansion the particular combination of A's and a's species the genotype, and the portion involving q's and numerical values gives

the frequencies. For example, if n = 4, one term of the expansion will be

$$4q_1^2q_2(1-q_2)q_3(1-q_3)(1-q_4)^2\ A_1A_1A_2a_2A_3a_3a_4a_4 \quad .$$

This term indicates that the genotype, $A_1A_1A_2a_2A_3a_3a_4a_4$, will be expected in the frequency $4q_1^2q_2(1-q_2)q_3(1-q_3)(1-q_4)^2$. There will be a term in the expansion for each of the possible genotypes.

Of some special interest is the frequency of the genotype homozygous for all of a set of desired genes. Clearly this will be $q_1^2\ q_2^2\ \ldots\ldots\ q_n^2$. If $q_1 = q_2 = \ldots\ldots = q_n$ as in the F_2 of a cross of homozygous lines this becomes q^{2n}.

Linkage between loci prevents independent segregation. In the event of linkage between loci the foregoing relative frequencies of genotypes may be regarded as equilibrium values that will be approached over a span of generations if the actual frequencies in a random breeding population are for any reason out of equilibrium. For example, the joint distribution of genotypes at linked loci may be expected to be out of equilibrium in a population descended from a recent cross of contrasting genotypes. How rapidly equilibrium is approached will depend on the closeness of linkage . (see

The Distribution of Genotypes in Inbred Populations.

Inbreeding may be defined as the mating of related individuals. It is a form of non-random mating. The closest or most intense sort of inbreeding is self-fertilization. Because related individuals are more likely to possess the same gene at any given locus than are non-related individuals the progeny of matings involving inbreeding will on the average be homozygous at more loci than the progeny of random matings.

if $f = .4$ the heterozygote of origin pop. is decreased by .4
the average decrease in hetz per ind.

for self fertilization f' inbreeding in previous gen.

$$f = \tfrac{1}{2}(1 + f')$$

Generation	f'	f
1	0	.5
2	.5	.75

Brother-sister or Sib-mating

$$f = \tfrac{1}{4}\cdot(1 + 2f' + f'')$$

Generation	f'	f''	f
1	0	0	.25
2	.25	0	.375

$\frac{1}{8M} + \frac{1}{8F}$ = % decrease in heterozygocies of that present in preceding generation

Generation	% decrease of hetezygs.	H.
1	25%	75%
2		9/16

---

1 In an F_2 between homozyg. lines differing in genes at 4 loci. calc is the probabilty of genotypes with a maximum of 1 undesirable gen.

2. Compute f for the 1st 5 generations of
   1. selfing
   2. sib mating
   3 random mating among 2♂ and 2♀
   then " " 5♂ and 5♀.

3. If freq of red calves in Angus breed is .0081 if random mating is alternated with sib-mating and all red are disc from sib what is $\Sigma(pq)$ after 2nd cycle.

4

RR	Red	11,592	What is P R?
RW	Roan	9,538	are the data compatable with the hypothesis
WW	White	2,139	of (a) mono factorial
		23,269	(b) random mating

Mimeo. No. 7
R.E.Comstock:6:49

All sorts of inbreeding which result in progressive decrease in heterozygosis break a population into non-interbreeding lines. For example, a self-fertilized line can contain no more individuals in any one generation than can be produced from the seed of a single plant, the parent of that generation. As heterozygosis decreases the individuals belonging to such a line must become more and more alike in genotype. On the other hand different lines may become homozygous for different genes at a given locus and so while inbreeding reduces genetic variation within lines, it increases the genetic variation in a population of such lines.

Wright (16) devised a measure of inbreeding commonly called the coefficient of inbreeding. Its calculation is such that its magnitude is the expected decrease in heterozygosis in an inbred population as a percentage of the heterozygosis that would have been expected in the same population if there had been no inbreeding. Inbreeding as measured by Wright's coefficient will be symbolized as f. The distribution of genotypes involving two allelic genes in an inbred population will be as follows.

Genotype	Frequency
BB	$q^2 + fq(1-q)$
Bb	$2q(1-q)(1-f)$
bb	$(1-q)^2 + fq(1-q)$

Comparing this distribution with that expected under random mating it will be noted that the increase in homozygous loci is divided equally between the BB and bb types. As f approaches 1.0 the frequency of the BB type approaches q, and that of the bb type approaches (1-q). The gene frequency in the whole population is unaffected by inbreeding if the population is large. On the other hand the gene frequency in any one inbred lines of the population shifts, as inbreeding continues, from q to either 1.0 or zero. There is no directional force involved in this shift. It occurs, instead, because when a line is

carried from one generation to another by only a few gametes, the opportunity for random shift in gene frequency is greatly increased. This is most easily perceived in the case of self-fertilization where the line is carried from one generation to the next by but two gametes. Obviously if the line is heterozygous for a gene in one generation there is an even chance that it will be homozygous in the next. Thus the expected decrease in heterozygosis in any generation is fifty percent of that present in the preceding one.

The derivation of Wright's coefficient and some of the consequences of inbreeding will receive attention in a later section.

References:

16. Wright, Sewall (1922) Coefficients of Inbreeding and Relationship. Amer. Nat. 56:330-338.

17.

Mimeo. No. 7
R.E.Comstock:6:49

VII. Genotypic Variance Arising from Segregation of a Single Pair of Genes.

If both members of a gene pair (say B,b) are present in a population, three genotypes (with respect to the B locus) will appear; BB, Bb, and bb. Assuming random mating and equal viability of the different types of gametes and zygotes, their expected frequencies will be in the ratio $q^2:2q(1-q):(1-q)^2$. Let the average effect of bb on the organism be v, that of BB be (v+2u), and that of Bb be (v+u+au). Note that the effects of the three genotypes are not specified as constant but in terms of averages for the population. These effects would be constant only if there were no interaction of genotype at the B locus with either the remainder of the genotype or with environment. The situation is summarized in tabular form below.

Genotype	Frequency	y'	Y	X
BB	q^2	z+2u	u	2
Bb	$2q(1-q)$	z+u+au	au	1
bb	$(1-q)^2$	z	-u	0

y' is the phenotypic average for the specified genotype.
z, the value of y' for the bb genotype, is the sum of
Y and the average effect of the remainder of the genotype with respect to the population of environments involved.
y is obtained by coding y': $y = y' - (z+u)$
X is the number of B genes in the genotype.
a reflects the dominance involved in the action of the genes. When $a = 0$, the heterozygote is midway between the homozygotes, i.e. there is no dominance. If there is dominance of B, $a > 0.0$; if there is dominance of b, $a < 0.0$.

The total genotypic variance due to segregation of this pair of genes is

$$\sigma_y^2 = \frac{S(y^2) - (Sy)^2/N}{N}$$

Since N, the total frequency, = 1.0

$$\sigma_y^2 = Sy^2 - (Sy)^2 .$$

$$Sy = q^2u + 2q(1-q)au - (1-q)^2u$$

$$= (q^2 - 1 + 2q - q^2)u + 2q(1-q)au$$

$$= (2q-1)u + 2q(1-q)au \qquad (63)$$

$$Sy^2 = q^2u^2 + 2q(1-q)a^2u^2 + (1-q)^2u^2$$

and

$$\sigma_y^2 = \left[q^2 + (1-q)^2\right]u^2 + 2q(1-q)a^2u^2 - \left[(2q-1)u + 2q(1-q)au\right]^2$$

$$= \left[q^2 + 1-2q+q^2 - 4q^2+4q-1\right]u^2 - 4q(1-q)(2q-1)au^2$$

$$+ \left[2q(1-q) - 4q^2(1-q)^2\right]a^2u^2$$

$$= 2q(1-q)\left[1 + 2(1-2q)a + (1-2q+2q^2)a^2\right]u^2 \qquad (64)$$

The total genotypic variance, σ_y^2, is divisible into portions which are called (a) additive genetic variance, and (b) variance due to dominance deviations from the additive scheme [See Wright (18) and Lush (3), page 74].

The additive effect of the gene B may be defined as the regression of y' (or of y, which will be equivalent) on x, the number of B's in the genotype. This definition is couched in different words but has the same meaning as that given by Wright (18). The additive genetic variance (or the variance due to additive effects of the genes) is the variance in y that is accountable to linear regression on x. Remembering that the regression coefficient is computed in a manner that minimizes the variance due to deviations from regression, it will observed that the additive effect of B is defined such that as large a por tion of σ_y^2 as possible will be explained on the basis <u>additive gene effects</u>.

It happens that in a population where mating is random, the additive effect as defined above is the response that would be obtained from substituting B for b, averaged for all loci in the population at which b is present (and hence the substitution is theoretically possible). Thus, the concept of

additive effects applied when gene action is not of the simple additive sort is not so abstract as it may at first appear. Lush (3, page 73) defines the additive effect of a gene in terms of its substitution value.

The derivation of formulae for the additive effect of B; the additive genetic variance, and the variance due to dominance deviations is as follows.

$$Sx = 2q^2 + 2q(1-q) = 2q \qquad (65)$$

$$\sigma_x^2 = Sx^2 - (Sx)^2 \qquad \text{(remembering that N = 1)}$$

$$= 4q^2 + 2q(1-q) - 4q^2 = 2q(1-q) \qquad (66)$$

$$\sigma_{xy} = Sxy - (Sx)(Sy)$$

$$= 2q^2u + 2q(1-q)au - 2q\left[(2q-1)u + 2q(1-q)au\right]$$

$$\left[\text{see (63) for Sy.}\right]$$

$$= \left[2q^2 - 4q^2 + 2q\right]u + 2q(1-q)(1-2q)au$$

$$= 2q(1-q)\left[1 + (1-2q)a\right]u \qquad (67)$$

The additive effect of B is by our definition

$$b_{yx} = \sigma_{xy}/\sigma_x = \left[1 + (1-2q)a\right]u \qquad (68)$$

It is comparatively simple to verify that this same value will result from computation of the average substitution value of B. Substitution of B for b in the heterozygote increases y' by the amount (u-au). The frequency of heterozygotes is 2q(1-q) which is therefore the frequency of possible substitutions having the effect (u-au). Substitution of B for b in the bb genotype increases y' from -u to au, i.e. by the amount (u+au). Since the frequency of the bb genotype is $(1-q)^2$ and since there are two loci per individual of this genotype at which the substitution is possible, the frequency of possible substitutions having the effect (u+au) is $2(1-q)^2$. The average effect of all the possible substitutions is

$$\frac{2q(1-q)(u-au) + 2(1-q)^2(u+au)}{2q(1-q) + 2(1-q)^2}$$

$$= \frac{q(u-au) + (1-q)(u+au)}{q + (1-q)} = \left[1 + (1-2q)a\right]u$$

Mimeo. No. 7
R.E.Comstock:6:49

as obtained for the regression of y on x see (68) .

The additive genetic variance, to be symbolized hereafter as σ^2_g, was defined as the portion of σ^2_y due to regression on x. Hence

$$\sigma^2_g = \frac{\sigma^2_{xy}}{\sigma^2_x} = 2q(1-q)\left[1 + (1-2q)a\right]^2 u^2 \qquad (69)$$

[from (11), (66), and (67)]

Note now that when a = 0, i.e. when there is no dominance, $\sigma^2_g = \sigma^2_y$ [substitution of a = 0 in (64) and (69) leads to $2q(1-q)u^2$ in both cases]. This is logical since with no dominance, we have simple additive gene action and all genetic variance should be of the additive sort. On the other hand, whenever $a \neq 0$, $\sigma^2_g < \sigma^2_y$. The difference is the variance due to deviations from regression resulting from dominance and is termed variance due to dominance deviations (from the additive scheme). It will be symbolized hereafter as σ^2_d .

$$\sigma^2_d = \sigma^2_y - \sigma^2_g = 4q^2(1-q)^2 a^2 u^2 \qquad (70)$$

[from (64) and (69)]

Values of the Additive Effect and the Genetic Variances when a takes Values of Special Interest

1. When a = 0.0, i.e. there is no dominance.

   (a) $b_{yx} = u$ for all values of q.

   (b) $\sigma^2_y = \sigma^2_g = 2q(1-q)u^2$ with maximum value when q = .5. Values are listed below for several values of q.

q	$\sigma^2_g = \sigma^2_y$
.1	$.18u^2$
.2	$.32u^2$
.3	$.42u^2$
.4	$.48u^2$
.5	$.50u^2$
.6	$.48u^2$
.7	$.42u^2$
.8	$.32u^2$
.9	$.18u^2$

Note that for q = .2 to .8 the genetic variance is from 64 to 100 percent of its maximum.

(c) $\sigma^2_d = 0$

2. When a = 1.0, i.e. when B is completely dominant to b.

(a) $b_{yx} = 2(1-q)u$. It approaches 2u as q approaches zero and approaches zero as q approaches 1.0. Thus the additive effect of a dominant gene is high when its frequency is low, but low when its frequency is high.

(b) $\sigma^2_y = 4q(2-q)(1-q)^2u^2$ with maximum when q = .293 [For corresponding formula see Wright (19)].

(c) $\sigma^2_g = 8q(1-q)^3u^2$ with maximum when q = .25.

(d) $\sigma^2_d = 4q^2(1-q)^2u^2$ with maximum when q = .5.

The maximum for σ^2_d is at q = .5 for all values of a. [For formulae corresponding to (c) and (d) see Fisher (20)].

(e) The ratio of σ^2_g to σ^2_y is

$$\sigma^2_g / \sigma^2_y = 2(1-q)/(2-q)$$

which approaches one as q approaches zero and approaches zero as q approaches one, in harmony with the behaviour of the additive effect of the gene. Values of this ratio and its complement, σ^2_d / σ^2_y, are listed below for various values of q.

q	σ^2_g / σ^2_y	σ^2_d / σ^2_y
.2	.88	.12
.4	.75	.25
.5	.67	.33
.6	.57	.43
.7	.46	.54
.8	.33	.67
.9	.18	.82

3. When a > 1.0, i.e. when the heterozygote is superior to the superior homozygote, and $q = \frac{1+a}{2a}$.

(a) $b_{yx} = 0$

(b) $\sigma^2_g = 0$

(c) $\sigma^2_d = \sigma^2_y$

(d) $\bar{y} = (2q-1)u + 2q(1-q)au$ is at its maximum.

$\frac{1+a}{2a}$ is listed below for several values of a.

a	(1+a)/2a
1.0	1.0
1.5	.833
2.0	.75
3.0	.667
4.0	.625

Problems:

Plot b_{yx}, σ^2_y, σ^2_g, σ^2_d, and σ^2_g/σ^2_y against q for a = 0, .5, 1.0, 1.5, and 2.0.

References.

18. Wright, Sewall (1932). The Analysis of Variance and the Correlations between Relatives with Respect to Deviations from an Optimum. Jour. Gen. 30:243-256.

19. Wright, Sewall (1931). Evolution in Mendelian Populations. Genetics 16:97-159.

20. Fisher, R. A. (1718). The Correlation between Relatives on the Supposition of Mendelian Inheritance. Trans. Roy. Soc. Edinb. 52, Part 2, 399-433.

formula for figuring gain made by selection

$$\bar{y}_1 - \bar{y}_0 = k \sigma_p \frac{\sigma_g^2}{\sigma_p^2}$$

Regression of X on mean of Sire families

STATISTICAL CONCEPTS IN GENETICS

VIII. Genetic Variance Arising from Segregation of Two Independently Assorting Pairs of Genes.

Consider two pairs of genes (say B_1, b_1, and B_2, b_2) segregating independently in a random breeding population. Assuming equal viability of all types of gametes and zygotes, the distribution of genotypes may be represented in a two-way table as follows:

	B_1, b_1 genotype				
B_2, b_2 genotype	B_1B_1	B_1b_1	b_1b_1	Mean	Frequency
B_2B_2	q^2p^2	$2q(1-q)p^2$	$(1-q)^2p^2$		p^2
	Y_{22}	Y_{12}	Y_{02}	$Y_{.2}$	
B_2b_2	$2q^2p(1-p)$	$4q(1-q)p(1-p)$	$2(1-q)^2p(1-p)$		$2p(1-p)$
	Y_{21}	Y_{11}	Y_{01}	$Y_{.1}$	
b_2b_2	$q^2(1-p)^2$	$2q(1-q)(1-p)^2$	$(1-q)^2(1-p)^2$		$(1-p)^2$
	Y_{20}	Y_{10}	Y_{00}	$Y_{.0}$	
Mean	$Y_{2.}$	$Y_{1.}$	$Y_{0.}$	$Y_{..}$	
Frequency	q^2	$2q(1-q)$	$(1-q)^2$		1.0

q is the frequency of the gene B_1, and

p is the frequency of the gene B_2.

The expressions in p and q in the nine cells of the table are the frequencies of the nine genotypes resulting from the possible combinations of one of the B_1, b_1 genotypes (indicated on the upper border of the table) with one of the B_2, b_2 genotypes (indicated on the left border of the table). For example, q^2p^2 is the frequency of the $B_1B_1B_2B_2$ genotype. The genotypic values are symbolized by Y with subscripts identifying the genotype. The first subscript refers to the B_1 locus, the second to the B_2 locus, and the numerical value of the subscript specifies the number of favorable genes. Thus Y_{12} is the genotypic value of the genotype containing one B_1 and two B_2's, i.e. the genotype $B_1b_1B_2B_2$. The presence of a dot in place of a numerical subscript indicates a weighted average of the Y's for all genotypes of

the locus indicated. Thus, $Y_{2.}$ is the weighted mean for the genotypes, $B_1B_1B_2B_2$, $B_1B_1B_2b_2$, and $B_1B_1b_2b_2$, i.e.

$$Y_{2.} = q^2p^2Y_{22} + 2q^2p(1 - p)Y_{21} + q^2(1 - p)^2Y_{20}$$

$Y_{..}$ is the weighted mean for all genotypes for both loci, the general mean.

The total variance among the Y's (or in other words, the total genotypic variance) can be partitioned into three major parts,

(1) The portion between B_1, b_1 genotypes.

(2) The portion between B_2, b_2 genotypes.

(3) The remainder which arises from interaction among genotypes at the two loci.

The first two portions can be further sub-divided, in accordance with Section VII, into additive genetic variance and variance due to dominance deviations from segregation at the B_1 locus and the B_2 locus, respectively. It may be helpful to consider this partitioning of the total variance in terms of an analysis of variance table.

Source of variance	d.f.
B_1, b_1 genotypes	2
Regression on number of B_1 genes	1
Deviations from regression	1
B_2, b_2 genotypes	2
Regression on number of B_2 genes	1
Deviations from regression	1
Interactions among B_1, b_1 and B_2, b_2 genotypes	4
Total	8

Let us first consider the variance due to B_1, b_1 genotypes and its subdivision into additive genetic variance and variance due to dominance deviations arising from segregation at the B_1 locus. In terms of symbols established in Section VII,

$$Y_{2.} = z + 2u_1$$
$$Y_{1.} = z + u_1 + a_1u_1$$
$$Y_{0.} = z$$

(The subscript attached to a and u identifies them with respect to the locus for which they apply, the B_1 locus.) Obviously,

$$\left.\begin{aligned} u_1 &= (Y_{2\cdot} - Y_{0\cdot})/2 \\ a_1u_1 &= (2Y_{1\cdot} - Y_{2\cdot} - Y_{0\cdot})/2 \\ a_1 &= \frac{(2Y_{1\cdot} - Y_{2\cdot} - Y_{0\cdot})}{Y_{2\cdot} - Y_{0\cdot}} \end{aligned}\right\} \qquad (71)$$

These values could be substituted in equations (69) and (70) to furnish expressions for σ_q^2 and σ_d^2 in terms of the Y's, but the expressions become cumbersome and have not proven very useful. It has been found more convenient, when analyzing two locus genetic systems, to first assign numerical values to the Y's, making it possible to determine u and a as functions of q and p which can then be substituted into equations (69) and (70). The values assigned the Y's will of course depend on the genetic situation to be studied. Two examples are given near the end of this section.

The values of u_2, a_2u_2, and a_2 are obviously

$$\left.\begin{aligned} u_2 &= (Y_{\cdot 2} - Y_{\cdot 0})/2 \\ a_2u_2 &= (2Y_{\cdot 1} - Y_{\cdot 2} - Y_{\cdot 0})/2 \\ a^2 &= \frac{(2Y_{\cdot 1} - Y_{\cdot 2} - Y_{\cdot 0})}{Y_{\cdot 2} - Y_{\cdot 0}} \end{aligned}\right\} \qquad (72)$$

The total genotypic variance is

$$S(Y_{ij} - Y_{\cdot\cdot})^2 = q^2p^2Y_{22}^2 + 2q(1-q)p^2Y_{12}^2 + \cdots + (1-q)^2(1-p)^2Y_{00}^2 - Y_{\cdot\cdot}^2 \qquad (73)$$

As before the correction term is simply the squared mean because the total frequency is unity. Now, if we set

$$Y_{22} = Y_{\cdot\cdot} + (Y_{\cdot 2} - Y_{\cdot\cdot}) + (Y_{2\cdot} - Y_{\cdot\cdot}) + i_{22}$$

$$Y_{12} = Y_{\cdot\cdot} + (Y_{\cdot 2} - Y_{\cdot\cdot}) + (Y_{1\cdot} - Y_{\cdot\cdot}) + i_{12}$$

and in general

$$Y_{ij} = Y_{..} + (Y_{.j} - Y_{..}) + (Y_{i.} - Y_{..}) + i_{ij} \qquad (74)$$

We have defined a set of i-values such that

$$i_{ij} = Y_{ij} - (Y_{.j} - Y_{..}) - (Y_{i.} - Y_{..}) - Y_{..} \qquad (75)$$

If we substitute for the Y's in (73) in terms of (74), the following expression can be obtained

$$S(Y_{ij} - Y_{..})^2 = \left[p^2(Y_{.2} - Y_{..})^2 + 2p(1-p)(Y_{.1} - Y_{..})^2 + (1-p)^2(Y_{.0} - Y_{..})^2 \right]$$
$$+ \left[q^2(Y_{2.} - Y_{..})^2 + 2q(1-q)(Y_{1.} - Y_{..})^2 + (1-q)^2(Y_{0.} - Y_{..})^2 \right]$$
$$+ \left[q^2p^2i_{22}^2 + 2q(1-q)p^2i_{12}^2 + \cdots + (1-q)^2(1-p)^2i_{00}^2 \right]$$

The first of the three bracketed quantities is the sum of squares due to differences among means of the B_2, b_2 genotypes, the second is the sum of squares due to differences among means of the B_1, b_1 genotypes, and the third is the sum of squares arising from interactions among genotypes at the two loci. Remember that the total frequency is unity and that therefore the variance and the sum of squares are equal. Remembering, also, that the variance among means of the genotypes for a given locus is composed of additive genetic variance, σ^2_g, and variance due to dominance deviations, σ^2_d, the variance from segregation at two non-linked loci can be expressed as

$$\sigma^2_y = \sigma^2_{g_1} + \sigma^2_{g_2} + \sigma^2_{d_1} + \sigma^2_{d_2} + \sigma^2_i \qquad (75a)$$

where numerical subscripts identify the locus and σ^2_i is interaction variance. The i's may logically be called <u>interaction effects</u>. Lush (21) calls them epistatic deviations from the additive scheme. Interaction variance is absent when all i's have the value zero, i.e., when there are no interaction effects.

The analysis of variance table indicated but four degrees of freedom for interaction. This suggests that relationships exist among the i's such that all

could be expressed in terms of four of them. It is not difficult to show this to be the case. Note that

$$Y_{2\cdot} = p^2 Y_{22} + 2p(1-p)Y_{21} + (1-p)^2 Y_{20}$$

Substituting for the Y's in terms of equation (74), we get

$$\begin{aligned} Y_{2\cdot} &= p^2 \left[Y_{\cdot\cdot} + (Y_{\cdot 2} - Y_{\cdot\cdot}) + (Y_{2\cdot} - Y_{\cdot\cdot}) + i_{22} \right] \\ &\quad + 2p(1-p) \left[Y_{\cdot\cdot} + (Y_{\cdot 1} - Y_{\cdot\cdot}) + (Y_{2\cdot} - Y_{\cdot\cdot}) + i_{21} \right] \\ &\quad + (1-p)^2 \left[Y_{\cdot\cdot} + (Y_{\cdot 0} - Y_{\cdot\cdot}) + (Y_{2\cdot} - Y_{\cdot\cdot}) + i_{20} \right] \\ &= Y_{2\cdot} + p^2 i_{22} + 2p(1-p)i_{21} + (1-p)^2 i_{20} \end{aligned}$$

from which it is apparent that

$$p^2 i_{22} + 2p(1-p)i_{21} + (1-p)^2 i_{20} = 0 \quad .$$

Proceeding in the same way a total of six such expressions can be shown to hold. However, only five are required to specify all the relationships. The following is one set that embodies all the relationships among the nine i's.

$$\left.\begin{aligned} p^2 i_{22} + 2p(1-p)i_{21} + (1-p)^2 i_{20} &= 0 \\ p^2 i_{02} + 2p(1-p)i_{01} + (1-p)^2 i_{00} &= 0 \\ q^2 i_{22} + 2q(1-q)i_{12} + (1-q)^2 i_{02} &= 0 \\ q^2 i_{20} + 2q(1-q)i_{10} + (1-q)^2 i_{00} &= 0 \\ q^2 i_{21} + 2q(1-q)i_{11} + (1-q)^2 i_{01} &= 0 \end{aligned}\right\} \qquad (76)$$

Making use of these expersions all nine i's can be written in terms of only four of them.

When interaction among genotypes of the two loci is absent, i.e., when all i's equal zero,

$$\left.\begin{aligned} &Y_{22} - Y_{02} = Y_{21} - Y_{01} = Y_{20} - Y_{00} = Y_{2\cdot} - Y_{0\cdot} \\ \text{and} \\ &2Y_{12} - Y_{22} - Y_{02} = 2Y_{11} - Y_{21} - Y_{01} = 2Y_{10} - Y_{20} - Y_{00} = 2Y_{1\cdot} - Y_{2\cdot} - Y_{0\cdot} \end{aligned}\right\} \qquad (77)$$

These equalities hold only when all i's are zero. Now expanding equations (71) for u_1 and a_1u_1, we obtain

$$2u_1 = p^2(Y_{22} - Y_{02}) + 2p(1 - p)(Y_{21} - Y_{01}) + (1 - p)^2(Y_{20} - Y_{00})$$

and

$$2a_1u_1 = p^2(2Y_{12} - Y_{22} - Y_{02}) + 2p(1-p)(2Y_{11} - Y_{21} - Y_{01}) + (1-p)^2(2Y_{10} - Y_{20} - Y_{00})$$

From these expressions we see that when equations (77) hold (in other words, when interaction is absent) the values of u_1 and a_1u_1, which measure differences among the effects of the B_1, b_1 genotypes, are not dependent on p, the frequency of B_2. Of course, this is only a high-powered way of showing that interaction is absent when interaction is absent. Viewed in this way, the content of this paragraph seems a silly matter of going around in circles. However, the fact that, when interactions are present, u_1 and a_1u_1 vary with frequency of a gene other than B_1 should upon reflection prove illuminating with respect to just what the values u and a represent. It should be apparent that these quantities are not constants for any particular locus, but that they may have one value in one population, another value in another population. An obvious extension is the fact that the additive genetic variance and variance due to dominance deviations arising from segregation at a particular locus can be affected by the frequency of genes at other loci as well as by the frequency of genes at the locus in question.

The final point to be made is that in the absence of interactions among non-allelic genes the total genetic variance from the segregation of a number of gene pairs is simply the sum of the additive genetic variance for all pairs plus the sum of the variance due to dominance deviations for all pairs. Symbolically,

$$\sigma_y^2 = \sum \sigma_g^2 + \sum \sigma_d^2$$

when there are no non-allelic interactions.

Examples

It is instructive to study the composition of the variance from a pair of loci assuming specific interaction systems.

Case 1

Consider the classical case of complimentary action between two genes, in which either gene by itself has no effect but the presence of both results in an effect which does not depend in any way on whether either gene is present in the duplex or simplex condition. Examples of this sort of two factor inheritance are given and discussed by Sinnott and Dunn (22). The situation is adequately specified by setting

$$Y_{22} = Y_{21} = Y_{12} = Y_{11} = 1$$

and

$$Y_{20} = Y_{02} = Y_{10} = Y_{01} = Y_{00} = 0$$

Then $Y_{2\cdot} = Y_{1\cdot 0} = p^2 + 2p(1 - p) = p(2 - p)$,

$Y_{0\cdot} = 0$

and from (71)

$$u_1 = p(2 - p)/2 \qquad \text{and} \qquad a_1 = 1.0$$

From (69)

$$\sigma^2_{g_1} = 2q(1 - q)\left[1 + (1 - 2q)a_1\right]^2 u_1^2$$

$$= 2q(1 - q)^3 p^2 (2 - p)^2$$

and from (70)

$$\sigma^2_{d_1} = 4q^2(1 - q)^2 a_1^2 u_1^2 = q^2(1 - q)^2 p^2 (2 - p)^2$$

Proceeding in the same way

$$\sigma^2_{g_2} = 2p(1 - p)^3 q^2 (2 - q)^2$$

and

$$\sigma^2_{d_2} = p^2(1 - p)^2 q^2 (2 - q)^2$$

Mimeo. No. 8
R. E. Comstock:1:50

To obtain the total variance $Y_{..}$ must first be computed

$$Y_{..} = q^2p^2 + 2q(1 - q)p^2 + 2q^2p(1 - p) + 4q(1 - q)p(1 - p)$$

The total variance is then

$$\sigma_y^2 = q^2p^2 + 2q(1 - q)p^2 + 2q^2p(1 - p) + 4q(1 - q)p(1 - p)$$

$$- \left[q^2p^2 + 2q(1 - q)p^2 + 2q^2p(1 - p) + 4q(1 - q)p(1 - p)\right]^2$$

$$= pq(2 - p)(2 - q)\left[1 - pq(2 - p)(2 - q)\right]$$

From (75a) the interaction variance is

$$\sigma_i^2 = \sigma_y^2 - \sigma_{g_1}^2 - \sigma_{g_2}^2 - \sigma_{d_1}^2 - \sigma_{d_2}^2 \quad .$$

It turns out that

$$\sigma_i^2 = pq(2 - p)(2 - q)\left[1 - p(2 - p) - (1 - p)^2q(2 - q)\right] \quad .$$

Note that when either p or q equals either zero or one, $\sigma_i^2 = 0$. This is as it should be since in those cases the genotype at one of the loci is constant for the entire population, i.e., there is no variation at the one locus to interact with variation at the other.

Table 5 lists values of σ_Y^2, σ_g^2, σ_d^2, and σ_i^2 for various values of p and q. $\sigma_{g_1}^2$ and $\sigma_{g_2}^2$ are summed to give total additive genetic variance, σ_g^2, and $\sigma_{d_1}^2$ and $\sigma_{d_2}^2$ are summed to give total variance due to dominance deviations, σ_d^2.

Significant aspects of Table 5 are:

1. σ_i^2 is but a small fraction of σ_Y^2 except when both p and q are small.
2. σ_d^2 is small as a fraction of σ_Y^2 except when both p and q are large.
3. σ_g^2 is the major component of σ_Y^2 except when both p and q are either large or small.

Table 5. Components of genetic variance from action of two complimentary gene pairs.

p		q = .1	.3	.5	.7	.9
.1	σ^2_g	.0105	.0454	.0865	.1221	.1430
	σ^2_d	.0006	.0037	.0068	.0083	.0082
	σ^2_i	.0237	.0384	.0289	.0126	.0015
	σ^2_Y	.0348	.0875	.1222	.1430	.1527
.3	σ^2_g		.1071	.1483	.1803	.2022
	σ^2_d		.0229	.0411	.0480	.0453
	σ^2_i		.0624	.0468	.0204	.0025
	σ^2_Y		.1924	.2362	.2487	.2500
.5	σ^2_g			.1406	.1248	.1235
	σ^2_d			.0703	.0766	.0658
	σ^2_i			.0352	.0153	.0019
	σ^2_Y			.2461	.2167	.1912
.7	σ^2_g				.0626	.0385
	σ^2_d				.0703	.0499
	σ^2_i				.0068	.0009
	σ^2_Y				.1424	.0893
.9	σ^2_g					.0035
	σ^2_d					.0159
	σ^2_i					.0001
	σ^2_Y					.0195

The student should note that the genetic model just considered is related to the sort that Beadle(23) and his associates have found in control of physiological processes in Neurospora. For example, the synthesis of arginine by Neurospora dependson the presence of no less than seven genes. It appears that each step in

a chain of chemical reactions is catalyzed by a specific gene and that if one of these genes is absent the chain is blocked at the point of the reaction for which the absent gene is catalyst. Thus all seven (perhaps more) genes must be present or orginine cannot be formed. The difference between the model considered above and the situation in Neurospora lies in the fact that the cells of the vegetative portion of the life cycle of this mold contain only the haploid chromosome complement and hence dominance and recessiveness of genes is not a factor. However, a few instances are known of chemical reactions in humans under control of specific genes and hence there is reason to believe that, fundamentally, classical complementary gene action is an important genetic model. Consequently, it will receive special attention in a later section devoted to the effect of selection. At that point systems involving more than two gene pairs will be considered.

Case 2

Consider two pairs of genes to which the response of a character is completely additive (no dominance, no interactions of non-allelic genes). However, assume that the optimum for the character with respect to adaptability or selective value of the organism is not an extreme value but rather an intermediate value. This model has been considered by Wright (18). It is not difficult to visualize characters for which the model may apply. Corn can be too tall or too short and it is quite likely that the liver of a cow can be too large or too small for optimum balance with the rest of the animal. It is not impossible that the genes controlling height of corn and size of liver in cattle may act in an essentially additive fashion on height and size, respectively.[1/] However, because adaptability in these

[1/] The critical will point out that there is heterosis for height of corn and that therefore gene action cannot be additive. True; general vigor is reflected in height as well as in other traits. On the other hand, equally vigorous strains or lines (as measured by criteria other than height) may vary greatly in height. Thus we must admit genes which affect height independent of vigor and these may exert their effects in an additive fashion. In harmony with this, there are tall inbreds which when crossed produce still taller F_1's, and short inbreds of which F_1's exhibit heterosis for height but may be even shorter than certain inbreds.

cases is not a linear function of height or size their effects in terms of adaptability would not be completely additive.

As a specific example, assume that, measured in terms of adaptability

$$Y_{22} = Y_{00} = 0$$

$$Y_{21} = Y_{12} = Y_{10} = Y_{01} = 3$$

$$Y_{02} = Y_{20} = Y_{11} = 4$$

Note that the optimum genotype is one with two plus genes, regardless of the locus at which they are present; and that genotypes with more or less plus genes are not so favorable, the least favorable genotypes being those in which the number of plus genes deviates by two from the optimum number.

$$Y_{2.} = 6p(1 - p) + 4(1 - p)^2 = 4 - 2p - 2p^2$$

$$Y_{1.} = 3p^2 + 8p(1 - p) + 3(1 - p)^2 = 3 + 2p - 2p^2$$

$$Y_{0.} = 4p^2 + 6p(1 - p) = 6p - 2p^2$$

From equation (71)

$$u_1 = (Y_{2.} - Y_{0.})/2 = 2 - 4p = 2(1 - 2p)$$

$$a_1u_1 = (2Y_{1.} - Y_{2.} - Y_{0.})/2 = 1$$

$$a_1 = 1/2(1 - 2p)$$

Substituting in equations (69) and (70) and reducing

$$\sigma^2_{g_1} = 2q(1 - q)(3 - 2q - 4p)^2$$

$$\sigma^2_{d_1} = 4q^2(1 - q)^2$$

From the symmetrical nature of the model it is obvious that

$$\sigma^2_{g_2} = 2p(1 - p)(3 - 2p - 4q)^2$$

$$\sigma^2_{d_2} = 4p^2(1 - p)^2$$

Computing σ^2_y directly for different values of p and q, and obtaining σ^2_i by subtraction of σ^2_g and σ^2_d from σ^2_y, we obtain the values listed in Table 6.

The principle points to be noted from Table 6 at this time are:

(1) That it deviates in its composition from Table 5. The significance of this is that the consequences of gene interactions vary from one system to another.

(2) That when p and q equal .5 additive genetic variance is zero though total genetic variance is considerable. This situation is true at this particular

Table 6. Components of genetic variance from the action of two pairs of genes to which the primary response of a character is strictly additive, the value of the character in terms of total fitness of the organism being greatest at an intermediate expression of the character.

		q				
p		.1	.3	.5	.7	.9
.1	σ^2_q	2.08	2.14	1.40	.61	.23
	σ^2_d	.06	.21	.28	.21	.06
	σ^2_i	.13	.30	.36	.30	.13
	σ^2_y	2.27	2.65	2.04	1.12	.42
.3	σ^2_g		1.21	.39	.13	.61
	σ^2_d		.35	.42	.35	.21
	σ^2_i		.71	.84	.71	.30
	σ^2_y		2.27	1.65	1.19	1.12
.5	σ^2_g			.00	.39	1.40
	σ^2_d			.50	.42	.28
	σ^2_i			1.00	.84	.36
	σ^2_y			1.50	1.65	2.04
.7	σ^2_g				1.21	2.14
	σ^2_d				.35	.21
	σ^2_i				.71	.30
	σ^2_y				2.27	2.65
.9	σ^2_g					2.08
	σ^2_d					.06
	σ^2_i					.13
	σ^2_y					2.27

point relative to gene frequencies because the optimum number of plus genes was assumed to be two which is one half the total number possible. Had the optimum number of plus genes been one or three, there would have been additive genetic variance when p and q were one-half, but it would have been absent at certain other combinations of the gene frequencies.

Problems:

16. Given $Y_{00} = 0$, all other Y's equal 1.0. Determine σ^2_g, σ^2_d, σ^2_i, and σ^2_y for

    a. $p = q = .5$

    b. $p = q = .2$

    c. $p = q = .8$

    d. $p = .2$, $q = .8$

    e. $p = .8$, $q = .2$

17. Given $Y_{22} = 4$, $Y_{21} = Y_{12} = 3$, $Y_{20} = Y_{02} = Y_{11} = 2$, $Y_{10} = Y_{01} = 1$, $Y_{00} = 0$. Determine σ^2_g, σ^2_d, σ^2_i, and σ^2_y for the same pairs of values of p and q as in problem 16.

References:

21. Lush, J. L. (1943). Animal Breeding Plans, 2nd Ed. The Iowa State College Press. Ames, Iowa.

22. Sinnott, Edmund W. and L. C. Dunn (1939). Principles of Genetics, 3rd Ed. McGraw-Hill Book. Co., New York and London.

23. Beadle, G. W. (1946) Genes and the Chemistry of the Organism. Am. Scientist 34:31-53.

Mather's symbols p 53

$d = u$ $\qquad D = \sum \frac{u^2}{2}$

$h = au$

$a = \frac{h}{d}$ $\qquad H = \sum \frac{a^2 u^2}{4}$

Intra-sire Regression of offspring on dam