SOME BIOMETRICAL FORMULAE FOR THE ANALYSIS OF QUANTITATIVE INHERITANCE SYSTEMS INVOLVING TWO HAPLOID OR INBRED DIPLOID PARENTS¹

D. S. ROBSON

Cornell University, Ithaca, New York

Received January 14, 1957

THE general model underlying all methods of quantitative inheritance analysis asserts that in any genetic population consisting of a mixture of genotypes the population frequency distribution of phenotypes is a corresponding mixture of the individual phenotypic distributions associated with each of the genotypes. Thus, if G is a genetic population consisting of the n different genotypes $g_1, g_2, \dots, g_n$ mixed with the relative frequencies $P(g_1 | G), P(g_2 | G), \dots, P(g_n | G)$, then the relative frequency $f(x | G)$ of the phenotype x in the population G is

$$(1) \quad f(x | G) = P(g_1 | G) \cdot f(x | g_1) + P(g_2 | G) \cdot f(x | g_2) + \dots + P(g_n | G) \cdot f(x | g_n).$$

By analogy to qualitative inheritance analysis, the complete solution of a quantitative inheritance problem includes the determination of (a) all the genotypes g occurring in the system, (b) the phenotypic characterization of each genotype g by means of its distribution function $f(x | g)$, and (c) the relative frequency $P(g | G)$ of the genotype g in each genetic population G of the system. In view of the many complicating factors known to occur in qualitative inheritance analysis this solution to a quantitative problem appears unobtainable; examples of quantitative heritable characters have been found, however, for which the genetic system was sufficiently simple to permit the fulfillment of (a), (b), and (c). Two such examples are given by ALLARD (1956) and POWERS (1955). Furthermore, where a complete solution is unobtainable a partial or approximate solution may be possible and, for some practical purposes, sufficient.

This paper is directed toward a partial solution of quantitative inheritance problems involving two parents, haploids or inbred diploids, and the population series generated from their F_1 cross by further crossing and selfing. The partial solution aimed for is the determination of the number N of loci at which the two parents are genetically different (with respect to the character in question) and the estimation of phenotypic frequency distributions associated with particular *collections* of genotypes, as opposed to *individual* genotypes. The formulas developed here may be regarded as generalizations of certain aspects of the *partitioning method of genetic analysis* described by POWERS (1934).

BACKCROSSING IN HAPLOIDS

Haploid organisms, particularly the lower forms, are convenient genetic material for the study of quantitative inheritance. There is, for example, the obvious advantage of the simplicity of haploid inheritance, the relatively short time interval

¹ Paper No. 337 of the Department of Plant Breeding, Cornell University, Ithaca, New York.

between generations, the ability to produce large numbers of progeny, usually both sexually and asexually, and a life form which permits the experimental manipulation of the large numbers of individuals essential to a quantitative inheritance study. Backcrossing, starting from the cross of two parental genotypes, produces a series of generations in which the number of populations per generation increases geometrically, resulting in a rapid accumulation of the necessary data.

With respect to a single pair of alleles, say A_0 and A_1 , the genotypic distribution in the two parental lines and their F_1 cross is

$$P(A_0 | P_0) = 1$$

$$P(A_1 | P_1) = 1$$

$$P(A_1 | F_1) = \frac{1}{2} = P(A_0 | F_1).$$

The two first generation backcross populations, $B_0 = F_1 \times P_0$ and $B_1 = F_1 \times P_1$, contain the genotypes A_0 and A_1 in the proportions

$$P(A_1 | B_0) = \frac{1}{4} = 1 - P(A_0 | B_0)$$

$$P(A_1 | B_1) = \frac{3}{4} = 1 - P(A_0 | B_1).$$

If we let $\delta_1 = 0$ or 1 according as the first backcross is to parent P_0 or P_1 then we may denote a general first backcross population by

$$B_{\delta_1} = F_1 \times P_{\delta_1}$$

and its genotypic frequency distribution by

$$P(A_1 | B_{\delta_1}) = \frac{1 + 2\delta_1}{2^2} = 1 - P(A_0 | B_{\delta_1}).$$

Extending this notation, we designate a general second generation backcross population by the symbol

$$B_{\delta_1, \delta_2} = B_{\delta_1} \times P_{\delta_2}$$

and it is easy to verify that the genotypic distribution in this population may be written

$$P(A_1 | B_{\delta_1, \delta_2}) = \frac{1 + 2\delta_1 + 2^2\delta_2}{2^3} = 1 - P(A_0 | B_{\delta_1, \delta_2}).$$

In general, then, we denote a k th generation backcross population by

$$B_{\delta_1, \dots, \delta_k} = B_{\delta_1, \dots, \delta_{k-1}} \times P_{\delta_k}$$

where $\delta_1, \dots, \delta_k$ gives the complete ancestry of the population, and we have

$$P(A_1 | B_{\delta_1, \dots, \delta_k}) = \frac{1 + 2\delta_1 + 2^2\delta_2 + \dots + 2^k\delta_k}{2^{k+1}} = 1 - P(A_0 | B_{\delta_1, \dots, \delta_k}).$$

With this understanding we shall simplify the notation by writing simply B for $B_{\delta_1, \dots, \delta_k}$ and a for $P(A_1 | B_{\delta_1, \dots, \delta_k})$.

N independently segregating loci. The two parental genotypes will, in general, differ at some unknown number *N* of loci, and we denote these two genotypes by

$$P_0^{(N)} = A_{01}A_{02} \cdots A_{0N}$$

$$P_1^{(N)} = A_{11}A_{12} \cdots A_{1N}.$$

The F_1 and every backcross population each contain all of the 2^N different possible genotypes, and if segregation is independent at the *N* loci then the relative frequency of any given genotype in the population *B* is obtained simply as the coefficient of that genotype in the expansion of

$$\prod_{j=1}^N [(1-a)A_{0j} + aA_{1j}].$$

Upon expanding we find that every genotype carrying *v* genes from parent P_1 and $N-v$ genes from parent P_0 occurs with the same relative frequency, namely, $a^v(1-a)^{N-v}$. The number of such genotypes is

$$\binom{N}{v} = \frac{N(N-1) \cdots (N-v+1)}{v(v-1) \cdots 1};$$

hence, the probability that an individual randomly selected from the population *B* will carry *v* genes from parent P_1 and $N-v$ genes from parent P_0 is

$$\binom{N}{v} a^v (1-a)^{N-v}.$$

We shall use the notation $P_1^{(v)}P_0^{(N-v)}$ to denote this genetic subpopulation consisting of $\binom{N}{v}$ equally frequent genotypes, so we may write

$$P(P_1^{(v)}P_0^{(N-v)} | B) = \binom{N}{v} a^v (1-a)^{N-v}.$$

The fundamental assumption (1) then asserts that the phenotypic frequency distribution for the population *B* may be written as

$$(2) \quad f(x | B) = \sum_{v=0}^N \binom{N}{v} a^v (1-a)^{N-v} f(x | P_1^{(v)}P_0^{(N-v)}).$$

Thus, the theoretical phenotypic frequency distribution for each backcross population is expressible as a linear combination of the $N+1$ unknown distributions $f(x | P_1^{(v)}P_0^{(N-v)})$, $v = 0, 1, \dots, N$.

The observed phenotypic frequency distribution for population *B* does not satisfy equation (2), but is subject to sampling error. If *n* is the number of individuals in the experimental population *B* and $f_n(x | B)$ is the fraction of these individuals in a given phenotypic class *x* the error

$$e(x) = f_n(x | B) - f(x | B)$$

is subject to a sampling variance of

$$\sigma_{e(x)}^2 = \frac{f(x | B) [1 - f(x | B)]}{n}.$$

This random error is added to both sides of (2) to give

$$(3) \quad f_n(x|B) = f(x|B) + e(x) = \sum_{v=0}^N \binom{N}{v} a^v (1-a)^{N-v} f(x|P_1^{(v)} P_0^{(N-v)}) + e(x).$$

A genetic experiment containing all of the possible backcross populations through generation k produces $2^{k+1} + 1$ different equations of the form (3), including those for the two parents and the F_1 . When N is known and $N \leq 2^{k+1}$ the statistical solution to these equations represents a classical regression problem. The normalized equations obtained by dividing both sides of (3) by $\sigma_{e(x)}$,

$$(4) \quad \frac{f_n(x|B)}{\sigma_{e(x)}} = \sum_{v=0}^N \frac{\binom{N}{v} a^v (1-a)^{N-v}}{\sigma_{e(x)}} f(x|P_1^{(v)} P_0^{(N-v)}) + \frac{e(x)}{\sigma_{e(x)}},$$

satisfy the conditions of Markov's theorem on least squares, and the limiting distribution of the residual sum of squares,

$$\sum_B \left[\frac{f_n(x|B)}{\sigma_{e(x)}} - \sum_{v=0}^N \frac{\binom{N}{v} a^v (1-a)^{N-v}}{\sigma_{e(x)}} \hat{f}(x|P_1^{(v)} P_0^{(N-v)}) \right]^2,$$

where $\hat{f}(x|P_1^{(v)} P_0^{(N-v)})$ is the least squares solution to (4), is the chi square distribution with $2^{k+1} - N$ degrees of freedom. A comparison of the residual sum of squares with the tabulated chi square therefore constitutes a test of the model. Conventionally, when the weighting factors, in this case $\sigma_{e(x)}$, are unknown they are replaced by their estimates, here

$$(5) \quad \hat{\sigma}_{e(x)}^2 = \frac{f_n(x|B) [1 - f_n(x|B)]}{n}.$$

The model further implies that the parameters $f(x|P_1^{(v)} P_0^{(N-v)})$ being estimated here are non negative and that, when summed over all phenotypic classes x , each adds to unity. Since these conditions are not imposed upon the estimation procedure their fulfillment or nonfulfillment constitutes a further, informal test of the model.

This method of solution applies when N is known; but for a given choice of parental material, P_0 and P_1 , there is exactly one value of N which gives a solution to (2) for all k . When N is unknown, therefore, the integer $\hat{N}$ which, by trial and error, yields a minimum residual sum of squares and non negative solutions $\hat{f}(x|P_1^{(v)} P_0^{(N-v)})$ is an estimate of N . Trial and error, however, where each trial consists of a multiple regression analysis, is a tedious process. To facilitate the estimation of N we may compute the algebraic solutions to sets of $N + 1$ equations of the form (2) and so establish linear relationships among the phenotypic frequency distributions $f(x|B)$ which, again, hold for exactly one value of N . The two recurrent backcross population series are of particular interest in this connection because of the convenient form of the algebraic solutions for $f(x|P_0^{(N)})$ and $f(x|P_1^{(N)})$.

The population $B_{\delta_1, \dots, \delta_k}$ for which $\delta_1 = \delta_2 = \dots = \delta_k = \delta$ is called the k th recurrent backcross population where P_δ is the recurrent parent, $\delta = 0$ or 1 , and will be denoted by the symbol B_δ^k . The phenotypic frequency distribution for population B_δ^k has the simple form

$$(6) \quad f(x|B_\delta^k) = \sum_{v=0}^N \binom{N}{v} \left(\frac{1}{2^{k+1}} \right)^v \left(1 - \frac{1}{2^{k+1}} \right)^{N-v} f(x|P_{1-\delta}^{(v)} P_\delta^{(N-v)}).$$

The $N + 1$ equations for the successive generations $B_\delta^k, B_\delta^{k+1}, \dots, B_\delta^{k+N}$ are easily solved to give, for all $k \geq 0$,

$$(7) \quad f(x | P_\delta^{(N)}) = \frac{\sum_{v=0}^N (-1)^v \alpha_v(N) f(x | B_\delta^{k+N-v})}{\sum_{v=0}^N (-1)^v \alpha_v(N)}$$

where the constants $\alpha_v(N)$, $v = 0, 1, \dots, N$, are determined from the conditions

$$\alpha_0(N) = 1$$

$$\alpha_v(N) = \frac{1}{2^v} [\alpha_{v-1}(N-1) + \alpha_v(N-1)], \quad 0 < v < N$$

$$\alpha_N(N) = \frac{1}{\frac{N(N+1)}{2}}$$

$$\alpha_{N+i}(N) = 0, \quad i > 0.$$

For example,

$$\begin{aligned} \alpha_0(1) &= 1, & \alpha_1(1) &= \frac{1}{2} \\ \alpha_0(2) &= 1, & \alpha_1(2) &= \frac{3}{4}, & \alpha_2(2) &= \frac{1}{8} \\ \alpha_0(3) &= 1, & \alpha_1(3) &= \frac{7}{8}, & \alpha_2(3) &= \frac{7}{32}, & \alpha_3(3) &= \frac{1}{64}. \end{aligned}$$

As an alternative to equation (6) we have the recursion formula

$$(8) \quad f(x | B_\delta^{k+N+1}) = \sum_{v=0}^N (-1)^v 2^{v+1} \alpha_{v+1}(N+1) f(x | B_\delta^{k+N-v});$$

thus, if the two parents are differentiated by $N = 1$ gene then for all $k \geq 0$

$$f(x | B_\delta^{k+2}) = \frac{3}{2} f(x | B_\delta^{k+1}) - \frac{1}{2} f(x | B_\delta^k)$$

while for $N = 2$

$$(9) \quad f(x | B_\delta^{k+3}) = \frac{7}{4} f(x | B_\delta^{k+2}) - \frac{7}{8} f(x | B_\delta^{k+1}) + \frac{1}{8} f(x | B_\delta^k)$$

and for $N = 3$

$$(10) \quad f(x | B_\delta^{k+4}) = \frac{15}{8} f(x | B_\delta^{k+3}) - \frac{35}{32} f(x | B_\delta^{k+2}) + \frac{15}{64} f(x | B_\delta^{k+1}) - \frac{1}{64} f(x | B_\delta^k).$$

Formula (8), with f replaced by f_n , may then be used in place of (4) to obtain the estimate $\hat{N}$ by trial and error. The procedure would be to first see if the consecutive recurrent backcross populations fit the model with $N = 1$; if not, then try $N = 2$, and so on. Each relation (8) may be tested by comparing the statistic

$$(11) \quad t = \frac{f_n(x | B_\delta^{k+N+1}) - \sum_{v=0}^N (-1)^v 2^{v+1} \alpha_{v+1}(N+1) f_n(x | B_\delta^{k+N-v})}{\sqrt{\hat{\sigma}_{e(x)}^2(B_\delta^{k+N+1}) + \sum_{v=0}^N 4^{v+1} \alpha_{v+1}^2(N+1) \hat{\sigma}_{e(x)}^2(B_\delta^{k+N-v})}}$$

with the normal deviate $t_{.05} = 1.96$. The variance estimates $\hat{\sigma}_{\hat{\mu}_r'(B_\delta^m)}^2$ are computed from formula (5); formula (11) indicates that each experimental population B_δ^m contains the same number n of individuals, but this is merely a notational convenience and is not necessary.

Any linear relationship which holds among frequency distributions likewise holds for the moments of the frequency distributions. The r th moment of a frequency distribution $f(x)$, written $\mu_r'(f)$, is defined as

$$\mu_r'(f) = \sum_x x^r f(x).$$

Consequently, equation (6) implies that

$$(12) \quad \mu_r'(B) = \sum_{v=0}^N \binom{N}{v} a^v (1-a)^{N-v} \mu_r'(P_1^{(v)} P_0^{(N-v)}).$$

Since the r th sample moment for population B ,

$$\hat{\mu}_r'(B) = \sum_x x^r f_n(x | B),$$

is an unbiased estimate of $\mu_r'(B)$ with variance

$$\sigma_{\hat{\mu}_r'(B)}^2 = \frac{1}{n} \sum_x [x^r - \mu_r'(B)]^2 f(x | B)$$

then, writing $u(B) = \hat{\mu}_r'(B) - \mu_r'(B)$, the normalized equation corresponding to (4) becomes

$$\frac{\hat{\mu}_r'(B)}{\sigma_{\hat{\mu}_r'(B)}} = \sum_{v=0}^N \frac{\binom{N}{v} a^v (1-a)^{N-v}}{\sigma_{\hat{\mu}_r'(B)}} \cdot \mu_r'(P_1^{(v)} P_0^{(N-v)}) + \frac{u(B)}{\sigma_{\hat{\mu}_r'(B)}}$$

which, again, may be attacked by the multiple regression analysis, substituting for $\sigma_{\hat{\mu}_r'(B)}$ the estimate obtained from

$$\hat{\sigma}_{\hat{\mu}_r'(B)}^2 = \frac{1}{n} \sum_x [x^r - \hat{\mu}_r'(B)]^2 f_n(x | B).$$

Similarly, the relations (7) and (8) hold for the moments of the recurrent backcross populations; corresponding to (8) we have the recursion

$$\mu_r'(B_\delta^{k+N+1}) = \sum_{v=0}^N (-1)^v 2^{v+1} \alpha_{v+1} (N+1) \mu_r'(B_\delta^{k+N-v})$$

which may be tested by the criterion

$$t = \frac{\hat{\mu}_r'(B_\delta^{k+N+1}) - \sum_{v=0}^N (-1)^v 2^{v+1} \alpha_{v+1} (N+1) \hat{\mu}_r'(B_\delta^{k+N-v})}{\sqrt{\hat{\sigma}_{\hat{\mu}_r'(B_\delta^{k+N+1})}^2 + \sum_{v=0}^N 4^{v+1} \alpha_{v+1}^2 (N+1) \hat{\sigma}_{\hat{\mu}_r'(B_\delta^{k+N-v})}^2}}.$$

Thus, one method of determining the estimate $\hat{N}$ is by trial and error using the first one or two sample moments. For example, if the sample means $\bar{B}_\delta^k = \hat{\mu}_r'(B_\delta^k)$ satisfy the relations

$$\begin{aligned}\bar{B}_0^2 &= \frac{3}{2} \bar{B}_0^1 - \frac{1}{2} \bar{F}_1 & \bar{B}_1^2 &= \frac{3}{2} \bar{B}_1^1 - \frac{1}{2} \bar{F}_1 \\ \bar{B}_0^3 &= \frac{3}{2} \bar{B}_0^2 - \frac{1}{2} \bar{B}_0^1 & \bar{B}_1^3 &= \frac{3}{2} \bar{B}_1^2 - \frac{1}{2} \bar{B}_1^1 \\ &\vdots & &\vdots \\ \bar{B}_0^{k+2} &= \frac{3}{2} \bar{B}_0^{k+1} - \frac{1}{2} \bar{B}_0^k & \bar{B}_1^{k+2} &= \frac{3}{2} \bar{B}_1^{k+1} - \frac{1}{2} \bar{B}_1^k\end{aligned}$$

to within sampling error this would strongly suggest that $N = 1$ would give the optimum solution to (4). Or if one or more of these relations does not hold but

$$\begin{aligned}\bar{B}_0^3 &= \frac{7}{4} \bar{B}_0^2 - \frac{7}{8} \bar{B}_0^1 + \frac{1}{8} \bar{F}_1 & \bar{B}_1^3 &= \frac{7}{4} \bar{B}_1^2 - \frac{7}{8} \bar{B}_1^1 + \frac{1}{8} \bar{F}_1 \\ \bar{B}_0^4 &= \frac{7}{4} \bar{B}_0^3 - \frac{7}{8} \bar{B}_0^2 + \frac{1}{8} \bar{B}_0^1 & \bar{B}_1^4 &= \frac{7}{4} \bar{B}_1^3 - \frac{7}{8} \bar{B}_1^2 + \frac{1}{8} \bar{B}_1^1 \\ &\vdots & &\vdots \\ \bar{B}_0^{k+3} &= \frac{7}{4} \bar{B}_0^{k+2} - \frac{7}{8} \bar{B}_0^{k+1} + \frac{1}{8} \bar{B}_0^k & \bar{B}_1^{k+3} &= \frac{7}{4} \bar{B}_1^{k+2} - \frac{7}{8} \bar{B}_1^{k+1} + \frac{1}{8} \bar{B}_1^k\end{aligned}$$

were all satisfied to within sampling error then the conclusion would be that $N = 2$ should give the best solution to (4), and so on.

Two linked genes. If the parents P_0 and P_1 differ in exactly two genes controlling the quantitative character, both of which occur on the same chromosome, this fact is relatively easy to demonstrate by recurrent backcrossing. Letting $p/2$ denote the relative frequency of each parental genotype in the F_1 population we may write the phenotypic frequency distribution for the k th recurrent backcross population B_δ^k as

$$\begin{aligned}f(x | B_\delta^k) &= \left[1 - \frac{1}{2^k} + \left(\frac{p}{2} \right)^{k+1} \right] f(x | P_\delta^{(2)}) \\ &\quad + \frac{1 - p^{k+1}}{2^{k+1}} f(x | P_1^{(1)} P_0^{(1)}) + \left(\frac{p}{2} \right)^{k+1} f(x | P_{1-\delta}^{(2)}).\end{aligned}$$

The equations for the three populations $B_\delta^k, B_\delta^{k+1}, B_\delta^{k+2}$ then give the solution

$$(13) \quad f(x | P_\delta^{(2)}) = \frac{f(x | B_\delta^{k+2}) - \frac{1+p}{2} f(x | B_\delta^{k+1}) + \frac{p}{4} f(x | B_\delta^k)}{1 - \frac{1+p}{2} + \frac{p}{4}}$$

Alternatively, since $f(x | P_\delta^{(2)}) = f(x | P_\delta)$ can be estimated from the experimental population P_δ , we may regard (13) as a solution for p ; namely,

$$(14) \quad \frac{p}{2} = \frac{2f(x | B_\delta^{k+2}) - f(x | B_\delta^{k+1}) - f(x | P_\delta)}{2f(x | B_\delta^{k+1}) - f(x | B_\delta^k) - f(x | P_\delta)}$$

for all k . The recursion corresponding to (8) is then

$$(15) \quad f(x | B_\delta^{k+3}) = \frac{p+3}{2} f(x | B_\delta^{k+2}) - \frac{2+3p}{4} f(x | B_\delta^{k+1}) + \frac{p}{4} f(x | B_\delta^k).$$

Formulas (14) and (15) likewise hold with $f(x | B_\delta^m)$ replaced by the r th moment $\mu_r'(B_\delta^m)$.

If linkage occurs when $N > 2$ the problem becomes much more complicated. For example, if $N = 3$ genes are distributed over two chromosomes then the relations corresponding to (13) and (15) become, respectively,

$$(16) \quad f(x | P_\delta^{(3)}) = \frac{1}{24 - 18p + 3p^2} [64f(x | B_\delta^{k+4}) - 48(1 + p)f(x | B_\delta^{k+3}) \\ + (8 + 36p + 8p^2)f(x | B_\delta^{k+2}) - 6p(1 + p)f(x | B_\delta^{k+1}) + p^2f(x | B_\delta^k)]$$

$$(17) \quad f(x | B_\delta^{k+5}) = \frac{7 + 3p}{4} f(x | B_\delta^{k+4}) - \frac{14 + 21p + 2p^2}{16} f(x | B_\delta^{k+3}) \\ + \frac{4 + 21p + 7p^2}{32} f(x | B_\delta^{k+2}) - \frac{p(6 + 7p)}{64} f(x | B_\delta^{k+1}) + \frac{p^2}{64} f(x | B_\delta^k)$$

where p has the same meaning as before with respect to the two linked genes. Equation (16) is quadratic in p and, in general, will be of degree $N - 1$ in p if the N genes occur on $N - 1$ chromosomes. Also, we observe that the recursion (17) involves the frequency distributions for six different populations as compared to five in the corresponding expression (10) for the case of $N = 3$ independently segregating genes; this is due to the fact that all genotypes carrying v genes from parent P_1 and $N - v$ from parent P_0 no longer appear with equal frequency in the backcross populations.

BACKCROSSING AND SELFING IN DIPLOIDS

The formulas presented above for a haploid backcrossing system also apply, with appropriate changes in notation, to a diploid system of backcrossing and selfing starting with the cross of two homozygous parents. If the N factors by which the two parents differ (with respect to the quantitative character in question) segregate independently then in the population $B_{\delta_1, \dots, \delta_k, 0}$ the $\binom{N}{v}$ different genotypes which are heterozygous at exactly v loci and hence homozygous for P_0 genes at the remaining $N - v$ loci all appear with equal frequency; namely, $a^v(1 - a)^{N-v}$ where, again,

$$a = \frac{1 + 2\delta_1 + 2^2\delta_2 + \dots + 2^k\delta_k}{2^{k+1}}.$$

Hence, if we define $F_1^{(v)}P_0^{(N-v)}$ to be the subpopulation consisting of a uniform mixture of these $\binom{N}{v}$ different genotypes then the preceding formulas for haploids apply with $f(x | P_1^{(v)}P_0^{(N-v)})$ replaced by $f(x | F_1^{(v)}P_0^{(N-v)})$. Likewise, replacing $f(x | P_1^{(v)}P_0^{(N-v)})$ by $f(x | P_1^{(v)}F_1^{(N-v)})$ gives the formulas for the diploid backcross population $B_{\delta_1, \dots, \delta_k, 1}$. If the population $B_{\delta_1, \dots, \delta_k, 0}$ is selfed for r consecutive generations following the $(k + 1)$ th backcross then in the resulting population, denoted by $B_{\delta_1, \dots, \delta_k, 0 \otimes r}$, the relative frequency of any particular genotype is given by the coefficient of that genotype in the expansion of

$$\prod_{j=1}^N \left\{ \left[\left(1 - a \right) + \frac{a}{2} \left(1 - \frac{1}{2^r} \right) \right] A_{0j} A_{0j} + \frac{a}{2^r} A_{0j} A_{1j} + \frac{a}{2} \left(1 - \frac{1}{2^r} \right) A_{1j} A_{1j} \right\}.$$

Clearly, every genotype which is heterozygous at exactly v loci and homozygous for P_0 genes at exactly β loci (and hence homozygous for P_1 genes at exactly $N - v - \beta$ loci) appears with the same relative frequency; namely,

$$\left[\frac{a}{2^r}\right]^v \left[(1-a) + \frac{a}{2}\left(1 - \frac{1}{2^r}\right)\right]^\beta \left[\frac{a}{2}\left(1 - \frac{1}{2^r}\right)\right]^{N-v-\beta}.$$

The number of such genotypes is

$$\frac{N!}{v!\beta!(N-v-\beta)!}$$

so, letting $F_1^{(v)}P_0^{(\beta)}P_1^{(N-v-\beta)}$ denote the subpopulation consisting of these genotypes mixed with equal frequency, we may write

$$P(F_1^{(v)}P_0^{(\beta)}P_1^{(N-v-\beta)}) = \frac{N!}{v!\beta!(N-v-\beta)!} \left[\frac{a}{2^r}\right]^v \left[(1-a) + \frac{a}{2}\left(1 - \frac{1}{2^r}\right)\right]^\beta \left[\frac{a}{2}\left(1 - \frac{1}{2^r}\right)\right]^{N-v-\beta}.$$

The fundamental assumption (1) then gives the phenotypic distribution in population $B_{\delta_1, \dots, \delta_k, 0 \otimes r}$ as

$$f(x | B_{\delta_1, \dots, \delta_k, 0 \otimes r}) = \sum_{v=0}^N \sum_{\beta=0}^{N-v} P(F_1^{(v)}P_0^{(\beta)}P_1^{(N-v-\beta)}) f(x | F_1^{(v)}P_0^{(\beta)}P_1^{(N-v-\beta)}).$$

Thus, the theoretical phenotypic frequency distribution is expressed as a linear function of the $(N+1)(N+2)/2$ unknown distributions $f(x | F_1^{(v)}P_0^{(\beta)}P_1^{(N-v-\beta)})$. Regression estimation of these unknowns requires the inclusion of at least $(N+1)(N+2)/2$ populations of the form $B_{\delta_1, \dots, \delta_k, \delta_{k+1} \otimes r}$ in the genetic experiment, where N , again, may be estimated from the recursion (8) which holds also for diploids.

The selfing series offers the same advantages as the recurrent backcrosses in that the phenotypic frequency distribution of population $F_{k+1} = F_{k \otimes}$ is expressible as a linear function of $N+1$ unknown distributions, and the series $F_1, F_2, F_3, \dots$ satisfies the same recursion (8) as the recurrent backcross series. In F_{k+1} the relative frequency of each of the $\binom{N}{v}2^{N-v}$ genotypes which are heterozygous at exactly v loci is

$$\left(\frac{1}{2^k}\right)^v \left[\frac{1}{2}\left(1 - \frac{1}{2^k}\right)\right]^{N-v}.$$

Letting $F_1^{(v)}$ denote the uniform mixture of these genotypes we may then write the phenotypic distribution for F_{k+1} as

$$f(x | F_{k+1}) = \sum_{v=0}^N \binom{N}{v} \left(\frac{1}{2^k}\right)^v \left(1 - \frac{1}{2^k}\right)^{N-v} f(x | F_1^{(v)}).$$

The solution corresponding to (7) is then

$$f(x | F_1^{(0)}) = \frac{\sum_{v=0}^N (-1)^v \alpha_v(N) f(x | F_{k+N-v})}{\sum_{v=0}^N (-1)^v \alpha_v(N)}$$

and the recursion corresponding to (8) is

$$f(x | F_{k+N+1}) = \sum_{v=0}^N (-1)^v 2^{v+1} \alpha_{v+1} (N+1) f(x | F_{k+N-v}).$$

Linkage offers the same difficulties in diploid systems as in haploid systems; in fact, the recursions (15) and (17) derived for haploid recurrent backcrosses apply directly to diploid recurrent backcrosses. No simple recursion corresponding to (15) exists for the selfing series.

DISCUSSION

The fundamental relation (1) applies regardless of whether the phenotypes x are real numbers, ordered classes, or simply qualitative (unordered) classes; the existence of a natural ordering, however, sometimes permits the further partition of a distribution such as $f(x | F_1^{(v)} P_0^{(N-v)})$ into its component parts. It is fairly common to find that phenotypes are graded on a scale, say, from one to ten, and in this case we say a natural ordering of phenotypes exists but not a natural metric. When the estimated distribution $\hat{f}(x | F_1^{(v)} P_0^{(N-v)})$ has more than one mode with respect to the natural ordering of x then it is sometimes possible to visually pick out individual components of the distribution. Thus, in a diploid system involving two independently segregating loci the distribution $f(x | F_1^{(1)} P_0^{(1)})$ is a uniform mixture of two component distributions,

$$f(x | F_1^{(1)} P_0^{(1)}) = \frac{1}{2} f(x | A_0 A_0 B_0 B_1) + \frac{1}{2} f(x | A_0 A_1 B_0 B_0),$$

where $A_0 A_0 B_0 B_0$ is the genotype of parent P_0 . If these two component distributions do not overlap with respect to the ordering of x then by plotting $\hat{f}(x | F_1^{(1)} P_0^{(1)})$ it may be possible to separate them. For example, if phenotypes are graded one to ten and the estimated distribution $\hat{f}(x | F_1^{(1)} P_0^{(1)})$ were

$$\begin{array}{cccccccccc} x = & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \hat{f}(x | F_1^{(1)} P_0^{(1)}) = & .20 & .30 & .05 & 0 & 0 & 0 & .05 & .15 & .20 & .10 \end{array}$$

one would tentatively conclude that the two probability masses separated by an interval of no probability correspond to the two component distributions. Verification of this fact and identification of genotypes with component distributions could then be made by progeny tests; the statistical problems of progeny testing in a quantitative inheritance study, however, are not within the province of this paper. In its present stage of development, the *partitioning method* is dependent upon such visual decompositions for a complete solution to the inheritance problem; further statistical refinement of the method awaits the development of non parametric tests for linear combinations of distribution functions.

The successful application of the partitioning method and the general formulas given above requires data from a large number of individuals per population. POWERS (1955) has empirically arrived at the minimum figure of roughly 400 entries per segregating population. Another essential feature of the genetic experiment, as stressed by POWERS, is that all data be collected at one time, from one "location", and preferably from a completely randomized experimental design. This, of course, does not exclude

sequential experimentation in which at each stage of the sequence all previously derived populations are represented in the experiment. Complete randomization in the experimental design is the one basic assumption underlying the fundamental relation (1) and any restrictions upon the randomization such as the use of randomized blocks or even the use of "plots" of more than one individual per plot constitute violations of the model. Where these requirements lead to an experiment of unmanageable proportions because of an apparent large number of segregating genes the problem may be modified and simplified by selection to fix the genotype at some of the loci and so reduce the number N .

SUMMARY

Formulas are given which show some of the theoretic relationships among the phenotypic frequency distributions in a backcrossing and selfing system involving two haploid or homozygous diploid parents. The parents, say P_1 and P_0 , are assumed to differ genotypically at some unknown number N of loci, and one of the primary uses of the formulas is to determine N . Linkage is excluded with the exception of systems involving exactly one pair of linked genes, and other complicating factors such as sexual dimorphism and differential viability of gametes or zygotes are assumed nonexistent.

A device used in obtaining the formulas is the partitioning of the collection of all possible genotypes into subsets such that in any population derived by the breeding system every genotype of a given subset appears with the same theoretic relative frequency. In the case of haploids any backcross population is a mixture of the 2^N possible genotypes, but when segregation is independent at the N loci every genotype which carries exactly v genes of parent P_1 and $N - v$ of parent P_0 appears with the same theoretic relative frequency. We call this collection of $\binom{N}{v}$ equally frequent genotypes the subpopulation $P_1^{(v)}P_0^{(N-v)}$ and regard each backcross population as a mixture of these $N + 1$ subpopulations. The relative frequency of the subpopulation $P_1^{(v)}P_0^{(N-v)}$ in the backcross population B obtained from the series of crosses $F_1 \times P_{\delta_1} \times P_{\delta_2} \times \cdots \times P_{\delta_k}$, where $\delta_i = 1$ or 0 , is then $\binom{N}{v}a^v(1-a)^{N-v}$ where

$$a = \frac{1 + 2\delta_1 + 2^2\delta_2 + \cdots + 2^k\delta_k}{2^{k+1}}.$$

Theoretically, then, the relative frequency $f(x | B)$ of the phenotype x in the population B is

$$f(x | B) = \sum_{v=0}^N \binom{N}{v} a^v (1-a)^{N-v} f(x | P_1^{(v)}P_0^{(N-v)})$$

where the component $f(x | P_1^{(v)}P_0^{(N-v)})$ is the relative frequency of phenotype x in the subpopulation $P_1^{(v)}P_0^{(N-v)}$. Statistical techniques are suggested for estimating these component distributions and their moments, thus generalizing certain aspects of the *partitioning method of genetic analysis* described by POWERS.

The recurrent backcross population where $\delta_1 = \delta_2 = \cdots = \delta_k = \delta = 1$ or 0 is denoted by B_δ^k and this series is shown to satisfy

$$f(x | B_\delta^{k+N+1}) = \sum_{v=0}^N (-1)^v 2^{v+1} \alpha_{v+1} (N+1) f(x | B_\delta^{k+N-v})$$

where $\alpha_{v+1}(N+1)$, $v = 0, 1, \dots, N$, are known constants. For a given choice of parental material there is exactly one value of N such that this recursive relation holds for all k ; hence, by testing the various possible recursions on the experimental data N may be estimated. For the case of $N = 2$ linked genes this recursion becomes

$$f(x | B_{\delta}^{k+3}) = \frac{p+3}{2} f(x | B_{\delta}^{k+2}) - \frac{2+3p}{4} f(x | B_{\delta}^{k+1}) + \frac{p}{4} f(x | B_{\delta}^k)$$

where p is the unknown linkage parameter. With some slight changes in notation these same formulas are made to apply to the diploid system.

LITERATURE CITED

- ALLARD, R., 1956 Biometrical approach to plant breeding. Brookhaven Symposia in Biology, No. 9, 69-88.
- POWERS, L., 1955 Components of variance method and partitioning method of genetic analysis applied to weight per fruit of tomato hybrid and parental populations. U. S. D. A. Tech. Bull. 1131. 64 pp.
- 1934 The nature and interaction of genes differentiating habit of growth in a cross between varieties of *Triticum vulgare*. J. Agr. Research **49**: 573-605.