OPTIMUM FAMILY SIZE FOR THE ESTIMATION OF HETEROZYGOSITY IN PLANT POPULATIONS

A. H. D. BROWN,[1] B. S. WEIR[2] and D. R. MARSHALL[3]
Department of Agronomy, University of California, Davis, California

Received 3.vi.69

1. Introduction

The great majority of loci governing discontinuous morphological traits show complete dominance, at least with respect to their visual effects, and some form of progeny testing is required to obtain estimates of genotypic frequencies of such loci. In studies with predominantly inbreeding species where the expected frequency of recessives among the progeny of heterozygous parents approaches 0·25 (the expectation under self-fertilisation), the usual procedure has been to grow 10-16 plants per parent in progeny tests. This procedure is justified on the grounds that the probability of obtaining at least one recessive in a family from a heterozygous individual is slightly greater than 95 and 99 per cent. for progenies of 10 and 16 individuals, respectively. With this procedure, heterozygosity is estimated as the proportion of families among the dominants which contain at least one recessive individual. This estimate will be biased downward, although with progeny sizes greater than nine, the degree of bias is relatively small.

Since progeny testing is both time-consuming and expensive, it is necessary to make the most efficient use of the limited numbers of individuals that can be handled in such tests. Consequently, we undertook a detailed study of the effects of varying the number and size of progenies on the efficiency of estimating heterozygosity in plant populations.

2. Results

Consider a plant population polymorphic for a diallelic locus (alleles A, a) in which the genotypic frequencies are (D, H, R). If we score a random sample of individuals from this population for the dominant $(A-)$ and recessive (aa) classes, then the relative proportion of dominants in the sample will provide an estimate of $(D+H)$. The question we wish to consider here is: What is the optimum procedure for the estimation of the relative proportion of heterozygotes, $H^*(H^* = H/[H+D] = H/[1-R])$, in the dominant class? Initially we will consider a population which reproduces by complete selfing (outcrossing rate, $t = 0$), and then extend the results to species which practice mixed mating or random mating $(0 < t \leqslant 1)$.

(a) *Complete selfing*, $t = 0$

For a species which practises complete selfing, the probability of obtaining at least one double recessive individual in a progeny of size k from a heterozygous parent is

$$\pi = 1-(p)^k$$

[1] Present Address: Department of Biology, University of York, Heslington, York YO15DD.

[2] Present Address: Department of Mathematics, Massey University, Palmerston North, N.Z.

[3] Present Address: Division of Plant Industry, C.S.I.R.O., P.O. Box 109, Canberra City, A.C.T., Australia 2601.

where p is the probability any member of such a progeny will show the dominant phenotype. Under complete selfing p is a constant and equals 0·75.

If we grow n progenies from individuals with the dominant phenotype ($A-$) and score the number of progenies in which all individuals show the dominant phenotype (class 0 below) and those with at least one recessive (class 1), we have the following expectations:

Class	Observed number	Expected number
0	a_0	$n(1-\pi H^*)$
1	a_1	$n\pi H^*$
Total	n	n

Since there is one degree of freedom and one parameter to be estimated (it is assumed at this time that R is known), it follows, using a result due to Bailey (1951), that the maximum likelihood estimate of H^* can be obtained by equating either of the observed numbers with their expected values. This procedure yields:

$$H^* = a_1/n\pi = a_1/n[1-(p)^k] \tag{2}$$

with $p = 0{\cdot}75$.

The information per family is given by

$$i_f = \pi/H^*(1-\pi H^*) \tag{3}$$

and the variance of H^*

$$\mathrm{Var}(H^*) = 1/I = H^*(1-\pi H^*)/n\pi \tag{4}$$

where the total information $I = ni_f$ (Mather, 1957).

Given that we have the resources to grow a total of $N(= nk)$ plants, the problem reduces to one of determining what combination of n and k will minimise the variance of H^*. In other terms, since N is fixed, we wish to determine what combination of n and k will maximise the information per plant ($i_p = i_f/k$) with respect to H^*.

It is evident from (3) above that the optimum family size (k_0) is a function of H^*. However, it was not possible to obtain an explicit expression for k_0 in terms of H^*. Consequently, we determined the relationship between H^* and k_0, shown in fig. 1, numerically. It will be noted that progeny sizes of 10 or greater are optimal only if $H^* > 0{\cdot}86$ and further, that for all values of $0 < H^* \leqslant 0{\cdot}57$ the optimum family size is 1.

(b) *Mixed mating or random mating*, $0 < t \leqslant 1$

If the assumption $t = 0$ is relaxed, then the probability, p, that a progeny of a heterozygous parent will show the dominant phenotype is no longer a contant but a function of H, R, and t. If we assume that the outcrossing rate is homogeneous for all genotypes then, assuming there is no gametic selection, p is given by

$$p = \tfrac{1}{4}\{3+t[1-2R-H]\}. \tag{5}$$

Since p is a function of H (or H^*), the observed numbers of dominant and recessive individuals in each progeny will yield additional information about the magnitude of H^*. Therefore, if we wish to maximise the total information with respect to H^*, it is necessary to classify the families into $k+1$ classes

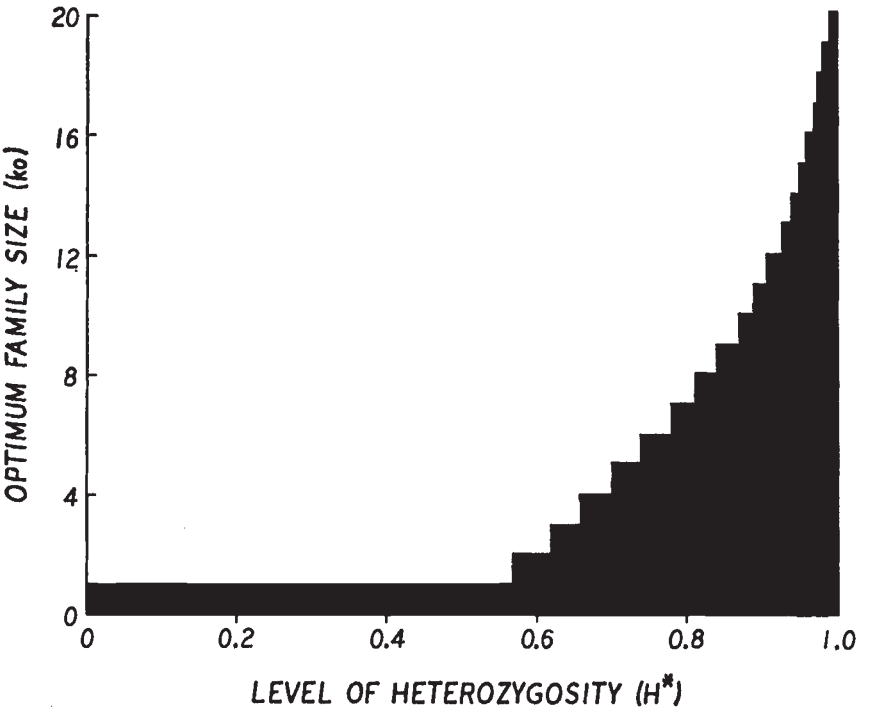

FIG. 1.—Relationship between optimum family size (k_0) and level of heterozygosity (H^*) for $t = 0$. It should be noted that for $H^* \geqslant 0{\cdot}99$, $k_0 \geqslant 20$ and as H^* approaches one, k_0 approaches infinity.

(*i.e.* those with 0, 1, 2, ..., k recessives) in contrast to the two classes used above. If, as before, we grow n families from individuals with the dominant phenotype we have the following expectations (assuming R and t are known):

Class	Observed number	Expected number
0	a_0	$nK_0H^*p^k+n(1-H^*)$
1	a_1	$nK_1H^*p^{k-1}(1-p)$
2	a_2	$nK_2H^*p^{k-2}(1-p)^2$
3	a_3	$nK_3H^*p^{k-3}(1-p)^3$
.	.	.
.	.	.
.	.	.
k	a_k	$nK_kH^*(1-p)^k$
Total	n	n

where $K_i = k!/i!(k-i)!$

From the logarithmic likelihood function we have the following equation for the estimation of H^*,

$$\frac{a_0\{H^*(1-R)tkp^{k-1}+4\pi\}}{4(1-\pi H^*)} - \frac{n-a_0}{H^*} + \frac{tk(n-a_0)(1-R)}{4p} - \frac{t(1-R)}{4p(1-p)}\sum_{i=1}^{k} ia_i = 0. \qquad (6)$$

For $t = 0$, (6) reduces to (1) above as expected. However, for $t>0$ there appears to be no simple algebraic solution to this equation, except when $k = 1$ (see (8) below). Nevertheless, for any specific set of data, estimates of H^* can easily be obtained by standard numerical methods.

The information per family is given by

$$i_f = \frac{\{\frac{1}{4}ktH^*(1-R)p^{k-1}+\pi\}^2}{1-\pi H^*} + \sum_{i=1}^{k} K_i \frac{(1-p)^{i-2}p^{k-i-2}}{H^*}\{p(1-p)+\tfrac{1}{4}H^*t(1-R)(i+kp-k)\}^2. \qquad (7)$$

While it is evident that, for $t>0$, the optimum family size is a function of H^*, R and t, as before, it was not possible to obtain an explicit algebraic expression for k_0. Consequently, we determined the optimum family size numerically for a wide range of values of the above parameters. The results are shown, in part, in table 1. It will be noted that, for a given $H^*>0{\cdot}50$, the optimum family size is a decreasing function of the outcrossing rate (t) and is a minimum in a population which is completely outcrossed. These results confirm our previous conclusions and indicate that, in predominantly inbreeding species, progeny size of ten or greater are optimal only if $H^*>0{\cdot}86$ In predominantly outcrossing species H^* must be even greater before such progeny sizes are optimal. For all other values of H^*, the most efficient procedure to use is the available resources to grow a larger number of families but fewer individuals per family. In fact, for the great majority of possible genotypic combinations the optimum family size is one. If the genotypic frequencies are close to those predicted by Wright's equilibrium law under inbreeding:

$$\begin{array}{l} AA : p-(1-F)pq \\ Aa : 2(1-F)pq \\ aa : q-(1-F)pq, \end{array} \qquad \text{where } F = \frac{1-t}{1+t}$$

a curvilinear relationship exists between H and R. From this relationship and the data given in table 1, it can be shown that R should exceed 0·90 for inbreeders and 0·50 for outbreeders before larger family sizes are considered. Thus only natural populations which largely consist of recessives require family size greater than one for optimal efficiency.

3. Discussion

The results obtained here indicate that it is possible to increase substantially the efficiency of progeny tests for the estimation of heterozygosity by careful experimental planning. For example, if $k_0 = 1$, an experiment based on single progeny families is from two to five times more efficient than one of equal size but with 16 individuals per family. In practice, of course, it will not be possible to determine accurately the optimum family size for a particular population since it is a function of the parameter (H^*) we wish to estimate. However, under most circumstances, some *a priori* knowledge of the order of magnitude of t will exist, the original sample from the population furnishes an estimate of R. From this information it is possible to obtain an approximate estimate of H by assuming genotypic frequencies obey Wright's equilibrium law. With these approximate values of t, R and H the

TABLE 1

Optimum family size (k_0) *for the estimation of* H* *in populations with various values of outcrossing* (t) *and genotype proportions* (H *and* R)

(*a*) Outcrossing rate, $t = 0{\cdot}10$

Heterozygosity	Frequency of recessive homozygote (R)										
(H)	0	0·05	0·10	0·20	0·30	0·40	0·50	0·60	0·70	0·80	0·85
0·10	1	1	1	1	1	1	1	1	1	1	3
0·20	1	1	1	1	1	1	1	1	3	∞	
0·30	1	1	1	1	1	1	1	5	∞		
0·40	1	1	1	1	1	3	6	∞			
0·50	1	1	1	2	4	8	∞				
0·60	2	2	3	5	8	∞					
0·70	3	4	6	9	∞						
0·80	6	8	10	∞							
0·90	11	14	∞								
0·95	14	∞									

(*b*) Outcrossing rate, $t = 0{\cdot}50$

Heterozygosity	Frequency of recessive heterozygote (R)										
(H)	0	0·05	0·10	0·20	0·30	0·40	0·50	0·60	0·70	0·80	0·85
0·10	1	1	1	1	1	1	1	1	1	1	2
0·20	1	1	1	1	1	1	1	1	2	∞	
0·30	1	1	1	1	1	1	1	3	∞		
0·40	1	1	1	1	1	1	4	∞			
0·50	1	1	1	1	1	5	∞				
0·60	1	1	1	1	6	∞					
0·70	1	1	1	7	∞						
0·80	2	5	8	∞							
0·90	9	13	∞								
0·95	13	∞									

(*c*) Outcrossing rate, $t = 0{\cdot}90$

Heterozygosity	Frequency of recessive homozygote (R)										
(H)	0	0·05	0·10	0·20	0·30	0·40	0·50	0·60	0·70	0·80	0·85
0·10	1	1	1	1	1	1	1	1	1	1	1
0·20	1	1	1	1	1	1	1	1	1	∞	
0·30	1	1	1	1	1	1	1	2	∞		
0·40	1	1	1	1	1	1	3	∞			
0·50	1	1	1	1	1	4	∞				
0·60	1	1	1	1	4	∞					
0·70	1	1	1	5	∞						
0·80	1	1	5	∞							
0·90	5	11	∞								
0·95	12	∞									

optimum family size is obtained by interpolation in table 1. Effective outcrossing, t^* (Allard and Workman, 1963), could be used in place of t, to adjust for possible heterotic selection.

In deriving the above results we have assumed that:

(i) R and t are known from independent estimates,
(ii) all families are of equal size, and
(iii) n and k can be varied without limit.

These assumptions will seldom be met in practice. If R and t are unknown, it simply means they must be estimated co-jointly with H. The procedure for the joint estimation of H, R and t for family size of one is given by Jain and Marshall (1967). When $k>1$, R can be derived from the proportion of recessives in the population. The parameters $\hat{t}$ and $\hat{H}$ are estimated simultaneously by the scoring method using the expectations given above, augmented by a further data class: the number of heterozygotes observed among m progeny grown from the recessive parents, which has expectation $[\frac{1}{2}mt(H+2R)]$. The variance of these estimates can be estimated from the information matrix derived by partial differentiation. The technique is straightforward although cumbersome.

Further, relaxation of the assumption of equal family size does not alter the problem in principle. It does, however, increase the difficulty in estimation of H, as it is necessary to derive an estimate of H for each family size and then combine the estimates.

Relaxation of the assumption that n and k can be varied without limit does alter the above conclusions. If n is limited, then the number of individuals per family should be increased to the limit of resources. This is the situation when the target population of dominant $(A-)$ plants is smaller than the experimental resources. If the biological maximum for k is less than k_0, the number of families of size k should be increased to the limit of the available recources. Such a situation might occur in a species with low fecundity.

It is also important to consider the relative effort implied by an increment in k compared with an increment in n in the above model. Our analysis defines " limited resources " solely in terms of plant numbers. In practice, however, an experiment in which $k = 1$ possesses a number of additional advantages. First, a much simpler experimental layout is required as the dominants can be grown as a bulk and families no longer need be kept separate. Second, the iteration procedure for the joint estimation of H and t is simpler than when $k > 1$. Third, when only an estimate of H is required, it follows from (2) and (5) above that $\hat{H}^*$ is given by:

$$\hat{H}^* = \frac{-[1-t(1-2R)]+\sqrt{[1-t(1-2R)]^2+16a_1(1-R)t/n}}{2(1-R)t}, \text{ given } t>0. \quad (8)$$

These advantages increase the variety of populational conditions under which an experimenter would commit his resources to growing single progeny families.

4. Summary

The efficiency of estimating heterozygosity at a dominant locus in plant populations by progeny testing, under limited experimental resources, was examined in terms of the amount of information per plant. The main findings were:

1. The optimum family size is a function of the relative proportions of heterozygotes (H) and double recessives (R) in the population, and the outcrossing rate (t).

2. In predominantly inbreeding species the usually accepted practice of growing progenies of 10 or more individuals per family is optimal only if

H^* ($= H/1-R$) $>0{\cdot}86$. In predominantly outcrossed species, H^* must be even greater before such family sizes are optimal.

3. In the great majority of situations likely to be encountered in practice, the most efficient procedure is to grow as many families as possible with only a single progeny per family. The applicability of this procedure and its advantages are discussed.

Acknowledgments.—This work was supported in part by grants from the National Science Foundation (GB-3246) and the National Institutes of Health (GM-10476). The authors wish to thank Drs R. W. Allard and S. K. Jain for their helpful criticisms of the manuscript.

5. References

ALLARD, R. W., AND WORKMAN, P. L. 1963. Population studies in predominantly self-pollinated species. IV. Seasonal fluctuations in estimated values of genetic parameters in lima bean populations. *Evolution, 17,* 470-480.

BAILEY, N. T. J. 1951. Testing the solubility of maximum likelihood equations in the routine application of scoring methods. *Biometrics, 7,* 268-274.

JAIN, S. K., AND MARSHALL, D. R. 1967. Genetic changes in a barley population analyzed in terms of some life cycle components of selection. *Genetica, 38,* 355-374.

MATHER, K. 1957. *The Measurement of Linkage in Heredity.* Methuen and Co. Ltd., London.