ESTIMATION OF EFFECTIVE NUMBER OF FACTORS FROM THE COMPONENTS OF GENETIC VARIANCES

By

S.N. SEN

Rajendra Agricultural University, Agricultural Research Institute, Patna-14

(Received in August, 1972 ; Accepted in January, 1974)

For estimation of the effective number of controlling genes for inheritance of character in the breeding population, the theory of the components of the genetic variances, employed by Wright (1931) and Mather (1949) is helpful. In case of back-cross, a simple formula for estimating the effective numbers of genes has been worked out (Sen, 1960) and the lower bound of the estimate is obtained. The generalised expression for the genetic variances when either linkage is present or absent have already been derived (Sen, 1966) and the effective number of factors is estimated from F_n progenies where the progenies at any generation can be raised by selfing the successive generations derived from a cross between two pure lines.

Let the additive model with dominance described by Fisher, Immer and Tedin (1932) and later expanded by Mather (1949) be considered. Let $2d_i$=the difference in effect between the homozygous genotypes at the ith locus, and hi=the deviation of the effect of the heterozygous genotypes from the mean effect of the two homozygotes. Thus, 'd' and 'h' are the additive and the non-additive increments respectively due to genes. If, P_1 and P_2 are pure lines whose genotypes with respect to the quantitative character under study differ at 'k' loci, then, $\overline{P_1}$ and $\overline{P_2}$ may be denoted as the phenotypic mean of P_1 and P_2 respectively. The, $1/2\ (\overline{P}_2-\overline{P_1})=\sum_{i=1}^{k} \delta_i d_i$, where $\delta_i=+1$ or -1, depending upon which of the two homozygous allellic pairs appears at the ith locus in the pure line.

If, $\overline{F_1}$=mean phenotype in F_1 population,

then, $\overline{F_1}=1/2\ (\overline{P}_1+\overline{P_2})+\sum_{t=1}^{k} h_t$

and also,

$$\overline{F_2} = 1/2\,(\overline{P_1}+\overline{P_2}) + 1/2\sum_{i=1}^{k} h_i \text{ etc.,}$$

and $$(\overline{F_n} = 1/2)(\overline{P_1}+\overline{P_2}) + \frac{1}{2\,(n-1)}\sum_{i=1}^{k} h_i$$

Again, $$B_1 = 1/2\,\overline{P_1}+\overline{P_2}) - 1/2\sum_{i=1}^{k} \delta_i d_i + 1/2\sum_{i=1}^{k} h_i$$

$$B_2 = 1/2\,(\overline{P_1}+\overline{P_2}) + 1/2\sum_{i=1}^{k} \delta_i d_i + 1/2\sum_{i=1}^{k} h_i$$

In the absence of linkage and ignoring the effect of the environmental variability, the expressions of the variances in the first few generations were already given as :

$$V\,(F_2) = 1/2\sum_{i=1}^{k} d_i^2 + 1/4\sum_{i=1}^{k} h_i^2$$

$$\overline{V\,(F_3)} = 1/4\sum_{i=1}^{k} d_i^2 + \frac{1}{8}\sum_{i=1}^{k} h_i^2$$

$$V\,(\overline{F_3}) = 1/2\sum_{i=1}^{k} d_i^2 + \frac{1}{16}\sum_{i=1}^{k} h_i^2$$

$$V\left[V(F_3)\right] = \frac{1}{16}\sum_{i=1}^{k} d_i^4 + \frac{1}{64}\sum_{i=1}^{k} h_i^4 + \frac{1}{16}\sum_{\substack{i,\,j=1 \\ (i\neq j)}}^{k} d_i^2 h_j^2$$

$$V(B_1) = 1/4\sum_{i=1}^{k} d_i^2 + 1/4\sum_{i=1}^{k} h_i^2 + 1/8\sum_{i=1}^{k} \delta_i d_i h_i$$

$$V(B_2) = 1/4\sum_{i=1}^{k} d_i^2 + 1/4\sum_{i=1}^{k} h_i^2 - 1/2\sum_{i=1}^{k} \delta_i d_i h_i$$

A method, readily available to the breeders for estimating the number of controlling factors from the Bi-parental backcross data, have been earlier derived (Sen, 1960).

Consider the inequality : $\sum_{i=1}^{k} x_i^2 - \frac{1}{k}\left(\sum_{i=1}^{k} x_i\right)^2 \geqslant 0.$

or $$\hat{K} \geqslant \left(\sum_{i=1}^{k} x_i^2\right)^2 / \left(\sum_{i=1}^{k} x_i\right), \quad (i=1 \ldots k).$$

This suggests that the estimate of the effective number of controlling genes from the backcross bi-parental data may be obtained as :

$$\hat{k}_b = \left(\bar{B}_2 - \bar{B}_1\right)^2 / 2\left[2V\left(F_2\right) - V\left(B_1\right) - V\left(B_2\right)\right],$$

which, after substitution of the values in terms of genetic components of variances, may correspond to the expression :

$$\hat{k}_b = \left(\sum_{i=1}^{k} \delta_i d_i\right)^2 \Big/ \left(\sum_{i=1}^{k} d_i^2\right).$$

The estimate of $\hat{k}_b$ may be used only for the backcross cases. It may be mentioned that the method of this estimation involves the heritable part of the respective second degree statistics. The effect of the environmental variability may be made at par by conducting the breeding trial under an identical condition with the same management.

In constructing the lower bounds of $\hat{k}_b$, the following inequality may be used :

$$r^2_{xy} = \frac{[\Sigma xy - 1/k\Sigma \times \Sigma y]^2}{[\Sigma x^2 - 1/k(\Sigma x)^2] \cdot [\Sigma y^2 - 1/k\,(\Sigma y)^2]} \leqslant 1,$$

which is equivalent to :

$$\frac{\Sigma x^2\,(\Sigma y)^2 + (\Sigma x)^2\,\Sigma y^2 - 2\Sigma x \Sigma y\,\Sigma xy}{\Sigma x^2\,\Sigma y^2 - (\Sigma xy)^2} \leqslant k.$$

By substituting in the above inequality, for the backcross case,

$$x = \delta d \text{ and } y = d,$$

the following inequality may be set up as follows:

$$\hat{k}_b = \frac{\Sigma\,(\delta\,d)^2\,(\Sigma\,d)^2 + (\Sigma\delta\,d)^2\,(\Sigma\,d^2) - 2\,\Sigma\delta\,d\,\Sigma\,d\,\Sigma\delta\,d^2}{\Sigma\,(\delta\,d)^2\,\Sigma\,d^2 - (\Sigma\delta\,d^2)^2} \leqslant k.$$

This expression of $\hat{k}_b$ may be further simplified by writing in terms of the moments. Since, for estimation of the values of either denominator or numerator, the separation of $(\delta^2 . d^2)$ and d^2 cannot be made, the denominator will be inflated in this estimate of $\hat{k}_b$.

Originally, Wright (1931) suggested the estimate of 'k' as:

$$\hat{k}_1 = \frac{(\bar{P}_1 - \bar{P}_2)^2}{8\,[V(F_2) - V(F_1)]},$$

which corresponds to: $\left(\sum_{i=1}^{k} \delta_i d_i\right)^2 \Big/ \sum_{i=1}^{k} d_i^2$.

When dominance and interaction effects have not been compensated by scaling, the genic effects include these components and the denominator is inflated while the numerator may be less so. The estimate of 'k' is thus further reduced.

Panse (1940) and Mather (1949) estimated the number of controlling genes as:

$$\hat{k}_2 = [\bar{V}(F_3)]^2 / [V(F_3)]$$

$$= \frac{(1/4\sum_i d_i^2 + \frac{1}{8}\sum_i h_i^2)^2}{\left(\frac{1}{16}\sum_i d_i^4 + \frac{1}{64}\sum_i h_i^4 + \frac{1}{16}\sum_{i,j} d_i^2 h_j^2\right)} \quad (i, j = 1, \ldots k;\ i \neq j)$$

It is evident that the estimate $\hat{k}_2$ is superior to the estimate $\hat{k}_1$ but $\hat{k}_1$ is readily workable by the breeders and the parameters up to F_3 generations are not necessary. The expression of $V[V(F_3)]$ is inflated with the sampling error which is equal to $2\,\bar{V}^2(F_3)/(n-1)$, where $\bar{V}^2(F_3)$=square of the mean variance of all the families of F_3 including an environmental component. When this correction (Mather, 1949) is not made, the estimate of 'k' is lower than that it should be (Panse, 1940).

Assuming that the covariances (d, h) are negligible, the genetic variance in F_n, for k pairs of genes, at the n^{th} generation, has been

shown (Sen, 1966) as:

$$V(F_n)=\left(1-\frac{1}{2^{n-1}}\right)\sum_{i=1}^{k} d_i^2 \pm 2\left[\frac{1-2p}{1+2p}-\frac{(1-2p)^n}{2^{n-1}(1+2p)}\right]$$

$$\sum_{\substack{i,\,(j=1)\\ i\neq j}}^{k} d_i\, d_j+\left(\frac{1}{2^{n-1}}-\frac{1}{4^{n-1}}\right)\sum_{i=1}^{k} h_i^2$$

$$\pm 2\left[\frac{(1-2p+2p^n)^{n-1}}{2^{n-1}}-\frac{1}{4^{n-1}}\right]\sum_{\substack{i,j=1\\ (i\neq j)}}^{k} h_i h_j$$

where p=the recombination fraction, which is assumed to be same for all the loci.

When $p=0$, *i.e.* the linkage is complete,

$$V(F_n)=\left(1+\frac{1}{2^{n-1}}\right)\left(\sum_{i=1}^{k} d_i\right)^2+\left(\frac{1}{2^{n-1}}-\frac{1}{4^{n-1}}\right)\left(\sum_{i=1}^{k} h_i\right)^2.$$

Again, when $p=\frac{1}{2}$, *i.e.* the linkage is absent,

$$V(F_n)=\left(1-\frac{1}{2^{n-1}}\right)\sum_{i=1}^{k} d_i^2+\left(\frac{1}{2^{n-1}}-\frac{1}{4^{n-1}}\right)\sum_{i=1}^{k} h_i^2,$$

which is Mather's case in absence of linkage.

Other functions of genetic variances in F_n progenies may be suitably worked out in case of absence of linkage.

Thus, the variance of means of F_n-progenies may be derived as:

$$V(\bar{F}_n)=\tfrac{1}{2}\sum_{i=1}^{k} d_i^2+\frac{1}{2^{2n-2}}\sum_{i=1}^{k} h_i^2,$$

and, mean variance of F_n progenies is given as:

$$\bar{V}(F_n)=\frac{1}{2^{n-1}}\left(\sum_{i=1}^{k} d_i^2+\tfrac{1}{2}\sum_{i=1}^{k} h_i^2\right).$$

The expression for the variance among "within F_n-families variances" may be derived as:

$$V\,[V\,(F_n)] = \frac{1}{2^{\cdot n-2}}\Big[\sum_{i=1}^{k} d_i^4 + \tfrac{1}{4}\sum_{i=1}^{k} h_i^4 + \sum_{\substack{i,\,j=1 \\ (i\neq j)}}^{k} d_i^2\, h_j^2.$$

Thus a more general formula for estimation of the effective number of controlling genes at any generation may be worked out from the expression of $\bar{V}\,(F_n)$ and $V\,[V\,(F_n)]$. In case of absence of linkage, the formula may be set up as :

$$\hat{k}_3 = [\bar{V}\,(F_n)]^2 / V\,[V\,(F_n)]$$

$$= \frac{\frac{1}{2^{2n-2}}\Big[\sum_{i=1}^{k} d_i^2 + \frac{1}{2}\sum_{i=1}^{k} h_i^2\Big]^2}{\frac{1}{2^{2n-2}}\Big[\sum_{i=1}^{k} d_i^4 + \frac{1}{4}\sum_{i=1}^{k} h_i^4 + \sum_{\substack{i,\,j=1 \\ (i\neq j)}}^{k} d_i^2\, h_j^2\Big]}.$$

This estimate is more general and can be applied for any generation, 'n' taking any value. It is clear that the methods for estimating $\hat{k}_2$ and $\hat{k}_3$ which involves the mean variance and the variance of the variance of F_n progenies assumes that $(\frac{1}{2}\, d_i^2 + \frac{1}{4}\, h_i^2)$ is constant over the loci.

In case of presence of linkage, Sen (1966) had already obtained the genetic variance for the polygenic case. After sufficient advancement of generations, when 'n' is very large, it then follows that,

$$V\left(F_n\right) \simeq \left(\sum_{i=1}^{k} d_i\right)^2 \text{ or } \sum_{i=1}^{k} d_i^2,$$

according as there is complete linkage or absence of linkage.

Thus the genetic variance becomes asymptotically a function of the additive increments only and the effective number of controlling genes act as the single unit in the presence of perfect linkage.

But, again, since $\left(\sum_{i=1}^{k} d_i\right)^2 \Big/ \left(\sum_{i=1}^{k} d_i^2\right) \leqslant \frac{1}{k}$,

an interesting result (Sen, 1966) follows in the asymptotic case. The genetic variance in the n^{th} generation, when 'n' is sufficiently large and linkage is complete is not more than $\frac{1}{k}$ times the genetic variance at the same generation when there is no linkage, '$\hat{k}$' being the effective number of controlling genes. Thus '$\hat{k}$' bears a relationship between the two variances in the limiting case.

SUMMARY

Wright (1931) and Mather (1949) had earlier suggested the formula for estimation of gene numbers from F_2 and F_3 progenies by construction of suitable functions of components of the genetic variability. An easy and quick formula for estimating the effective number of controlling genes is provided here in case backcross data are available. The lower bound of this estimate is also suggested by the author. Mather's formula for estimating gene numbers from F_3 progenies may be suitably extended for estimating from n^{th} generation derived from the initial parental stock where 'n' may take any value. The genetic variances, in the asymptotic case, for the presence of linkage and its absence, bear a simple relation depending on k, the effective number of controlling genes.

REFERENCES

1. Fisher, R.A., Immer, F.R. and Tedin, O. (1932) : The genetical interpretation of Statistics of the third degree in the study of quantitative inheritance : Genetics **17**, 107-124.

2. Mather, K. (1949) : Biometrical Genetics, New York : Dever Publications.

3. Panse, V.G. (1940) : A Statistical Study of Quantitative Inheritance : Ann. Eugen, **10**, 76-105.

4. Sen, S.N. (1960) : Estimation of gene numbers from backcross data : Prac. 47th session of Ind. Sci. Congress.

5. Sen, S.N. (1966) : The effect of linkage on genetic variance in the polygenic case : Calcutta Stat. Assoc. Bulletin, **15**, No. 60, 175-177.

6. Wright, S. (1931) : Statistical methods in Biology Jr. Amer. Stat. Assoc. 26, Suppl., 155-163.