MODIFICATIONS IN ESTIMATING THE NUMBER OF GENES FOR A QUANTITATIVE CHARACTER

C. CLARK COCKERHAM

Department of Statistics, North Carolina State University, Box 8203, Raleigh, North Carolina 27695-8203

Manuscript received October 26, 1985
Revised copy accepted June 19, 1986

ABSTRACT

In estimating the minimum number of genes contributing to a quantitative character, it is suggested that the squared difference between the means of the two parents be corrected for experimental variance and that the genetic variance stemming from differences in gene frequencies of the parents be estimated by least squares utilizing information on all entries.

In this note some statistical modifications are suggested for WRIGHT'S (1968) and LANDE'S (1981) methods of estimating the number of genes contributing to a quantitative character. The original method of WRIGHT (in CASTLE 1921) utilized two inbred parental lines and their F1 and F2 populations. WRIGHT later (1968) included backcrosses. LANDE (1981) generalized the formulations to accommodate two heterogeneous parent populations and their crosses.

The basic theoretical formula for estimating the minimum number, M, of loci affecting a quantitative character is

\[ M = \frac{(\mu_1 - \mu_2)^2}{8\sigma_i^2}, \]

where \( \mu_1 \) and \( \mu_2 \) are the means of the two parental populations, and \( \sigma_i^2 \) is the additive genetic variance stemming from differences in gene frequencies of the parental populations. The usual estimate of \( M \), hats denote estimators, is

\[ \hat{M} = \frac{(\hat{\mu}_1 - \hat{\mu}_2)^2 - C}{8\hat{\sigma}_i^2}, \]

but without the correction \( C \). Since \( \hat{\mu}_1 \) and \( \hat{\mu}_2 \) are estimates, the expectation, denoted by \( \mathbb{E} \), of the squared quantity is

\[ \mathbb{E}((\hat{\mu}_1 - \hat{\mu}_2)^2) = (\mu_1 - \mu_2)^2 + \sigma_{\hat{\mu}_1}^2 + \sigma_{\hat{\mu}_2}^2, \]

where \( \sigma_{\hat{\mu}_1}^2 \) and \( \sigma_{\hat{\mu}_2}^2 \) are variances of the parental sample means. Consequently, a correction factor would appear to be in order,

\[ C = \hat{\sigma}_{\hat{\mu}_1}^2 + \hat{\sigma}_{\hat{\mu}_2}^2, \]

where \( \hat{\sigma}_{\hat{\mu}}^2 \) is the estimated variance of \( \hat{\mu}_i \).

The above modification eliminates bias only in the numerator. As pointed out by CARSON and LANDE (1984), $\delta M$ does not exist, since $\delta^2_i$ may be negative and its distribution overlaps zero.

The other suggestion has to do with the manner in which $\sigma^2_\epsilon$ is estimated. There are several ways of estimating $\sigma^2_\epsilon$, and LANDE (1981) gives four. Instead of four interrelated, and often conflicting, estimates, it is prudent to combine all the information into one estimate. This can be done by least squares.

A full experiment is considered consisting of the two parents, $P_1$ and $P_2$, and their $F_1$, $F_2$ and first backcross, $B_1$ and $B_2$, populations. Let their estimated variances be indexed accordingly

$$Y^T = (\sigma^2_{P_1}, \sigma^2_{P_2}, \sigma^2_{F_1}, \sigma^2_{F_2}, \sigma^2_{B_1}, \sigma^2_{B_2})$$

($T$ indicates transpose of a vector or matrix). When the parental populations are segregating, each will have genetic variance, $\sigma^2_{\epsilon_i}$, in addition to the environmental variance $\sigma^2_\epsilon$. Let

$$\beta^T = (\sigma^2_{\epsilon_1}, \sigma^2_{\epsilon_2}, \sigma^2_{\epsilon_3}, \sigma^2_{\epsilon_4}).$$

Then, the model is

$$Y = X\beta + \epsilon,$$

where $\epsilon$ is a vector of errors and

$$X = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 1 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & 1 \end{bmatrix}$$

The unweighted least squares solution for the $\beta$’s is

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

($I$ denotes inverse).

Consider first the case of inbred parental lines for which $\sigma^2_{\epsilon_1} = \sigma^2_{\epsilon_2} = 0$, and the first two columns of the $X$ matrix are eliminated. Then, for the reduced matrix,

$$X^T X = \begin{bmatrix} 1.5 & 2 \\ 2 & 6 \end{bmatrix} \quad (X^T X)^{-1} = \begin{bmatrix} 1.2 & -0.4 \\ -0.4 & 0.3 \end{bmatrix}$$

and

$$\hat{\delta}^2_i = 0.2(4\hat{\delta}^2_{\epsilon_2} + \hat{\delta}^2_{\epsilon_3} + \hat{\delta}^2_{\epsilon_4}) - 0.4(\hat{\delta}^2_{\epsilon_1} + \hat{\delta}^2_{\epsilon_2} + \hat{\delta}^2_{\epsilon_3})$$

$$\hat{\sigma}^2_\epsilon = 0.3(\hat{\delta}^2_{\epsilon_1} + \hat{\delta}^2_{\epsilon_2} + \hat{\delta}^2_{\epsilon_3}) + 0.1(\hat{\delta}^2_{\epsilon_1} + \hat{\delta}^2_{\epsilon_2} - \hat{\delta}^2_{\epsilon_3}).$$

For the more general case of heterogeneous parents, neither $\sigma^2_{\epsilon_1}$ nor $\sigma^2_{\epsilon_2}$ is zero, and the $X^T X$ matrix is singular. This is because

$$x_{i1} + x_{i2} = x_{i4},$$
where the \( x \)'s are elements in the \( X \) matrix indexed according to column in the \( i \)th row. We substitute \( x_{i1} + x_{i2} \) for \( x_{i4} \), which eliminates the fourth column, and collect together the coefficients of the remaining \( x \)'s, which provides the functions to be estimated,

\[
\beta' = (\sigma_1^2, \sigma_2^2, \sigma_r^2)
\]

where

\[
\sigma_1^2 = \sigma_{g1}^2 + \sigma_r^2, \quad \sigma_2^2 = \sigma_{g2}^2 + \sigma_r^2.
\]

For this reduced matrix,

\[
X'X = \begin{bmatrix}
2.125 & 0.875 & 1 \\
0.875 & 2.125 & 1 \\
1 & 1 & 1.5
\end{bmatrix}, \quad (X'X)' = \begin{bmatrix}
0.7 & -0.1 & -0.4 \\
-0.1 & 0.7 & -0.4 \\
-0.4 & -0.4 & 1.2
\end{bmatrix}
\]

and \( \sigma_r^2 \) is estimated by the same function of variances as for inbred parents,

\[
\hat{\sigma}_r^2 = 0.2(4\hat{\sigma}_{g1}^2 + \hat{\sigma}_{g2}^2) - 0.4(\hat{\sigma}_{g1}^2 + \hat{\sigma}_{g2}^2 + \hat{\sigma}_{F}^2).
\]

LANDE (1981), assuming normally distributed populations, formulated an approximate variance, \( \hat{V} \), of \( \hat{\sigma}_r^2 \),

\[
\hat{V}(\hat{\sigma}_r^2) = \hat{\sigma}_r^2 \left[ \frac{4(\hat{\sigma}_{g1}^2 + \hat{\sigma}_{g2}^2)}{(\hat{\mu}_1 - \hat{\mu}_2)^2} + \hat{V}((\hat{\sigma}_r^2)) \right],
\]

where \( \hat{\sigma}_{g1}^2 = \hat{\sigma}_{g1}^2/N_{p1} \), and \( N \) denotes the number of individuals sampled. For the unweighted least squares estimator,

\[
\hat{V}(\hat{\sigma}_r^2) = 0.08 \left[ \frac{16\hat{\sigma}_{F2}^4}{N_{F2}} + \frac{\hat{\sigma}_{g1}^4}{N_{p1}} + \frac{\hat{\sigma}_{g2}^4}{N_{p2}} \right] + 0.32 \left[ \frac{\hat{\sigma}_{F1}^4}{N_{p1}} + \frac{\hat{\sigma}_{F2}^4}{N_{p2}} + \frac{\hat{\sigma}_{F}^4}{N_{p1}} \right].
\]

This \( \hat{V}(\hat{\sigma}_r^2) \) ignores the correction factor suggested for the numerator, which, if taken into account, unduly complicates the variance. In any case, \( \hat{V}(\hat{\sigma}_r^2) \) should be viewed as a rough guide for the variance.

With very large \( N \)'s, the suggested correction for the numerator will be negligible. Aside from the confusion of having several estimates, a combined estimate of \( \sigma_r^2 \) containing all the information is preferable. A comparison of the variances, \( V(\hat{\sigma}_r^2) \), for the various methods can be made by substituting the model variances for the entry variances and collecting the coefficients for functions of the model variances. For example, with inbred lines, substitute \((\sigma_r^2 + \sigma_r^2)^2\) for \( \hat{\sigma}_{F2}^4 \), \((\frac{1}{2}\sigma_r^2 + \sigma_r^2)^2\) for \( \hat{\sigma}_{g1}^4 \), and so on, in \( V(\hat{\sigma}_r^2) \), and collect together the coefficients for \( \sigma_r^4 \), \( \sigma_r^2 \sigma_r^2 \) and \( \sigma_r^4 \). This is done in Table 1 for the four methods considered by LANDE (1981) (his formulas 8a–d), denoted as a–d, and the least squares estimator, denoted as LS. All \( N \)'s for entries are assumed to be equal.

Strictly speaking, for unweighted least squares to be appropriate, the variances of the estimated variances in \( Y \) should be the same. This means that \( \sigma_r^2 \) in \( Y \) are all equal and that \( N \) is the same for all entries, neither of which generally holds in practice. Since \( V(\hat{\sigma}_r^2) \) for each method in Table 1 is given by
### Table 1

Coefficients of functions of the parametric variances in $V(\hat{\sigma}^2)$ for the four methods of estimating $\hat{\sigma}^2$ by Lande (1981) denoted as a, b, c, d and the least squares method denoted as LS

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2/N$</th>
<th>$\sigma^2/N$</th>
<th>$\sigma^2 \sigma^2/N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>b</td>
<td>2.75</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>c</td>
<td>12</td>
<td>9</td>
<td>20</td>
</tr>
<tr>
<td>d</td>
<td>7</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>LS</td>
<td>2.4</td>
<td>1.32</td>
<td>2.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma^2/N$</th>
<th>$\sigma^2/N$</th>
<th>$\sigma^2 \sigma^2/N$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>b</td>
<td>0.75</td>
<td>1.25</td>
<td>2</td>
</tr>
<tr>
<td>c</td>
<td>3.25</td>
<td>5.5</td>
<td>10</td>
</tr>
<tr>
<td>d</td>
<td>2.25</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>LS</td>
<td>0.77</td>
<td>0.86</td>
<td>1.36</td>
</tr>
</tbody>
</table>

In practice, all $N$'s are not equal, there generally being less observations on the less variable populations, $P_1$, $P_2$ and $F_1$, than on the variable populations, $F_2$, $B_1$ and $B_2$. A weighted least squares method, which is a maximum likelihood solution for normal distributions, has no explicit solution and requires iteration. The variance covariance matrix, $W$, of the $Y$'s is diagonal; $W = \begin{bmatrix} 2\sigma_1^4 \frac{1}{N_{P_1}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 2\sigma_2^4 \frac{1}{N_{P_2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 2(\frac{1}{2}\sigma_1^2 + \frac{1}{2}\sigma_2^2) \frac{1}{N_{F_1}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 2(\frac{1}{2}\sigma_1^2 + \frac{1}{2}\sigma_2^2 + \sigma_1^2) \frac{1}{N_{F_2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 2(\frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\sigma_1^2) \frac{1}{N_{B_1}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 2(\frac{1}{4}\sigma_1^2 + \frac{1}{4}\sigma_2^2 + \frac{1}{2}\sigma_1^2) \frac{1}{N_{B_2}} \end{bmatrix}$

The sum of products of the coefficients and the components, which takes into account differences in $E\hat{\sigma}^2$, the unweighted least squares estimator is much superior to the other four if all $N$'s are equal.
The weighted least squares solution to be iterated is
\[ \hat{\beta} = (X^T W^2 X)^T X^T W^2 Y, \]
with values obtained for \( \hat{\beta} \) substituted into \( W \) for the next round of iteration. The initial \( \hat{\beta}'s \) can be found by unweighted least squares. The variance covariance matrix of the \( \hat{\beta}'s \) is estimated as
\[ \hat{V}(\hat{\beta}) = (X^T \hat{W}^2 X)', \]
with the final \( \hat{\beta}'s \) substituted for their parametric counterparts in \( W \) to obtain \( \hat{W} \). The diagonal value in \( \hat{V}(\hat{\beta}) \) corresponding to \( \hat{\sigma}^2 \) is \( \hat{V}(\hat{\sigma}^2) \), to be substituted into the formula for \( \hat{V}(\hat{M}) \). Unfortunately, no explicit theoretical comparison of \( V(\hat{\sigma}^2) \) with those of the other methods can be made.

COMMENTS

Ideally, one wants to utilize numbers of observations for the entries, restricted to some total number of observations, that minimize the variance of \( \hat{M} \). Finding these \( N's \) involves \( (\mu_1 - \mu_2)^2 \) and the parametric components of variance that are to be estimated. Just minimizing \( V(\hat{\sigma}^2) \) illustrates some of the problems. Consider two inbred lines and their \( F_2 \) generation with
\[ \hat{\sigma}^2_P = \hat{\sigma}^2_{P_1} = \sigma^2_e, \quad \hat{\sigma}^2_P = \sigma^2_e + \sigma^2_i, \]
\[ \hat{\sigma}^2 = \hat{\sigma}^2_{P_2} = (\hat{\sigma}^2_{P_1} + \hat{\sigma}^2_{P_2})/2. \]
Then, assuming normally distributed observations,
\[ V(\hat{\sigma}^2) = \frac{2(\sigma^2_e + \sigma^2_i)^2}{N_{F_2}} + \frac{\sigma^4_e}{2N_{P_1}} + \frac{\sigma^4_i}{2N_{P_2}}. \]
The total number of observations is \( N_{F_2} + N_{P_1} + N_{P_2} = N \). Letting \( N_{P_1} = N_{P_2} = N_p \) for obvious reasons,
\[ V(\hat{\sigma}^2) = \frac{2(\sigma^2_e + \sigma^2_i)^2}{N_{F_2}} + \frac{2\sigma^4_i}{(N - N_{F_2})}. \]
The number of \( F_2 \) observations that minimize this variance is
\[ N_{F_2} = \frac{N}{2 - h}, \quad h = \frac{\sigma^2_i}{\sigma^2_e + \sigma^2_i}, \]
where \( h \) is the heritability. Thus, something over one-half of the observations are allotted to the \( F_2 \), and less than one-fourth to each of the parents. If the \( F_1 \) is included with \( N_{F_1} = N_p \), since the \( F_1 \) has the same variance as each parent, the allocation to the \( F_2 \) is the same, with the remaining observations distributed equally among \( P_1, P_2 \) and \( F_1 \). There is no advantage in including \( F_1 \) for the estimation of \( \sigma^2_i \). There is actually a disadvantage if the number of observations on parental means is reduced, since the variances of parental means also contribute directly to \( V(\hat{M}) \). The inclusion of backcrosses at the expense of \( F_2 \) observations also appears to increase \( V(\hat{\sigma}^2) \), since they provide less information about \( \sigma^2_i \) than the \( F_2 \). Variable parents would not alter the qualitative nature of these conclusions.
For assumptions and limitations of this procedure for estimating the number of genes, the reader is referred to Wright (1968) and Lande (1981). Wright (1968) illustrates the use of transformations to obtain additivity with further corrections for dominance. Carson and Lande (1984) extend Lande's (1981) extension to variable parental populations to the utilization of variances within full-sib families and only one parental population.

While the estimate of the number of genes is purported to be a minimum, the modification suggested for the numerator will always reduce the estimate. The modifications suggested for estimating $\sigma^2_a$ have the aim of obtaining a single estimate with small variance, with the weighted least squares being, presumably, the best in all situations.

This note should not be construed as an endorsement of this procedure of estimating the number of genes but, rather, as an attempt at improving the procedure if it is to be used.

Paper 10163 of the Journal Series of the North Carolina Agricultural Research Service, Raleigh. This investigation was supported in part by National Institutes of Health research grant GM 11546 from the National Institute of General Medical Sciences.

LITERATURE CITED


Lande, R., 1981 The minimum number of genes contributing to quantitative variation between and within populations. Genetics 99: 541–553.


Communicating editor: W. J. Ewens