Indian J. Genet., **61**(2): 87-91 (2001)

A bootstrap method of estimating heritability from varietal trial data

A. R. Rao and V. T. Prabhakaran

Indian Agricultural Statistics Research Institute, New Delhi 110 012

(Received: March, 2000; Revised: January, 2001; Accepted: March, 2001)

Abstract

The broad sense heritability, useful in the selection of promising parents in vegetatively propagated crops, is estimated as a function of variance components in the analysis of variance of multilocation varietal data. A bootstrap procedure for obtaining robust estimates of these components has been outlined in this paper. The estimates of broad sense heritability for different combinations of varieties and locations have been given along with their mean squared errors.

Key words : Broad sense heritability, multi-location trials, bootstrap procedure, simulation.

Introduction

The proportion of genetic variance to the total phenotypic variance is termed as heritability (h^2). It indicates how much of the phenotypic variation among genotypes can be carried over to the progeny. in vegetatively propagated crops the entire genetic variability is fixable. Therefore, in such crops (eg. Sugarcane) broad sense heritability is more relevant and is being estimated for important characters. Characters, which show high heritability, contribute to higher genetic advance so much so that characters with high heritability are very useful for rapid genetic improvement of these crops. When genotype x environment interaction is present the estimation must be based on multilocation. In this context the following observation by Hill [2] is worth noting, "A plant breeder pins his hopes for crop improvement upon the evidence of genetic variation for the character being selected. Accurate estimates of genetic variance will be obtained only if such estimates are unbiased by variation due to GE interactions". Singh *et al.* [1] considered the estimation of heritability using varietal trial data. Although h^2 is defined in the interval [0, 1], it sometimes assumes negative values or values exceeding unity due to sampling fluctuations. None of the traditional methods of estimation can avoid the occurrence of such estimates, termed as inadmissible estimates.

It appears that very little work has been done in estimating heritability from multi-location trial data involving genotype-environment interaction when data are generated over years. Here we propose a procedure for empirical estimation of broad sense heritability, which can also take care of the problem of inadmissible estimates. The procedure is based on bootstrap technique of Efron [3, 4] which has been modified suitably to fit in the multi-location trial situation. The quality of the estimates under different experimental situations is assessed on the basis of mean squared error and percentage root mean square error.

The method

Simulation of varietal trial data

Consider a series of trials conducted in a randomized complete block design with t genotypes and r replications, in s environments over c years. Let h^2 be the heritability of trait Y on which we have observed the response, y_{ijkl} of ith genotype (i = 1, 2, ..., t), at the jth environment (j = 1, 2, ..., s), in the kth year (k = 1, 2, ..., c) for the lth block (l = 1, 2, ..., r). The model for $y_{ijkl}$is

$$y_{ijkl} = \mu + g_i + e_j + y_k + (ge)_{ij} + (gy)_{ik} + (ey)_{jk} + (gey)_{ijk} + \beta_{jkl} + \varepsilon_{ijkl} \qquad \text{... (1)}$$

where μ is the general mean, g_i is the effect of ith genotype, e_j is the effect of jth environment, y_k is the effect of kth year, $(ge)_{ij}$ is the interaction effect of ith genotype and the jth environment, $(gy)_{ik}$ is the interaction effect of ith genotype and kth year, $(ey)_{jk}$ is the interaction effect of jth environment and kth year, $(gey)_{ijk}$ is the interaction effect of ith genotype, jth environment and kth year, β_{jkl} is the effect of lth block within jth environment within kth year and ε_{ijkl}

is the error component. The effects of g_i, e_j, y_k, $(ge)_{ij}$, $(gy)_{ik}$, $(ey)_{jk}$, $(gey)_{ijk}$ and ε_{ijkl} are assumed to be independently distributed with zero means and variances σ^2_g, σ^2_s, σ^2_y, σ^2_{ge}, σ^2_{gy}, σ^2_{ey}, σ^2_{gey} and σ^2_e respectively. The analysis of variance for the model in (1) is as given in Table 1.

Table 1. Analysis of variance for multilocation data

Source	*d.f.*	MSS	E(MSS)
Blocks/environments/years	$sc(r-1)$	Not relevant	
Genotypes (G)	$n_1 = (t-1)$	M_1	$\theta_1 = \sigma^2_e + r\sigma^2_{gey} + rs\sigma^2_{gy} + rc\sigma^2_{ge} + rsc\sigma^2_g$
Environments (E)	$n_2 = (s-1)$	M_2	$\theta_2 = \sigma^2_e + r\sigma^2_{gey} + rt\sigma^2_{ey} + rc\sigma^2_{ge} + rct\sigma^2_s$
Years (Y)	$n_3 = (c-1)$	M_3	$\theta_3 = \sigma^2_e + r\sigma^2_{gey} + rs\sigma^2_{gy} + rt\sigma^2_{ey} + rst\sigma^2_y$
G × E	$n_4 = (t-1)(s-1)$	M_4	$\theta_4 = \sigma^2_e + r\sigma^2_{gey} + rc\sigma^2_{ge}$
G × Y	$n_5 = (t-1)(c-1)$	M_5	$\theta_5 = \sigma^2_e + r\sigma^2_{gey} + rs\sigma^2_{gy}$
E × Y	$n_6 = (s-1)(c-1)$	M_6	$\theta_6 = \sigma^2_e + r\sigma^2_{gey} + rt\sigma^2_{ey}$
G × E × Y	$n_7 = (t-1)(s-1)(c-1)$	M_7	$\theta_7 = \sigma^2_e + r\sigma^2_{gey}$
Error	$n_8 = sc(t-1)(r-1)$	M_8	$\theta_8 = \sigma^2_e$
Total	$tscr - 1$		

$$h^2_b = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_s + \sigma^2_y + \sigma^2_{ge} + \sigma^2_{gy} + \sigma^2_{ey} + \sigma^2_{gey} + \sigma^2_e} \quad \text{... (2)}$$

The simulation of data based on the model in (1) has been undertaken as follows.

Suppose we want to generate varietal trial data with heritability 0.1, we choose

$$\sigma^2_g = 96.1, \sigma^2_s = 10.0, \sigma^2_y = 210.0,$$

$$\sigma^2_{ge} = 100.0, \sigma^2_{gy} = 50.0, \sigma^2_{ey} = 200.0, \sigma^2_{gey}$$

$$= 190.0$$

and σ^2_e = 105.0 or any other numerical value, so that h^2 should come to be 0.1. Several such combinations are possible, from which any combination, giving the required heritability value can be considered

The model for simulating data will be,

$$y_{ijkl} = \mu + \sigma_g g'_i + \sigma_s e'_j + \sigma_y y'_k + \sigma_{ge} (ge)'_{ij}$$

$$+ \sigma_{gy} (gy)'_{ij} + \sigma_{gy} (ey)'_{jk} + \sigma_{gey} (gey)'_{ijk} +$$

$$\beta_{jkl} + \sigma_e \varepsilon'_{ijkl} \quad \text{... (3)}$$

where g'_i, e'_j, y'_k, $(ge)'_{ij}$, $(gy)'_{ik}$, $(ey)'_{jk}$, $(gey)'_{ijk}$ and ε'_{ijkl} are standard normal variates while μ and β_{jkl} are any fixed constants. The standard normal variates are generated by Box-Muller transformation as described in Kennedy and Gentle [5].

Estimation of h^2 through improved estimation of variance components

The quality of the estimates of heritability depends on the quality of the estimates variance components obtained from the ANOVA, as outlined in table 1, for multilocation data. The quality of these components can be improved with the help of bootstrap technique. Let MSE be the parameter of interest. Let MSE be the estimator based on the simulated master sample data. Let $M\hat{S}E^*(1)$, $M\hat{S}E^*(2)$, ..., $M\hat{S}E^*(N)$ be the estimates from N bootstrap samples. Then the bootstrap estimate of MSE, namely $M\hat{S}E^*(\cdot)$ is given by

$$M\hat{S}(E)^*(\cdot) = \frac{1}{N} \sum_{i=1}^{N} M\hat{S}E^*(i)$$

The bias and sample variance are estimated as

$$\hat{\beta}_{M\hat{S}E} = M\hat{S}E^*(\cdot) - M\hat{S}E$$

$$\text{and } \hat{V}^*_{MSE(BOOT)} = \frac{1}{N-1} \sum_{i=1}^{N} [M\hat{S}E^*(i) - M\hat{S}E(\cdot)]^2$$

The variances of different mean squares (m_1, m_2, ..., m_g of Table 1) taking into account their distributional properties have been used in estimating the variance components and hence heritability. Obviously,

$$\frac{n_1M_1}{\theta_1}, \frac{n_2M_2}{\theta_2}, \frac{n_3M_3}{\theta_3}, \frac{n_4M_4}{\theta_4}, \frac{n_5M_5}{\theta_r}, \frac{n_6M_6}{\theta_6}, \frac{n_7M_7}{\theta_7}$$

and $\frac{n_8M_8}{\theta_8}$, where θ_i's (i = 1, 2, ..., 8) are the expectations of the mean squares, are independently distributed as χ^2 with $n_1, n_2, n_3, n_4, n_5, n_6, n_7$ and n_8 degrees of freedom respectively. The master sample, for bootstraping, has been generated with the help of simulation model as defined in Eq (3). This master sample is treated as if drawn from infinite population with a given heritability (parameter) value. Now bootstrapping technique (Efron and Tibshirani [6]) is applied both at genotypic and environmental level and the year and replication responses attached to the selected genotype- environment combination are automatically choosen and a new bootstrap sample is thus obtained. Likewise 200 bootstrap samples are drawn from a single master sample. From these bootstrap samples the variance of mean squares components are equated to their expectations based on their distributions to allow for the robust estimation of $\sigma_g^2, \sigma_s^2, \sigma_y^2, \sigma_{ge}^2, \sigma_{gy}^2\ \sigma_{ey}^2, \sigma_{gey}^2$ and σ_e^2 and in turn the broad sense heritability which is a function of these components.

Denoting the variance estimates of $M_1, M_2, M_3, M_4, M_5, M_6, M_7$ and M_8 by $V(M_1^*), V(M_2^*), V(M_3^*), V(M_4^*), V(M_5^*), V(M_6^*), (M_7^*)$ and $V(M_8^*)$ respectively, we can write :

$$V(M_1^*) = \frac{2\,\theta_1^2}{n_1},\quad V(M_2^*) = \frac{2\,\theta_2^2}{n_2},\quad V(M_3^*) = \frac{2\,\theta_3^2}{n_3},$$

$$V(M_4^*) = \frac{2\,\theta_4^2}{n_4};\quad V(M_5^*) = \frac{2\,\theta_5^2}{n_5},\quad V(M_6^*) = \frac{2\,\theta_6^2}{n_6},$$

$$V(M_7^*) = \frac{2\,\theta_7^2}{n_7} \text{ and } V(M_8^*) = \frac{2\,\theta_8^2}{n_8} \qquad \text{... (4)}$$

It immediately follows that

$$\theta_1^* = \frac{[n_1\ V(M_1^*)]^{1/2}}{2},\quad \theta_2^* = \frac{[n_2\ V(M_2^*)]^{1/2}}{2}$$

$$\theta_3^* = \frac{[n_3 V(M_3^*)]^{1/2}}{2},\quad \theta_4^* = \frac{[n_4 V(M_4^*)]^{1/2}}{2},$$

$$\theta_5^* = \frac{[n_5\ V(M_5^*)]^{1/2}}{2},\quad \theta_6^* = \frac{[n_6 V(M_6^*)]^{1/2}}{2},$$

$$\theta_7^* = \frac{[n_7 V(M_7^*)]^{1/2}}{2} \text{ and}$$

$$\theta_8^* = \frac{[n_8 V(M_8^*)]^{1/2}}{2} \qquad \text{... (5)}$$

Now the different variance components can be estimated as

$$\sigma_e^2 = \theta_8 \qquad \text{... (6)}$$

$$\sigma_{gey}^{2^*} = (\theta_7 - \theta_8)/r \qquad \text{... (7)}$$

$$\sigma_{ey}^{2^*} = (\theta_6 - \theta_7)/rt \qquad \text{... (8)}$$

$$\sigma_{gy}^{2^*} = (\theta_5 - \theta_7)/rs \qquad \text{... (9)}$$

$$\sigma_{ge}^{2^*} = (\theta_4 - \theta_7)/rc \qquad \text{... (10)}$$

$$\sigma_y^{2^*} = ((\theta_3 - \theta_5) - rt\,(\sigma_{ey}^{2^*}))/rst \qquad \text{... (11)}$$

$$\sigma_s^{2^*} = ((\theta_2 - \theta_4) - rt\,(\sigma_{ey}^{2^*}))/rct \qquad \text{... (12)}$$

$$\sigma_g^{2^*} = ((\theta_1 - \theta_5) - rc\,(\sigma_{ge}^{2^*}))/rsc \qquad \text{... (13)}$$

Using these bootstrap estimates the estimate of heritability is computed as :

$$\hat{h}_b^2 = \frac{\sigma_g^{2^*}}{\sigma_g^{2^*} + \sigma_s^{2^*} + \sigma_y^{2^*} + \sigma_{ge}^{2^*} + \sigma_{gy}^{2^*} + \sigma_{ey}^{2^*} + \sigma_{gey}^{2^*} + \sigma_e^{2^*}}. \qquad (14)$$

The h^2 estimates thus obtained are based on a single bootstrap replication consisting of 200 bootstrap samples. Using different 'seed' values a few more bootstrap replications are obtained and we take the average heritability estimates over these replicates to get a robust estimate of heritability. A programme in 'C' language has been developed to get the robust heritability estimate.

Results and discussion

The simulation of master samples was carried out by considering three populations with heritability values 0.10, 0.25 and 0.50. Master samples with different number of genotypes (t = 10, 15, 20, 25), environment (s = 7, 10 and 15), with 3 years and 3 replications were considered. In the present study, we have not taken into account the experimental situation involving

Table 2. Precision of bootstrap estimates of heritability (h^2 = 0.1) from multilocation trial daya over years

Genotypes, environments, years, replications (g, e, y, r)	Population value	Estimate	Bias	Sampling variance	Meansquare error	% Root meansquare error
(10, 7, 3, 3)	0.1	0.09	–0.026	0.0023	0.0029	62.7
(15, 7, 3, 3)	0.1	0.09	–0.013	0.0021	0.0023	54.8
(20, 7, 3, 3)	0.1	0.11	–0.013	0.0021	0.0023	40.8
(25, 7, 3, 3)	0.1	0.10	–0.011	0.0011	0.0012	33.8
(10, 10, 3, 3)	0.1	0.08	–0.032	0.0014	0.0025	58.8
(15, 10, 3, 3)	0.1	0.08	–0.024	0.0005	0.0006	30.7
(20, 10, 3, 3)	0.1	0.09	–0.019	0.0001	0.0005	24.1
(25, 10, 3, 3)	0.1	0.10	–0.008	0.0002	0.0003	17.4
(10, 15, 3, 3)	0.1	0.10	–0.013	0.0015	0.0017	38.3
(15, 15, 3, 3)	0.1	0.09	0.010	0.0003	0.0004	22.9
(20, 15, 3, 3)	0.1	0.09	0.008	0.0003	0.0004	22.5
(25, 15, 3, 3)	0.1	0.09	–0.002	0.0002	0.0002	15.8

Table 3. Precision of bootstrap estimates of heritability (h^2 = 0.25) from multilocation trial data over years

Genotypes, environments, years, replications (g, e, y, r)	Population value	Estimate	Bias	Sampling variance	Meansquare error	% Root meansquare error
(10, 7, 3, 3)	0.25	0.17	–0.082	0.0058	0.0126	64.8
(15, 7, 3, 3)	0.25	0.22	–0.052	0.0045	0.0073	38.8
(20, 7, 3, 3)	0.25	0.24	–0.035	0.0014	0.0027	21.8
(25, 7, 3, 3)	0.25	0.24	–0.031	0.0010	0.0019	18.7
(10, 10, 3, 3)	0.25	0.20	–0.073	0.0049	0.0104	50.2
(15, 10, 3, 3)	0.25	0.21	–0.064	0.0024	0.0066	38.5
(20, 10, 3, 3)	0.25	0.20	–0.033	0.0004	0.0015	19.3
(25, 10, 3, 3)	0.25	0.23	–0.025	0.0008	0.0014	16.4
(10, 15, 3, 3)	0.25	0.21	–0.045	0.0024	0.0045	31.2
(15, 15, 3, 3)	0.25	0.25	–0.025	0.0022	0.0029	21.5
(20, 15, 3, 3)	0.25	0.22	–0.019	0.0010	0.0014	17.5
(25, 15, 3, 3)	0.25	0.23	–0.011	0.0007	0.0008	12.9

Table 4. Precision of bootstrap estimates of heritability (h^2 = 0.5) of multilocation trial data over years

Genotypes, environments, years, replications (g, e, y, r)	Population value	Estimate	Bias	Sampling variance	Meansquare error	% Root meansquare error
(10, 7, 3, 3)	0.5	0.40	–0.19	0.0017	0.0137	29.1
(15, 7, 3, 3)	0.5	0.44	–0.097	0.0018	0.0113	23.9
(20, 7, 3, 3)	0.5	0.44	–0.086	0.0019	0.0093	21.8
(25, 7, 3, 3)	0.5	0.52	–0.036	0.0009	0.0022	6.5
(10, 10, 3, 3)	0.5	0.40	–0.095	0.0038	0.0128	28.3
(15, 10, 3, 3)	0.5	0.41	–0.089	0.0040	0.0120	27.0
(20, 10, 3, 3)	0.5	0.48	–0.083	0.0029	0.0098	20.4
(25, 10, 3, 3)	0.5	0.47	–0.027	0.0008	0.0015	8.3
(10, 15, 3, 3)	0.5	0.47	–0.092	0.0085	0.0170	27.7
(15, 15, 3, 3)	0.5	0.52	–0.080	0.0110	0.0174	25.1
(20, 15, 3, 3)	0.5	0.52	–0.047	0.0014	0.0037	11.8
(25, 15, 3, 3)	0.5	0.50	–0.021	0.0007	0.0012	7.1

more than 3 years, as in practice varieties are grown for two years as initial varietal trials (IVT) and one year as uniform varietal trials (UVT), and the number of replications rarely exceeds three. Ten bootstrap replications were obtained under each combination (h^2, t, s, c, r) by changing the random 'seed' values. Two hundred bootstrap samples from each of these replicates were generated and used to estimate the parameter of interest. The precision of the estimates are computed by simulating 200 master samples from each of the combination (h^2, t, s, c, r). Average of the bootstrap estimates over the master samples is taken as the expected value of the estimator, and deviation of this from the population value, provides the bias in the estimate. The variance of the estimates is also determined by considering the estimates from all the master samples. The mean square error (MSE) is worked out from the relationship, MSE ($\hat{\theta}$) = Var ($\hat{\theta}$) + Bias^2 ($\hat{\theta}$). The percentage root mean square error is considered here as a measure of reliability of the estimation procedure. The estimates along with the measures of their precision are given in Tables 2 to 4. The results reveal that the limitation in size of the experiment is not going to affect the robustness of the estimate to any significant extent. In situations where the performance of a large number of genotypes (t = 20 to 25) is tested data even from a smaller number of environments (s = 7) would give a very good estimate of heritability. However, in situations involving fewer genotypes, this has to be compensated by considering much higher number of environments. The first situation, with a large number of genotypes and fewer environments is more commonly encountered in varietal testing. As can be appreciated, robust estimates of heritability, in this case, can be obtained by basing the computation on data coming from seven or more environments. The reliability of the estimates is quite evident from the low or more environments. The reliability of the estimates is quite evident from the low values of percentage root mean square appearing in Tables 2 to 4. It may however, be noted that the robustness of the estimates is much higher in the cases of $h^2 \geq 0.25$ as compared to $h^2 \geq 0.1$.

References

1. **Singh M, Ceccarelli S and Hamblin J.** 1993. Estimation of heritability from varietal trials data. Theor. Appl. Genet., **86**: 473-41.
2. **Hill J.** 1975. Genotype - environment interaction - a challenge for plant breeding. J. Agric. Sci., Camb., **85**: 477-93.
3. **Efron B.** 1979. Bootstrap methods: Another look at Jackknife. Ann. Statist., **7**: 1-26.
4. **Efron B.** 1982. The Jackknife, the Bootstrap and other resampling plans. CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph, **38**, Philadelphia: SS/M.
5. **Kennedy Jr., W. J. and Gentle J. E.** 1980. Statistical computing. Marcel Dekker Inc., New York and Basel.
6. **Efron B and Tibshirani R.** 1993. An Introduction to the Bootstrap. Chapman and Hall.