A Bootstrap Look on the Precision of Heritability Estimators in Unbalanced Data

V.K. Bhatia and J. Jayasankar[1]
Indian Agricultural Statistics Research Institute, New Delhi, India

SUMMARY

This paper attempts at modifying of the bootstrap procedure for precision estimation of heritability estimates obtained under different family structures in unbalanced data sets. Parametric bootstrapping is compared against exact values in variances and bias estimation. The non-parametric bootstrap confidence intervals are compared with certain intervals available in literature in the case of half sib and parent offspring. The influence functions of sire effect in the half sib case are estimated by using jackknife after bootstrap and further the standard error of bootstrap estimates are obtained. Results obtained showed overwhelming performance of parametric and non-parametric bootstrap procedure for various precision parameters of heritability estimates. It is seen that in comparison to parametric bootstrapping, the non-parametric bootstrapping is not lagging behind in terms of information about the precision. The performance of non-parametric bootstrapping in the case of worst samples is even better than the theoretically proven good procedure of parametric bootstrapping. Both parametric and non-parametric bootstrapping perform better than the approximate procedures of Taylor series and Smith. Through influence function it is seen that variance estimation is the more informative procedure of bootstrap.

Key words : Parametric and non-parametric bootstrapping, Heritability, Jackknife after bootstrap.

1. Introduction

Heritability is one of the most important genetic parameters widely used in plant and animal breeding studies. However, it is hardly possible to get to its real value in practical circumstances. It is a well known fact that any estimate devoid of any measure of its precision is not of much use. Lately attempts have been made to arrive at the distribution of heritability by using the distribution of individual variance components. The logic followed is to express the mean squares which are quadratic forms involving the vector of phenotypic values as a linear combination of mutually independent chi square variables with one degree of freedom each. It can be seen that heritability estimator turns out to be a function of ratio of chi squares whose distribution function can

---

1   Present Address : C.S.W.R.I. (ICAR), Avikanagar, Malpura, Rajasthan.

be got as a sum of infinite series involving incomplete beta function (Singh [16], [17]). Using some approximations based on the distribution function, the exact value of variance of heritability in the one-way data was arrived at by Singh [17]. Other approximations like those of Smith's and Taylor series were widely discussed and compared by Smith [18] and Singh [17]. The confidence interval estimation also followed almost the same path as that of variance estimation as reported by Searle [14] and Graybill [10]. Tai [19] derived confidence intervals of heritability for balanced data wherein the between sires mean square follows a simple chi square distribution. Complications however arise, once we deal with an unbalanced data. In such cases the between sires mean squares has to be expressed as difference between two chi-squares. Exact confidence intervals for one-way data were discussed in Harville and Fenech [12] wherein after the derivation of exact interval for practical computing purposes subsequent approximations were also mentioned. Satterthwait type approximations were recommended for two-way nested unbalanced (Full sib) case by Sen *et al.* [15]. However in practical situation such procedures are the near impossibility of calculating the exact confidence intervals and the coverage probabilities.

In addition to the above, one of the constant nagging problem in all such estimation procedures is their over dependency on the normality assumptions and the influence of the sample on the final values. Though in the model assumptions, the various effects of the components like sires effect and dam effect are supposed to be random but in a particular data of finite number of sires/dams a few are expected to be highly influential which is also a worthy area to concentrate upon. Thus a distribution studying procedure which is moderately robust to the sample fluctuations will be useful for such situations. The Bootstrap method is one such method which can be attempted for such problems. After bootstrapping one can also get jackknife after bootstrap estimates for the precision.

Although the concept of bootstrapping originated with an one sample problem in mind but there are many situations which warranted their applicability in linear models. The applications of this technique on regression models is available in Efron [7], Hall [11], and the major attempt to apply bootstrapping in a typical random model case is discussed by Aastveit [1] and Bhatia *et al.* [3]. A specific algorithm mentioned by Aastveit is used to estimate the variance and confidence intervals of heritability by Bhatia *et al.* [3] and Bhatia and Jayasankar [2]. One of the short comings of that algorithm is that it cannot analyse different bootstrap samples arising from unbalanced data for estimation of heritability and its precision on account of the change in total number of observations. Also that algorithm is not suitable for the estimation of heritability from a dam-daughter data.

The present study therefore attempts at validating parametric bootstrapping against exact values and comparing the performance of non-parametric and parametric bootstrap estimates for variance and bias estimation. The non-parametric bootstrap confidence intervals are further compared with certain intervals available in literature in the half sib data case. The non-parametric bootstrap is also used to estimate bias, variance and confidence intervals in the case of parent-offspring data structures. The influence functions of sire effect in the half sib case is also estimated by using jackknife after bootstrap and thus the standard error of bootstrap estimates are obtained.

2. *Methodology Used*

2.1. *The naive bootstrap*

Bootstrap, a resampling procedure, shot into limelight through the path-breaking paper of Efron [5]. The simplicity of the logic, which is based on the empirical distribution of any distribution backed by sound theoretical justification in the form of Edgeworth expansion, multinomial distribution by Efron ([6], [7]) and Hall [11], has caught the imagination of applied statisticians throughout the world. The best part of the bootstrap logic is the approximation which it admits by the Monte Carlo algorithm (Efron [7]) and is highly computer intensive.

In a simple one sample situation let the observed data $x' = (x_1, x_2, \ldots x_n)$ be obtained by iid sampling from an unknown distribution F,

$$F \xrightarrow{\text{iid}} (x_1, x_2 \ldots x_n) = x'$$

The usual nonparametric estimate of F is $\hat{F}$, the empirical probability distribution, putting probability $1/n$ on each point x_i,

$$\hat{F}: \text{probability } 1/n \text{ on } x_i, \ i = 1, 2, \ldots, n$$

A bootstrap sample $x^{*'} = (x_1^*, x_2^*, \ldots x_n^*)$ is a random sample of size n drawn from $\hat{F}$,

$$\hat{F} \xrightarrow{\text{iid}} (x_1^*, x_2^*, \ldots, x_n^*) = x^{*'}$$

is used in finding out $g^* = g(x^*)$ which is the bootstrap replication of any statistic involving smooth function $g = g(x)$. The practical Monte Carlo algorithm for getting the replicate is to draw independently large number (B) of bootstrap samples, evaluating $g^{*b} = g(x^{*b})$ for $b = 1, 2, \ldots B$ and estimating precision parameters of $g(x)$.

The bias is estimated by

$$\text{bias}_{\text{boot}}(g) = g^* - g(x)$$

where $g^* = 1/B \ \Sigma \ g^{*b}$

and the sampling variance by

$$SV_{\text{boot}}(g) = (1/(B-1)) \ \Sigma \ (g^{*b} - g^*)^2$$

The confidence intervals are estimated by the percentile intervals,

$$[g^{*(0.05)}, g^{*(0.95)}]$$

wherein $g^{*(\alpha)}$ denotes the 100α th percentile of the empirical distribution of the B bootstrap replications g^{*b}.

2.2. Model based bootstrapping

Efron [7] and Hall [11] described bootstrapping based on residuals of regression models

i.e. for the model

$$y = X \beta + e$$

$r = y - X \beta$ is calculated and resample vectors r^* are taken from r and the bootstrap set is got by $y^* = X \beta + r^*$

The same idea was extended to balanced random models by Aastveit [1] and Bhatia et al. [3] wherein for the model

$$y = \mu \ 1 + X \ \tau + e$$

y^* is constructed by selecting families at random and repeating all the observations of a particular sire. But this can not be worked out in the case of unbalanced data as the total data size will vary in the resamples. Thus there is a need to use bootstrapping in such a way that the family and within family effect are selected at random without disturbing the data structure. The parametric bootstrapping has the assumption about the underlying distribution and can be the best tool in case the distribution is known correctly. Herein the sample is only used to estimate the parametric values and not the distribution. We can generate data from a known distribution, the parametric bootstrapping can be used as a control against whose performance, the non-parametric bootstrap can be compared.

2.3. *Bootstrapping for unbalanced data*

Let us consider the half sib data analysis.

The model is

$$y_{ij} = \mu + s_i + e_{ij}$$

$$i = 1, 2, \ldots . s$$

$$j = 1, 2, \ldots n_i \text{ and } \Sigma n_i = N$$

where

y_{ij} = Phenotypic value of the jth half sib of ith sire

μ = general mean

s_i = effect of ith sire

e_{ij} = error effect of $(i, j)^{th}$ half sib (iid)

Then

$$\hat{h}^2 = \frac{4 \hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_e^2}$$

For typical inferences it is assumed that $s_i \sim N(0, \sigma_s^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. In other words irrespective of the normality assumption $\underline{\hat{s}}' = (\hat{s}_1, \hat{s}_2, \hat{s}_3, \ldots \hat{s}_s)$ is a random sample of size s from the population of sire effects and $\underline{\hat{e}}' = (\hat{e}_{11}, \hat{e}_{12}, \hat{e}_{13} \ldots \hat{e}_{1n1}, \ldots \hat{e}_{sns})$ is a random sample of size N from the population of error effects. Under this set up, bootstrapping can be thought of with the resample vectors being

$$\underline{s}^{*'} = (s_1^*, s_2^*, \ldots s_s^*)$$

and

$$\underline{e}^{*'} = (e_{11}^*, e_{12}^* \ldots, e_{1n1}^* \ldots, e_{sns}^*)$$

thus getting the bootstrap phenotypic values as

$$y_{ij}^* = \hat{\mu} + s_i^* + e_{ij}^*$$

$$\text{with} \quad i = 1, 2, \ldots, s$$
$$j = 1, 2 \ldots, n_i$$

where $\hat{\mu} = \overline{y}_{..}$ the mean of the sample observations. The estimates of sire effects can be obtained as

$$\hat{s}_i = \overline{y}_{i.} - \overline{y}_{..} \qquad \text{where } \overline{y}_{i.} = (1 / n_i) \, \Sigma_j \, y_j$$

and those of error effects by

$$\hat{e}_{ij} = y_{ij} - \overline{y}_{i.} - \overline{y}_{..}$$

The estimates of heritability obtained from y^* are the bootstrap replicates for h^2. The same thing can be extended for full sib data wherein the model is a two way nested model.

But in the case of parent-offspring data the algorithm will be different. The model of analysis is

$$P_{ij} = \mu_p + g_{pi} + e_{pij}$$

$$O_{ij} = \mu_o + a \, g_{pi} + \sqrt{(1 - a^2)} \, g_{oi} + e_{oij}$$

$$\text{with} \qquad i = 1, 2 \ldots s$$

$$j = 1, 2 \ldots p_i$$

where P_{ij} is the phenotypic value of j-th dam mated to i-th sire and O_{ij} is the phenotypic value of the offspring of j-th dam-daughter pair. a is the coefficient of relationship which is 1/2 in parent-offspring case. From this model it is clear that this amount of genotypic correlation has to be maintained in the bootstrap samples. With an aim to achieve this, bootstrap method is proposed as follows.

Two samples of size s each, are taken from the samples $g_p = (g_{p1}, g_{p2}, \ldots, g_{ps})$ and $g_o = (g_{o1}, g_{o2}, \ldots g_{os})$ respectively. Let the resamples be $g_p^{*'}$ and $g_o^{*'}$. The error resamples are taken as usual to give e_p^* and e_o^*. The new bootstrap sample is

$$P_{ij}^* = \mu_p + g_{pi}^* + e_{pij}^*$$

$$O_{ij}^* = \mu_o + a \cdot g_{pi}^* + \sqrt{(1 - a^2)} \, g_{oi}^* + e_{oij}^*$$

The main difference here is g_{pi}^* is the same for both P_{ij}^* and O_{ij}^*. The selection of various effects at random is done by the aid of the built-in random (0, 1) generator.

2.4. *Parametric bootstrapping*

Apart from the above mentioned non-parametric bootstrapping, parametric bootstrapping is also done to estimate the bias and central moments of heritability estimators in the sib analysis case. According to Efron and Tibshirani [9] for the one sample situation, the bootstrap samples are generated from the estimated population F_{PAR} which is the original distribution but with parametric values estimated from sample.

$$F_{PAR} \longrightarrow (x_1^*, x_2^*, \ldots x_n^*)$$

These samples are taken B times and from the B replicates the required measures are obtained in the same way as mentioned in the non-parametric case. For the sub data from the original sample the various variance components like $\sigma_s^2, \sigma_d^2, \sigma_e^2$ are estimated. Using the estimates as the parametric values, half sib and full sib data are generated according to the models given by Ronningen [13] with the assumption that the random effects follow normal distribution with mean 0 and corresponding variances. If the real parametric values are known, this parametric bootstrapping reduces to repeated simulation to data sets of a given size from the same population.

2.5. *Identification of influential sires*

Towards satisfying the need to study the influence of influential sires which incidentally turn out to be a sample of size s from the population of sires, the empirical influential functions of sire effects are calculated for a given sample on the lines of Efron [7]. To achieve this, jackknifing is done on the B bootstrap replicates without any additional data generation. From the set of B replicates, the replicates which are got without the ith sire effect are regrouped and from each of the s such groups the various parameters like bias, variance, 90% confidence intervals etc. are calculated. From the s such values of each parameter the influence $U_i (\hat{\tau})$ and the relative influence $U_i^+ (\hat{\tau})$ are calculated.

where

$$U_i (\hat{\tau}) = (s - 1) (\overline{\tau}_{()} - \overline{\tau}_{(i)})$$

$$\text{where } \overline{\tau}_{()} = (1 / s) \Sigma \, \hat{\tau}_{(i)}$$

and

$$U_i^+ (\hat{\tau}) = U_i (\hat{\tau}) / \{ \Sigma U_i (\hat{\tau}) / s - 1) \}^{1/2}$$

where

$\hat{\tau}$ is any precision parameter under study (Bias, variance etc.) Efron [7]. These influence values also help in getting a jackknife variance estimate of the bootstrap estimates.

2.6. *Length, normalised length and shape*

The various other parameters concentrated apart from bias, variance and confidence interval of the estimates of heritability are higher order moments, length, normalized length and shape of the curve as discussed in Efron [7].

Thus if U and L are the upper and lower 90% confidence limits of estimate of parameter h^2, then

$$\text{Length} = U - L$$

$$\text{Normalised Length} = (U - L) / 2 \times 1.645 \; \sqrt{\; Var_{BOOT}}$$

and $$\text{Shape} = \log_e \{ (U - \text{Median}) / (\text{Median} - L) \}$$

3. Results and Discussion

The precision of heritability estimators for the different family structures have been obtained by using bootstrap technique on the simulated data of various sizes. The parametric, non-parametric bootstrapping and as well as exact theoretical results are obtained and they are as under :

3.1. *Half sib family structure*

To compare the precision of the bootstrap estimators with those of the theoretical estimators, half sib data sets of size 15 and 25 were generated with three and five sires and at three levels of unbalancedness. A simulated data set is used in the case of non-parametric bootstrapping with the intention to knowing the actual value of heritability. With these actual values parametric bootstrapping was also done. The results for bias estimation were compared with the actual bias values calculated by Singh [17] and are presented in Table 1 for the case of three sires and in Table 2 in the case of five sires.

Table 1. True, parametric & non-parametric bootstrapping bias of heritability for different values of h^2 under different magnitudes of unbalancedness in case of three sires

(n_1 n_2 n_3)	Degree of unbalancedness	True bias	Parametric	Non-parametric
$h^2 = 0.0$				
5 5 5	0	–0.0438	–0.0234	–0.0116
3 3 9	12	–0.0842	–0.1068	–0.1619
1 1 13	48	–0.8271	–0.8705	–0.8464
$h^2 = 0.4$				
5 5 5	0	–0.1285	–0.0920	–0.1809
3 3 9	12	–0.1741	–0.1979	–0.4225
1 1 13	48	–0.9166	–0.8770	–1.0010

Table 2. Parametric and non-parametric bootstrapping bias of heritability for different values of h^2 under different magnitudes of unbalancedness in case of five sires

(n_1	n_2	n_3	n_4	n_5)	Degree of unbalancedness	Parametric	Non-parametric
$h^2 = 0.0$							
5	5	5	5	5	0	0.0015	−0.0461
3	3	7	7	5	4	−0.0191	−0.0319
1	1	2	2	19	61.5	−0.1587	−0.3920
$h^2 = 0.4$							
5	5	5	5	5	0	−0.0770	−0.0578
3	3	7	7	5	14	−0.0726	−0.0522
1	1	2	2	19	61.5	−0.2771	−0.4946

From Table 1, where exact bias value is available, it can be noticed that bias is increasing with the increase in unbalancedness. The degree of unbalancedness is calculated as follows :

$$\text{Degree of unbalancedness} = N (n - K)$$

where $N = \Sigma n_i$ is the total data size and $\bar{n} = N / s$

and

$$K = (1 / (s - 1)) [n - \Sigma n_i^2 / N]$$

The parametric bootstrap values are closer to the exact values due to the fact that the parametric values are known. The non-parametric bootstrap values are also not lagging far behind. Both follow the same trend i.e. increase with the increasing unbalancedness. In the five sire case (Table 2) also, the trend followed is same but the overall bias values are reduced due to the fact that the family size and number of sires have increased. One more noteworthy observation is that lower the degree of unbalancedness the lower are the bias values both in parametric or non-parametric bootstrapping.

The variance estimates of heritability estimator for the half sib data are given in the Table 3 for the case of three sires and in Table 4 for five sires. As the original data is a simulated one we can compare the non-parametric and parametric bootstrap performances with the exact values and Smith and Taylor series approximations at the backdrop.

Table 3. Exact, Smith, Taylor series, parametric and non-parametric variances for different values of heritability in 3 sires

$(n_1$	n_2	$n_3)$	Degree of unbalanced-ness	Exact	Smith	Taylor series	Parametric	Non-para-metric
$h^2 = 0.0$								
5	5	5	0	0.5799	0.7467	1.1017	0.6068	1.0042
3	3	9	12	0.7781	1.0582	1.6265	0.7841	0.9846
1	1	13	48	4.6751	5.7613	6.9984	4.6241	4.3073
$h^2 = 0.4$								
5	5	5	0	0.8375	1.1854	1.7649	0.8552	0.9479
3	3	9	12	1.0366	1:6959	2.7104	1.0276	1.4458
1	1	13	48	4.8238	5.4688	6.7560	4.6359	5.0193

Table 4. Smith, Taylor series, parametric and non-parametric variances for different values of heritability in 5 sires

$(n_1$	n_2	n_3	n_4	$n_5)$	Degree of unbalancedness	Smith	Taylor series	Parametric	Non-parametric
5	5	5	5	5	0	0.3840	0.4975	0.5029	0.4943
3	3	7	7	5	4	0.6427	0.9784	0.5296	0.4600
1	1	2	2	19	61.50	1.6434	2.8260	1.3165	1.2597

From the above two Tables 3 and 4 it can be seen that as the degree of unbalancedness increases, the variance also shows up considerably. The parameter bootstrap variances are the closest to the exact variances followed by the non-parametric bootstrap variances. The parametric bootstrap is obtained using the population values where as the non-parametric bootstrapping is done on a finite sample generated with a particular h^2. Moreover in some cases the sample may give inadmissible estimate also, which can not be used for parametric bootstrapping. Even the approximate variances are calculated based upon the population values of σ_s^2 and σ_e^2. But in practice one can make use of the sample estimates only, which means a further fall in the precision of the approximation as discussed in Table 5.

Table 5. Comparison of parametric and non-parametric bootstrapping.

h^2	$\hat{h}^2$	Approximation based on population		Approximation based on sample		Non-parametric bootstrap variance
		Smith	Taylor	Smith	Taylor	
0.0	-0.9420	0.4098	0.5294	0.0774	2.1814	0.4095
0.4	-0.4006	0.3840	0.4975	0.1670	0.2240	0.4943
0.8	-0.3834	0.7963	0.9629	0.1753	0.2349	0.4927

It can be seen from Table 5 that even with every bad samples the bootstrap particularly non-parametric extracts better information about the variance. One more point worth noting is the highly inflated variance values given by the Taylor series approximations especially in the extremely unbalanced cases.

The confidence interval estimation always happens to be a ticklish problems in the case of heritability because there is a danger of getting inadmissible estimates. While the often suggested remedy to avoid the inadmissible estimates is to replace the inadmissible ones by the suitable permissible extremes (0 & 1 in the case of heritability). The soundness of such intervals is under clouds because of the restrictions. A comparative study was done to evaluate the performance of non-parametric bootstrap percentile intervals with those suggested as best by Donner & Walls [4] which is a symmetric interval based on Smith variance and Harville and Fenech [12] who have given approximate versions for the exact intervals. For this comparison, 90% intervals are chosen.

Table 6. Confidence intervals and coverage probability for different values of heritability in the case of 3 sires.

n_1	n_2	n_3	Harville & Fenech	C.P.	Smith	Non-parametric	C.P.
$h^2 = 0.0$							
5	5	5	[−0.1074, 1.8760]	40	[-1.4151, 1.8273]	[−0.8863, 2.3702]	90
3	3	9	[−0.4271, 2.8100]	54	[- 1.6745, 0.7881]	[−1.1503, 2.0706]	92
1	1	13	[−1.6027, 2.4148]	55	[-7.1081 , 1.7941]	[−4.4042, 2.5609]	90
$h^2 = 0.4$							
5	5	5	[−0.1531, 2.0476]	54	[-1.4197, 1.5567]	[−0.8461, 2.2371]	88
3	3	9	[0.2321, 0.9850]	20	[-1.1617, 3.5291]	[−0.9675, 2.8729]	87
1	1	13	[−0.8927, 1.1816]	31	[-5.2584, 2.8464]	[−4.0429, 3.0636]	90

Table 6. shows that as the degree of unbalancedness increases the length of the interval increases. The coverage probability of Harville-Fenech approximate intervals is very badly affected by the sample values though the length is shorter. The coverage of the non-parametric bootstrap estimator is pretty good with the value lingering around the 90% mark. Though the Smith interval is expected to give very good coverage, it is badly affected by the sample by giving unusually longer left tails with rise of unbalancedness.

One more drawback of the Smith intervals is that it assumes the tails to be equidistant from the sample estimate, which is not so in the case of heritability estimates. The following discussion throws some light in the nature of the distribution.

The cases of three and five sires are considered with population heritability values at 0.4 and 0.8. The percentile intervals alongwith the length, normalized length and shape statistics are obtained and are given in Table 7.

Table 7. Length, normalised length and shape for bootstrap estimates of heritability

n_1	n_2	n_3	n_4	n_5	90% Percentile interval	Length	Normalised length	Shape
3 Sires								
$h^2 = 0.4$								
5	5	5			[−0.8461, 2.2371]	3.0832	0.9625	0.4468
3	3	9			[−0.9675, 2.8729]	3.8404	0.9708	−0.1006
1	1	13			[−4.0429, 3.0636]	7.1065	0.9641	−0.6350
5 Sires								
$h^2 = 0.4$								
5	5	5	5	5	[−0.7079, 1.5675]	2.2754	0.9837	0.2126
3	3	7	7	5	[−0.6980, 1.5712]	2.2692	1.0169	0.4022
1	1	2	2	19	[0.0335, 3.6882]	3.6527	0.9892	−1.6356
$h^2 = 0.8$								
5	5	5	5	5	[−0.6335, 1.6267]	2.2602	0.9738	0.3673
3	3	7	7	5	[−0.2362, 2.4735]	2.7097	0.9920	−0.0484
1	1	2	2	19	[−1.4051, 2.6037]	4.0088	1.0154	−0.1925

From the Table 7 it can be observed that in general the length of the confidence interval increases with increase in the degree of unbalancedness. Moreover the shape factor denotes that as the data gets highly unbalanced, the

length of the right tail with respect to median gets shortened in comparison with the left tail. Regarding normalized length which is the ratio of the actual length and that of symmetric interval, where in the estimates are assumed to be normally distributed, is almost near unity for all the cases. This implies that percentile intervals are behaving in a similar way as that of classical confidence intervals following normal theory.

The changes undergone by the two estimates of variances components viz. $\hat{\sigma}_s^2, \hat{\sigma}_e^2$ and the estimate of heritability due to change in unbalancedness for a three sire case are shown in Figure 1, 2 and 3 by way of histograms.

From the Figures it can be noticed that the distribution of $\hat{\sigma}_e^2$ is almost symmetric and almost unaffected by this type of data. But the distribution of $\hat{\sigma}_s^2$ is skewed with a longer right tail. The most important impact is on the distribution of $\hat{h}^2$ wherein the heavy left tail slowly vanishes with the creation of heavy right tail as the degree of unbalancedness increases. The spread of the $\hat{h}^2$ distribution gives us the idea as to why the Harville-Fenech's confidence interval is giving a low percentage coverage. It also sufficiently shapes up the symmetry assumption of the Smith confidence interval estimate.
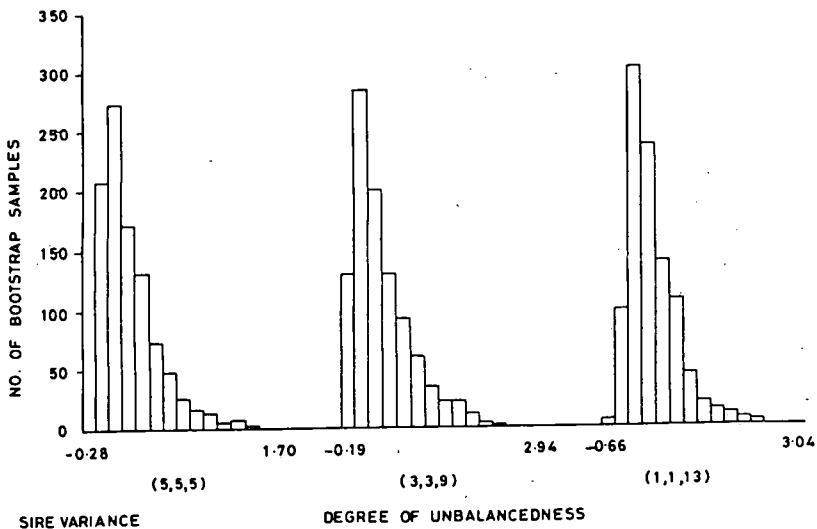

Fig. 1. Distribution of estimates of sire variance for varying degree of unbalancedness
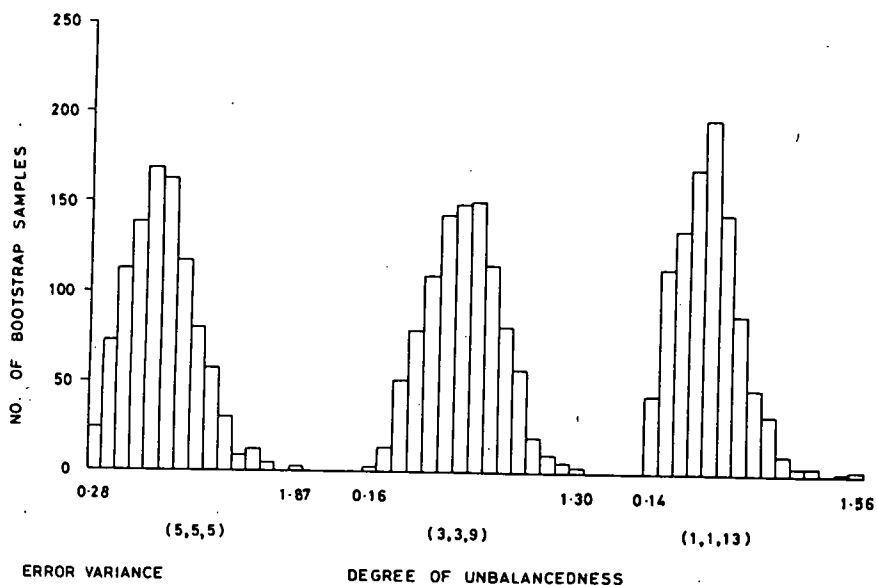
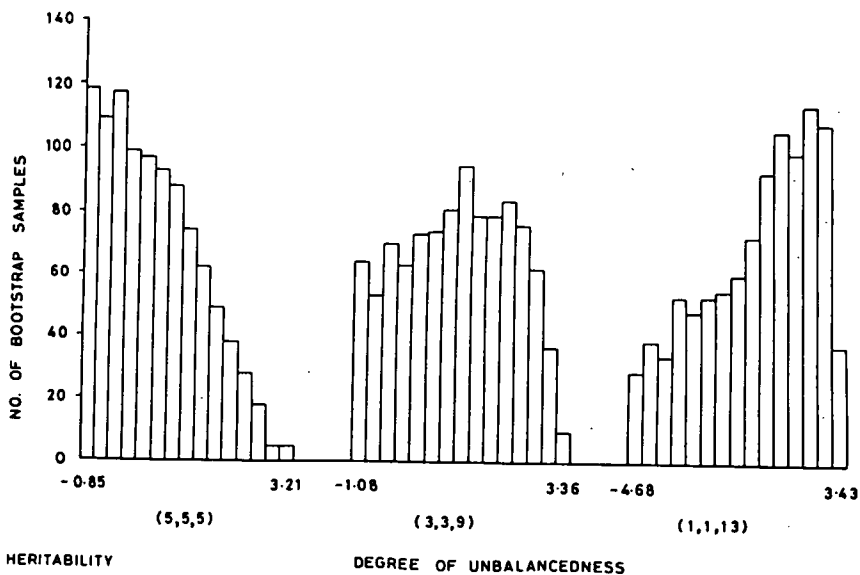**Fig. 2.** Distribution of estimates of error variance for varying degree of unbalancedness

Fig. 3. Distribution of estimates of heritability for varying degree of unbalancedness

3.2. *Parent offspring family structure*

Using the modified bootstrap algorithm, samples generated with five sires each having different number of dam-daughter pairs and with population heritability values 0.0, 0.4 and 0.8 were bootstrapped and the results are given in Table 8.

Table 8. Bias, variance and 90% confidence interval for different dam-daughter pairs for given heritability

No. of pairs/sires					Bias	Variance	90% Confidence interval
$h^2 = 0.0$							
5	5	5	5	5	0.0025	0.1490	[−0.6409, 0.6499]
						(0.1363)	
3	3	7	7	5	0.0015	0.2937	[−0.8605, 0.8815]
						(0.2773)	
1	1	2	2	19	0.0047	0.2467	[−0.7879, 0.7954]
						(0.2351)	
$h^2 = 0.4$							
5	5	5	5	5	0.00245	0.1086	[−0.5351, 0.5362]
						(0.1020)	
3	3	7	7	5	0.000755	0.1543	[−0.6394, 0.6174]
						(0.1119)	
1	1	2	2	19	0.003657	0.1523	[−0.6210, 0.6607]
$h^2 = 0.8$						(0.1330)	
5	5	5	5	5	0.002309	0.3122	[−0.9170, 0.9245]
						(0.2769)	
3	3	7	7	5	0.008681	0.3187	[−0.9212, 0.8900]
						(0.2455)	
1	1	2	2	19	0.000487	0.1470	[−0.6378, 0.6000]
						(0.1185)	

Figures in the parentheses indicate the theoretical values of the estimate of variance of heritability estimator.

It may be noticed from Table 8 that there is an almost zero bias and closeness of bootstrap variance estimates with the theoretical ones. This implies that intra-sire regression method gives unbiased estimate of heritability. The variation in the heritability and the length of the confidence interval is very

less due to change in the number of pairs per sire. One more noteworthy feature is the lying of the upper limit in the admissible range whereas the lower limit expands to the left as the heritability value increases. This advocates that if the data on dam-daughters are available then one should always go for intra-sire regression method.

3.3. Influence of sire effects

Taking clue from Efron [7] an attempt was made to study the influence of sire effects which are originally considered to be random upon the various precision estimates. The approximate jackknife influence values and the relative influence values for bias, variance, length of 90% percentile interval and shape are calculated and are given in Table 9. The case taken was a 10 sire half sib data. The original bootstrap estimates are given in Table 9 alongwith the jackknife-after-bootstrap estimate of standard error for the various statistics. The latter is obtained as

$$\hat{Se}_{jack} \, (\hat{\tau}) = [\, \Sigma \, \tilde{U}_i \, (\hat{\tau})^2 \, / \, s \, (s-1) \,]^{1/2}$$

where $\tilde{U}_i \, (\tau)$ is the jackknife influence function and s is the number of sires.

Table 9. Influence function of different statistics of precision

Sire No.	Bias	Variance	Length of 90% confidence interval	Shape
1	0.1389	−0.0751	−0.3642	−0.3569
2	0.2312	0.2174	0.1749	−0.0292
3	0.1109	−0.3662	−0.8682	−0.3749
4	−2.2381	1.1831	2.8110	0.6404
5	0.2882	−0.1651	−0.3318	−0.2021
6	−0.3733	0.1715	0.9660	1.4864
7	0.5258	−0.2569	−0.7971	−0.5135
8	0.2288	−0.4162	−0.7782	−0.5918
9	0.6302	0.1157	0.0561	0.5882
10	0.4574	−0.4063	−0.8682	−1.3604
Bootstrap estimate	0.0566	0.3932	2.0874	0.1096
Se_{jack}	0.2636	0.1512	0.3630	0.2525
C.V.	4.6500	0.3845	0.1738	2.3038

The table indicates the presence of influential sires in the term of sire no. 4 which has a definite influence on the bias and length apart from moderate influence on variance and shape. Of the four precision estimates, the jackknife after bootstrap variance of the bootstrap estimates showed least variance for bootstrap variance estimate followed by shape. The coefficient of variation (C.V.) laid more emphasis on length in comparison to variance. Wherever there is positive influence on bias there is a negative influence on variance in most of the cases.

4. Conclusion

The different results indicate the overwhelming performance of the parametric bootstrap in the precision of heritability estimates. The non-parametric bootstrap also does not lag far behind. Both give better information about the sampling distribution of h^2. The variance estimation is better performed by both these methods than the Smith and Taylor series approximations. The performance of percentile confidence interval is better in terms of coverage and are the shortest possible to give that coverage probability. For fewer sires the distribution is distinctly asymmetric with the degree of influence having a tilting influence on it. The algorithm used in parent-offspring to do non-parametric bootstrapping performs well producing almost zero bias and variance estimates close to the expected values. The influence function estimated by means of jackknifing after bootstrapping is able to bring out influential sire effect and the related level of influence on various precision statistics. It also gives us an idea about the precision of the various bootstrap precision estimates.

Though the parametric bootstrapping is the best method to be followed under normality assumptions, it fails to take off intake of negative estimates of heritability. In such situations non-parametric bootstrapping become inevitable. The level of performance of non-parametric bootstrapping in the case of worse samples is also very good. So given a practical situation the best method to adopt is the non-parametric bootstrapping.

REFERENCES

[1]   Aastveit, A.H., 1990. Use of bootstrapping for estimation of standard deviation and confidence intervals of genetic variance and covariance component. *Biom. J.*, **34**, 515-517.

[2]   Bhatia, V.K. and Jayasankar, J., 1996. On use of bootstrapping for estimation of confidence intervals. *J. Indian Soc. Agric. Statist.*, **48(2)**, 217-223.

[3]   Bhatia, V.K., Jayasankar, J. and Wahi, S.D., 1994. Use of bootstrap technique for variance estimation of heritability estimators. *Ann. Agric. Res.*, **15**, 446-480.

[4]   Donner, A. and Wells, G., 1986. A comparison of confidence interval methods for the intra class correlation coefficient. *Biometrics,* **42,** 401-412.

[5]   Efron, B., 1979. Bootstrap methods : Another look at Jackknife. *Ann. Statist.,* **1,** 1-26.

[6]   Efron, B., 1982. The Jackknife, the bootstrap and other resampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics, Monograph,* **38,** Philadelphia.

[7]   Efron, B., 1992. Jackknife after bootstrap; standard errors and influence functions. *J. Roy. Statist. Soc.,* **54,** 83-127.

[8]   Efron, B. and Tibshirani, R., 1986. Bootstrap measures for standard errors, confidence intervals and other measures of statistical accuracy. *Statist. Sci.,* **1,** 54-77.

[9]   Efron, B. and Tibshirani, R., 1993. *An Introduction to the Bootstrap.* Chapman and Hall, London.

[10]   Graybill, F.A., 1976. *Theory and Application of the Linear Model.* Duxbury.

[11]   Hall, P., 1991. *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York.

[12]   Harville, D.A. and Fenech, A.P., 1985. Confidence intervals for a variance ratio, or the heritability in an unbalanced mixed linear model. *Biometrics,* **41,** 137-152.

[13]   Ronnigen, K., 1973. Monte Carlo simulation of statistical biological models which are of interest in animal breeding. *Acta Agric. Scand.,* **24,** 135-142.

[14]   Searle, S.R., 1972. *Linear Models.* John Wiley and Sons, New York.

[15]   Sen, B., Graybill, F.A. and Ting, N., 1992. Confidence intervals on ratios of variance components in unbalanced two factor nested model. *Biom. J.,* **34,** 259-274.

[16]   Singh, B., 1989. A comparison of variance components estimators under unbalanced situation. *Sankhya,* **51,** 323-330.

[17]   Singh, B., 1992. On the estimation of heritability from unbalanced data. *Biom. J.,* **35,** 823-831.

[18]   Smith, C.A.B., 1956. On the estimation of intra class correlation. *Ann. Hum. Gen.,* **21,** 363-373.

[19]   Tai, G.C.C., 1983. Confidence intervals for genetic parameters estimated from progeny test experiments in plant breeding. *Canad. J. Cytology,* **25,** 502-508.