Theor Appl Genet (1985) 69:523–530

THEORETICAL AND APPLIED GENETICS
TAG

Effects of data imbalance on estimation of heritability*

R. F. Caro[1], M. Grossman[2,]** and R. L. Fernando[3]

[1] Department of Agronomy, [2] Department of Dairy Science, [3] Department of Animal Science;
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Received July 1, 1984
Communicated by L. D. Van Vleck

Summary. Effects of data imbalance on bias, sampling variance and mean square error of heritability estimated with variance components were examined using a random two-way nested classification. Four designs, ranging from zero imbalance (balanced data) to "low", "medium" and "high" imbalance, were considered for each of four combinations of heritability ($h^2 = 0.2$ and 0.4) and sample size ($N = 120$ and 600). Observations were simulated for each design by drawing independent pseudo-random deviates from normal distributions with zero means, and variances determined by heritability. There were 100 replicates of each simulation; the same design matrix was used in all replications. Variance components were estimated by analysis of variance (Henderson's Method 1) and by maximum likelihood (ML). For the design and model used in this study, bias in heritability based on Method 1 and ML estimates of variance components was negligible. Effect of imbalance on variance of heritability was smaller for ML than for Method 1 estimation, and was smaller for heritability based on estimates of sire-plus-dam variance components than for heritability based on estimates of sire or dam variance components. Mean square error for heritability based on estimates of sire-plus-dam variance components appears to be less sensitive to data imbalance than heritability based on estimates of sire or dam variance components, especially when using Method 1 estimation. Estimation of heritability from sire-plus-dam components was insensitive to differences in data imbalance, especially for the larger sample size.

Key words: Unbalanced data – Heritability – Variance components

Introduction

Heritability, the proportion of phenotypic variance associated with variance among additive genetic effects, is needed in devising efficient selection programs. Heritability can be estimated from experimental or field data as the ratio of estimated components of additive genetic variance and phenotypic variance using analysis of variance techniques, on the assumption of a mixed linear model. Numbers of observations in subclasses are commonly unequal (unbalanced data), to a greater extent in field data than in experimental data.

Traditional methods to estimate variance components with unbalanced data have been Methods 1, 2 and 3 (Henderson 1953). These methods are translation invariant quadratic unbiased estimators. More recently, other methods such as minimum variance quadratic unbiased estimation (MIVQUE), maximum likelihood (ML) and restricted maximum likelihood (REML) have become available (see Searle 1979).

One question of interest in genetics is to what extent data imbalance affects bias, sampling variance and mean square error (MSE) of estimates of variance components and heritability. Data imbalance will result from selection. Harville (1968) examined biases in variance components estimated by Methods 1 and 3 with unbalanced data for a two-way random-effects model with interaction. He modeled a selection process not independent of some of the unobservable random effects; thus subclass numbers were random and associated with these random effects. Expected values of estimators were insensitive to number of levels of the effects and expected value of subclass numbers. Under certain conditions, Method 3 estimators were less biased than Method 1.

* Supported by grants from the Illinois Agricultural Experiment Station and the University of Illinois Research Board. Charles Smith, H. W. Norton and D. Gianola contributed valuable suggestions

** To whom reprint requests should be addressed

Rothschild et al. (1979) investigated effects of data imbalance on variance and covariance components estimated by Method 1 and ML, and on heritability and correlation estimates, for a two-trait, one-way random model. Fifty percent of the data on trait two was selected randomly or by truncation on trait one. Subclass numbers for trait one were fixed, whereas for trait 2 they were random variables, independent of other random variables in the model (random selection) or associated with trait 1 (truncation selection). Method 1 and ML estimates had similar MSE when data were selected at random. With truncation selection, however, Method 1 estimates had higher MSE than ML.

Corbeil and Searle (1976) compared variance component estimators for balanced and unbalanced data with a two-way mixed model with no interaction. In comparisons for unbalanced data, subclass numbers were either zero or one observation per cell with 10, 30 or 60% of cells empty, and were not associated with variables in the model. ML had greater efficiency under the range of experimental conditions.

This study examined effects of data imbalance on variance components estimated by Method 1 and ML, and on heritability estimates, for a random two-way nested classification. Four designs with increasing levels of data imbalance due to random loss of observations from an optimally structured experiment were considered. Each design had fixed subclass numbers; thus they were not associated with random variables in the model.

The model and the designs

In a two-way nested classification, observations are assumed to follow the random linear model:

$$P_{ijk} = \mu + S_i + D_{ij} + E_{ijk}$$

where P_{ijk} is the observation for progeny k of dam j mated to sire i, μ is a fixed effect common to all observations, S_i is the effect of sire i, D_{ij} is the effect of dam j mated to sire i, and E_{ijk} is a residual associated with progeny k of dam j mated to sire i. There are $i = 1, \ldots, s$ sires; $j = 1, \ldots, d_i$ dams per sire i, and $k = 1, \ldots, n_{ij}$ progeny per dam i j. Also, $\sum d_i = d.$, the total number of dams; $\sum n_{ij} = n_{i.}$, the number of progeny of sire i; and $\sum\sum n_{ij} = N$, the total number of progeny. In the case of equal numbers of observations in the subclasses, $d_i = d$ and $n_{ij} = n$. Effects S_i, D_{ij} and E_{ijk} are assumed to be mutually uncorrelated random variables with zero means and variances σ^2_S, σ^2_D and σ^2_E, so that $\sigma^2_P = \sigma^2_S + \sigma^2_D + \sigma^2_E$.

Four designs, ranging in imbalance from zero (balanced data or equal subclass numbers) to "low", "medium", and "high" imbalance were used for each of four combinations of heritability ($h^2 = 0.2$ and 0.4) and sample size (N = 120 and 600). The balanced designs were chosen to have optimum structure so as to estimate heritability from both sire and dam components with approximate minimum variance. Below is the optimal number of sires (s), dams per sire (d = 2) and progeny per dam (n) for each balanced design, for combinations of h^2 and N (Grossman and Norton 1981).

Heritability (h^2)	Balanced structure	Total no. progeny (N)	
		120	600
0.2	s	10	30
	d	2	2
	n	6	10
0.4	s	15	75
	d	2	2
	n	4	4

The three unbalanced designs were chosen from among 20 designs generated at random. Numbers of sires and of dams per sire in the random designs were fixed and were the same as for the balanced design, for each combination of h^2 and N. The number of progeny per dam, however, were generated using a Poisson pseudo-random number generator (subroutine GGPOS, International Mathematics and Statistics Library).

Numbers of progeny per dam were assumed to follow a Poisson distribution, if each dam had the same probability of producing progeny (Cavalli-Sforza and Bodmer 1971). A random variable x has a Poisson distribution if its probability density function is of the form

$$f(x, \lambda) = \lambda^x e^{-\lambda}/x!, \quad \text{for} \quad x = 0, 1, 2, \ldots; \quad \text{and} \quad \lambda > 0;$$

where e is the base of natural logarithms. The mean and the variance of the Poisson distribution are each equal to λ. In this study, λ is taken to be the number of progeny per dam (n_{ij}, dam-family size) given by the balanced design for each combination of h^2 and N. The number of dam families (2 s), and the mean and variance of dam-family size (λ) are summarized below:

h^2	N	No. of dam families (2 s)	Mean and variance of dam-family size (λ)
0.2	120	20	6
	600	60	10
0.4	120	30	4
	600	150	4

To quantify the degree of imbalance, the coefficient of variation (CV) of dam-family size (the n_{ij}'s) was chosen as:

$$CV(n_{ij}) = 100\,\hat{\sigma}_n/\hat{\mu}_n$$

where $\hat{\mu}_n$ and $\hat{\sigma}^2_n$ are estimates of the mean and the variance of dam-family size. Designs that are more unbalanced have larger coefficients of variation for dam-family size. Within each combination of h^2 and N, larger $CV(n_{ij})$'s have larger estimated variance of dam-family size, the "estimated" mean dam-family size being the same and equal to λ.

Because of unequal numbers of progeny among dam families, sire family sizes will usually be unequal. Imbalance among sire families was measured by $CV(n_{i.}) = 100\,\hat{\sigma}/2\,\hat{\mu}_n$, the coefficient of variation of sire-family size, where σ^2 is the estimate of the variance of sire-family size.

For each combination of h^2 and N, the design with lowest $CV(n_{ij})$, from among the 20 designs randomly generated, represented "low" imbalance; the design with the highest $CV(n_{ij})$, "high" imbalance; and a design with intermediate $CV(n_{ij})$, "medium" imbalance.

Bias, sampling variance and mean square error of heritability estimated by a ratio of variance component estimates could not be calculated directly, with unbalanced data, and therefore were estimated by computer simulation. Observations were simulated for each design by drawing independent

Table 1. Analysis of variance for estimating heritabilities from unbalanced data

Source	df	MS	E (MS)
Sires (S)	$s-1$	S	$\sigma_E^2 + k_2\,\sigma_D^2 + k_3\,\sigma_S^2$
Dams (D)/S	$\sum_i (d_i - 1)$	D	$\sigma_E^2 + k_1\,\sigma_D^2$
Progeny/D/S	$\sum_i \sum_j (n_{ij} - 1)$	E	σ_E^2

$$k_1 = \frac{1}{\sum_i (d_i - 1)} \left[N - \sum_i \frac{(\sum_j n_{ij}^2)}{n_{i.}} \right]$$

$$k_2 = \frac{1}{s-1} \left[\sum_i \frac{(\sum_j n_{ij}^2)}{n_{i.}} - \frac{\sum_i \sum_j n_{ij}^2}{N} \right]$$

$$k_3 = \frac{1}{s-1} \left[N - \frac{\sum_i n_{i.}^2}{N} \right]$$

pseudo-random deviates from normal distributions with zero means, and with variances determined by heritability. There were 100 replications of each simulation; the same design matrix was used in all replications. Variance components were estimated by analysis of variance (Method 1; Henderson 1953) and by maximum likelihood (ML), using PROC NESTED and PROC VARCOMP in SAS®[1] (SAS Institute Inc. 1982), respectively.

Estimates of variance components by Method 1 were computed by equating each mean square to its expected value and solving the resulting system of equations for the variance components. From the analysis of variance in Table 1,

$$\hat{\sigma}_E^2 = E\,,$$

$$\hat{\sigma}_D^2 = (D - E)/k_1\,,$$

$$\hat{\sigma}_S^2 = [k_1\,S - k_2\,D + (k_2 - k_1)\,E]/k_1\,k_3\,,$$

$$\hat{\sigma}_P^2 = \hat{\sigma}_S^2 + \hat{\sigma}_D^2 + \hat{\sigma}_E^2\,.$$

For equal numbers of progeny per dam ($n_{ij} = n$) and dams per sire ($d_i = d$), the coefficients for variance components are $k_1 = k_2 = n$ and $k_3 = n\,d$.

Variances and covariances of the variance components were computed according to Searle (1971). These analytical variances and covariances are referred to in this study as parameter values to distinguish them from estimated variances and covariances from simulated observations.

Three estimates of heritability were obtained (Falconer 1981): from the sire component of variance, $\hat{h}_S^2 = 4\,\hat{\sigma}_S^2/\hat{\sigma}_P^2$; from the dam component, $\hat{h}_D^2 = 4\,\hat{\sigma}_D^2/\hat{\sigma}_P^2$; and from the sire-plus-dam components, $\hat{h}_{S+D}^2 = 2\,(\hat{\sigma}_S^2 + \hat{\sigma}_D^2)/\hat{\sigma}_P^2 = (\hat{h}_S^2 + \hat{h}_D^2)/2$. Assuming only additive genetic effects, $\sigma_S^2 = \sigma_D^2 = \sigma_A^2/4$, where σ_A^2 is the additive genetic variance. For traits showing non-additive genetic effects, this assumption is not valid.

Approximate large-sample expectation and bias of heritabilities for parametric values were computed as approximate expectation (E) and bias (B) of ratios of variance components:

$$E\left(\frac{X}{Y}\right) = \frac{E(X)}{E(Y)} \left[1 + \frac{V(Y)}{E^2(Y)} - \frac{C(X,Y)}{E(X)\,E(Y)} \right], \quad \text{for} \quad Y > 0$$

and

$$B\left(\frac{X}{Y}\right) = E\left(\frac{X}{Y}\right) - \frac{E(X)}{E(Y)} = \frac{E(X)}{E(Y)} \left[\frac{V(Y)}{E^2(Y)} - \frac{C(X,Y)}{E(X)\,E(Y)} \right]$$

where X and Y are random variables (e.g., $X = 4\,\hat{\sigma}_S^2$ and $Y = \hat{\sigma}_P^2$) with expectations E(X) and E(Y), variances V(X) and V(Y), and covariance C(X, Y). Large-sample variances and covariance of heritabilities for parametric values were computed as approximate variances and covariances of ratios of variance components:

$$V\left(\frac{X}{Y}\right) = \frac{E^2(X)}{E^2(Y)} \left[\frac{V(X)}{E^2(X)} + \frac{V(Y)}{E^2(Y)} - \frac{2\,C(X,Y)}{E(X)\,E(Y)} \right]$$

and

$$C\left(\frac{U}{V}, \frac{X}{Y}\right) = \frac{E(U)\,E(X)}{E(V)\,E(Y)} \left[\frac{C(U,X)}{E(U)\,E(X)} - \frac{C(V,X)}{E(V)\,E(X)} - \frac{C(U,Y)}{E(U)\,E(Y)} + \frac{C(V,Y)}{E(V)\,E(Y)} \right],$$

where U, V, X and Y are random variables (Pearson 1897).

Maximum likelihood estimates of variance components were computed for each combination, except for $h^2 = 0.4$ and $N = 600$ because of expense, using the method of Hemmerle and Hartley (1973) as described by SAS. Initial estimates of the components were computed using MIVQUE(0). The procedure iterated until the change in the log-likelihood objective function was less than 1×10^{-8}.

The mean, sampling variance and mean square error of 100 replicate variance components and heritabilities were computed for Method 1 and ML estimates. Mean square error (MSE) combines the effects of bias and sampling variance (see Kendall and Stuart 1979). For example, for the sire component of variance,

$$\text{MSE}\,(\hat{\sigma}_S^2) = E\,(\hat{\sigma}_S^2 - \sigma_S^2)^2 = V\,(\hat{\sigma}_S^2) + B^2\,(\hat{\sigma}_S^2)$$

which is estimated as

$$\widehat{\text{MSE}}\,(\hat{\sigma}_S^2) = \sum_{i=1}^{100} (\hat{\sigma}_{S_i}^2 - \sigma_S^2)^2/100\,,$$

where $\sigma_S^2 = 0.05$ when $h^2 = 0.2$ or $\sigma_S^2 = 0.10$ when $h^2 = 0.4$.

Results and discussion

Coefficients for variance components and CV's of dam families and of sire families appear in Table 2. CV's for sire- and dam-family sizes did not always increase together. Expectation and variance of variance components estimated by Method 1 (Table 3), showed that variances were larger for the smaller sample size and generally smaller for the higher heritability. Data imbalance did not affect accuracy of the estimates of variance components (as expected, Method 1 yields unbiased estimates); however, precision of the estimates generally decreased with increasing imbalance. Variances of sire and of dam components increased with increasing imbalance; this was more so for the smaller sample size and higher heritability.

Approximate bias of heritabilities (Table 4), based on Method 1 estimates of variance components, showed that bias was larger for smaller sample size and

[1] SAS® is the registered trademark of SAS Institute Inc., Cary, NC, USA

Table 2. Coefficients for variance components (k_1, k_2 and k_3) and coefficients of variation for dam-family sizes, CV(n_{ij}), and sire-family sizes, CV($n_{i.}$), ranked by CV(n_{ij}) within combination

h^2	N	s	n	k_1	k_2	k_3	CV($n_{i.}$)	CV(n_{ij})
0.2	120	10	6	6	6	12	0	0
				5.68	6.27	11.90	29.13	37.07
				5.13	6.84	11.96	18.84	43.26
				5.15	6.79	11.88	32.16	50.15
	600	30	10	10	10	20	0	0
				9.68	10.31	19.98	17.37	24.63
				9.60	10.39	19.97	20.38	28.34
				9.43	10.56	19.97	20.21	30.81
0.4	120	15	4	4	4	8	0	0
				3.70	4.28	7.96	26.31	37.72
				3.48	4.49	7.95	30.62	48.69
				3.32	4.66	7.96	28.74	50.86
	600	75	4	4	4	8	0	0
				3.54	4.46	7.99	33.70	47.86
				3.50	4.49	7.99	35.47	51.24
				3.48	4.51	7.98	39.93	53.88

Table 3. Expectation (E) and variance (V) for variance components, based on Method 1 estimates of variance components, ranked by CV(n_{ij}) within combination

h^2	N	CV(n_{ij})	E($\hat{\sigma}_S^2$)	V($\hat{\sigma}_S^2$)	E($\hat{\sigma}_D^2$)	V($\hat{\sigma}_D^2$)	E($\hat{\sigma}_E^2$)	V($\hat{\sigma}_E^2$)
			(× 100)					
0.2	120	0	5	0.70	5	0.84	90	1.62
		37.07	5	0.76	5	0.92	90	1.62
		43.26	5	0.86	5	1.08	90	1.62
		50.15	5	0.87	5	1.08	90	1.62
	600	0	5	0.13	5	0.14	90	0.30
		24.63	5	0.14	5	0.14	90	0.30
		28.34	5	0.14	5	0.14	90	0.30
		30.81	5	0.14	5	0.14	90	0.30
0.4	120	0	10	1.19	10	1.29	80	1.42
		37.72	10	1.33	10	1.45	80	1.42
		48.69	10	1.43	10	1.58	80	1.42
		50.86	10	1.52	10	1.70	80	1.42
	600	0	10	0.23	10	0.26	80	0.28
		47.86	10	0.27	10	0.31	80	0.28
		51.24	10	0.27	10	0.32	80	0.28
		53.88	10	0.28	10	0.32	80	0.28

for higher heritability. With increasing imbalance, bias increased for heritability estimated from the sire component and from sire-plus-dam components of variance, more so for the smaller sample size; bias did not change for heritability estimated from the dam component. Approximate variances and covariance of heritabilities (Table 5), based on Method 1 estimation, were larger for smaller sample size and for higher heritability, and increased with increasing imbalance (more so for the higher heritability).

Table 4. Approximate bias (B) for heritabilities, based on Method 1 estimates of variance components, ranked by CV(n_{ij}) within combination

h^2	N	CV(n_{ij})	B($\hat{h}_S^2$)	B($\hat{h}_D^2$)	B($\hat{h}_{S+D}^2$)
0.2	120	0	− 0.0084	− 0.0034	− 0.0058
		37.07	− 0.0090	− 0.0034	− 0.0062
		43.26	− 0.0090	− 0.0033	− 0.0062
		50.15	− 0.0094	− 0.0034	− 0.0064
	600	0	− 0.0019	− 0.0008	− 0.0014
		24.63	− 0.0020	− 0.0008	− 0.0014
		28.34	− 0.0020	− 0.0008	− 0.0014
		30.81	− 0.0020	− 0.0008	− 0.0014
0.4	120	0	− 0.0157	− 0.0054	− 0.0106
		37.72	− 0.0170	− 0.0054	− 0.0112
		48.69	− 0.0178	− 0.0054	− 0.0112
		50.86	− 0.0180	− 0.0053	− 0.0112
	600	0	− 0.0028	− 0.0011	− 0.0020
		47.86	− 0.0032	− 0.0011	− 0.0021
		51.24	− 0.0032	− 0.0011	− 0.0022
		53.88	− 0.0033	− 0.0011	− 0.0022

Table 5. Approximate variances and covariance for heritabilities, based on Method 1 estimates of variance components, ranked by CV(n_{ij}) within combination

h^2	N	CV(n_{ij})	V($\hat{h}_S^2$)	V($\hat{h}_D^2$)	V($\hat{h}_{S+D}^2$)	C($\hat{h}_S^2$, $\hat{h}_D^2$)
0.2	120	0	0.1079	0.1331	0.0267	− 0.0671
		37.07	0.1177	0.1456	0.0270	− 0.0769
		43.26	0.1337	0.1709	0.0274	− 0.0976
		50.15	0.1347	0.1704	0.0277	− 0.0971
	600	0	0.0202	0.0209	0.0047	− 0.0111
		24.63	0.0211	0.0219	0.0048	− 0.0120
		28.34	0.0214	0.0222	0.0048	− 0.0122
		30.81	0.0218	0.0228	0.0048	− 0.0127
0.4	120	0	0.1751	0.1987	0.0396	− 0.1076
		37.72	0.1957	0.2244	0.0407	− 0.1286
		48.69	0.2119	0.2456	0.0413	− 0.1461
		50.86	0.2253	0.2651	0.0414	− 0.1624
	600	0	0.0338	0.0397	0.0077	− 0.0214
		47.86	0.0404	0.0482	0.0080	− 0.0283
		51.24	0.0411	0.0490	0.0080	− 0.0290
		53.88	0.0416	0.0495	0.0081	− 0.0294

Among the 100 simulated replicates of each combination, the number of negative estimates of sire and dam variance components from balanced data decreased with increasing heritability and sample size (Table 6). None was significantly different from expectation. There was some trend towards increased number of negative estimates with increasing data imbalance, consistent with Gill and Jensen (1968). The number of replicates that failed to converge for maximum likelihood estimation increased with increasing imbalance; this was more so for the smaller sample size and for the lower heritability.

Table 6. Number of negative estimates for sire (σ^2_S) and dam (σ^2_D) variance components by Method 1, and number of replicates that failed to converge by maximum likelihood among 100 replicates

h^2	N	CV(n_{ij})	Method 1		Maximum likelihood
			$\hat{\sigma}^2_S$	$\hat{\sigma}^2_D$	
0.2	120	0	24	31	5
		37.07	22	26	9
		43.06	20	40	15
		50.15	39	39	16
	600	0	8	8	0
		24.63	2	7	0
		28.34	13	6	0
		30.81	7	7	0
0.4	120	0	16	18	2
		37.72	16	15	4
		48.69	17	26	2
		50.86	16	27	3
	600	0	1	1	–
		47.86	1	5	–
		51.24	4	1	–
		53.88	5	4	–

Table 7. Means of replicate variance components estimated by Method 1 and maximum likelihood, ranked by CV(n_{ij}) within combination

h^2	N	CV(n_{ij})	Method 1			Maximum likelihood		
			$\hat{\sigma}^2_S$	$\hat{\sigma}^2_D$	$\hat{\sigma}^2_E$	$\hat{\sigma}^2_S$	$\hat{\sigma}^2_D$	$\hat{\sigma}^2_E$
			(× 100)					
0.2	120	0	5.58	4.93	89.63	5.03	5.48	88.80
		37.07	5.19	6.96	89.76	5.60	7.13	88.10
		43.06	6.43	3.58	89.24	5.73	5.08	87.81
		50.15	3.82	5.84	90.55	5.11	6.75	87.91
	600	0	4.49	5.03	89.64	4.27	4.93	89.59
		24.63	4.99	5.46	89.21	4.61	5.47	89.20
		28.34	4.63	5.67	89.50	4.51	5.41	89.50
		30.81	4.76	5.20	89.23	4.41	5.27	89.20
0.4	120	0	10.19	9.66	78.00	9.35	9.40	77.60
		37.72	11.48	11.04	78.48	10.90	10.54	78.44
		48.69	10.23	8.82	80.22	8.83	10.62	78.98
		50.86	10.43	8.72	78.87	10.03	9.87	77.63
	600	0	10.37	10.03	80.06	–	–	–
		47.86	8.80	11.13	79.80	–	–	–
		51.24	10.35	9.13	80.71	–	–	–
		53.88	11.21	8.76	79.75	–	–	–

Means of replicate variance components (Table 7), estimated by Method 1 and ML (when converged), indicated average estimates of variance components generally were close to parameter values (Table 3). There did not appear to be a trend in means of variance components associated with data imbalance when they were estimated by ML; Method 1 is known to be unbiased (Henderson 1953). Variances of replicate variance components estimated by Method 1 (Table 8) followed patterns similar to their parameter values (Table 3). Variances tended to increase with increasing imbalance; this was less so when components of variance were estimated by ML or for the larger sample size. Estimated mean square errors (MSE) for variance components (Table 9) were smaller for ML than for Method 1 estimation and smaller for

Table 8. Variances of replicate variance components estimated by Method 1 and maximum likelihood, ranked by CV(n_{ij}) within combination

h^2	N	CV(n_{ij})	Method 1			Maximum likelihood		
			$\hat{V}(\hat{\sigma}^2_S)$	$\hat{V}(\hat{\sigma}^2_D)$	$\hat{V}(\hat{\sigma}^2_E)$	$\hat{V}(\hat{\sigma}^2_S)$	$\hat{V}(\hat{\sigma}^2_D)$	$\hat{V}(\hat{\sigma}^2_E)$
			(×100)					
0.2	120	0	0.59	0.96	1.96	0.33	0.52	1.80
		37.07	0.74	1.17	1.36	0.48	0.75	1.15
		43.06	0.82	0.95	1.32	0.31	0.28	1.24
		50.15	1.01	1.27	1.85	0.39	0.72	1.58
	600	0	0.13	0.18	0.27	0.09	0.14	0.27
		24.63	0.10	0.10	0.22	0.10	0.10	0.22
		28.34	0.17	0.18	0.34	0.13	0.15	0.34
		30.81	0.13	0.17	0.25	0.11	0.14	0.24
0.4	120	0	1.24	1.04	1.75	0.82	0.70	1.71
		37.72	1.68	1.40	1.18	0.93	0.82	1.07
		48.69	1.06	1.48	0.92	0.65	0.85	0.90
		50.86	1.45	1.98	1.03	0.89	1.42	0.92
	600	0	0.21	0.22	0.28	–	–	–
		47.86	0.22	0.36	0.22	–	–	–
		51.24	0.30	0.32	0.22	–	–	–
		53.88	0.30	0.26	0.25	–	–	–

Table 9. Estimated mean square error ($\widehat{\text{MSE}}$) for variance components estimated by Method 1 and maximum likelihood, ranked by CV(n_{ij}) within combination

h^2	N	CV (n_{ij})	Method 1			Maximum likelihood		
			$\widehat{\text{MSE}}$ ($\hat\sigma^2_S$)	$\widehat{\text{MSE}}$ ($\hat\sigma^2_D$)	$\widehat{\text{MSE}}$ ($\hat\sigma^2_E$)	$\widehat{\text{MSE}}$ ($\hat\sigma^2_S$)	$\widehat{\text{MSE}}$ ($\hat\sigma^2_D$)	$\widehat{\text{MSE}}$ ($\hat\sigma^2_E$)
			(× 100)					
0.2	120	0	0.59	0.95	1.94	0.33	0.52	1.80
		37.07	0.74	1.20	1.35	0.48	0.78	1.18
		43.06	0.83	0.96	1.31	0.32	0.28	1.28
		50.15	1.01	1.26	1.84	0.39	0.74	1.61
	600	0	0.13	0.18	0.27	0.10	0.14	0.27
		24.63	0.10	0.10	0.23	0.10	0.10	0.23
		28.34	0.17	0.19	0.34	0.13	0.15	0.34
		30.81	0.12	0.17	0.25	0.11	0.13	0.24
0.4	120	0	1.23	1.03	1.78	0.81	0.70	1.75
		37.72	1.68	1.40	1.19	0.93	0.81	1.08
		48.69	1.05	1.47	0.91	0.66	0.84	0.90
		50.86	1.44	1.98	1.03	0.88	1.40	0.96
	600	0	0.21	0.22	0.28	–	–	–
		47.86	0.28	0.37	0.22	–	–	–
		51.24	0.30	0.32	0.22	–	–	–
		53.88	0.31	0.28	0.25	–	–	–

Table 10. Means of replicate heritabilities, estimated by Method 1 and maximum likelihood variance components, ranked by CV(n_{ij}) within combination

h^2	N	CV (n_{ij})	Method 1			Maximum likelihood		
			$\hat h^2_S$	$\hat h^2_D$	$\hat h^2_{S+D}$	$\hat h^2_S$	$\hat h^2_D$	$\hat h^2_{S+D}$
0.2	120	0	0.2147	0.1974	0.2060	0.1974	0.2203	0.2089
		37.07	0.1968	0.2667	0.2318	0.2160	0.2721	0.2440
		43.06	0.2477	0.1437	0.1957	0.2272	0.2039	0.2155
		50.15	0.1419	0.2204	0.1812	0.1992	0.2535	0.2263
	600	0	0.1802	0.2005	0.1904	0.1720	0.1970	0.1845
		24.63	0.1990	0.2189	0.2089	0.1841	0.2202	0.2022
		28.34	0.1842	0.2264	0.2053	0.1803	0.2168	0.1985
		30.81	0.1898	0.2079	0.1989	0.1766	0.2113	0.1939
0.4	120	0	0.3961	0.3994	0.3978	0.3740	0.3910	0.3825
		37.72	0.4292	0.4354	0.4323	0.4247	0.4111	0.4179
		48.69	0.4007	0.3395	0.3701	0.3467	0.4206	0.3836
		50.86	0.4151	0.3447	0.3799	0.3988	0.3832	0.3910
	600	0	0.4108	0.3959	0.4034	–	–	–
		47.86	0.3508	0.4439	0.3974	–	–	–
		51.24	0.4104	0.3621	0.3862	–	–	–
		53.88	0.4436	0.3527	0.3981	–	–	–

the larger sample size, but they were larger for the higher heritability. $\widehat{\text{MSE}}$ increased generally with increasing imbalance for components of variance for sire and for dam, especially for the smaller sample size with Method 1.

For Method 1 and for ML estimation, the effect of data imbalance on $\widehat{\text{MSE}}$ was similar to the effect of imbalance on estimated variances. For Method 1 estimation, mean square error is expected to be equal to the variance because variance components estimated by Method 1 are unbiased. For ML estimation, mean square error is expected to be larger than the variance because variance components estimated by ML are biased; however, $\widehat{\text{MSE}}$ of variance component estimates (Table 9) were close to (sometimes even smaller than) their estimated variances (Table 8), more so for sire and for dam components of variance. This relation between $\widehat{\text{MSE}}$ and estimated variance indicates bias in

Table 11. Variances and covariance of replicate heritabilities estimated by Method 1 and maximum likelihood variance components, ranked by CV(n_{ij}) within combination

h^2	N	CV(n_{ij})	Method 1				Maximum likelihood			
			$\hat{V}(\hat{h}^2_S)$	$\hat{V}(\hat{h}^2_D)$	$\hat{V}(\hat{h}^2_{S+D})$	$\hat{C}(\hat{h}^2_S, \hat{h}^2_D)$	$\hat{V}(\hat{h}^2_S)$	$\hat{V}(\hat{h}^2_D)$	$\hat{V}(\hat{h}^2_{S+D})$	$\hat{C}(\hat{h}^2_S, \hat{h}^2_D)$
0.2	120	0	0.0879	0.1433	0.0291	− 0.0574	0.0468	0.0745	0.0207	− 0.0192
		37.07	0.1061	0.1619	0.0280	− 0.0779	0.0608	0.0936	0.0267	− 0.0237
		43.06	0.1289	0.1603	0.0271	− 0.0905	0.0473	0.0436	0.0193	− 0.0068
		50.15	0.1422	0.1847	0.0354	− 0.0927	0.0544	0.0914	0.0315	− 0.0099
	600	0	0.0200	0.0274	0.0050	− 0.0138	0.0146	0.0212	0.0048	− 0.0083
		24.63	0.0146	0.0164	0.0037	− 0.0081	0.0149	0.0159	0.0036	− 0.0082
		28.34	0.0261	0.0294	0.0056	− 0.0165	0.0202	0.0225	0.0054	− 0.0105
		30.81	0.0193	0.0268	0.0053	− 0.0124	0.0168	0.0214	0.0049	− 0.0093
0.4	120	0	0.1741	0.1743	0.0365	− 0.1012	0.1112	0.1155	0.0331	− 0.0472
		37.72	0.2382	0.2226	0.0449	− 0.1406	0.1258	0.1125	0.0420	− 0.0353
		48.69	0.1641	0.2295	0.0373	− 0.1223	0.0957	0.1199	0.0293	− 0.0491
		50.86	0.2145	0.3025	0.0415	− 0.1755	0.1145	0.1621	0.0469	− 0.0446
	600	0	0.0296	0.0318	0.0085	− 0.0137	–	–	–	–
		47.86	0.0408	0.0558	0.0080	− 0.0324	–	–	–	–
		51.24	0.0447	0.0475	0.0081	− 0.0300	–	–	–	–
		53.88	0.0413	0.0434	0.0083	− 0.0258	–	–	–	–

Table 12. Estimated mean square error ($\widehat{MSE}$) for heritabilities, estimated by Method 1 and maximum likelihood variance components, ranked by CV(n_{ij}) within combination

h^2	N	CV(n_{ij})	Method 1			Maximum likelihood		
			$\widehat{MSE}(\hat{h}^2_S)$	$\widehat{MSE}(\hat{h}^2_D)$	$\widehat{MSE}(\hat{h}^2_{S+D})$	$\widehat{MSE}(\hat{h}^2_S)$	$\widehat{MSE}(\hat{h}^2_D)$	$\widehat{MSE}(\hat{h}^2_{S+D})$
0.2	120	0	0.0872	0.1418	0.0288	0.0463	0.0741	0.0206
		37.07	0.1504	0.1647	0.0288	0.0603	0.0978	0.0284
		43.06	0.1299	0.1619	0.0268	0.0475	0.0432	0.0194
		50.15	0.1441	0.1833	0.0354	0.0537	0.0931	0.0318
	600	0	0.0202	0.0271	0.0050	0.0152	0.0210	0.0050
		24.63	0.0145	0.0166	0.0038	0.0150	0.0162	0.0036
		28.34	0.0260	0.0288	0.0056	0.0204	0.0225	0.0054
		30.81	0.0192	0.0266	0.0052	0.0172	0.0213	0.0049
0.4	120	0	0.1724	0.1726	0.0362	0.1108	0.1144	0.0330
		37.72	0.2367	0.2216	0.0455	0.1251	0.1115	0.0418
		48.69	0.1625	0.2309	0.0378	0.0976	0.1191	0.0293
		50.86	0.2126	0.3026	0.0415	0.1133	0.1607	0.0465
	600	0	0.0294	0.0315	0.0084	–	–	–
		47.86	0.0428	0.0572	0.0079	–	–	–
		51.24	0.0443	0.0485	0.0082	–	–	–
		53.88	0.0428	0.0452	0.0082	–	–	–

ML estimates of variance component is negligible, perhaps because there is only one degree of freedom for fixed effects.

Means of replicate heritabilities (Table 10), estimated by Method 1 and ML variance components, indicated average estimates of heritabilities generally were close to parameter values, especially for h^2_{S+D}. Data imbalance did not appear to affect bias. Variances and covariances of replicate heritabilities (Table 11), estimated by Method 1 variance components, generally were close to parameter values (Table 5). Variances were smaller for ML estimation, and larger for smaller sample size and for higher heritability. Effect of imbalance on variance of heritability was smaller for ML than for Method 1 estimation, and was smaller for $\hat{h}^2_{S+D}$ than for $\hat{h}^2_S$ or $\hat{h}^2_D$.

$\widehat{MSE}$ for heritabilities (Table 12) generally were smaller for ML than for Method 1 estimation and smaller for $\hat{h}^2_{S+D}$ than for $\hat{h}^2_S$ or $\hat{h}^2_D$. $\widehat{MSE}$ were larger generally for smaller sample size and for higher heritability. $\widehat{\mathrm{MSE}}$ generally increased with increasing imbalance, for Method 1 more so than for ML estimation and for $\hat{h}_S^2$ and $\hat{h}_D^2$ more so than for $\hat{h}_{S+D}^2$. As with variance components, $\widehat{\mathrm{MSE}}$ and variances of heritabilities were close indicating that for Method 1 and ML estimation bias is negligible.

For the design and model used in this study, bias in heritability based on Method 1 and ML estimates of variance components is negligible. Mean square error for heritability based on estimates of sire-plus-dam variance components appears to be less sensitive to data imbalance than heritability based on estimates of sire or dam variance components, especially when using Method 1 estimation. Estimation of heritability from sire-plus-dam components is insensitive to differences in data imbalance, especially for the larger sample size.

References

Cavalli-Sforza LL, Bodmer WF (1971) The genetics of human populations. Freeman, San Francisco

Corbeil RR, Searle SR (1976) A comparison of variance component estimators. Biometrics 32:779–791

Falconer DS (1981) Introduction to quantitative genetics. Longman, New York

Gill JL, Jensen EL (1968) Probability of obtaining negative estimates of heritability. Biometrics 24:517–526

Grossman M, Norton HW (1981) An approximation of the minimum-variance estimator of heritability based on variance component analysis. Genetics 98:417–426

Harville DA (1968) Statistical dependence between subclass means and the numbers of observations in the subclasses for the two-way completely-random classification. J Am Stat Assoc 63:1484–1494

Hemmerle WJ, Hartley HO (1973) Computing maximum likelihood estimators for the mixed A.O.V. model using the W-transformation. Technometrics 15:819–832

Henderson CR (1953) Estimation of variance and covariance components. Biometrics 9:226–252

Kendall M, Stuart A (1979) The advanced theory of statistics, vol 2. Macmillan, New York, p 21

Pearson K (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc London, Ser B 60:489–498

Rothschild MF, Henderson CR, Quaas RL (1979) Effects of selection on variances and covariances of simulated first and second lactations. J Dairy Sci 62:996–1002

SAS Institute Inc (1982) SAS user's guide: statistics. SAS Institute Inc, Cary, North Carolina

Searle SR (1971) Linear models. Wiley and Sons, New York, p 475

Searle SR (1979) Notes on variance component estimation: a detailed account of maximum likelihood and kindred methodology. Mimeo BU-673-M Biometrics Unit, Cornell University, Ithaca, New York