ESTIMATION OF VARIANCE AND COVARIANCE COMPONENTS*

C. R. HENDERSON

Cornell University

INTRODUCTION

The theory of variance component analysis has been discussed recently by Crump (1946, 1951) and by Eisenhart (1947). These papers and, indeed, most of the published works on estimating variance components deal with the one-way classification, with "nested" classifications, and with factorial classifications having equal subclass numbers. Also most papers on this subject are concerned with what Eisenhart (1947) has called Model II; that is, all elements of the linear model save μ are regarded as random variables. In the above cases, estimation of variance components is usually accomplished by computing the mean squares in the standard analysis of variance, equating these mean squares to their expectations, and solving for the unknown variances. These techniques are described in many statistical textbooks.

Unfortunately, research workers in some of those fields in which much use is made of variance component estimates are unable to obtain data which have the above described characteristics. This is particularly true in those fields in which survey data must be used or where, even in a well-planned experiment, the subclasses are of quite unequal size due, for example, to differences in litter numbers. Also,

*Presented at North Carolina Summer Statistics Conference June 24, 1952.

Model II is sometimes not appropriate. Instead the data more appropriately correspond to what Eisenhart called the Mixed Model. For example, the data may represent several different years, and the year effects should be regarded as fixed rather than as random variables.

It is the purpose of this paper to describe some methods for estimating variance components in the non-orthogonal case and to illustrate the methods with a small sample of butterfat records made by cows resulting from an artificial breeding program. The three methods described are:

1. Compute sums of squares as in the standard analysis of variance of corresponding orthogonal data. Equate these sums of squares to their expectations obtained under the assumption of Model II and solve for the unknown variances.
2. Obtain least squares estimates of fixed effects, "correct" the data according to these estimates of the fixed effects, and then using the corrected data in place of the original data, proceed as in Method 1.
3. Compute mean squares by a conventional least squares analysis of non-orthogonal data (method of fitting constants, weighted squares of means, e.g.). Equate these mean squares to their expectations and solve for the unknown variances.

These three methods henceforth called Method 1, Method 2, and Method 3 vary greatly in computational labor. Method 1 is the simplest. Method 2 in many cases is only slightly more difficult. Method 3 is usually much the most laborious. Method 1, however, leads to biased estimates if certain elements of the model are fixed or if some of them are correlated. Estimates obtained by Method 2 are free of the first of these biases, but not of the second. Method 3 yields unbiased estimates, but the computations required may be prohibitive. The relative sizes of the sampling variances of estimates obtained by these three methods are not known.

DESCRIPTION AND ILLUSTRATION OF METHODS OF ESTIMATION

The Data

In New York State most artificial breeding of dairy cows is accomplished with semen supplied by the New York Artificial Breeders' Cooperative, Inc. This cooperative organization has approximately 60 bulls in service. The operations of the organization are conducted in such a manner that it is largely a matter of chance to which bull's semen a particular cow is bred. This fact as well as the large number of daughters sired by each bull make the production records of these daughters particularly suitable for studying the genetic differences among bulls and for estimating the magnitudes of other sources of variation in milk production records. Good estimates of these variances are needed in designing efficient testing and selection programs.

The difficulties in estimating the pertinent variance components are typical of those faced by research workers in animal breeding and in other fields as well. The difficulties in the present example are due to the following causes:

1. Several years' data are involved and time trends are known to be important.
2. The two major classifications of the data are sire and herd. The number of sires exceeds 100 and the number of different herds exceeds 2000.
3. The number of observations per herd-sire subclass varies; the majority being 0.

We have estimated from these data the pertinent variances. Both Method 1 and Method 2 have been employed and have yielded estimates essentially the same. A small sample of records is presented in this paper and the three methods of estimation are illustrated.

Table 1 shows the number of first lactation butterfat records in each of the year $\times$ herd $\times$ sire subclasses and also the sum of the records for each of these subclasses.

TABLE 1

Herd	Sire	Year				Total
		1	2	3	4	
1	1	3–1414	2– 981			5–2395
1	2		4–1766	2– 862		6–2628
1	3				5–1609	5–1609
2	1	1– 404	3–1270			4–1674
2	2			5–2109		5–2109
2	3			4–1563	2– 740	6–2303
3	1		3–1705			3–1705
3	2		4–2310	2–1134		6–3444
4	1	3–1113	5–1951			8–3064
4	3			3–1291	6–2457	9–3748
Total		7–2931	21–9983	16–6959	13–4806	57–24679

The Linear Model

Let y_{hijk} denote the record made in the h-th year by the k-th daughter of the j-th sire in the i-th herd. Suppose that the appropriate linear model representing these observations is

$$y_{hijk} = \mu + a_h + h_i + s_j + (hs)_{ij} + e_{hijk}$$

$$h = 1, \cdots, p \qquad k = 1, \cdots, n_{hij}$$

$$i = 1, \cdots, q \qquad \sum_h \sum_i \sum_j n_{hij} = N$$

$$j = 1, \cdots, r \qquad \text{Total number of filled subclasses} = s$$

μ is common to all observations. a_h is common to all observations in the h-th year, h_i to all observations in the i-th herd, and s_j to all records made by daughters of the j-th sire; $(hs)_{ij}$ is peculiar to all records made by the daughters of the j-th sire in the i-th herd. Peculiar to each record is a random element e_{hijk} which is assumed to have mean zero and variance σ_e^2. The assumptions made concerning the other elements of the model are described for each estimation method.

Method 1*

Method 1 can be used only if it is assumed that, except for μ, all elements of the model are uncorrelated variables with means zero and variances σ_a^2, σ_h^2, σ_s^2, σ_{hs}^2, or σ_e^2. This is, of course, the Eisenhart Model II.

The following quantities are computed:

$$T = \sum_h \sum_i \sum_j \sum_k y_{hijk}^2 \qquad H = \sum_i \frac{y_{\cdot i \cdot\cdot}^2}{n_{\cdot i \cdot}}$$

$$A = \sum_h \frac{y_{h\cdots}^2}{n_{h\cdot\cdot}} \qquad S = \sum_j \frac{y_{\cdot\cdot j \cdot}^2}{n_{\cdot\cdot j}}$$

$$HS = \sum_i \sum_j \frac{y_{\cdot ij \cdot}^2}{n_{\cdot ij}} \qquad CF = \frac{y_{\cdots\cdot}^2}{N}$$

Dots in the subscripts denote summation. For example,

$$y_{h\cdots} = \sum_i \sum_j \sum_k y_{hijk} \, .$$

Next the expectations of the above quantities are computed. Under the assumptions of Model II, the coefficients of μ^2 and the variances in these expectations are as shown in Table 2.

---

*This method was first suggested to me by Dr. S. Lee Crump.

TABLE 2

	μ^2	σ_a^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2
T	N	N	N	N	N	N
A	N	N	K_1	K_2	K_3	p
HS	N	K_4	N	N	N	s
H	N	K_5	N	K_6	K_6	q
S	N	K_7	K_8	N	K_8	r
CF	N	K_9	K_{10}	K_{11}	K_{12}	1

N, p, q, r, s in the above table were defined in the statement of the linear model. K_1, K_2, $\cdots$, K_{12} must be computed as follows:

$$K_1 = \sum_h \frac{\sum_i n_{hi\cdot}^2}{n_{h\cdot\cdot}} \qquad K_7 = \sum_j \frac{\sum_h n_{h\cdot j}^2}{n_{\cdot\cdot j}}$$

$$K_2 = \sum_h \frac{\sum_j n_{h\cdot j}^2}{n_{h\cdot\cdot}} \qquad K_8 = \sum_j \frac{\sum_i n_{\cdot ij}^2}{n_{\cdot\cdot j}}$$

$$K_3 = \sum_h \frac{\sum_i \sum_j n_{hij}^2}{n_{h\cdot\cdot}} \qquad K_9 = \sum_h n_{h\cdot\cdot}^2/N$$

$$K_4 = \sum_i \sum_j \frac{\sum_h n_{hij}^2}{n_{\cdot ij}} \qquad K_{10} = \sum_i n_{\cdot i\cdot}^2/N$$

$$K_5 = \sum_i \frac{\sum_h n_{hi\cdot}^2}{n_{\cdot i\cdot}} \qquad K_{11} = \sum_j n_{\cdot\cdot j}^2/N$$

$$K_6 = \sum_i \frac{\sum_j n_{\cdot ij}^2}{n_{\cdot i\cdot}} \qquad K_{12} = \sum_i \sum_j n_{\cdot ij}^2/N$$

If the data were orthogonal, the sums of squares in the analysis of variance would be

Among Years $= A - CF$
Among Herds $= H - CF$
Among Sires $= S - CF$
Herds $\times$ Sires $= HS - H - S + CF$
Error $= T - A - HS + CF$

If these same quantities are computed in spite of the non-orthogonality and are equated to their expectations, unbiased estimates of the variances can be obtained by solving the resulting equations. The necessary expectations are derived from Table 2. To illustrate, E (Among Years) $= E(A - CF) = E(A) - E(CF)$.

Computation of the K's is facilitated by constructing from Table 1 the following two-way tables of subclass numbers (Tables 3, 4, 5).

TABLE 3

Herd	Year				Total
	1	2	3	4	
1	3	6	2	5	16
2	1	3	9	2	15
3	0	7	2	0	9
4	3	5	3	6	17
Total	7	21	16	13	57

TABLE 4

Sire	Year				Total
	1	2	3	4	
1	7	13	0	0	20
2	0	8	9	0	17
3	0	0	7	13	20
Total	7	21	16	13	57

TABLE 5

Herd	Sire			Total
	1	2	3	
1	5	6	5	16
2	4	5	6	15
3	3	6	0	9
4	8	0	9	17
Total	20	17	20	57

Also certain totals are computed from Table 1.

Year	Herd	Sire
1. 2931	1. 6632	1. 8838
2. 9983	2. 6086	2. 8181
3. 6959	3. 5149	3. 7660
4. 4806	4. 6812	
Total 24,679	Total 24,679	Total 24,679

Using the above totals and the totals in Table 1,

$$A = \frac{2931^2}{7} + \cdots + \frac{4806^2}{13} = 10{,}776{,}451$$

$$HS = \frac{2395^2}{5} + \cdots + \frac{3748^2}{9} = 10{,}970{,}369$$

$$H = \frac{6632^2}{16} + \cdots + \frac{6812^2}{17} = 10{,}893{,}666$$

$$S = \frac{8838^2}{20} + \frac{8181^2}{17} + \frac{7660^2}{20} = 10{,}776{,}278$$

$$CF = \frac{24{,}679^2}{57} = 10{,}685{,}141$$

The expectations of these quantities are presented in Table 6. The computations of these entries proceed as follows:

From Table 3, $$K_1 = \frac{3^2 + 1^2 + 3^2}{7} + \cdots + \frac{5^2 + 2^2 + 6^2}{13} = 19.51$$

From Table 4, $$K_2 = \frac{7^2}{7} + \frac{13^2 + 8^2}{21} + \frac{9^2 + 7^2}{16} + \frac{13^2}{13} = 39.22$$

From Table 1, $$K_3 = \frac{3^2 + 1^2 + 3^2}{7} + \cdots + \frac{5^2 + 2^2 + 6^2}{13} = 15.10$$

From Table 1, $$K_4 = \frac{3^2 + 2^2}{5} + \cdots + \frac{3^2 + 6^2}{9} = 37.35$$

From Table 3, $$K_5 = \frac{3^2 + 6^2 + 2^2 + 5^2}{16} + \cdots + \frac{3^2 + 5^2 + 3^2 + 6^2}{17} = 21.49$$

From Table 5, $$K_6 = \frac{5^2 + 6^2 + 5^2}{16} + \cdots + \frac{8^2 + 9^2}{17} = 24.04$$

From Table 4, $K_7 = \dfrac{7^2 + 13^2}{20} + \dfrac{8^2 + 9^2}{17} + \dfrac{7^2 + 13^2}{20} = 30.33$

From Table 5, $K_8 = \dfrac{5^2 + 4^2 + 3^2 + 8^2}{20} + \cdots + \dfrac{5^2 + 6^2 + 9^2}{20} = 18.51$

From Table 3, $K_9 = \dfrac{7^2 + 21^2 + 16^2 + 13^2}{57} = 16.05$

From Table 3, $K_{10} = \dfrac{16^2 + 15^2 + 9^2 + 17^2}{57} = 14.93$

From Table 4, $K_{11} = \dfrac{20^2 + 17^2 + 20^2}{57} = 19.11$

From Table 1, $K_{12} = \dfrac{5^2 + 6^2 + \cdots + 9^2}{57} = 6.19$

TABLE 6

	μ^2	σ_a^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
T	57	57.	57.	57.	57.	57	11,124,007
A	57	57.	19.51	39.22	15.10	4	10,776,451
HS	57	37.35	57.	57.	57.	10	10,970,369
H	57	21.49	57.	24.04	24.04	4	10,893,666
S	57	30.33	18.51	57.	18.51	3	10,776,278
CF	57	16.05	14.93	19.11	6.19	1	10,685,141

The equations to be solved are presented in Table 7. The first equation reads: $40.95\ \sigma_a^2 + 4.58\ \sigma_h^2 + 20.11\ \sigma_s^2 + 8.91\ \sigma_{hs}^2 + 3\ \sigma_e^2 =$ 91,310.

TABLE 7

	σ_a^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
$A - CF$	40.95	4.58	20.11	8.91	3	91,310
$H - CF$	5.44	42.07	4.93	17.85	3	208,525
$S - CF$	14.28	3.58	37.89	12.32	2	91,137
$HS - H - S + CF$	1.58	−3.58	−4.93	20.64	4	−14,434
$T - A - HS + CF$	−21.30	−4.58	−20.11	−8.91	44	62,328

The solution to these equations is $\sigma_a^2 = 763$, $\sigma_h^2 = 4531$, $\sigma_s^2 = 1587$, $\sigma_{hs}^2 = -164$, $\sigma_e^2 = 2950$. If σ_{hs}^2 is set equal to 0, the solution is $\sigma_a^2 = 756$, $\sigma_h^2 = 4468$, $\sigma_s^2 = 1542$, $\sigma_e^2 = 2952$. These estimates, of course, have no practical value for p, q, r, and s are much too small for accurate estimation of the corresponding variances. The illustration of their computation does, however, show that even with many different classes the computations are relatively simple. We have successfully adapted most of these computations to International Business Machines operations.

A difficulty with Method 1 is that it may be inappropriate to regard the year effects as random variables. If these effects actually are fixed, the estimates of σ_h^2, σ_s^2, and σ_{hs}^2 are biased. The estimate of σ_e^2 may or may not be biased depending on how it is estimated. This estimate is biased if obtained from the equations of Table 7. If, however, σ_e^2 had been estimated from

$$T - \sum_h \sum_i \sum_j \frac{y_{hij\cdot}^2}{n_{hij}},$$

the within year $\times$ herd $\times$ sire subclass sum of squares, the estimate would be unbiased regardless of the assumptions concerning the a_h.

It might be well at this point to state briefly a convenient procedure for finding the expected values of quantities like H, S, etc. Substitute for the y's their corresponding linear models, and then remembering the assumptions concerning the elements of the model proceed to write out the expectations. For example,

$$E \sum_i \sum_j \frac{y_{\cdot ij\cdot}^2}{n_{\cdot ij}} = \sum_i \sum_j E \frac{y_{\cdot ij\cdot}^2}{n_{\cdot ij}}$$

$$= \sum_i \sum_j E[n_{\cdot ij}\mu + n_{1ij}a_1$$

$$+ \cdots + n_{pij}a_p + n_{\cdot ij}h_i + n_{\cdot ij}s_j + n_{\cdot ij}(hs)_{ij}$$

$$+ \sum_h \sum_k e_{hijk}]^2/n_{\cdot ij}$$

$$= \sum_i \sum_j E[n_{\cdot ij}^2\mu^2 + n_{1ij}^2a_1^2$$

$$+ \cdots + n_{pij}^2a_p^2 + n_{\cdot ij}^2h_i^2 + n_{\cdot ij}^2s_j^2 + n_{\cdot ij}^2(hs)_{ij}^2$$

$$+ \sum_h \sum_k e_{hijk}^2 + \text{cross products all having zero expectation}]/n_{\cdot ij}$$

$$= \sum_i \sum_j [n^2_{\cdot ij}\mu^2 + n^2_{1ij}\sigma^2_a + \cdots + n^2_{pij}\sigma^2_a + n^2_{\cdot ij}\sigma^2_h + n^2_{\cdot ij}\sigma^2_s + n^2_{\cdot ij}\sigma^2_{hs}$$

$$+ \sum_h \sum_k \sigma^2_e]/n_{\cdot ij}$$

$$= N\mu^2 + \sum_i \sum_j \frac{\sum_h n^2_{hij}}{n_{\cdot ij}} \sigma^2_a + N(\sigma^2_h + \sigma^2_s + \sigma^2_{hs}) + s\sigma^2_e$$

Method 2

The bias in estimating variance components due to the assumption that fixed elements of the model are random variables can be eliminated by using Method 2. At the same time the relative simplicity of Method 1 can be retained. Method 2 involves estimating the fixed effects by least squares, correcting the data in accordance with these estimates, and then applying Method 1 to the "corrected" data.

This method was used by Hazel and Terrill (1945) on data which were orthogonal except for the fixed effects. Their estimates were biased for they assumed for computational purposes that, except for fixed effects, the expectations of sums of squares of corrected data are the same as the expectations of the corresponding sums of squares of the uncorrected data. Method 2 enables one to appraise this bias and to correct for it.

Before we apply this method to our example, let us consider the general case. Suppose the linear model is

$$(1) \qquad y_a = \sum_{i=1}^{p} b_i x_{ia} + e_a \qquad a = 1, \cdots, N$$

The x's are known. The e's are uncorrelated with mean $= 0$ and variance $= \sigma^2_e$.

If the b's are all fixed, the least squares equations for estimating them are as shown in (2). The b's are, in fact, not all fixed in the variance components estimation problem, but they can be estimated by least squares as a matter of expediency.

$$(2) \qquad \begin{aligned} \sum_{i=1}^{p} C_{1i}\hat{b}_i &= Y_1 \qquad & C_{ij} &= \sum_{a=1}^{N} x_{ia}x_{ja} \\ \sum_{i=1}^{p} C_{2i}\hat{b}_i &= Y_2 \qquad & Y_i &= \sum_{a=1}^{N} x_{ia}y_a \\ &\vdots \qquad \vdots & & \\ \sum_{i=1}^{p} C_{pi}\hat{b}_i &= Y_p & & \end{aligned}$$

It is sometimes necessary to impose one or more linear restrictions on the estimates in order to obtain a solution to equations (2).

Now suppose that $b_1, \cdots, b_s$ are fixed and also that for all $i = 1, \cdots, s$

$$E(\hat{b}_i - b_i)^2 = K_i\sigma_e^2 \tag{3}$$

It is not true that all least squares estimates have this property. For example, in our butterfat production example described in Method 1

$$E(\hat{\mu} - \mu)^2 \neq K_\mu\sigma_e^2$$

Instead

$$E(\hat{\mu} - \mu)^2 = \frac{1}{p}\sigma_a^2 + \frac{1}{q}\sigma_h^2 + \frac{1}{r}\sigma_s^2 + \lambda\sigma_{hs}^2 + K_\mu\sigma_e^2$$

Method 2 applies only to correcting data by least squares estimates for which (3) applies. It is not difficult to determine which $\hat{b}$'s qualify.

Now the data are corrected as follows (in practice only certain linear functions of the observations need to be corrected):

$$z_a = y_a - \sum_{i=1}^{s} \hat{b}_i x_{ia} \tag{4}$$

Suppose that for $i, j = s + 1, \cdots, r \leq p$ all $C_{ij} = 0$ when $i \neq j$. Let

$$Z_i = \sum_a x_{ia}z_a .$$

Then compute (5). Note that

$$Z_u = Y_u - \sum_{i=1}^{s} \hat{b}_i C_{ui}$$

$$\sum_{i=s+1}^{r} Z_i^2/C_{ii} \tag{5}$$

It is found that, except for σ_e^2, the expectation of (5) is the same as the expectation of (6) with $b_1, \cdots, b_s$ assumed $= 0$.

$$\sum_{i=s+1}^{r} Y_i^2/C_{ii} \tag{6}$$

The coefficient of σ_e^2 in the expectation of (5) is increased over that of (6) by the quantity.

$$\sum_{i=1}^{s}\sum_{j=1}^{s} C^{ij}P_{ij} , \qquad \text{where} \tag{7}$$

C^{ii} are elements of the matrix inverse to the matrix of C_{ij} $(i, j = 1, \cdots, p)$, and

$$P_{uv} = \sum_{i=s+1}^{r} C_{iu}C_{iv}/C_{ii}$$

Computation of (7) is simple if s is small and if least squares equations (2) can be rewritten as (8). This can be done in many cases.

$$\text{(8)} \qquad \sum_{j=1}^{p} C_{ij}b_j = Y_i \qquad i = 1, \cdots, s$$

$$\sum_{j=1}^{s} C_{ij}b_j + C_{ii}b_i = Y_i \qquad i = s+1, \cdots, p$$

Note in these equations that for all $i, j = s + 1, \cdots, p$, $C_{ij} = 0$ when $i \neq j$. When equations (8) prevail, C^{ii} and $\hat{b}_i$ $(i, j = 1, \cdots, s)$ can be computed from equations (9).

$$\text{(9)} \qquad \sum_{j=1}^{s} C'_{ij}\hat{b}_j = Y'_i \qquad (i = 1, \cdots, s)$$

In equations (9)

$$C'_{uv} = C_{uv} - \sum_{i=s+1}^{p} C_{iu}C_{iv}/C_{ii}$$

$$Y'_u = Y_u - \sum_{i=s+1}^{p} C_{iu}Y_i/C_{ii}$$

The least squares estimates of $b_1, \cdots, b_s$ are the solution to equations (9), and C^{ii} $(i, j = 1, \cdots, s)$ are the elements of the matrix inverse to the matrix of coefficients in (9).

Let us illustrate Method 2 with the data of Table 1. We shall now assume that the a's are fixed. First the least squares estimates of the a's are computed. This is done most simply by estimating them jointly with $d_{ij} = \mu + h_i + s_j + (hs)_{ij}$. Thus the equations are reduced to the form of equations (8). Looking at Table 1 these equations are

$$7\hat{a}_1 + 3\hat{d}_{11} + \hat{d}_{21} + 3\hat{d}_{41} = 2931$$

and similarly for the other "a" equations

$$3\hat{a}_1 + 2\hat{a}_2 + 5\hat{d}_{11} = 2395$$

and similarly for the other "d" equations
Then by (9) these equations reduce to the ones shown in Table 10.

TABLE 10

$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	
3.825	−3.825	0.	0.	− 73.500
−3.825	6.492	−2.667	0.	101.500
0.	−2.667	6.000	−3.333	41.333
0.	0.	−3.333	3.333	− 69.333

One restriction must be imposed before a solution is obtainable. A convenient one is $\hat{a}_4 = 0$. Then the solution is $\hat{a}_1 = 12.08$, $\hat{a}_2 = 31.30$, $\hat{a}_3 = 20.80$, $\hat{a}_4 = 0$.

Inverting the matrix of coefficients of Table 10 with fourth row and column deleted, the C^{ii} pertaining to the a's are obtained. These are presented in Table 11.

TABLE 11

a_1	a_2	a_3
.936438	.675000	.300000
.675000	.675000	.300000
.300000	.300000	.300000

Now the data can be corrected for the $\hat{a}$'s. For example, the corrected total for the subclass pertaining to herd 1 × sire 1 is $2395 - 3(12.08) - 2(31.30) = 2296.16$. The corrected subclass and class totals are shown in Table 12.

TABLE 12

Herd	Sire			Total
	1	2	3	
1	2296.16	2461.20	1609.00	6366.36
2	1568.02	2005.00	2219.80	5792.82
3	1611.10	3277.20		4888.30
4	2871.26		3685.60	6556.86
Total	8346.54	7743.40	7514.40	23,604.34

Using the totals of Table 12 in conjunction with the subclass and class numbers of Tables 1 and 4, the corrected sums of squares are computed. Thus, $H' = (6366.36)^2/16 + \cdots + (6556.86)^2/17$. These quantities are:

$$HS' = 10{,}016{,}791 \qquad S' = 9{,}833{,}620$$

$$H' = 9{,}954{,}295 \qquad CF' = 9{,}774{,}822$$

Next the amounts by which the coefficients of σ_e^2 are increased in the corrected as compared to the uncorrected sums of squares are needed. Using (7), the P_{ij} pertaining to HS' are computed. Looking at Table 1,

$$P_{11} = \frac{3^2}{5} + \frac{1^2}{4} + \frac{3^2}{8} = 3.175$$

$$P_{12} = \frac{3(2)}{5} + \frac{1(3)}{4} + \frac{3(5)}{8} = 3.825$$

Table 13 presents the complete set of P's for HS'.

TABLE 13

a_1	a_2	a_3
3.175	3.825	0.
3.825	14.508	2.667
0.	2.667	10.000

The sum of products of corresponding entries in Table 11 and Table 13 is

$$.936438(3.175) + \cdots + .300000(10.000) = 22.53$$

Therefore, the coefficient of σ_e^2 in $E(HS') = 10 + 22.53 = 32.53$.

The P_{ij} for H' can be computed by reference to Table 3 thus

$$P_{11} = \frac{3^2}{16} + \frac{1^2}{15} + \frac{3^2}{17} = 1.159$$

$$P_{12} = \frac{3(6)}{16} + \frac{1(3)}{15} + \frac{3(5)}{17} = 2.207$$

Table 14 presents the complete set of P_{ij} for H'.

TABLE 14

a_1	a_2	a_3
1.159	2.207	1.504
2.207	9.765	4.988
1.504	4.988	6.624

Multiplying these values by those of Table 11, the addition to σ_e^2 in $E(H') = 16.54$.

The P_{ij} for S' are shown in Table 15. Referring to Table 4,

$$P_{11} = \frac{7^2}{20}, \qquad P_{12} = \frac{7(13)}{20}, \qquad \text{etc.}$$

TABLE 15

a_1	a_2	a_3
2.450	4.550	0.
4.550	12.215	4.235
0.	4.235	7.215

Then the addition to σ_e^2 in $E(S') = 2.450\ (.936438) + \cdots = 21.39$. Finally the P_{ij} for CF' are

$$P_{11} = \frac{7^2}{57}, \qquad P_{12} = \frac{7(21)}{57}, \qquad \text{etc. as shown in Table 16.}$$

TABLE 16

a_1	a_2	a_3
.860	2.579	1.965
2.579	7.737	5.895
1.965	5.895	4.491

Multiplying and summing corresponding entries of Tables 11 and 16, the addition to the coefficient of σ_e^2 in $E(CF') = 15.57$.

Table 17 shows the corrected sums of squares and their expectations.

TABLE 17

	μ^2	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
HS'	57.	57.	57.	57.	32.53	10,016,791
H'	57.	57.	24.04	24.04	20.54	9,954,295
S'	57.	18.51	57.	18.51	24.39	9,833,620
CF'	57.	14.93	19.11	6.19	16.57	9,774,822

The equations to be solved are presented in Table 18

TABLE 18

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
$H' - CF'$	42.07	4.93	17.85	3.97	179,473
$S' - CF'$	3.58	37.89	12.32	7.82	58,798
$HS' - H' - S' + CF'$	−3.58	−4.93	20.64	4.17	3,698

The estimate of σ_e^2 can be obtained readily from the residual sum of squares after estimating the a's and d's, that is from

$$\sum_h \sum_i \sum_j \sum_k y_{hijk}^2 - \text{Reduction } (a_h , d_{ij}).$$

$$\sum_h \sum_i \sum_j \sum_k y_{hijk}^2 = 11{,}124{,}007$$

$$\text{Reduction } (a_h , d_{ij}) = 12.08\ (-73.500) + 31.30\ (101.500) + 20.80\ (41.333) + HS = 3149 + 10{,}970{,}369 = 10{,}973{,}518$$

The residual sum of squares is therefore 11,124,007 − 10,973,518 = 150,489, with expectation 44 σ_e^2 . Consequently $\hat{\sigma}_e^2 = 150{,}489/44 = 3{,}420$. Substituting $\sigma_e^2 = 3{,}420$ in the equations of Table 18 and solving, the estimates of the variances are $\sigma_h^2 = 3792$, $\sigma_s^2 = 409$, $\sigma_{hs}^2 = 243$. It is not surprising that these estimates are different from the estimates obtained by Method 1. The sampling variances must be extremely large in both cases.

Adapting Method 2 to a model with covariates is easy to accomplish. In some problems this would simplify the computations. For example, if many years were involved, the number of fixed elements in the model could be reduced by fitting a quadratic or cubic to years instead of estimating individual yearly effects as was done in this example. If the years exhibit no trend, the simplest procedure is to regard a's as random variables and then to apply Method 1.

Method 3

When it is computationally feasible, Method 3 is the most satisfactory of the three methods for estimating variance components. For one thing it gets around the difficulty of fixed elements in the model. For another, it yields unbiased estimates even though certain elements of the model are correlated. The manner in which interference by these correlations is eliminated is described subsequently.

Unfortunately Method 3 is not likely to be computationally feasible in the non-orthogonal case unless the number of different classes is small or unless the design incorporates planned non-orthogonality and consequently the mean squares of the analysis of variance can be computed without solving least squares equations. In these two cases the expectations of the mean squares are easy to compute. For example, the analysis of the balanced incomplete block design is simple and so is the writing of the expectations of the mean squares. The basic facts needed for employing Method 3 are stated below.

Let the linear model describing y_a, the a-th observation be

$$y_a = \sum_{i=1}^{p} b_i x_{ia} + e_a \tag{10}$$

The x's are known. The e's have mean zero, are uncorrelated, and have common variance σ_e^2. For the present we shall not specify which b's are fixed and which distributed.

Now if $b_{q+1}, \cdots, b_p$ are set $= 0$, the least squares estimates of $b_1, \cdots, b_q$ are the solution to equations (11).

(11)
$$\hat{b}_1 C_{11} + \hat{b}_2 C_{12} + \cdots + \hat{b}_q C_{1q} = Y_1$$
$$\hat{b}_1 C_{21} + \hat{b}_2 C_{22} + \cdots + \hat{b}_q C_{2q} = Y_2$$
etc.

In equation (11)

$$C_{ij} = \sum_{a=1}^{N} x_{ia} x_{ja}$$

$$Y_i = \sum_{a=1}^{N} x_{ia} y_a$$

The reduction in sum of squares due to $\hat{b}_1, \cdots, \hat{b}_q$ is

$$R(b_1, \cdots, b_q) = \sum_{i=1}^{q} \hat{b}_i Y_i \tag{12}$$

But since

$$\hat{b}_i = \sum_{j=1}^{q} C^{ij} Y_j ,$$

where C^{ij} are elements of the matrix inverse to the C_{ij} matrix $(i, j = 1, \cdots, q)$,

$$R(b_1, \cdots, b_q) = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} Y_i Y_j \tag{13}$$

Using (13), the expectation of $R(b_1, \cdots, b_q)$ is easy to write, but not necessarily easy to compute.

$$E[R(b_1, \cdots, b_q)] = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} E(Y_i Y_j) \tag{14}$$

Use will be made of the fact that (14) can be written

$$E[R(b_1, \cdots, b_q)] = \sum_{i=1}^{q} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + \sum_{i=q+1}^{p} \sum_{j=1}^{q} C_{ij} E(b_i b_j) + \sum_{i=q+1}^{p} \sum_{j=q+1}^{p} \lambda_{ij} E(b_i b_j) + q' \sigma_e^2 , \tag{15}$$

where

$$\lambda_{uv} = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} [C_{iu} C_{jv} + C_{iv} C_{ju}] \qquad \text{when } u \neq v,$$

and

$$\lambda_{uu} = \sum_{i=1}^{q} \sum_{j=1}^{q} C^{ij} C_{iu} C_{ju} \qquad \text{(See 14)}.$$

In most variance components problems $Eb_i b_j = 0$ $(i \neq j)$. Thus only the λ_{ii} need to be computed. q' refers to the number of independent equations in (11).

It is easy to verify that the expectation of the uncorrected total sum of squares is

$$E \sum_{a=1}^{N} y_a^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + N \sigma_e^2 . \tag{16}$$

Now it becomes clear why the residual mean square has expectation σ_e^2 regardless of the assumptions concerning the b's. Making use of (15) it is seen that

$$E[R(b_1, \cdots, b_p)] = \sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij} E(b_i b_j) + p' \sigma_e^2 . \tag{17}$$

Therefore, the expectation of the residual sum of squares is

$$(18)\qquad E[\sum_a y_a^2 - R(b_1, \cdots, b_p)]$$

$$= \left[\sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij}E(b_ib_j) + N\sigma_e^2\right] - \left[\sum_{i=1}^{p} \sum_{j=1}^{p} C_{ij}E(b_ib_j) + p'\sigma_e^2\right]$$

$$= (N - p')\sigma_e^2$$

Suppose that $b_{q+1}, \cdots, b_p$ are independently distributed with means = 0 and common variance σ^2. This variance can be estimated by equating $R(b_1, \cdots, b_p) - R(b_1, \cdots, b_q)$ to its expectation. The expectation of this difference is seen by reference to (15) and (17) to be

$$(19)\qquad \sum_{i=q+1}^{p} \sum_{j=q+1}^{p} (C_{ij} - \lambda_{ij})E(b_ib_j) + (p' - q')\sigma_e^2$$

$$= \sum_{i=q+1}^{p} (C_{ii} - \lambda_{ii})\sigma^2 + (p' - q')\sigma_e^2$$

Then using the estimate of σ_e^2 arising from (18) an unbiased estimate of σ^2 can be obtained by equating $R(b_1, \cdots, b_p) - R(b_1, \cdots, b_q)$ to (19). It will be noted that the assumptions made concerning $b_1, \cdots, b_q$ are of no consequence.

Now we shall illustrate Method 3 with our data of Table 1. If one were to carry out the usual tests of hypotheses by least squares, the following sums of squares would be computed.

Among Years = R(years, herd $\times$ sire subclasses) $-$ R(herd $\times$ sire subclasses)

Among Herds = R(years, herds, sires) $-$ R(years, sires)

Among Sires = R(years, herds, sires) $-$ R(years, herds)

Herds $\times$ Sires = R(years, herd $\times$ sire subclasses) $-$ R(years, herds, sires)

Residual = $\sum_h \sum_i \sum_j \sum_k y_{hijk}^2$ $-$ R(years, herd $\times$ sire subclasses)

The last four of these quantities can also be used to estimate σ_h^2, σ_s^2, σ_{hs}^2, and σ_e^2. If years were regarded as random variables, the first would be used to estimate σ_a^2. Our present assumption is that the year effects are fixed, however.

According to (15) we need not be concerned with μ and the a's in the expectations since the expectation of each of the above reductions contains the following quantity,

$$N\mu^2 + 2 \sum_h n_{h..}\mu a_h + \sum_h n_{h..}a_h^2 .$$

This expression vanishes in the sums of squares, which are differences between two reductions.

Aside from μ and a_h terms, the expectations of the pertinent reductions are those shown in Table 19.

TABLE 19

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2
$\sum_h \sum_i \sum_j \sum_k y_{hijk}^2$	N	N	N	N
R(years, herd × sire subclasses)	N	N	N	$p + s - 1$
R(years, herds, sires)	N	N	K_1	$p + q + r - 2$
R(years, herds)	N	K_2	K_3	$p + q - 1$
R(years, sires)	K_4	N	K_5	$p + r - 1$

The entries other than the K's in Table 19 are derived from (15) and (16). The K's are computed by (14). The expectations of the sums of squares of the analysis of variance are presented in Table 20.

TABLE 20

S.S.	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2
Among Herds	$N - K_4$	0	$K_1 - K_5$	$q - 1$
Among Sires	0	$N - K_2$	$K_1 - K_3$	$r - 1$
Herds × Sires	0	0	$N - K_1$	$s - q - r + 1$
Residual	0	0	0	$N - p - s + 1$

The computations of the various reductions proceed as follows in our example. R(years, herd × sire subclasses) = 10,973,518 as was shown in the description of Method 2. To obtain R(years, herds, sires) the equations of Table 21 need to be solved. In these equations μ is estimated jointly with each a_h , while h_4 and s_3 are set = 0. These three or some other set of three restrictions on the estimates are necessary.

TABLE 21

	a_1	a_2	a_3	a_4	h_1	h_2	h_3	s_1	s_2	
a_1	7	0	0	0	3	1	0	7	0	2931
a_2	0	21	0	0	6	3	7	13	8	9983
a_3	0	0	16	0	2	9	2	0	9	6959
a_4	0	0	0	13	5	2	0	0	0	4806
h_1	3	6	2	5	16	0	0	5	6	6632
h_2	1	3	9	2	0	15	0	4	5	6086
h_3	0	7	2	0	0	0	9	3	6	5149
s_1	7	13	0	0	5	4	3	20	0	8838
s_2	0	8	9	0	6	5	6	0	17	8181

The solution is

$$a_1 = 414.77 \qquad h_1 = 6.13 \qquad s_1 = 2.48$$

$$a_2 = 419.83 \qquad h_2 = -8.15 \qquad s_2 = 15.09$$

$$a_3 = 412.39 \qquad h_3 = 143.05$$

$$a_4 = 368.59$$

$$R(\text{years, herds, sires}) = 414.77(2931) + \cdots + 15.09(8181)$$

$$= 10{,}921{,}107$$

In order to compute R(years, herds) the s_1 and s_2 rows and columns of Table 21 are deleted and the resulting equations solved. The solution is

$$a_1 = 414.76 \qquad h_1 = 11.35$$

$$a_2 = 422.99 \qquad h_2 = -6.35$$

$$a_3 = 418.32 \qquad h_3 = 150.16$$

$$a_4 = 366.30$$

$$R(\text{years, herds}) = 414.76(2931) + \cdots + 150.16(5149)$$

$$= 10{,}919{,}698$$

The reduction due to years and sires requires solution of the equations of Table 21 with the h_1, h_2, h_3 rows and columns deleted. The solution is

$$a_1 = 425.46 \qquad a_3 = 407.72$$

$$a_2 = 461.13 \qquad a_4 = 369.69$$

$$s_1 = -6.75 \qquad s_2 = 48.38$$

$$R(\text{years, sires}) = 425.46(2931) + \cdots + 48.38(8181)$$
$$= 10{,}800{,}679$$

The computations of the K's in Table 19 require inversions of certain matrices. To obtain K_1, the inverse of the matrix of coefficients in Table 21 is needed. This inverse matrix is presented in Table 22. The entries to the left of the diagonal are omitted since the matrix is symmetric.

TABLE 22

	a_1	a_2	a_3	a_4	h_1	h_2	h_3	s_1	s_2
a_1	.64297	.41685	.15893	.03469	−.07895	−.02813	−.05580	−.46226	−.22447
a_2		.42381	.16668	.03183	−.06608	−.04169	−.09296	−.38257	−.21929
a_3			.19176	.02719	−.03647	−.08558	−.04440	−.13108	−.12625
a_4				.10925	−.06501	−.04759	−.04686	−.00004	.02410
h_1					.14349	.06382	.09075	.00834	−.05104
h_2						.14976	.07769	−.02062	−.02907
h_3							.23943	.00581	−.07214
s_1								.46163	.25050
s_2									.28088

Now we need the coefficients of σ^2_{hs} in the expectations of squares and products of the right members of the equations of Table 21. These computations are facilitated by setting up Table 23.

TABLE 23

Herd × sire subclasses

Right members	11	12	13	21	22	23	31	32	41	43
$y_{1..}$	3	0	0	1	0	0	0	0	3	0
$y_{2..}$	2	4	0	3	0	0	3	4	5	0
$y_{3..}$	0	2	0	0	5	4	0	2	0	3
$y_{4..}$	0	0	5	0	0	2	0	0	0	6
$y_{.1.}$	5	6	5	0	0	0	0	0	0	0
$y_{.2.}$	0	0	0	4	5	6	0	0	0	0
$y_{.3.}$	0	0	0	0	0	0	3	6	0	0
$y_{..1}$	5	0	0	4	0	0	3	0	8	0
$y_{..2}$	0	6	0	0	5	0	0	6	0	0

The coefficients of σ_{hs}^2 ir the squares and products of right members are the squares or products of appropriate rows in Table 23. For example, the coefficient of σ_{1s}^2 in $E\ y_1^2 ...$ is $3^2 + 1^2 + 3^2 = 19$. That in $E(y_1 ... y_2 ...)$ is $3(2) + 1(3) + 3(5) = 24$. The complete set of coefficients is presented in Table 24, with entries to the left of the diagonal omitted due to the symmetry of the matrix.

TABLE 24

	a_1	a_2	a_3	a_4	h_1	h_2	h_3	s_1	s_2
a_1	19	24	0	0	15	4	0	43	0
a_2		79	16	0	34	12	33	71	48
a_3			58	26	12	49	12	0	49
a_4				65	25	12	0	0	0
h_1					86	0	0	25	36
h_2						77	0	16	25
h_3							45	9	36
s_1								114	0
s_2									97

Multiplying and summing corresponding entries of Tables 22 and 24,

$$K_1 = 19(.64297) + 2(24)(.41685) + \cdots + 97(.28088)$$
$$= 38.29$$

Calculation of K_2 requires the inverse of the matrix of coefficients of Table 21 with s_1 and s_2 columns and rows deleted. This inverse is presented in Table 25.

TABLE 25

	a_1	a_2	a_3	a_4	h_1	h_2	h_3
a_1	.17524	.03588	.03770	.03027	−.06049	−.04552	−.03629
a_2		.10581	.05361	.03375	−.06365	−.06022	−.09421
a_3			.13358	.03635	−.05523	−.09823	−.07138
a_4				.10523	−.05576	−.04461	−.03433
h_1					.12204	.05734	.06178
h_2						.14663	.06867
h_3							.20025

Also needed in the computation of K_2 are the coefficients of σ_s^2 in the expectations of squares and products of right members of the least squares equations. Table 26 facilitates this computation.

TABLE 26

Right members	Sires		
	1	2	3
$y_{1}...$	7	0	0
$y_{2}...$	13	8	0
$y_{3}...$	0	9	7
$y_{4}...$	0	0	13
$y_{.1..}$	5	6	5
$y_{.2..}$	4	5	6
$y_{.3..}$	3	6	0

The coefficient of σ_s^2 in $E(y_1^2...)$ is $7^2 = 49$, in $E(y_1...y_2...)$ is $7(13) = 91$, etc. The complete set is shown in Table 27.

TABLE 27

	a_1	a_2	a_3	a_4	h_1	h_2	h_3
a_1	49	91	0	0	35	28	21
a_2		233	72	0	113	92	87
a_3			130	91	89	87	54
a_4				169	65	78	0
h_1					86	80	51
h_2						77	42
h_3							45

$$\text{Now } K_2 = 49\ (.17524) + 2(91)\ (.03588) + \cdots + 45\ (.20025)$$
$$= 42.29$$

K_3 is obtained from Table 24 and 25, thus

$$K_3 = 19\ (.17524) + 2\ (24)\ (.03588) + \cdots + 45\ (.20025)$$
$$= 31.71.$$

In a similar manner K_4 is found to be 22.57 and K_5 to be 22.57 (the equality of K_4 and K_5 is only a coincidence.)

Table 28 presents the pertinent reductions and their expectations (excluding μ and a_h terms)

TABLE 28

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
$\Sigma\Sigma\Sigma\Sigma\, y_{hijk}^2$	57	57	57	57	11,124,007
R(years, herd × sire subclasses)	57	57	57	13	10,973,518
R(years, herds, sires)	57	57	38.29	9	10,921,107
R(years, herds)	57	42.29	31.71	7	10,919,698
R(years, sires)	22.57	57	22.57	6	10,800,679

Then the equations to be solved are those of Table 29.

TABLE 29

	σ_h^2	σ_s^2	σ_{hs}^2	σ_e^2	
Among Herds	34.43	0	15.72	3	120,428
Among Sires	0	14.71	6.58	2	1,409
Herds × Sires	0	0	18.71	4	52,411
Residual	0	0	0	44	150,489

The solution to these equations is $\sigma_h^2 = 2255$, $\sigma_s^2 = -1295$, $\sigma_{hs}^2 = 2070$, and $\sigma_e^2 = 3420$.

ESTIMATION OF COMPONENTS OF COVARIANCE

The same general principles described in Methods 1, 2, and 3 for estimating variances can be employed to estimate covariances. To illustrate, suppose an observation is made on each of the progeny resulting from single crosses among inbred lines. If y_{ijk} is the observation on the k-th progeny of the i-th male line by the j-th female line cross, a model which might reasonably be assumed is

$$y_{ijk} = \mu + g_i + g_j + m_j + s_{ij} + e_{ijk}$$

where g_i is the general combining ability of the i-th line, g_j is the general combining ability of the j-th line, m_j is the maternal ability (exclusive of the genes transmitted to the progeny) of the j-th line, s_{ij} is peculiar to crosses $i \times j$ and of $j \times i$, and e_{ijk} is a random error. Suppose further that the elements of the model fit Eisenhart's Model II except that $E(g_i m_i) = \sigma_{gm}$.

The problem is to estimate the variances and σ_{gm}. If Method 3 is used, unbiased estimates of the variances are obtained. If Method 1 is used, the estimates are biased due to the presence of σ_{gm}. If the least squares estimates of g_i and m_i are computed, an unbiased estimate of σ_{gm} can be derived from $\sum_i \hat{g}_i \hat{m}_i$, the expectation of which is $(p-1)\sigma_{gm} + \sum_i k_i \sigma_e^2$, where p is the number of lines and $k_i \sigma_e^2$ is the covariance between $\hat{g}_i$ and $\hat{m}_i$, assuming that g_i and m_i are fixed.

A more frequently occurring type of covariance estimation problem in animal breeding arises in connection with estimation of genetic and phenotypic correlations. Observations are taken on two or more traits in some population. The following linear models characterize such observations on two traits.

$$y_a = \sum_{i=1}^{p} b_i x_{ia} + e_a \tag{20}$$

$$y_a' = \sum_{i=1}^{p} b_i' x_{ia} + e_a'$$

b_i and b_i' are fixed for $i = 1, \cdots, q$

b_i and b_i' are random variables for $i = q + 1, \cdots, p$

So far as the random variables are concerned,

$$\begin{aligned} Eb_i &= Eb_i' = 0 & E(e_a')^2 &= \sigma_{e'}^2 \\ E(b_i)^2 &= \sigma_i^2 & E(b_i b_i') &= \sigma_{ii'} \\ E(b_i')^2 &= \sigma_{i'}^2 & E(e_a e_a') &= \sigma_{ee'} \\ E(e_a)^2 &= \sigma_e^2 & \text{All other covariances} &= 0 \end{aligned}$$

Now in place of least squares reductions like

$$\sum_i \hat{b}_i Y_i \qquad \text{or} \qquad \sum_i Y_i^2/C_{ii}$$

we substitute

$$\sum_i \hat{b}_{i'} Y_i \qquad \text{or} \qquad \sum_i Y_i Y_{i'}/C_{ii},$$

where $Y_{i'}$, refers to a right member of the least squares equations for the second trait.

Then the expectations of these reductions are exactly the same as described for estimation of variance components except that $\sigma_{ii'}$, is substituted for σ_i^2 and $\sigma_{ee'}$, is substituted for σ_e^2. Therefore, any of the three methods for estimating variances can be used equally well for estimating covariances when (20) is the model.

REFERENCES

Crump, S. L. The estimation of variance components in analysis of variance. *Biom.* *2*: 7–11, 1946.

Crump, S. L. The present status of variance component analysis. *Biom.* *7*: 1–16, 1951.

Eisenhart, C. The assumptions underlying analysis of variance. *Biom.* *3*: 1–21, 1947.

Hazel, L. N. and C. E. Terrill. *Heritability of weaning weight and staple length in range Rambouillet lambs*, 1945.