Methods of variance component estimation

D. Rasch, O. Mašata

Biometric Unit, Research Institute of Animal Production, Prague-Uhříněves, Czech Republic

ABSTRACT: Estimation of variance components is a method often used in population genetics and applied in animal breeding. Even experienced population geneticists nowadays feel lost if confronted with the huge set of different methods of variance component estimation. Especially because there exists no uniformly best method, a decision which method should be used is often difficult to take. This paper gives a complete overview of methods existing in the simplest case of a one-way lay-out and demonstrates some of them by a numerical example for which the true situation is known. Of course, the one-way lay-out is of limited practical interest but can best be used to explain animal scientists the basic principles of the methods. These basic principles are principally also valid for higher classifications. Advantages and disadvantages of the methods are discussed. The symbols used are the standard biometric symbols as given in Rasch et al. (1994). We can say that all the methods offered by SPSS can be recommended.

Keywords: one-way ANOVA; ANOVA-method; MINQUE; MIVQUE; Bayesian approach; Gibbs sampling; numerical example

INTRODUCTION

We describe methods of variance component estimation for the simplest case, the one-way ANOVA, and demonstrate most of them by some data sets. For this purpose we assume that a sample of a levels of a random factor **A** has been drawn from the universe of factor levels which is assumed to be large. Not to be too abstract let us assume that the levels are sires. From the i-th sire a random sample of n_i daughters is drawn and their milk yield y_{ij} recorded. The scheme of the

$$N = \sum_{i=1}^{a} n_i$$

observations is given in Table 1. This case is called balanced if for each of the sires the same number of daughters has been selected. Balancedness in higher nested classification means equal subclass numbers as well as equal numbers of levels of nested factors within each of the levels of factors of a higher order in the hierarchy. Some of the methods described below differ from others only in the case of unbalanced data, which means in the one-way ANOVA that not for all the sires the number of daughters is the same. Therefore we simulated an unbalanced data set.

Not for all methods formulae will be given and only the principle is explained. The reason is that we wrote this paper mainly for non-mathematicians who often use several methods and need some basic understanding of what they are doing in applying special methods. Advantages and disadvantages of the methods are discussed.

According to Rasch et al. (1999) the model equation is given in (1)

$$y_{ij} = E(\boldsymbol{y}_{ij}) + \boldsymbol{e}_{ij} = \mu + \boldsymbol{a}_i + \boldsymbol{e}_{ij}$$
$$i = 1, \ldots, a; j = 1, \ldots, n_i \quad (1)$$

where: a_i = the main effects of the levels Ai. They are random variables
eij = the errors, also random
μ = constant = the overall mean

Model (1) is completed by the assumptions

($E()$ = expectation of = mean of; $V()$ = variance of)

$E(a_i) = 0$, $V(a_i) = \sigma^2_a$, $E(e_{ij}) = 0$, $V(e_{ij}) = \sigma^2$; all components on the right side of (1) are independent (2)

σ and σ^2 are called variance components.

The total number of observations is always denoted by N, in the balanced case we have $N = an$. In the sequel we give the formulae for the balanced case, generalization can be found in the references.

Let us assume that all the random variables in (1) are normally distributed even if this is not needed for all the methods.

Table 1. Observations in a one-way layout of ANOVA – unbalanced case

Level A_1	Level A_2	...	Level A_a
y_{11}	y_{21}	...	y_{a1}
y_{12}	y_{22}	.	y_{a2}
.	.	.	.
.	.		.
y_{1n_1}	y_{2n_2}	...	y_{an_a}

The normal distribution (Gauss-distribution) has the density (or likelihood) function

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \; e^{\frac{1}{2\sigma^2}(y-\mu)^2} \qquad (\sigma > 0) \qquad (3)$$

We say that y is $N(\mu; \sigma^2)$-distributed. If we have a random sample $y^T = (y_i, ...,y_n)$, its density is equal to the product of the densities of its components.

From (1), (2) and the normality assumption it follows that the $\boldsymbol{a}_i$ and the $\boldsymbol{e}_{ij}$ are independently of each other $N(0;\sigma_a^2)$- and $N(0;\sigma^2)$-distributed, respectively. The y_{ij} are not independently of each other $N(\mu; \sigma^2+ \sigma_a^2)$-distributed. The dependence exists between variables within the same factor level because

$$\text{cov}(\boldsymbol{y}_{ij}, \boldsymbol{y}_{il}) = \text{cov}(\mu + a_i + \boldsymbol{e}_{ij}, \mu + \boldsymbol{a}_i + \boldsymbol{e}_{il}) =$$
$$= \text{cov}(\boldsymbol{a}_i, \boldsymbol{a}_i) = \text{var}(\boldsymbol{a}_i) = \sigma_a^2 \qquad \text{if } j \neq 1$$

A standardized measure of this dependence is the intra-class correlation coefficient

$$\rho_{IC} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma^2} \qquad (4)$$

The analysis of variance (ANOVA) table is that of Table 2.

In Table 2 and the sequel we use standard notation with SS = sum of squares, MS = mean squares, df = degrees of freedom, res = residual, T = total and $E()$ = expectation of. Further a dot in place of a suffix means summation over that suffix and an additional bar above the y means dividing by the number of summands. Especially we have

$$\overline{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

The column "expected mean squares – E(MS)" in Table 2 is helpful to evaluate variance component estimators by the ANOVA method.

Some of the methods are developed for unbalanced data. Therefore we also assume a generalization of (1) by a general linear model for one random factor in (5).

$$Y = X\alpha + U\beta + e \qquad (5)$$

with random vectors Y and e of length N, design matrices X (Nxq) and U (Nxm), vector α of length q with fixed effects and random vector β of length m. We complete the model (5) by the following assumptions:

$\text{rank}(X) = q$; $\text{rank}(U) = m$.

The vectors e and β are independently normally distributed with expectation vector zero and covariance matrix $\sigma^2 I_N$ and $\sigma_a^2 I_m$, respectively.

For more details and application in genetics see Sorensen and Gianola (2002).

The numerical example

By random number generation we received a data set with a = 100 sires with n_i daughters as given in Table 3. We had in mind milk yields of heifers during the full first lactation with an assumed heritability coefficient.

Table 2. ANOVA table of one-way ANOVA model II in the balanced case

Source of variation	***SS***	*df*	***MS***	*E*(*MS*)
Factor A	$\boldsymbol{SS}_A = \sum_{i=1}^{a}\sum_{j=1}^{n} (\overline{\boldsymbol{y}}_i - \overline{\boldsymbol{y}})^2$	$a-1$	$\boldsymbol{MS}_A = \frac{\boldsymbol{SS}_A}{a-1}$	$\sigma^2 + n\sigma_\alpha^2$
Residual	$\boldsymbol{SS}_{\text{Res}} = \sum_{i=1}^{a}\sum_{j=1}^{n} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}}_{i.})^2$	$a(n-1)$	$\boldsymbol{MS}_{\text{Res}} = \frac{\boldsymbol{SS}_A}{a(n-1)}$	σ^2
Total	$\boldsymbol{SS}_T = \sum_{i=1}^{a}\sum_{j=1}^{n} (\boldsymbol{y}_{ij} - \overline{\boldsymbol{y}})^2$	$an-1$		

Table 3. Numbers of daughters of 100 sires

Sire 1–30	$n_i = 30$	Sire 71–75	$n_i = 23$
Sire 31–41	$n_i = 29$	Sire 76–80	$n_i = 22$
Sire 42–50	$n_i = 28$	Sire 81–85	$n_i = 21$
Sire 51–56	$n_i = 27$	Sire 86–90	$n_i = 20$
Sire 57–60	$n_i = 26$	Sire 91–95	$n_i = 19$
Sire 61–65	$n_i = 25$	Sire 96–100	$n_i = 18$
Sire 66–70	$n_i = 24$		

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma^2} = 0.4$$

Because

$$\sigma_\alpha^2 = \frac{1}{4}\sigma_g^2,$$

this can be verified for instance by $\sigma^2 = 6$ and $\sigma_\alpha^2 = 1$. Without loss of generality the overall mean was put equal to $\mu = 7\,000$. By (4) we get

$$\rho_{IC} = \frac{1}{1+6} = 0.14283$$

The data were generated in two steps. For both steps we used the pseudo-random-number generator of SPSS (transform – compute), which generates normally distributed random numbers with given mean and variance. At first we generated the 100 sire means distributed as N(7 000, 1). The numbers were repeated n_i times for each of the sires according to Table 3. In the second step to all of these values we added the effect of random error (environmental effect) distributed as N(0, 6). In this way we got a simulated value for each daughter.

In Figure 1 a part of the data is shown. The number of the sires can be found in the first column.

In Figure 2 the SPSS output of simple descriptive statistics over all data is shown.

Figure 2 shows that our random number generator works well. Mean (7 000) and total variance (7) is represented quite well and normality is reached because the estimated values of skewness and kurtosis are negligible (these parameters are zero in normal distributions). We therefore can use our

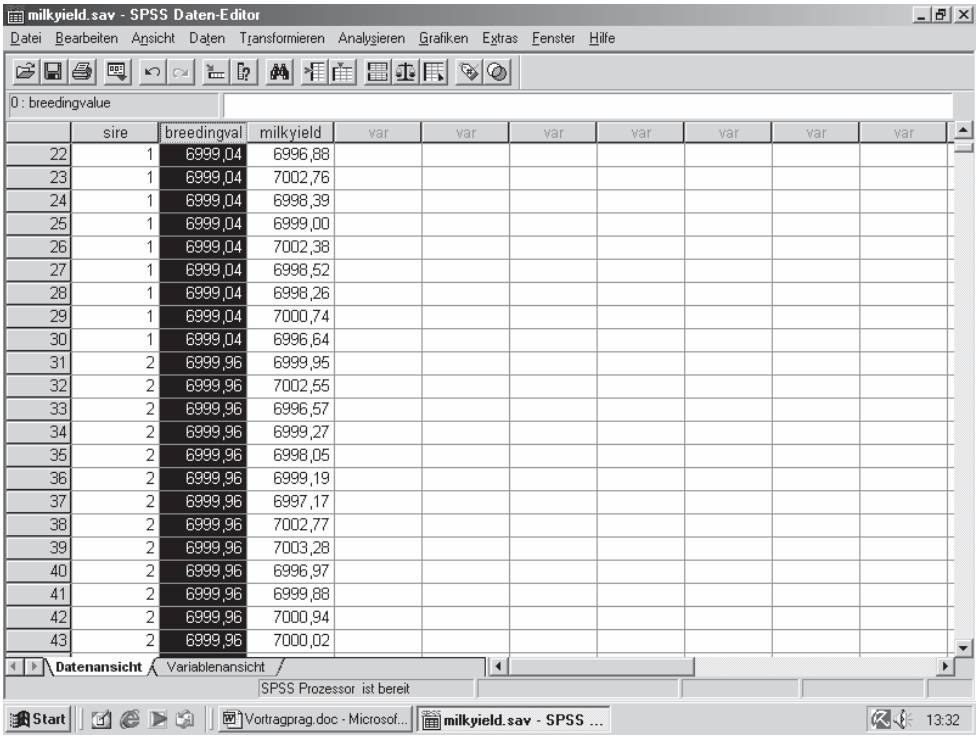

Figure 1. A part of the generated data (first lactation milk yield of 2 597 daughters)

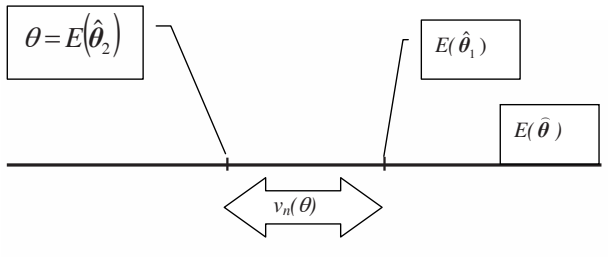

Figure 2. The parameter value θ estimated by $\hat{\theta}_1$ and $\hat{\theta}_2$

generated data set to compare the different methods of variance component estimation for the one-way-ANOVA. Methods giving estimates near $\sigma_a^2 = 1$ and $\sigma^2 = 6$ as used in the data generation are acceptable.

Properties of estimators

Before we discuss the different methods of estimating the variance components, we will introduce some theory about estimators. The following random variables can also be vectors – it follows from the context when this is the case.

Definition 1: The estimator $\hat{\theta} = \hat{\theta}(y)$ is a mapping of a random sample $y = (y_1,..., y_n)^T$ (T means a transposed vector) of size n on the parameter space of parameter θ of the distribution of this sample. The realization of the estimator is called an estimate $\hat{\theta}(y)$.

Definition 2: The estimator $\hat{\theta} = \hat{\theta}(y)$ is unbiased if $E\hat{\theta} = \theta$. The difference $E(\hat{\theta}) - \theta = v_n(\theta)$ is the bias of the estimator. The bias of an unbiased estimator is zero. The expression

$$E[\{\hat{\theta} - \theta\}^2] = V(\hat{\theta}) + v_n^2(\theta)$$

is called the mean squared error of $\hat{\theta} = \hat{\theta}(y)$ (MSE). The mean squared error of an unbiased estimator equals its variance.

In Figure 2 $\hat{\theta}_1$ is biased and $\hat{\theta}_2$ is unbiased.

Example 1: The estimator

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$$

of the mean μ of the components of a random sample $y_T = (y_1, ..., y_n)$ is unbiased.

Definition 3: The estimator $\hat{\theta} = \hat{\theta}(y)$ of θ is called minimum variance (or best) unbiased linear (MVUL) or quadratic (MVUQ) estimator if its variance is minimum amongst all unbiased linear and quadratic estimators, respectively.

Definition 4: The estimator $\hat{\theta} = \hat{\theta}(y)$ of θ is called minimum mean squared error estimator (MMSE) if its mean squared error MSE is minimum.

The estimator

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{m} = \bar{y}$$

of the mean μ in example 1 is a minimum variance (or best) unbiased linear estimator, shortly a BLUE.

Definition 5: Let $f(y, \theta)$ be the density function of θ, the parameter of the distribution of a random variable y. This is a function of two variables, θ and y. If we call it density function, we consider it as a function of y for fixed θ. But the same function $f(y, \theta)$ as a function of θ for fixed y is called likelihood function. If we estimate θ so that it will maximize the likelihood function in the parameter space of θ, we call the corresponding estimator Maximum-Likelihood Estimator (MLE).

Definition 6: The estimator is called minimum norm quadratic unbiased invariant estimator – MINQUE, if it is a quadratic form of Y in (5), unbiased and invariant against the translation of fixed effect $X\alpha$ in (5) and minimizing a matrix norm. For more details see Rao (1971), the first paper on MINQUE.

Before we discuss some of the existing methods of estimation, let us make a general remark. In (1) the model equation of the random variables y is given. Its realisations as well as the data observed are denoted by the non-bold letter y. Mathematical operations as minimizing (least squares) or maximizing (likelihood) can be performed for the realizations only. The result of such an optimisation is the estimate, the function of the y. If this is an explicit formula, we obtain the estimator by replacing the y-s in that formula by the random variables y. In implicit formulae the estimator cannot be represented in closed form but nevertheless some (often asymptotic) properties can be derived.

Estimation of the variance components

We first present methods for the so-called frequency approach. In this approach we assume that the parameters of the distribution and by this especially the overall mean and the variance components are fixed but unknown real values or vectors. In the sequel we use for all the methods the same symbols s^2 and s_α^2 for the estimates of the

Table 4. Descriptive statistics of 2 597 milk yields

Character	Number of measurements	Mean	Variance	Skewness	Kurtosis
Milk yield	2 597	6 999.9	6.944	0.002	–0.004

corresponding variance components σ^2 and σ_a^2, respectively. The estimators are given by printing all random variables bold.

Frequency approach

Analysis of variance method (ANOVA-method)

The oldest and simplest method of estimating the two variance components is due to Fisher (1925). In this method we replace in the column $E(MS)$ of the ANOVA table the variance components any σ by its estimate s and put the resulting expressions equal to the observed MS and finally we solve the resulting equations. In our special case of the balanced one-way classification this leads to

$$s^2 + ns_\alpha^2 = MS_A$$

$$s^2 = MS_{res}$$

Solving these equations gives:

$$s^2 = \mathrm{MS}_{\mathrm{res}} \tag{6}$$

$$s_\alpha^2 = \frac{1}{n}(MS_A - MS_{res}) \tag{7}$$

Properties of the estimators

Minimum variance unbiased quadratic estimator for any continuous distribution with existing first two moments (milk yield, body weight and so on) and Minimum variance quadratic estimator for normal distributions. These are not always estimators because (7) can become negative, which means that it is not a mapping into the parameter space (the positive real line) as necessary for the estimator according to definition 1. The probability of a negative estimator is given by

$$P(s_\alpha^2 < 0) = P\left[F(a-1, a(n-1)) < \frac{1}{1 + n\frac{\sigma_a^2}{\sigma^2}}\right] \tag{8}$$

In (8) $F(a-1, a(n-1))$ is a random variable with an F–distribution with $a-1$ and $a(n-1)$ d.f. Tables for these probabilities are given by Verdooren (1982).

If in our simulated example we assume approximately a mean sample size of 26, formula (8) gives

$$P(s_\alpha^2 < 0) = P\left[F(99.2497) < q = \frac{1}{1+26\frac{1}{6}} = \frac{1}{5.33} = 0.1875\right] = 0$$

For some other values of σ_a^2 and σ^2 the corresponding probabilities can be found in Table 4.

To have a real chance to find negative estimates, the environmental variance should be for instance 30 instead of 6 or the degrees of freedom must be smaller (this means, we expect negative estimates only for insufficient sample sizes). Therefore negative estimates should not be expected in our simulation study.

Quasi-maximum-likelihood method (QML)

A maximum likelihood estimate is obtained by minimizing the likelihood function of the sample under the restriction that the solution lies in the parameter space. Without this restriction some variance components may be negative. Because we then do not have an estimator in the sense of definition 1, we call the method quasi-maximum-likelihood method. If we replace in this minimum the realizations of the random variables by the random variables, we obtain the quasi-maximum likelihood estimator. Because the family of normal distributions is a two-parametric exponential family with the set of complete minimal sufficient statistics (see Rasch, 1995)

$$y.. = \frac{1}{an}\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij},\ SS_{res},\ SS_A$$

the likelihood function of the data in Table 1 can be written as

$$f(y_{11}, \ldots, y_{ab}) = \frac{e^{\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}SS_{res} + \frac{1}{\sigma^2+n\sigma^2}SS_A + \frac{1}{\sigma^2+n\sigma^2}(y-\mu)^2\right]\right\}}}{(2\pi)^{\frac{an}{2}}(\sigma^2)^{\frac{a(n-1)}{2}}(\sigma^2+\sigma_a^2)^{\frac{a}{2}}}$$

Deriving the logarithm[2] of this likelihood function with respect to μ and the two variance components without side conditions to restrict the solutions to the parameter space leads to the following estimators:

Table 5. Probabilities of negative estimates for $f_1 = 99$, $f_2 = 2\,497$, $n = 26$ and several values of σ_a^2

Value of σ_a^2/σ^2	h^2	q	Probability
6	0.40	0.19	0.000000000
20	0.17	0.43	0.000000233
25	0.14	0.49	0.000005901
30	0.12	0.54	0.000052454
40	0.09	0.61	0.000790015

$$\hat{\mu} = \bar{y}$$
$$s^2 = \boldsymbol{MS}_{res} \quad (9)$$
$$s_\alpha^2 = \frac{1}{n}\left(\frac{a-1}{a}\boldsymbol{MS}_A - \boldsymbol{MS}_{res}\right) \quad (10)$$

Equations (6) and (9) are identical, s_α^2 in (10) is biased and has a higher probability of becoming negative than the estimator in (7). The estimator of μ is needed to replace the μ in the equations stemming from the derivations with respect to the variance components only.

Maximum-likelihood method (ML)

Herbach (1959) derived a real maximum likelihood estimator (solutions restricted to the parameter space).

$$s^2 = \min\left[\boldsymbol{MS}_{res}, \frac{\boldsymbol{SS}_{res} + \boldsymbol{SS}_A}{an}\right] \quad (11)$$

$$s_a^2 = \frac{1}{n}\max\left\{\left[\frac{a-1}{a}\boldsymbol{MS}_A - \boldsymbol{MS}_{RES}\right];0\right\} \quad (12)$$

Both estimators are biased.

Restricted maximum-likelihood method (REML)

Anderson and Bancroft (1952) introduced a restricted ML method. This method uses a translation invariant restricted likelihood function depending on the variance components to be estimated only and not on the fixed effects like μ. This restricted likelihood function as a function of the sufficient statistics for the variance components. The latter is then derived with respect to the variance components under the restriction that the solutions are non-negative. The solutions are:

$$s^2 = \min\left[\boldsymbol{MS}_{res}, \frac{\boldsymbol{SS}_{res} + \boldsymbol{SS}_A}{an-1}\right] \quad (13)$$

$$s_a^2 = \frac{1}{n}\max\left\{\left[\boldsymbol{MS}_A - \boldsymbol{MS}_{RES}\right];0\right\} \quad (14)$$

Modified maximum-likelihood method (MML)

Stein (1964) and Kotz et al. (1969) found modified ML-estimators which have uniformly smaller mean squared error compared with the ML estimators of Herbach in (11) and (12). They are given by:

$$s^2 = \min\left[MS_{res}, \frac{\boldsymbol{SS}_{res} + \boldsymbol{SS}_A}{an+1}, \frac{\boldsymbol{SS}_{res} + \boldsymbol{SS}_A + an\bar{y}^2}{an+2}\right] \quad (15)$$

$$s_a^2 = \frac{1}{n}\min\left\langle\max\left\{\left[\frac{a-1}{a+1}MS_A - MS_{res}\right];0\right\}; \max\right.$$
$$\left.\left\{\left[\frac{SS_A + an\bar{y}^2}{an+2} - MS_{res}\right];0\right\}\right\rangle \quad (16)$$

Federer's estimator

From the Bayesian aspect it could be shown that some of the truncated estimators above are inadmissible. The following non-truncated and nonnegative estimators could not be proved to be inadmissible but they were not proved to be admissible either (there is no uniformly better estimator). The advantage of the estimators proposed by Federer (1968) is that they and their distribution function can be expressed in an analytical form.

$$s^2 = \boldsymbol{MS}_{res} \quad (17)$$

$$s_a^2 = \frac{1}{n}\left[\boldsymbol{MS}_A - \boldsymbol{MS}_{res}\left(1 - e^{-\delta MS_A}\right)\right] \quad (18)$$

For δ Federer proposed to choose a value in the interval

$$\delta \in \left[0, \frac{1}{SS_{res}}\right]$$

Minimum-norm and minimum-variance quadratic unbiased estimators (MINQUE)

We now consider model equation (5) with its side conditions. This model is especially useful if we consider higher classifications and mixed models. In this case the estimator is called invariant if it is

not influenced by the translation of non-random elements of the model. Rao (1971a, 1972) developed methods to estimate linear combinations of all the variance components in model (5). We will not go into details because for an understanding some knowledge of matrix algebra is needed. The basic idea is to find unbiased quadratic estimators which are invariant and minimize some matrix norm. Unfortunately, the solution in the most interesting cases depends on the unknown variance components. If they are replaced by estimates from the data, the solution is neither unbiased nor quadratic any longer. In SPSS the MINQUE procedure is the default method. MINQUE can result in negative estimates and the method to avoid this leads to neither unbiased nor quadratic estimators again.

Minimum-variance and minimum-mean squared error quadratic estimator

These procedures were developed mainly by Rao (1971b) and LaMotte (1973). The basic idea is to determine the matrix A of a quadratic form $Y^T AY$ of the vector Y in (5) so that an unbiased estimator with minimum variance (Rao, 1971b) or an estimator with minimum mean squared error (LaMotte, 1973) is obtained. On principle, the same problems arise as with the MINQUE procedure.

There are further proposals for estimators but we consider those discussed above as the most important ones from the practical point of view. For more details and variation of the above methods see Sarhai and Ojeda (2004, 2005).

Bayesian and empirical Bayesian approach and Markov chain Monte Carlo (MCMC) methods

We now only shortly mention methods for the so-called Bayesian approach. In this approach we assume that the parameters of the distribution and by this especially the variance components are random variables with some prior knowledge of their distribution. This prior knowledge is sometimes given by an a priori distribution, sometimes by data from an earlier experiment. This prior distribution is combined with the likelihood of the sample resulting in a posterior distribution. The estimator is used in such a way that it minimizes the so-called Bayes risk. If we use the squared error loss as this risk, the mean of the marginal posterior distribution is the Bayes estimator. The problem is the selection of a posterior distribution which should cover the user's knowledge (or imagination) of the unknown parameters to be estimated. So-called non-informative prior distribution is often used which in the one-way lay-out leads exactly to the Quasi-Maximum-Likehood method (QML). For more details see Gelman et al. (1995).

Method of Tiao and Tan

We explain the Bayes method by an approach of Tiao and Tan (1965). We assume like in (QML) that the random variables on the left side in (1) are normally distributed with the likelihood function given in QML

$$f(y_{11}, ..., y_{ab}) = \frac{e^{\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2} SSres + \frac{1}{\sigma^2 + n\sigma^2} SS_A + \frac{1}{\sigma^2 + n\sigma^2}(y-\mu)2\right]\right\}}}{(2\pi)^{\frac{an}{2}}(\sigma^2)^{\frac{a(n-1)}{2}}(\sigma^2 + \sigma_a^2)^{\frac{a}{2}}}$$

For the three parameters μ, σ^2, σ_a^2 the prior distribution is chosen as:

$$f_{prior}\left(\frac{1}{\sigma^2(\sigma^2 + \sigma_a^2)}\right),\ 0 < \sigma^2 < \sigma^2 + \sigma_a^2 \qquad (19)$$

By combining the prior with the likelihood function gives the posterior distribution proportional to

$$f_{posterior}(\mu, \sigma^2, \sigma_a^2 | y_{11}, ..., y_{ab}) \propto$$

$$\propto \frac{e^{\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2} SSres + \frac{1}{\sigma^2 + n\sigma^2} SS_A + \frac{1}{\sigma^2 + n\sigma^2}(y-\mu)2\right]\right\}}}{(2\pi)^{\frac{an}{2}}(\sigma^2)^{\frac{a(n-1)}{2}}(\sigma^2 + \sigma_a^2)^{\frac{a}{2}}}$$

Integrating over μ yields in the marginal distribution of the two variance components.

We now obtain the marginal distribution of each of the variance components by integrating over the other one. This leads to integral equations which can be solved by some computer programs.

For a squared error loss as the Bayes risk Klotz et al. (1969) found the Bayes estimators.

Gibbs sampling

Gibbs sampling is mainly used in the Bayesian context. Shortly spoken it is an iterative improvement of prior information.

Before we discuss the Gibbs sampling, we will define what is understood by data augmentation.

Let a posterior density be given by

$$p(\theta|y) = \int_z p(\theta|y,z)p(z|y)dz,$$

where: $p(\theta|y)$ = denotes the posterior density of the parameter θ given the observation y

$p(z|y)$ = denotes the predictive density of some so-called latent data z

$p(\theta|y,z)$ = the conditional density of θ given the so-called augmented data $x = (y,z)$

Then the predictive density is given by

$$p(z|y) = \int_z p(z|\phi,y)p(\phi|y)d\phi,$$

where: $p(z|\phi,y)$ = the conditional predictive density

After substitution and interchanging the order of integration we obtain

$$g(\theta) = \int K(\theta,\phi)g(\phi)d\phi,$$

with

$$K(\theta,\phi) = \int p(\theta|y,z)p(z|\phi)dz \qquad (22)$$

and

$$g(\theta) = p(\theta|y)$$

Starting with some initial

$g_0(\theta)$ we solve (22) by iteration via

$$g_{i+1}(\theta) = \int K(\theta,\phi)g_i(\phi)d\phi \qquad (i = 0, 1,...)$$

Gibbs sampling is based on so-called chained data augmentation (the chain is a Markov chain and Gibbs sampling is a multivariate special type of MCMC = Markov chain Monte Carlo).

When the (considered as a random variable) unknown parameter is a vector – as it is usually the case in variance component estimation – then the procedure works as follows. At first we have to know or to assume the conditional densities of each of the components of

$$\theta^T = (\theta_1, ..., \theta_p)$$

given the other components. We start with some initial value $\theta_0^T = (\theta_1^0, ..., \theta_p^0)$ of $\theta^T = (\theta_1, ..., \theta_p)$ and then iterate where each iteration is defined by the following loop:

Sample θ_1^{i+1} from $p(\theta_1|\theta_2^i, ... , \theta_p^i, y)$;
Sample θ_2^{i+1} from $p(\theta_2|\theta_1^{i+1}, \theta_3^i, ..., \theta_p^i, y)$;
.
.
Sample θ_p^{i+1} from $p(\theta_p|\theta_1^{i+1}, ..., \theta_{p-1}^{i+1}, y)$

As a good tool for those who want to be deeply engaged in Gibbs sampling we recommend two Fortran programs described in Reinsch (1996) and for the theoretical background to Gamerman (1997).

Results of some methods of variance component estimation

Results of the methods ANOVA, QML, REML and MINQUE could be obtained via the SPSS menu

Analyze
General linear model
Variance components

and then selecting the methods by the button "options".

Federer's method was not used because it assumes balanced data which need not really occur in animal breeding. MIVQUE was calculated by SAS.

The results of the methods applied to our generated data set are shown in Table 6.

The methods differ only slightly from each other. By all methods the variance component between the sires is a little bit overestimated while the residual variance component is a bit underestimated. The total variance (known to be 7) is estimated quite well by all methods. And all methods seem to work quite well in unbalanced one-way ANOVA models. But this may be quite different for higher classifications and covariates in the models.

Such conclusions can be drawn only if we apply the different methods to a data set for which the parameters are known like in our simulated data. Otherwise we can only see differences between the methods but we do not know which of them is good.

Table 6. Estimates of the variance components from the data set by different methods

Method	Estimate of σ_a^2	Estimate of σ^2	Total variance
ANOVA	1.060	5.894	6.954
QML	1.071	5.896	6.967
REML	1.084	5.896	6.980
MINQUE	1.096	5.895	6.991
MIVQUE	1.027	5.927	6.954

REFERENCES

Anderson R.L., Bancroft T.A. (1952): Statistical Theory in Research. McGraw-Hill, New York.

Federer W.T. (1968): Non-negative estimators for components of variance. Appl. Stat., *17*, 171–174.

Fisher R.A. (1925): Statistical Methods for Research Workers, Oliver and Boyd.

Gamerman D. (1997): Markov Chain Monte Carlo. Chapman and Hall, New York.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (1995): Bayesian Data Analysis. Chapman and Hall, New York.

Herbach L.H. (1959): Properties of model II type analysis of variance tests A: Optimum nature of the *F*-test for model II in balanced case. Ann. Math. Statist., *30*, 939–959.

Klotz J.H., Milton R.C., Zacks S. (1969): Mean square efficiency of estimators of variance components. J. Am. Stat. Assoc., *64*, 1383–1402.

LaMotte L.R. (1973): Quadratic estimation of variance components. Biometrics, *29*, 311–330.

Rao C.R. (1971a): Estimation of variance and covariance components: MINQUE theory. J. Multivar. Anal., *1*, 257–275.

Rao C.R. (1971b): Minimum variance quadratic unbiased estimation of variance components. J. Multivar. Anal., *1*, 445–456.

Rao C.R. (1972): Estimation of variance and covariance components in linear models. J. Am. Stat. Assoc., *67*, 112–115.

Rasch D. (1995): Mathematische Statistik. Joh. Ambrosius Barth and Wiley, Berlin, Heidelberg. 851 p.

Rasch D., Tiku M.L., Sumpf D. (1994): Elsevier's Dictionary of Biometry. Elsevier, Amsterdam, London, New York.

Rasch D., Verdooren L.R., Gowers J.I. (1999): Fundamentals in the Design and Analysis of Experiments and Surveys – Grundlagen der Planung und Auswertung von Versuchen und Erhebungen. Oldenbourg Verlag, München, Wien.

Reinsch N. (1996): Two Fortran programs for the Gibbs Sampler in univariate linear mixed models. Dummerstorf. Arch. Tierz., *39*, 203–209.

Sarhai H., Ojeda M.M. (2004): Analysis of Variance for Random Models, Balanced Data. Birkhäuser, Boston, Basel, Berlin.

Sarhai H., Ojeda M.M. (2005): Analysis of Variance for Random Models, Unbalanced Data. Birkhäuser, Boston, Basel, Berlin.

Sorensen A., Gianola D. (2002): Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics. Springer, New York.

Stein C (1969): Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. Ann. Inst. Statist. Math. (Japan), *16*, 155–160.

Tiao G.C., Tan W.Y. (1965): Bayesian analysis of random effects models in the analysis of variance I: Posterior distribution of variance components. Biometrika, *52*, 37–53.

Verdooren L.R. (1982): How large is the probability for the estimate of a variance component to be negative? Biom. J., *24*, 339–360.

Received: 2005–11–02
Accepted after correction: 2006–03–13

Corresponding Author

Ing. Ondřej Mašata, Biometric Unit, Research Institute of Animal Production, P.O. Box 1, 104 01 Prague-Uhříněves, Czech Republic
Tel. +420 267 009 527, fax +420 267 710 779, e-mail: masata.ondrej@vuzv.cz