# 7. Estimation of Genetic Parameters

**Julius van der Werf**

## Learning objectives

On completion of this topic you should:

- Understand the importance of estimation of genetic parameters in animal breeding
- Know when estimation of genetic parameters may be required
- Understand the principles of estimation of variance components
- Estimate heritability from sib analysis
- Estimate heritability from parent-offspring regression
- Be aware of the different methods that can be used to estimate genetic parameters

## Key terms and concepts

Genetic parameters
Analysis of variance
Variance components
Half Sib analysis
Offspring-parent regression
Animal model
Maximum likelihood and restricted maximum likelihood

## Introduction to the topic

The estimation of genetic parameters is an important issue in animal breeding. First of all, estimating additive genetic and possible non-additive genetic variances contributes to a better understanding of the genetic mechanism. Secondly, estimates of genetic and phenotypic variances and covariances are essential for the prediction of breeding values (selection index and BLUP) and for the prediction of the expected genetic response of selection programs. Parameters that are of interest are heritability, genetic and phenotypic correlation and repeatability, and those are computed as functions of the variance components.

Estimation of heritability is based on methods that determine resemblance between genetically related animals. Roughly, there are two methods that can be used:

1) the resemblance between parents and offspring. If we plot the observations on offspring against the values of their parents, (either sires, or dams, or their average), we can perform Offspring-Parent regression. The slope of the regression line reflects how much of the phenotypic differences that we find in parents are passed on to their offspring. The expected value of the regression line is $b_{OP} = 0.5h^2$ (or $h^2$ when regression is on mid-parent mean). Offspring-parent regression is not often used in practice. It requires data on 2 generations, and uses only this data. It is also not able to utilise genetic relationships among parents. However, the method is robust against selection of parents.

2) The estimation of variance components (within and between family components). If the variation within families is large relative to differences between families, the trait must be lowly heritable. Variance components are attributed to specific effects. For example, the (paternal) half-sib variance is due to differences between sires. The variance component represents the sire variance, which is a quarter of the additive genetic variance.

Estimation of variance components is easier to generalise, and this method is generally used to estimate genetic parameters. This topic will therefore mostly deal with variance component estimation.

In analysing data, we are promptly faced with variances. With each set of data we assume a (mixed) model that explains the observations. In this, we make a distinction between **fixed effects**, that determine the level (expected means) of observations, and **random effects** that determine variance. A model at least exists of one fixed (mean) and one random effect (residual error variance). If observations are also influenced by a genetic contribution of the animals, then a genetic variance component exists as well. In that situation, we have two components contributing to the total variance of the observations: a genetic and a residual variance component. If we calculate responses in breeding programs, we make use of those parameters. In predicting breeding values, a model can be applied which use both fixed and random effects and Best Linear Unbiased Prediction (**BLUP**). Variances and covariances are assumed to be known.

## 7.1 When to estimate variance components?

In general, the estimation of variances and covariances has to be based on a sufficient amount of data. Depending on the data structure and the circumstances during measuring, estimations can be based on some hundreds (selection experiments) or more than 10,000 observations (field recorded data). It is obvious that we are not interested in estimating variance components from every data set. The information in literature is in many cases even better than estimations based on a small data set. In general, we have to estimate variance if:

- we are interested in a new trait, from which no parameters are available;
- variances and covariances might have changed over time
- considerable changes have occurred in a population e.g. due to recent importations.

Mostly it is assumed that variances and covariances, and especially the ratio of both of them (to calculate heritability and correlations), are based on particular biological rules, which do not rapidly change over time. However, it is well known that the genetic variance changes as consequence of selection. Changes are especially expected in situations with short generation intervals, high selection intensities or high degrees of inbreeding or in a situation in which a trait is determined by only a few genes. Secondly, the circumstances under which measurements are taken can change. If conditions are getting more uniform over time, the environmental variance decreases, and consequently the heritability increases. Thirdly, the biological interpretation of a trait can change as a consequence of a changed environment; feed intake under limited feeding is not the same as feed intake under ad-lib feeding. In conclusion, there are sufficient reasons for regular estimation of (co-)variance components.

## 7.2 Principle of estimating variance components

What are variance components? Measures of extent of differences or variability are indicated with

$$var(y) \ = \ \sum_{i=1}^{n}(y_i - y)^2 \ / (n-1)$$

variance. Variance is always related to a particular effect that has an impact on observations. When we want to compute the variance on n observations (vector **y**), then an estimator for the variance is: The statistical model that describes those observations is:

$y_i = \mu + e_i$

An estimator for $\mu$ is the average of y. The differences between an observation and $\mu$, ($y_i$ -y.), are the random deviations as a consequence of the residual (or error) effect ($e_i$). In this situation, the variance of y is equal to the variance of only the random component in the model (var(y) = var(e), is the residual variance). In the numerator, the estimator of var(e) contains the **Sum of Squares** that can be ascribed to the residuals. The **expectation** of the sum of squares is equal to the multiplication of a coefficient times the variance component. In this situation it is equal to the degrees of freedom, which remain for the residual effect. Therefore, the variance is an average of the squared differences as a consequence of the effect concerned.

In a situation with more random effects, we are able to estimate more variance components. For this, we first have to quantify the contribution of each random effect. Afterwards we can compute the sum of squares for each of them. The widely used test and estimation procedure is **ANOVA**. In balanced data, it is rather simple to estimate variance components, by setting the "Mean Squares" equal to their expectations. Those expectations are linear functions of the variance components.

As an example, we can take a simple model with one main sire effect ($a_i$):

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

Assume **N** observations, with **s** sires, where N/s=**n** is the number daughters per sire. Then, the **ANOVA** table is as shown below (Table 7.1).

**Table 7.1 The ANOVA table.**

| Source | df | Sum of Squares | Mean Squares | EMS |
|---|---|---|---|---|
| Mean | 1 | SSM | SSM | |
| Sires | s-1 | SSA | SSA/(s-1) | $n\,\sigma_s^2 + \sigma_e^2$ |
| Error | N-s | SSE | SSE/(N-s) | $\sigma_e^2$ |
| Total | N | SST | | |

Where the total sums of squares (SST) is

$$SST = \sum_{i=1}^{s}\sum_{j=1}^{n} y_{ij}^2 \quad 1$$

$y_{ij}$ is an observation on the j$^{th}$ daughter of the i$^{th}$ sire. The SST is therefore the sum of each of the observations squared.

The mean sum of squares is N times the means squared.

$$SSM = N * \overline{y}_{..}^2$$

The sum of squares due to a particular effect (e.g. the sire effect) is the sum over all observations of the estimated (sire) effect in each observation squared (in balanced data this is the difference between the progeny group mean of a sire and the overall mean).

$$SSA = n\sum_{i=1}^{s}(\overline{y_{i.}} - \overline{y_{..}})^2$$

The sum of squares due to the residual (error) is the sum over all observations of the residual effect in each observation squared (this is the difference between the observation and its group mean).

$$SSE = \sum_{i=1}^{s}\sum_{j=1}^{n}(y_{ij} - \overline{y_{i.}})^2$$

From the ANOVA table we can calculate estimates of variance components as

$$\hat{\sigma}_e^2 = SSE / (N - s)$$

and $$\hat{\sigma}_s^2 = [(SSA / (s-1)) - \hat{\sigma}_e^2] / n$$

Notice that the sum of squares for the main effect (SSA) is the sum of all the squared estimates of $a_i$, because in a balanced data set the estimate of $a_i$ is equal to ($y_{i.}$ -y). In **balanced data**, it is rather simple to form the expectations for each sum of squares, because the number of observations per class of a is constant (n).

Originally, in **unbalanced data**, the same technique was applied, using for each class a weighted average. Henderson (1953) developed analogue techniques for unbalanced data. Because of the use of vector notation those techniques became popular for use in computer programmes, like Harvey and SAS. In essence techniques are the same as in balanced data, using an ANOVA table with the sum of squares for the different effects and their expectations.

## 7.3 Examples of estimating variance components

Similarity within a group can also be applied to repeated measures on animals. If repeated measurements on the same animals are more similar, the effect of temporary effects (think of *measurement error*) is not large. If repeated measurements are similar, we have a high *repeatability*, and the importance of repeatability has been outlined in lecture 5.
First we will illustrate the estimation of repeatability.

What is important in this chapter is to understand the concepts. You do not need to memorize exactly how to do the calculations.

### Example of resemblance between groups: repeatability

To get a feel for variance-between and variance- within groups, and how the first relates to resemblance (covariance) between observed values within a group , we look at an example.
Consider 3 repeated measures on each of five animals. Hence, observations are grouped by animal, and we look at resemblance between repeated measurements.

Five animals are shown, each with 3 measurements.

Example data set 1:

| **Sheep No.:** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **Day 1** | 21 | 24 | 27 | 20 | 27 |
| **Day 2** | 22 | 26 | 30 | 19 | 24 |
| **Day 3** | 20 | 25 | 30 | 18 | 27 |
| **Means:** | 21 | 25 | 29 | 19 | 26 |

Example data set 2:

| **Sheep No.:** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|
| **Day 1** | 17 | 21 | 25 | 22 | 24 |
| **Day 2** | 20 | 29 | 28 | 16 | 22 |
| **Day 3** | 23 | 28 | 34 | 16 | 32 |
| **Means:** | 20 | 26 | 29 | 18 | 26 |

By simply looking at the data we can already observe that

- the variation of observed values on the same animal is larger in data set 2.
- the observations on the same animal are more ‘alike’ in data set 1.
- the variation of the means is slightly larger in data set 2.

This ‘gut-feel’ about the data can be formally quantified with an analysis of variance.
We will perform this ANOVA based the example data sets. We expect to find

- more random error in data set 2.
- a lower repeatability in data set 2.

**Some detail about analysis of variance in the next 2 pages (reference only)**

***What do the true effects look like?***

To help understand an analysis of variance, the following Tables demonstrate ‘knowledge of the underlying parameters’ In reality we do not have this knowledge, but the example shows that larger effects means more variance. It also shows how variance components reflect ‘similarity’ (repeatability) of repeated performances.

Observed phenotypes P for each measure are the sums of permanent ($P_p$) and temporary ($P_t$) effects: $P = P_p + P_t$

We can call de temporary effects ‘measurement error’. We look again at the data, but now with the underlying effects. The actual measurements (P) are shown in **bold**, the other numbers are the underlying effects, and the means.

Consider again example data set 1, where the measurement error ($P_t$) is low:

| ANIMAL: | 1<br>Pp + Pt = P | 2<br>Pp + Pt = P | 3<br>Pp + Pt = P | 4<br>Pp + Pt = P | 5<br>Pp + Pt = P |
|---|---|---|---|---|---|
| 3 Measurements each | 22 -1 **21**<br>22 0 **22**<br>22 -2 **20** | 24 0 **24**<br>24 2 **26**<br>24 1 **25** | 27 0 **27**<br>27 3 **30**<br>27 3 **30** | 19 1 **20**<br>19 0 **19**<br>19 -1 **18** | 25 2 **27**<br>25 -1 **24**<br>25 2 **27** |
| MEANS: | 22 -1 21 | 24 1 25 | 27 2 29 | 19 0 19 | 25 1 26 |

Now consider example data set 2 where the measurement error ($P_t$) is high:

| ANIMAL: | 1<br>Pp + Pt = P | 2<br>Pp + Pt = P | 3<br>Pp + Pt = P | 4<br>Pp + Pt = P | 5<br>Pp + Pt = P |
|---|---|---|---|---|---|
| 3 Measurements each | 22 -5 **17**<br>22 -2 **20**<br>22 1 **23** | 24 -3 **21**<br>24 5 **29**<br>24 4 **28** | 27 -2 **25**<br>27 1 **28**<br>27 +7 **34** | 19 3 **22**<br>19 -3 **16**<br>19 -3 **16** | 25 -1 **24**<br>25 -3 **22**<br>25 7 **32** |
| MEANS: | 22 -2 20 | 24 2 26 | 27 2 29 | 19 -1 18 | 25 1 26 |

We want to quantify differences between $P_p$: variance **B**etween groups $\sigma^2_B$

and differences between $P_t$ : variances **W**ithin groups $\sigma^2_w$

### 7.3.1 Analysis of variance

We want to get an ***estimate*** of the variance between groups, as we cannot measure the $P_p$ effects directly (we cannot see Pp values - just phenotypic values P.
The variance of observed group means (i.e. means per animal) (21, 25, 29, 19 and 26 in example data 1) is made up of the variance of mean permanent effects ($s^2_B$: 22, 24, 27, 19, 25) plus one $n^{th}$ of the variance of mean temporary effects ($s^2_W/n$ , -1, 1, 2, 0, 1). This variance of observed

group means is determined by taking the squared differences of the group means (as a deviation from the overall mean), leading to the *sums of squares due to group effects.* We divide these SS groups by the degrees of freedom for groups (equal to the number of comparisons we can make between groups). We expect the *means squares for groups* to contain 3 times (because of 3 values per group) the variance due to groups means, i.e. 3*($\sigma_B^2 + \frac{\sigma_W^2}{n}$) = $\sigma_w^2 + 3\sigma_B^2$.

We can estimate the contribution of the variance of temporary effects within groups by taking all deviations within groups (we estimate the group mean and take the deviation of each record from each group mean. These deviations are called *residual effects* and if we square all these within group deviations, we obtain the *residual sums of squares*. If the residual sums of squares are divided by the number of residuals that we can compare (this is the *degrees of freedom* for the residual) than we obtain an estimate of the residual variance: $s^2_{W.}$ For this example:

| Effect | Degr. of Free. | Sums of Squares | Mean Squares | Expected Mean Squares |
|---|---|---|---|---|
| Mean | 1 | 8640 | | |
| Group effect (Between groups) | 4 | 192 | 48 | $\sigma_w^2 + 3\sigma_B^2$ |
| Residual (Within groups) | 10 | 18 | 1.8 | $\sigma_w^2$ |
| Total | 15 | 8850 | | |

*Analysis of variance Example data* set *1.*

Here is how these figures are calculated …

1) sums of squares due to means: 15 * $24^2$ = 8640

2) sums of squares due to group differences:
3 * ($21^2 + 25^2 + 29^2 + 19^2 + 26^2$) = 8832
corrected for mean:
3 * ($(21-24)^2 + (25-24)^2 + (29-24)^2 + (19-24)^2 + (26-24)^2$) = 192
or directly: 8832-8640 = 192

3) total sums of squares
$21^2 + 22^2$ + … <all individual weightings squared> … + $26^2$ = 8850

4) residual sums of squares
total SS – SS groups = 8850 – 8832 = 18
notice that also: $(-1)^2 + 0 + (-2)^2 + \ldots + (2)^2 = 18$

<u>The estimated variance components for example data set 1:</u>

Between groups $\sigma_B^2$ = 15.4

Within groups $\sigma_w^2$ = 1.8

Total variance is $s_B^2 + s_W^2$ = 17.2

Repeatability = intra-class correlation = 15.4/17.2 = 0.895

Variance of the group means $\sigma_B^2 + \frac{\sigma_W^2}{n}$ = 15.4 + 1.8/3 = 16

*Analysis of Variance example data* set *2:*

| Effect | Degr.of Free. | Sums of Squares | Mean Squares | Expected Mean Squares |
|---|---|---|---|---|
| Mean | 1 | 8640 | | |
| Group effect (Between groups) | 4 | 254.4 | 63.6 | $\sigma^2_w + 3\sigma^2_B$ |
| Residual (Within groups) | 10 | 178 | 17.8 | $\sigma^2_w$ |
| Total | 15 | 8929 | | |

The estimated variance components for example data set 2:

Between groups $\sigma^2_B = 15.3$

Within groups $\sigma^2_w = 17.8$

Total variance is $\sigma^2_B + \sigma^2_w = 33.1$

Repeatability = intra-class correlation = 15.3/33.1 = 0.46

Variance of the group means $\sigma^2_B + \frac{\sigma^2_W}{n}$ = 15.3 + 17.8/3 = 21.2

Comparing the two data sets:

The values of Pt are much larger in data set 2: we have larger measurement errors → The repeatability is lower
The group means are nearly the same (essentially, we have the same animals), they are only changed due to more variation in measurement error. The variance of the group means is a bit higher in data set 2.

Summary of the example

It is not critical to be able to do all these sums, they serve more as an illustration.

What is important is to get the concept:

- The extent of differences between the groups and differences within the groups can be quantified. We call these *variance components*.
- The variance components provide information about "how much alike" different observations within a group are. If differences between groups are large in relation to the differences we observe within groups, than observations within the same groups are very much 'alike'. If the variance between groups is large, the observations within the group have more *covariance*.
- Since the covariance among related animals is due to genetic components, the between group (full-sibs or half sibs) variance component can be used to determine genetic variance.

### 7.3.2 Estimation of heritability - by analysis of variance

Similarly to the example of repeatability, we can order observations into family groups, either half sib families or full sib families. We can do again an Analysis of Variance, and estimate the components of variance. Subsequently, we have to attach a meaning to the observed variance components. This depends on how we defined the groups, and what we think are the causes of common effects for the observed values within a group.

We expect half sibs to be less alike than full sibs (genetically), and therefore, the variance among half-sib groups is to a lesser extent determined by additive genetic variation.

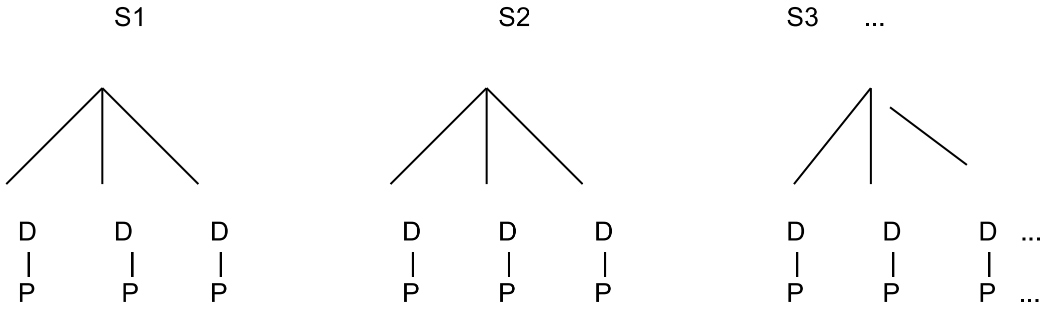
In addition, there are often other (than additive genetic) sources of variation that cause full sib groups to differ from each other. Those are dominance effects and an ‘environmental effects common to all full sibs within a family’.

*Table 7.1 Interpreting variance components, depending on the grouping made*

| **V(observed group means)**<br>**variance** | $\sigma2_B$<br>**between groups** | **+** | $\sigma^2_W/n$<br>**within groups** |
|---|---|---|---|
| V(observed HS family means) | $\frac{1}{4}\ V_A$ | + | $\frac{.75V_A + V_D + V_{Ec} + V_{Ew}}{n}$ |
| V(observed FS family means) | $\frac{1}{2} V_A + \frac{1}{4} V_D + V_{ec}$ | + | $\frac{\frac{1}{2}\ V_A + .75V_D + V_{Ew}}{n}$ |

## *Analysis of half-sib families*

Generally we have *paternal* half-sib groups, i.e. animals are only related through their sires. Each individual has a different dam:
One Progeny/dam:



From the previous chapter:

The variance between half-sib groups is equal to the covariance between half-sib individuals

$$\sigma_B^2 = \mathrm{Cov}_{HS} = V(\text{due to sires}) = \frac{1}{4}V_A$$

From analysis of variance we obtain estimates for between and within half-sib family

variance: $\hat{\sigma}_B^2$ and $\hat{\sigma}_W^2$

The correlation between 2 half sibs is \s\up6(^ = $\frac{\hat{\sigma}_B^{\ 2}}{\hat{\sigma}_B^{\ 2}+\hat{\sigma}_w^{\ 2}} = \frac{\frac{1}{4}\hat{V}_A}{\hat{V}_P} = \frac{1}{4}\hat{h}^2$ .

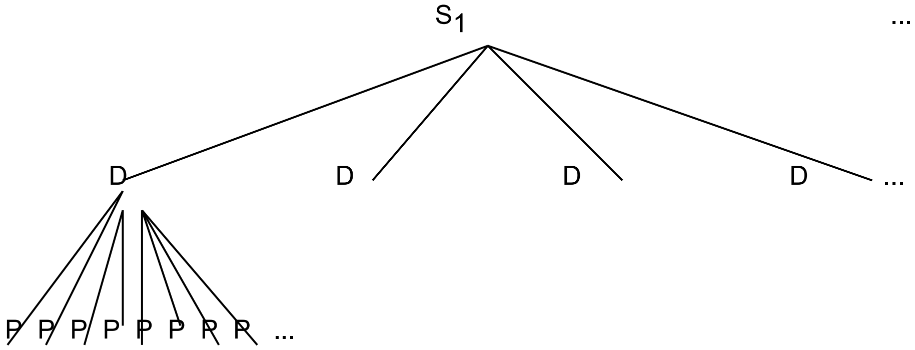
This is an intra-class correlation between half sibs.

The heritability is estimated as \s\up6(^² = $\frac{\hat{V}_A}{\hat{V}_P} = \frac{4\hat{\sigma}_s^{\ 2}}{\hat{\sigma}_s^{\ 2}+\hat{\sigma}_w^{\ 2}}$ .

or \s\up6(^² = 4 x intra-class correlation between half sibs..

## *Analysis of full-sib families*

A common structure of the data is that we have observations on full sib families, where each sire is mated to more dams, and each dam has more than one offspring. Hence, we have full sib families within half sib families and we can make groups of sires and dam groups, but dams are different for each sire. (This is called a Nested or Hierarchical design: dams are nested within sires).



In the analysis we don't have two, but three variance components: the variance between sires, the variance between dams (within sires) and the variance within dams (within full sib families). This example gives the following AoV table :

| SOURCE | EXPECTED MEAN SQUARE |
|---|---|
| Sires | $\sigma^2_w + 8\sigma_d^2 + 32\sigma_s^2$ |
| Dams | $\sigma^2_w + 8\sigma_d^2$ |
| Progeny | $\sigma^2_w$ |

Note: 8 progeny per dam
32 progeny per sire

And the expected value of the variance components is:

| Variance due to | Component | Expectation/ Interpretation |
|---|---|---|
| Sires | $\sigma_s^2$ | ¼ $V_A$ |
| Dams within sire | $\sigma_d^2$ | ¼ $V_A$+¼ $V_D$+$V_{Ec}$ |
| Progeny within dam | $\sigma_w^2$ | ½ $V_A$+.75$V_D$+$V_{Ew}$ |
| Total | $\sigma_P^2$ | $V_A$+$V_D$+$V_{Ec}$+$V_{Ew}$ |
| Sires +Dams | $\sigma_s^2+\sigma_d^2$ | ½ $V_A$+¼ $V_D$+$V_{Ec}$ |

*$V_A$ = Additive genetic variance*
*$V_D$ = Dominance variance.*
*$V_{Ec}$ = common environmental variance for full-sibs.*
*$V_{Ew}$ = environmental variance specific for each individual*

The intraclass correlation between full sibs is the between group (full sib family) proportion of $V_P$:

$$t_{FS} = \frac{(\hat{\sigma}_s^{\,2} + \hat{\sigma}_d^{\,2})}{\hat{\sigma}_P^{\,2}} = \frac{\tfrac{1}{2}\hat{V}_A + \tfrac{1}{4}\hat{V}_D + \hat{V}_{Ec}}{\hat{V}_P} \geq \tfrac{1}{2}\ h^2 \ \text{So \textbackslash s\textbackslash u}\hat{\text{p}}\text{5(}\hat{}\ \leq 2t_{FS}.$$

Since Full sibs have more in common that just genetic effects, their intra-class correlation will overestimate heritability. Only the half-sib correlation can give an unbiased estimate of heritability, since that contains genetic effects only.

**Assumptions in such ANOVA estimates of heritability:**

1. Randomly chosen sires.
   Since the estimate is based on the variance among sires. The variance among a selected group of sires will be smaller. An estimate of heritability based on progeny of selected sires will be biased downward.
2. Randomly allocated dams
3. Equal environment for each progeny group

### 7.3.3 Estimation of heritability - by regression.

## 1. Regression of Offspring on **one** parent.

What is the covariance between the performance of a sire, and the performance of its offspring? We expect there will be some covariance, because a sire and its offspring are genetically related. Of all the variation we observe between performances of sires (i.e. the phenotypic variance) we expect the sire only to transfer its genetic effects to its offspring. The random environmental effects of the sire and the offspring are assumed unrelated. Since the sire has only half of its genes in common with its offspring, we expect only half of the genetic variance in common between sires and offspring. Therefore, the theoretical expectation between performances of sires, and performances of their offspring is expected to be equal to half the additive genetic variance. The regression of the performances of offspring on performances of their parents is therefore

Regression of offspring on parent: $b_{OP} = \dfrac{Cov(parent, offspring)}{Var(parent)} = \dfrac{\tfrac{1}{2}V_A}{V_p} = \tfrac{1}{2}h^2$

Therefore, if we calculate the regression of offspring on parents, we know that, based on our quantitative genetic model, this regression this regression should be equal to ½$h^2$. We can use this knowledge to estimate heritability based on data. We can also use it to predict differences between offspring of two parents. If the parents differ an amount of 40 (say in mature weight) , we expect their offspring to differ an amount to ½$h^2$*40.

The regression of performance of offspring on performance of parent is equal to ½$h^2$

## 2. Regression of Offspring on mean of two parents.

$$b_{O\text{\textbackslash s\textbackslash up14(\_}} = \frac{\text{Cov(O, } \tfrac{1}{2}P_m + \tfrac{1}{2}P_f)}{V(\tfrac{1}{2}P_m + \tfrac{1}{2}P_f)} = \frac{\tfrac{1}{2}(\tfrac{1}{2}V_A + \tfrac{1}{2}V_A)}{\tfrac{1}{2}V_P} = \frac{V_A}{V_P} = h^2$$

Recall Definition: Heritability is the regression of offspring on the mean of parents!!

### 7.3.4 Accuracy Of Methods

The most common design to estimate heritability is livestock is from half-sib designs. For such designs, we can derive some formula to approximate the standard error on heritability estimates, using Falconer & Mackay (Chap 10):

The accuracy of heritability estimates can be measured as the standard error of the estimate. In Fig 7,1 you can see a plot of the distribution of true values for heritability, for a given estimate of 0.3 (equally you can see this as a distribution where the estimated value could end up given the true value of 0.3)

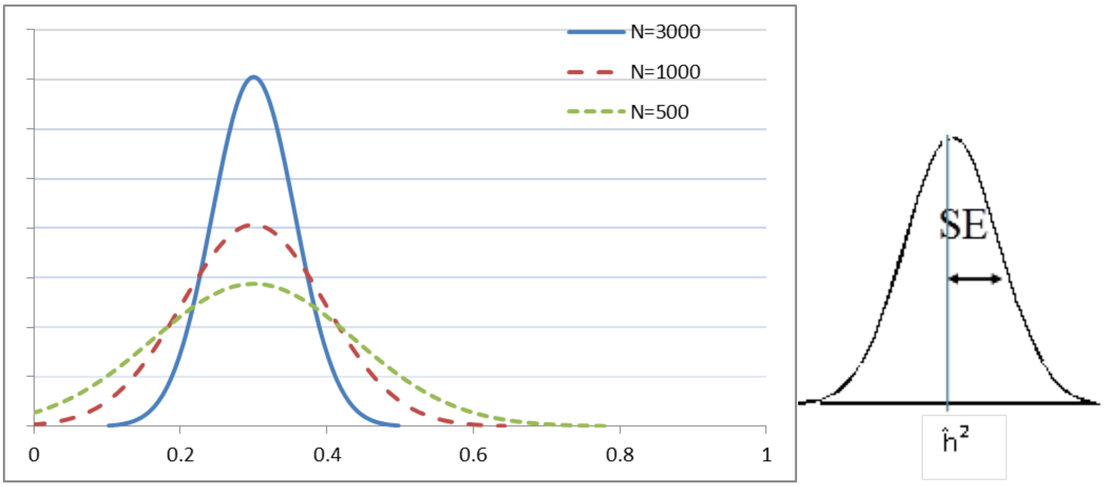


*Fig 7.1 Distribution of heritability value given an estimate of 0.3, for data sets of various size (N).*

Firstly, in HS desgns, the optimal number of progeny per sire is equal to optimal family size ≈ $4/h^2$

i.e. $n_{opt}$ =40 for $h^2 = 0.1$
and $n_{opt}$ = 8 for $h^2 = 0.5$

where $n_{opt}$ is the number of progeny per sire.
But note that the loss is greater if n is too low, so the conclusion is that optimal family sizes are about 20-40 progeny per sire (Fig 7.2)

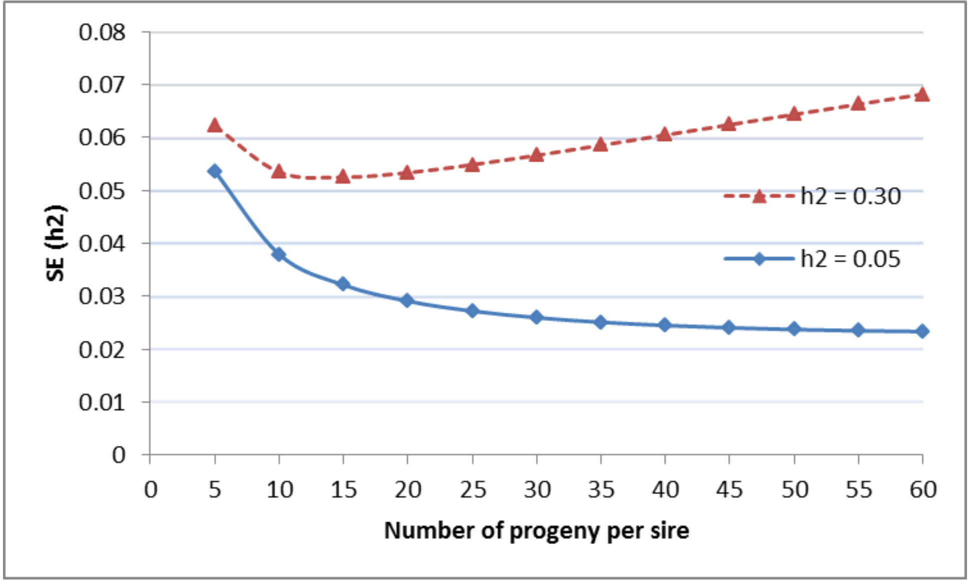


*Fig 7.2. Standard error of heritability depending on size of progeny group in a half sib design of a dataset of 3000 records. The optimum number of progeny is 40 (75 sires) for h2 = 0.1 and 13.33 (225 sires) for h2 = 0.30.*

Furthermore Falconer showed that var($h^2$) ≈ $32h^2/T$

i.e. a table for SE($h^2$) looks as follows (T is total number)

*Table 7.2 Standard error of heritability ($h^2$) estimate depending on total number (T) measured and true value of $h^2$*

| T | h2 | | |
|---|---|---|---|
| | 0.1 | 0.3 | 0.5 |
| 1000 | 0.056 | 0.098 | 0.126 |
| 2000 | 0.040 | 0.069 | 0.089 |
| 4000 | 0.028 | 0.049 | 0.063 |

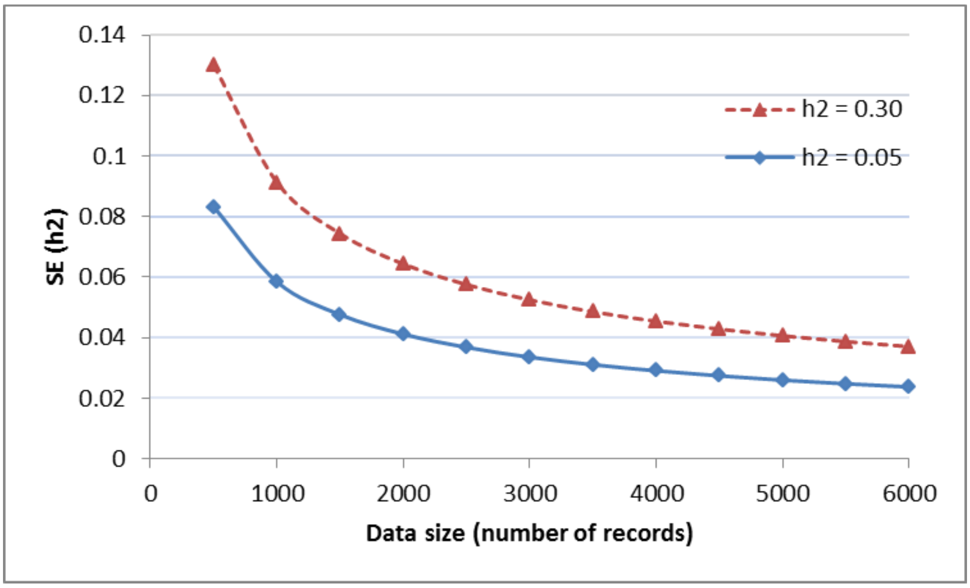


*Figure 7.3. Standard error of heritability, depending on total number of progeny. The number of progeny per sires is assumed to be optimal ($4/h^2$).*

For the estimation of heritability, we really want to be in the region of T=4000.
Note that in practice we will loose some information because of estimation of fixed effects and variation in progeny group size, so a total of T=5000 should be a target.

## 7.4 Methods used for estimating variance components

The methods of **Henderson** use **Least Squares (LS) equations** and variance components are estimated from certain quadratics (usually differences between quadratics) and their expectations. The variance of the estimates is not minimised (i.e. the estimation is not the most accurate) because sums of squares and expectations are not dependent on the variance-covariance structure of the data but rather on LS equations. Estimates of variances are unbiased but can fall outside the parameter space (e.g. they can be negative). Estimates are also not unique because when there are several random effects, sums of squares due to random effects can be computed in several ways, i.e. corrected for several combinations of other effects.

**ML** (**M**aximum **L**ikelihood) estimators maximise the likelihood of the parameters given the **density functions** and the **data**. Estimates are not unbiased but they have smaller variance then the unbiased estimators. A more detailed explanation is given in section 7.5.

**REML** (**Re**stricted **ML**) estimators maximise the likelihood of the parameters **after correcting for the fixed effects** (formally: in the space orthogonal to the fixed effects). In ML methods the loss in degrees of freedom due to correction for fixed effects is not taken into account. In REML we account for this loss in degrees of freedom. Different quadratic forms are calculated **based on the mixed model equations**

In most algorithms to obtain REML estimates, iteration is used. This process starts with a certain set of variance components and stops when the set of variance components which results in the highest likelihood of fitting data to a particular model is found. REML estimators are within the parameters space by definition but therefore they are biased. There are several algorithms to compute REML and in practice some algorithms even give negative estimates (therefore are not formally REML).

Choice of the best method to estimate variances is not obvious. One could choose for unbiasedness but in the practice of estimating variances accuracy (minimal variance) is usually preferred. It is also important to notice that unbiased methods use least squares equations and therefore cannot correct for selection in animal breeding data, e.g. through the relationships matrix or by using correlated traits. In animal breeding, data used to estimate variance components frequently originates from selection experiments or livestock improvement schemes, which involve continuous culling of animals on the basis of their performance or breeding values. In that case, ANOVA estimators, which assume that data are randomly sampled, tend to be subject to selection bias. Under certain conditions (RE)ML will account for selection, because it makes use of the mixed model equations. This very important feature has made REML the method of choice for most animal breeding applications.

## Genetic parameters

Variance components provide us with genetic parameters such as:

$$\text{Heritability } (h^2) = VA / VP = \frac{\text{additive genetic variance}}{\text{total phenotypic variance}}$$

$$\text{or} \quad \frac{\text{4*sire variance}}{\text{phenotypic variance}}$$

In an *animal model* we fit the additive genetic effect of the animal and the variance of this term gives us the additive genetic variance. In a sire model we fit the effect of sire, we estimate *sire variance* and this needs to be multiplied by 4 to get additive genetic variance.

$$\text{Genetic Correlation} = r_g = \frac{\text{Cov (A1,A2)}}{\text{sqrt}(V_{A1}{*}V_{A2})}$$

= genetic covariance divided by the product of genetic SD.

Repeatability = $(V_A + V_{ep}) / VP$

= sum of VA and Permanent Environmental Variance
divided by Phenotypic Variance

## Models of analysis

Developments in variance component estimation specific to animal breeding have been closely linked with advances in the genetic evaluation of animals by Best Linear Unbiased Prediction. Early REML applications were generally limited to models largely equivalent to those in corresponding ANOVA type analysis, considering one random effect only and estimating genetic variances from paternal half sib covariances (so-called sire model).

Recently, the Animal Model (AM) has come to dominate genetic evaluation schemes, allowing information on all known relationships between animals to be incorporated in the analysis. With the introduction of the AM, expanded models that are more accurate were described, e.g. models with permanent environmental, cytoplasmic or dominance effects. These effects are fitted as additional random effects. Maximum likelihood based methods appeared to be most flexible to accommodate such models. In terms of (RE)ML, estimation of variance components has changed thinking from the expectation of mean squares and the interpretation of observational components of variance in genetic terms (e.g. variance between and within half sib families) to a more direct approach of calculating a likelihood of a data vector for a given model with a given set of parameters, and maximising this likelihood. Such models indeed can be complicated with several random effects and covariances amongst the levels of each random effect to be specified e.g. additive genetic or dominance relationships.

There is obviously an advantage in using (RE)ML methods that are more flexible in handling animal breeding data on several (overlapping) generations (and possibly several random effects). However, the use of such methods has a danger in the sense that we do not need to think explicitly anymore about **data structure**. To estimate, as an example, additive genetic variance, we need to have a data set that contains a certain family structure that allows us to separate differences between families from differences within families. Or in other words, we need to separate genetic and residual variance. ANOVA methods require more explicit knowledge about such structure, since the data has to be ordered according to family structures (e.g. by half sib groups). Such ordering is not necessary in Likelihood Estimation. Some REML packages may even allow estimation based on data that have single records per animal and no family structure. Obviously, such data does not allow estimation of heritability.

In these notes we assume a mixed model with one random effect only. This could be either a sire effect or an animal effect. Different derivations and algorithms are easier to follow for models with one random effect only, but we will bear in mind that most methods, and particularly (restricted) maximum likelihood has been extended and applied to more complicated models.

# Appendix:

# Methods used for estimating genetic parameters

Note: This section gives more detail on some methods to estimate genetic parameters. It is mostly meant as a *future reference*, e.g. when you continue as a postgraduate student.. Try to understand the main concepts, not the detail in formulae, and details will not be examined.

*You may skip this section as an undergraduate student*

## Henderson's method 3

Genetic parameters have been estimated for many years using analysis of variance (ANOVA) or analogous methods. The ANOVA method has been popular because standard software like SAS provides such estimates.

In general, these methods require that individuals can be assigned to groups with the same degree of relationship for all members. Family structures considered most often are paternal half-sib groups or full-sib groups. In the case of paternal half-sib group all offspring of one sire are treated as one group and offspring of different sires are allocated to different groups.

Using ANOVA, the covariance among members of a family or group of relatives is usually determined as the variance component between groups. For example, in case of a sire model, the variance between sires $\sigma_s^2$ and variance within sires $\sigma_e^2$. As shown earlier, the sire variance $\sigma_s^2 = ¼\sigma_a^2$ while the variance within sires is $0.75\sigma_a^2+\sigma_e^2$.

Calculating the variance between groups, involves partitioning the sum of squared observations (SS) due to different sources of variation in the model of analysis, groups of relatives being one of them, and equating the corresponding mean squares. Mean squares are derived as the SS divided by the associated degrees of freedom, to their expectations. The same principle applies for multivariate analyses but considering sums of cross-products between traits instead of SS. For balanced data, the partial SS's are orthogonal and their expected values are simple linear combinations of the variance components between groups so that calculations are straightforward, even for multiple cross-classifications, and estimators are unique.

Data arising from animal genetics are usually not balanced but methods analogous to the ANOVA have been developed for unbalanced data. In particular, Henderson's (1953) method 3 of 'fitting constants' has found extensive use. This approach replaces the Sums of squares (SS) in the balanced ANOVA by quadratic forms involving the least squares solutions of effects for which variances are to be estimated. Its widespread application was greatly aided by the availability of a 'general' least-squares computer program tailored towards applications commonly arising in animal breeding (Harvey, 1977). Henderson's method 3 is also implemented in the statistical package SAS.

Consider a mixed linear model for one trait, represented by

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

with **y**, **b**, **u** and **e**
representing the vectors of observations, fixed effects, random effects (e.g. sire) and residual errors, respectively, and **X** and **Z** the corresponding design matrices. Assume all levels of **u** pertain to the same source of variation, for example sires, and that V(**u**)= $\sigma_u^2$ **I**, V(**e**)= $\sigma_e^2$ **I** and cov(**u**,**e**')=0.
The Least Squares equations are:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix}$$

Absorbing the fixed effects reduces the equations to $\mathbf{Z'MZ}\,\hat{u} = \mathbf{Z'\ M\ y}$

with $\mathbf{M} = \mathbf{I} - \mathbf{X}'(\mathbf{X'X})^{-1}\mathbf{X}'$.

When the inverse of $(\mathbf{X'X})$ does not exist, a generalized inverse can be used in its place.

Method 3 estimates of variance components are then:

$$\hat{\sigma}_e^2 = \frac{(\mathrm{y'y} - \hat{\mathrm{u}}'\mathrm{Z'y} - \hat{\mathrm{b}}'\mathrm{X'y})}{(N - r(\mathrm{X}) - r(\mathrm{Z}) + 1)}$$

Is residual sum of squares divided by the residual degrees of freedom

$$\hat{\sigma}_u^2 = \frac{(\hat{\mathrm{u}}'\mathrm{Z'M\,y} - (r(\mathrm{Z}) - 1)\hat{\sigma}_e^2)}{tr(\mathrm{Z'MZ})}$$

with r(**X**) and r(**Z**) denoting the column rank of **X** and **Z**, respectively, N the number of observations, and the trace operator. In this method any covariances between levels of **u** (i.e. relations between sires) are ignored.
You see again a sum of squares due to the effect captured by “Z”, e.g. the sire effect, a component of error variance is subtracted, as in the simple ANOVA. The denominator contains the total amount of information, as remember that the diagonals contain basically the number of observations of each level (sire).

An extension of method 3 to account for relationships between **u** has been considered by Sørensen and Kennedy (1984).

{Some analogy with the earlier ANOVA methods can be seen as follows:

The expression **y’y** is a vector notation for ‘total sum of squares’. Expressions like:

$\hat{b}'X'y$ (solution multiplied by right hand side)

can also be written as:

$$y'X(X'X)^{-1}X'y$$

since:

$\hat{b} = (X'X)^{-1}X'y$ (we ignore now random effects).

Since X’y contains the class totals, and X’X contains the number of observations per class, the expressions:

$$y'X(X'X)^{-1}X'y$$

gives the sum of the class totals squared, divided by the number of observations per class. This is exactly the Sum of squares due to the *b*-effect. The expression:

$$y'y - \hat{u}'Z'y - \hat{b}'X'y$$

is therefore equal to the residual sum of squares.}

## Restricted maximum likelihood

General interest in Maximum Likelihood estimators of variance components has been propelled by their desirable statistical properties: they are consistent, asymptotically normal and efficient (i.e. all information is used). Harville (1977) has given an extensive review of ML estimation. Furthermore, the ML framework provides a great deal of flexibility, allowing for designs and models for analysis which cannot be accommodated by ANOVA type of estimators. Initial interest in ML, to estimate both genetic parameters and fixed effects, was stimulated by concern about bias due to selection. A number of simulation studies have illustrated that selection can be accounted for by REML (van der Werf and de Boer, 1990) when the complete mixed model is used with all genetic relationships and all data used for selection included.

Restricted Maximum Likelihood is a ML method that accounts for the loss of degrees of freedom due to fitting fixed effects. Patterson and Thompson (1971) formally described REML. The procedure requires that **y** have a multivariate normal distribution although various authors have indicated that ML or REML estimators may be an appropriate choice even if normality does not hold (Meyer, 1990).

Over the last decade, extensive research effort has been directed towards the development of specialized and efficient algorithms for particular classes of models. These procedures will be discussed on the following pages. Before starting with that, let me give a brief introduction. In ML and REML the aim is to find the set of parameters which maximizes the likelihood of the data. The likelihood of the data for a given model can be written as a function. From calculus we know that we can find the maximum of a function by taking the first derivative and set that equal to zero. Solving that would result in the desired parameters (assuming that we did not find the minimum, this can be checked using second derivatives). The first and second derivatives of the likelihood function are complicated formulas. Different algorithms have been developed which try to circumvent this problem. An overview of different methods is given by Meyer (1990).

## Principle of maximum likelihood

Suppose we have a variable y with mean $\mu$ and standard deviation $\sigma$. The normal distribution of this variable can be represented as $y = N(\mu, \sigma^2)$ . A mathematical representation of a density function for a normally distributed variable is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}$$

This is called the **Probability Density Function** (PDF) of y.
A density function gives the density of finding certain values for y given the parameters. For example, if $\mu = 4$ and $\sigma = 1$, the density of the value 0.5 will be low, and the density will be highest for the mean: $y = \mu = 4$. Probabilities can be derived from densities, e.g. the probability of finding a y-value lower than 2. This is calculated as an integral, or a cumulative density. For example, $P(y<2)$ = the cumulative density (the integral) of $-\infty$ to 2. The probability of finding a y-value between –infinity and + infinity is equal to one, which is the integral of this function from $-\infty$ to $+\infty$.

If the probability density function (PDF) gives the value of an outcome, given the parameters, we can also work the other way around. If we have observed a number of outcomes, we can work out which parameters have the highest likelihood to match the data. For example, if we have observed the fleece weights 3, 4 and 5, and we assume a simple normal distribution, the likelihood that the mean = $\mu = 4$ is higher than for $\mu = 8$. Similarly, it is more likely that the variance is around 1 than that it will be around 10, or around 0.1. The likelihood is formally calculated using the PDF, hence we can calculate the density of finding a value 3, both for $\mu = 4$ as for $\mu = 8$, and sum these likelihoods over all observed values. The parameter set that gives the highest likelihood for all data is the maximum likelihood estimate.

The advantage of likelihood methods is that they are very flexible, as it can accommodate complex models. For example, it would be hard to accommodate related sires in the ANOVA method, but this is no problem in likelihood approaches.

A  function for a multidimensional normal distribution **y**=N(**Xb**, **V**) is:

$$f(y) \;=\; \frac{1}{2\pi^{\frac{1}{2}N}|\mathrm{V}|^{\frac{1}{2}}}\; e^{-\frac{1}{2}(\mathrm{y\text{-}Xb})'\mathrm{V}^{-1}(\mathrm{y\text{-}Xb})}$$

where N is the length of **y** and **|V|** is the determinant of **V**, the matrix describing the variances and covariances among the observations. The function f(**y**) is called a density function of **y**. However, the function can also be seen as the likelihood of a certain **y** given the parameters. The parameters are the means in **Xb** ("*location parameters*") and the variances in **V** ("*dispersion parameters*"). When the data **y** is known, f(**y**) is a likelihood function and this function can be maximised in the parameters, i.e. we want to find the parameters for which f(**y**) has the highest value. The log function is easier to work out than the actual likelihood function itself.

Instead of maximising f(**y**) we maximise the $^e$log of f(y); L(**b,V**|**X, y**), which is the log likelihood function:

$$L(b,V|X,y) = -\tfrac{1}{2}N\log(2\pi) - \tfrac{1}{2}\log(V) - \tfrac{1}{2}(y - Xb)'V^{-1}(Y - Xb) \qquad [1]$$

This function gives the likelihood of the unknown parameters **b** and **V** given the observed data **y** and the design matrix **X**. The matrix **V** depends on the variance components we are interested in. The maximum likelihood estimates of the parameters are obtained by maximising the likelihood function.

In Restricted Maximum Likelihood, as suggested by Patterson and Thompson (1971), the likelihood function of the data is maximised '*in the space of error contrasts*'. In other words, the density function is maximised after correcting all observations first for the fixed effects.

Methods available to get REML estimates can be divided in the following groups:

1) Methods using first derivatives of the likelihood function.
2) Methods using first and second derivatives of the likelihood function.
3) Derivative free methods.

For models with more random factors it is more difficult to find the maximum and it is also more difficult to construct derivatives. In categories 1 and 2, the derivatives can be calculated exactly but in most methods, approximations are used.

Methods which use both first and second derivatives, i.e geometrically speaking information on slope and curvature of the function, have been found to converge quickest (Meyer, 1989).  However, even for simple models, calculation of actual or expected second derivatives has proven to be computationally highly demanding if not prohibitive. Therefore, many REML applications are based on the so-called Expectation-Maximization (EM) algorithm. This requires, implicitly, first derivatives of the likelihood to be evaluated. The resulting estimators then have the form of quadratics in the vector of random effects solutions, obtained by BLUP for the assumed values of variances to be estimated, which are equated to their expectations. For the mixed model equations:

The
$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \alpha \mathrm{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'Y \end{bmatrix}$$

REML estimates of variance components using the EM algorithm can be obtained as:

$$\sigma_a^2 = \left[\; \hat{\mathrm{a}}'\mathrm{A}^{-1}\mathrm{a} + tr(\mathrm{A}^{-1}\mathrm{C})\sigma_e^2 \;\right] / \; q$$
$$\sigma_e^2 = \left[\; \mathrm{y'y} - \hat{\mathrm{b}}'\mathrm{X'y} - \hat{\mathrm{a}}'\mathrm{Z'y} \;\right] / \left(N - r(\mathrm{X})\right)$$

where N is the number of observations, q is the number of random genetic effect levels and **C** is the part of the inverse of the mixed model equations that corresponds with the random effects. REML can be applied to animal effects and, therefore, **a** is used to denote the vector of genetic effects (REML can be applied to a sire model as well).

The EM algorithms have the property of always yielding positive estimates as long as prior values (values which are used to start the calculations) are positive (Harville, 1977). The EM algorithm is not very difficult to program, because all elements which are needed can be derived from the mixed model equations. What is needed for each round of iteration is the solutions to the mixed model equations and the trace of the inverse of the random part of the coefficient matrix. This last element is computationally the most difficult part. Iterative methods can be used to obtain estimates for the fixed and random effect but the EM algorithm requires the direct inverse of a matrix of size equal to the number of levels of the random effects, in each round of iteration. This imposes restrictions on the kind of analyses feasible, especially for multivariate analyses.

The EM algorithm is an iterative procedure to get estimates. One starts the process with solving the equations for a given (prior) value of the variance components. These values are used in estimating the effects of the model ($\alpha$ depends on the assumed levels of the variance components). This results in a new value for the variance components and the corresponding value of $\alpha$. In an iterative process, the old values and the new value of the next iteration round are becoming closer, and ultimately converge (when the difference is very small) to a solution

## Derivative free REML (DFREML)

In the development of algorithms to compute REML an approach that did not make use of derivatives proved to be particularly successful to compute variance components from an animal model. This approach is called a derivative free approach and was first introduced by Smith and Graser (1986) and Graser et al. (1987). The maximum is found by comparing likelihood values of different parameter values.

The likelihood function from [1] can be re-written. First it is written after eliminating ('correcting for') the fixed effects. This is called the Restricted Maximum Likelihood. Secondly, it is re-written in terms of elements that relate to the mixed model equations:
where **W** is the coefficient matrix of the mixed model equations. The log **|A|** is a constant which does

$$\log L = -\frac{1}{2}\left[const \ + \ q\log\sigma_a^2 \ + \ N\log\sigma_e^2 + \mathrm{y}'\mathrm{Py} \ + \ \log|\mathrm{W}| \ + \ \log|\mathrm{A}|\right]$$

not depend on the parameters of interest (genetic relations between animals are constant for a given data set) and does not have to be evaluated. The matrix term **y'Py** represents the sum of squares of residuals and with the log determinant of the coefficient matrix (log **|W|**) it can be evaluated simultaneously by augmenting **W** by the vector of right hand sides and the total SS (**y'y**) and absorbing all rows and columns into the latter (Smith and Graser, 1986; Graser et al.,1987). The augmented mixed model array is:

$$\begin{bmatrix} \mathrm{y}'\mathrm{y} & \mathrm{y}'\mathrm{X} & \mathrm{y}'\mathrm{Z} \\ \mathrm{X}'\mathrm{y} & \mathrm{X}'\mathrm{X} & \mathrm{X}'\mathrm{Z} \\ \mathrm{Z}'\mathrm{y} & \mathrm{Z}'\mathrm{X} & \mathrm{Z}'\mathrm{Z}+\lambda\mathrm{A}^{-1} \end{bmatrix}$$

Absorption, which is also referred to as Gaussian Elimination, is used to calculate the quantities **y'Py** and log **|W|**. The residual variance can be estimated as **y'Py**/(N-r(X)) so that log L can be maximised with respect to one parameter only, the variance ratio $\lambda$ (=$\sigma_e^2/\sigma_a^2$), estimating the genetic variance $\sigma_a^2$. subsequently as $\sigma_e^2/\lambda$. This principle has been extended to models including additional random effects, such as environmental effect due to litters or a maternal genetic effect, and to multivariate analyses (Meyer, 1989).

The derivative free algorithm has been applied in the DFREML programmes that are written and distributed by Karin Meyer. These programs can be used for uni- and multivariate analysis and for

models with several random effects.
Groeneveld (1991) presented a second package for estimating variance components using a derivative free approach. This programme is distributed under the name VCE.

A more robust and efficient algorithm analysis is ‘Average Information REML’, now applied by the DFREML package. A very powerful program for parameter estimation is the ASREML package (Gilmour et al., 1996).

## The average information algorithm for REML estimation

Various techniques for solving ML / REML equations have been introduced (e.g. the Newton-Rhapson algorithm, Fisher’s scoring method and DF algorithm). The most efficient method nowadays is the Average Information algorithm. The algorithm uses the average of the second derivative (termed “Newton-Rhapson) and the expectation of the second derivative (termed Fischer scoring). Further details about this method are beyond the scope of this unit.

## Readings

1. Gilmour, A.R., Luff, A.F., Fogarty, N.M. and Banks, R. 1994, 'Genetic parameters for ultrasound fat depth and eye muscle measurements in live Poll Dorset sheep', *Australian Journal of Agricultural Research*, vol. 45, pp. 1281-1291.
2. Meyer, K. 1989, ‘Estimation of genetic parameters’, Chapter 23 in *Evolution and Animal Breeding: Reviews on Molecular and Quantitative Approaches in Honour of A. Robertson.*, eds W.G. Hill and T.F.M. McKay, Oxford University Press, Oxford, pp. 161-167.
3. Safari, A. and Fogarty, N.M. 2003, 'Genetic Parameters for Sheep Production Traits: Estimates from the Literature', Technical Bulletin 49, NSW Agriculture, Orange, Australia.
4. Thompson, R. 2002 ‘A review of genetic parameter estimation’, *Proceedings 7th World Congress On Genetics Applied to Livestock Production,* Communication N° 17-01.

## Summary

Genetic parameters (heritability, correlations) are important in animal breeding programs to estimate EBV, predict genetic change, and optimise breeding programs. They need to be estimated on a regular basis because populations as well as environments change.

Generally, large datasets with more than 1000 records are needed to estimate heritability, and 2000-4000 are needed to have a reasonable estimate of genetic correlation

ANOVA and Henderson III methods are being surpassed by Restricted Maximum Likelihood as the method of choice for estimating genetic parameters. REML is based on the mixed model and can take all family relationships into account. It also can account for selection.

Several REML algorithms exist, with the Average Information Algorithm being the most efficient.

## References

Falconer, D.S. & MacKay, T.F.C. (1996) Introduction to Quantitative Genetics (4th Ed.), Longman.

Gilmour, A. R., Thompson, R. and Cullis, B. R. 1995, 'Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models', *Biometrics,* vol. 51, pp. 1440-1450.

Gilmour, A.R., Thompson, R., Cullis, B.R. and Welham, S. 1996, ASREML – Biometrics Bulletin 3, NSW Agriculture 90pp.

Graser, H.U., Smith, S.P. and Tier, B. 1987, 'A derivative free approach for estimating variance components in animal models by restricted maximum likelihood', *Journal of Animal Science*, vol. 64, pp. 1362-1370.

Groeneveld, E. 1991, 'Simultaneous REML estimation of 60 covariance components in an animal model with missing values using the simplex algorithm', *42$^{nd}$ Annual Meeting of the EAAP*, Berlin.

Harvey, W.R. 1977, *Users Guide for LSML76. Mixed model least squares and maximum likelihood computer program*. U.S. Depart. of Agric., ARS.

Harville, D.A.. 1977, 'Maximum Likelihood approaches to variance component estimation and to related problems', *Journal of the American Statistical Association,* vol. 72, pp. 320-338.

Henderson, C.R. 1953, 'Estimation of variance and covariance components', *Biometrics,* vol. 9, pp. 226-252.

Johnson, D. L and Thompson, R. 1995, 'Restricted Maximum Likelihood Estimation of variance components for univariate animal models using sparse matrix techniques and average information', *Journal of Dairy Science,* vol. 78, pp. 449-456.

Lynch, M. and Walsh, B. 1998, 'Variance component estimation', in *Genetics and analysis of quantitative traits,* eds Lynch. M and Walsh. B. Sinauer Associaties, Inc. Sunderland, Massachusetts, U.S.A. pp. 779-803.

Meyer, K. 1989, 'Restricted maximum likelihood to estimate variance components for animals with several random effects using derivative-algorithm', *Genetics Selection Evolution,* vol. 21, pp. 317-340.

Meyer, K. 1990, 'Present status of knowledge about statistical procedures and algorithms to estimate variance and covariance components', *Proceedings of the 4th World Congress on Genetics Applied to Livestock Production,* Edinburgh, pp. 403.

Patterson, H.D. and Thompson, R. 1971, 'Recovery of inter-block information when block sizes are unequal', *Biometrika,* vol. 58, pp. 545-554.

Simm, G. 2000, *Genetic improvement of cattle and sheep*, Farming Press, Miller Freeman, UK.

Smith, S.P. and Graser, H.U. 1986, 'Estimating variance components in a class of mixed models by restricted maximum likelihood', *Journal of Dairy Science,* vol. 69, pp. 1156-1165.

Sorensen, D.A. and Kennedy, B.W. 1984, 'Estimation of genetic variance from selected and unselected populations', *Journal of Animal Science,* vol. 59, pp. 1213.

van der Werf, J.H.J. and de Boer, I.J.M. 1990, 'Estimation of additive genetic variance when base populations are selected', *Journal of Animal Science*, vol. 68, pp. 3124-3132.

## Glossary of terms

| | |
|---|---|
| Additive genetic variance[1] | The variance in a trait due to the combined effects of genes with additive action |
| Offspring-parent regression | Shows how much of the phenotypic differences of the parents are retrieved in the offspring |
| Fixed effects | Effects for which the defined classes comprise all the possible levels of interest, eg. sex, age, breed, contemporary group. Effects can be considered as fixed when the number of levels are relatively small and is confined to this number after repeated sampling. |
| Random effects | Effects which have levels that are considered to be drawn from an infinite large population of levels. Animal effects are often random. In repeated experiments there maybe other animals drawn from the population. |
| Repeatability[1] | The correlation between repeated records from the same animal |
| Variance components | A measure of the extent of differences or variability attributed to specific effects |
| Gaussian Elimination | A method to eliminate equations, similar to the 'elimination' if one has two equations with two unknowns. |
| Density Function | A function describing the likelihood of finding a certain value for a variable, given the distribution and distribution parameters of that variable. |
| Column rank | The number of independent column in a matrix |
| Trace operator | Sums the diagonal elements of a matrix |
| REML | Restricted Maximum Likelihood.is ML estimator based on the data after correction for fixed effects |
| ML | Maximum Likelihood: an estimation method that gives parameter values that are most likely, given the data and the model (defined by a density function) |

[1] Glossary terms taken from Simm (2000).