Estimation of Genetic Parameters

Guillaume Ramstein & Peter Sørensen

2022-03-04

Contents

Learning objective:

This section introduces the basic concepts of estimating genetic parameters such as:

- basic principles of estimating genetic parameters
- use of genetic relationships for estimating genetic parameters
- different methods, data sources and experimental designs for estimating genetic parameters
- importance of estimation of genetic parameters in breeding
- knowing when estimation of genetic parameters is required

1 Introduction

The estimation of genetic parameters is an important issue in animal and plant breeding. First of all, estimating additive genetic and possible non-additive genetic variances contributes to a better understanding of genetic mechanisms. Second, estimates of genetic and phenotypic (co)variances are essential for the prediction of breeding values. Third, these estimated (co)variances are used to predict the expected genetic response to selection by the breeders equation (presented later in the course). Genetic parameters of interest

are heritability, genetic and phenotypic correlation and repeatability. Genetic parameters are estimated using information on phenotypes and genetic relationships for individuals in the breeding population. In this section we will illustrate how different phenotypic sources and genetic relationships are used for estimating genetic parameters.

1.1 Genetic model

As introduced previously the phenotype for a quantitative trait is the sum of both genetic and environmental factors. In general the total genetic effect for an individual is the sum of both additive and non-additive effects. However, only the additive genetic effects are passed on to the offspring and therefore contribute to the breeding value. Therefore we only consider the additive genetic model as the basis for estimation of genetic parameters. The model for the phenotype (y) become:

$$y = \mu + a + e$$

where μ is the population mean, a is the additive effect, and $e = d + i + \epsilon$ is the pooled error in the model, which consists of non-additive genetic effects as well as environmental deviations. We assume that the additive genetic effect, a, and the residual term, e, are normally distributed which means that the observed phenotype is also normally distributed

$$\begin{aligned} a &\sim N(0, \sigma_a^2), \\ e &\sim N(0, \sigma_e^2), \\ y &\sim N(\mu, \sigma_a^2 + \sigma_e^2) \end{aligned}$$

where σ_a^2 is the additive genetic variance, σ_e^2 is the residual variance, and (σ_y^2) is the total phenotypic variance.

1.2 Genetic parameters

Heritability and genetic correlation are the key genetic parameters used in animal and plant breeding. They are defined in terms of the variance components (σ_a^2 and σ_e^2) defined in the previous section.

Heritability estimates the degree of variation in a phenotypic trait in a population that is due to genetic variation between individuals in that population. It measures how much of the variation of a trait can be attributed to variation of genetic factors, as opposed to variation of environmental factors. The narrow sense heritability is the ratio of additive genetic variance (σ_a^2) to the overall phenotypic variance ($\sigma_y^2 = \sigma_a^2 + \sigma_e^2$):

$$h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2) \tag{1}$$

A heritability of 0 implies that no genetic effects influence the observed variation in the trait, while a heritability of 1 implies that all of the variation in the trait is explained by the genetic effects. In general the amount of information provided by the phenotype about the breeding value is determined by the narrow sense heritability. Note that heritability is population-specific and a heritability of 0 does not necessarily imply that there is no genetic determinism for the trait. The trait might be highly influenced by genetic factors (e.g., the number of arms in a human population). Yet, the observed variation for the trait might not be due to genetic factors, because all alleles contributing to the trait are fixed, and there are no segregating causal alleles for the trait, in the population. Therefore, observed variation may only be due to environmental factors (e.g., accidents resulting in severed arms), and the heritability in that population might be 0.

Genetic correlation is the proportion of variance that two traits share due to genetic causes. Genetic correlations are not the same as heritability, as it is about the overlap between the two sets of influences and not the absolute magnitude of their respective genetic effects; two traits could be both highly heritable but not be genetically correlated, or they could have small heritabilities and be completely correlated (as long as

the heritabilities are non-zero). Genetic correlation (ρ_a) is the genetic covariance between two traits divided by the product of genetic standard deviation for each of the traits:

$$\rho_{a_{12}} = \frac{\sigma_{a_{12}}}{\sqrt{\sigma^2_{a_1}\sigma^2_{a_2}}} \tag{2}$$

where $\sigma_{a_{12}}$ is the genetic covariance and $\sigma^2_{a_1}$ and $\sigma^2_{a_2}$ are the variances of the additive genetic values for the two traits in the population. A genetic correlation of 0 implies that the genetic effects on one trait are independent of the other, while a correlation of 1 implies that all of the genetic influences on the two traits are identical. Thus in order to estimate the heritability and genetic correlation we need to estimate the variance component defined above (σ^2_a and σ^2_e), for each trait, in addition to the genetic covariance between traits.

1.3 Data required for estimating genetic parameters

Information on phenotypes and genetic relationships for individuals in a breeding population are, in combination with appropriate statistical models, used for accurate estimation of genetic parameters and breeding values of individuals.

Phenotypes for traits of economic importance need to be recorded accurately and completely. All individuals within a production unit (herd, flock, ranch, plot) should be recorded, as sampling bias results in samples that are not representative of the population under study. Individuals should not be selectively recorded. Data includes the dates of events when traits are observed, factors that could influence an individual's performance, and an identification of contemporaries that are raised and observed in the same environment under the same management regime. Observations should be objectively measured, if at all possible.

Genetic relationships for the individuals in the breeding population are required. Genetic relationships can be inferred from a pedigree or, alternatively, computed from genetic markers. Individuals and their parents need to be uniquely identified in the data.

Information about development (e.g., birth dates in animals, sowing date and flowering time in plants), breed composition, and genotypes for various markers could also be stored. If individuals are not uniquely identified, then genetic analysis of the population may not be possible at the individual level. In aquaculture species, for example, individual identification may not be feasible, but family identification (father and mother) may be known.

Prior information about the traits is useful. Read the literature. Most likely other researchers or breeders have already made analyses of the same species and traits. Even though their study populations are not the same as yours, their models could be useful starting points for further analyses. Their parameter estimates could result in useful predictions. The idea is to avoid the pitfalls and problems that other researchers have already encountered.

1.4 Statistical models and variance components

For estimating genetic parameters we need to specify a statistical model that describes the genetic and non-genetic factors that may affect the trait phenotypes. Often the non-genetic factors are referred to as systematic effect such as age, parity, litter size, days open, sex, herd, year, season, management, etc. (in animals) and sowing date, flowering time, location, block, etc. (in plants):

$$\text{phenotype} = \text{mean} + \text{systematic effect} + \text{genetic effect} + \textit{residual}$$

Here we make a distinction between fixed effects, that determine the level (expected mean) of observations, and random effects that determine variance. A model consists of at least one fixed effect (i.e. mean) and one random effect (the residual error variance). If observations also are influenced by a genetic contribution of the individuals, 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.

The statistical model is a formal representation of our quantitative genetic theory, but it is important to realize that all models are simple approximations of how genetic and non-genetic factors influence a trait. The goal of the statistical analysis is to find the best practical model that explains the most variation in the data. Statistical knowledge is required. The methods used for estimating genetic parameters is based on statistical concepts such as random variables, multivariate normal theory and linear (mixed) models. These concepts and their use will be explained in the following sections.

2 Methods for estimation of genetic parameters

In general, estimation of heritability and genetic correlation is based on methods that determine resemblance between genetically related individuals. Close (compared to distant) relatives share more alleles and, if the trait is under genetic influence, they will therefore share phenotypic similarities. Here we will present three methods for estimating heritability: parent-offspring regression, analysis of variance (ANOVA) for family data (e.g., half-sib/full-sib families) and restricted maximum likelihood (REML) analysis for a general pedigree. These methods are increasingly more complex, but they are also increasingly more flexible. While REML can analyze any type of relationships and structures, ANOVA can only analyze groups of individuals with similar relationships (e.g., half-sib, or full-sib families), and regression analysis can only analyze pairs of individuals with similar relationships (e.g., pairs of parent and respective offspring, or pairs of monozygotic twins). These methods can also be used for estimation of genetic correlation, but this will not be covered in these notes.

2.1 Estimating heritability using parent - offspring regression

The simplest method for estimating genetic parameters is based on regression analysis. Heritability may be estimated by comparing phenotypes for traits recorded in parent and offspring. Parent-offspring regression compares trait values in parents (y_p) to trait values in their offspring (y_o). Estimation of heritability is based on a linear regression model:

$$y_o = y_p b_{o|p} + e_o.$$

The slope of the regression line ($b_{o|p}$) is used to estimate the heritability of the trait when offspring values are regressed against the average phenotypic value of the parents (mid-parent regression) or the phenotypic values of one of the parents (single-parent regression). If only one parent's value is used then heritability is twice the slope. Therefore, the expected value of the regression line is $b_{o|p} = 0.5h^2$ (in single-parent regression) or h^2 (in mid-parent regression).

To better understand the parent-offspring regression method, consider a situation where we have collected phenotypes on a number of father-offspring families. From standard regression theory the slope is:

$$b_{o|f} = \frac{Cov(y_f, y_o)}{Var(y_o)}$$

where $Cov(y_f, y_o)$ is the covariance between the phenotypes of the father and the offspring and $Var(y_o)$ is the variance of the offspring's phenotypes.

The phenotypes of the father (y_f) and the offspring (y_o) can be expressed as:

$$\begin{aligned} y_f &= \mu + a_f + e_f \\ y_o &= \mu + 0.5a_m + 0.5a_f + a_{mendelian} + e_o \end{aligned}$$

where μ is the population mean, a_m and a_f are the additive genetic effect for the mother and the father, $a_{mendelian}$ is the mendelian deviation in the offspring (the residual part of the offspring's genetic value that is due to random chromosome segregation), and e_f and e_o are the residual effect for the father and the offspring, respectively.

The offspring gets half of the genes from each parent (i.e., genetic relationship is 0.5) and therefore the breeding value for the offspring is the average of the parents' breeding values plus the Mendelian deviation:

$$a_{\text{offspring}} = \frac{1}{2}a_{\text{father}} + \frac{1}{2}a_{\text{mother}} + a_{\text{mendelian}}$$

(a = additive genetic value = breeding value) The term $a_{mendelian}$ is necessary, because two full-sibs i and j both having parents $father$ and $mother$ receive different random sets of parental alleles. Hence the breeding values a_i and a_j of full-sibs i and j are not going to be the same. Furthermore we assume that the breeding values are normally distributed:

$$\begin{aligned} a_{father} &\sim N(0, \sigma_a^2) \\ a_{mother} &\sim N(0, \sigma_a^2) \\ a_{mendelian} &\sim N(0, 0.5\sigma_a^2) \end{aligned}$$

An expression for the covariance between the phenotypes of the parent and the offspring can be derived as:

$$\begin{aligned} Cov(y_f, y_o) &= Cov(a_f + e_f, 0.5a_m + 0.5a_f + a_{mendelian} + e_o) \\ &= Cov(a_f, 0.5a_f) \\ &= 0.5Cov(a_f, a_f) \\ &= 0.5\sigma_a^2 \end{aligned}$$

This derivation illustrate an important concept in which the phenotypic covariance between related individuals can be expressed in terms of the genetic covariance ($Cov(a_f, 0.5a_f)$). The genetic covariance between related individuals can be expressed by their genetic relationship (which in this example equals 0.5 because each offspring gets half of the genes form its parent) and the genetic variance (σ_a^2), assumed to be equal among parents (generation n) and among offspring (generation n+1).

The variance of the offspring phenotypes is:

$$\begin{aligned} Var(y_o) &= Var(0.5a_m + 0.5a_f + 0.5a_{mendelian} + e_o) \\ &= Var(0.5a_m) + Var(0.5a_f) + Var(a_{mendelian}) + Var(e_o) \\ &= 0.25Var(a_m) + 0.25Var(a_f) + Var(a_{mendelian}) + Var(e_o) \\ &= 0.25\sigma_a^2 + 0.25\sigma_a^2 + 0.5\sigma_a^2 + \sigma_e^2 \\ &= \sigma_a^2 + \sigma_e^2 \end{aligned}$$

Therefore the expected value of the regression coefficient for a father-offspring analysis is:

$$\begin{aligned} b_{o|f} &= \frac{0.5\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \\ &= 0.5\frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \\ &= 0.5h^2 \end{aligned}$$

Similar relationships can be derived for other types of parent-offspring regression analyses (mother-offspring or mid-parent-offspring regression). The heritability can therefore be estimated from the regression coefficient based on:

$$\begin{aligned} h^2 &= 2b_{o|m} \quad \text{(mother-offspring regression)} \\ h^2 &= 2b_{o|f} \quad \text{(father-offspring regression)} \\ h^2 &= b_{o|mf} \quad \text{(mean parent-offspring regression)} \end{aligned}$$

Parent-offspring regression is not often used in practice. It requires data on 2 generations, and uses only this data. It is based on the genetic relationship between parent and offspring, but it is not possible to utilize genetic relationships among parents. However, the method is robust against selection of parents.

2.2 Estimating heritability for family data using ANOVA

Genetic parameters have been estimated for many years using analysis of variance (ANOVA). This method requires that individuals can be assigned to groups with the same degree of genetic relationship for all members. Family structures considered most often are half-sib groups (in animals, paternal half-sibs' in plants, maternal half-sibs) or full-sib groups. In the case of half-sib group, all offspring of one parent are treated as one group and offspring from different parents are allocated to different groups.

2.2.1 Linear model

Estimation of heritability using ANOVA is based on a linear model. Consider a situation where we have phenotypic observations for for n_f families each with n_o offspring (half-sib or full-sib). The total number of observations is $n = n_f n_o$. A simple linear model for the phenotypic observation for the jth offspring in the ith family include the population mean (μ) and a family effect (f_i):

$$y_{ij} = \mu + f_i + e_{ij} \tag{3}$$

where e_{ij} is the residual error resulting from dominance, epistasis and environmental contributions. We assume that the e_{ij} are uncorrelated with each other and have common variance σ_e^2.

2.2.2 Assumption for the parameters in the model

The variance among family effects (the between-family, or among-family, variance) is denoted by σ_f^2. We further assume that the random factors (f and e) are uncorrelated with each other (i.e., $Cov(f_i, e_{ij}) = 0$). Therefore the analysis of variance partitions the total phenotypic variance into the sum of the variances from each of the contributing factors. Thus, the total phenotypic variance equals the variance due to family plus the residual variance:

$$\sigma_y^2 = \sigma_f^2 + \sigma_e^2 \tag{4}$$

2.2.3 Estimation of variance components

The total sum of squares (SST) around the overall mean ($\bar{y}$) is the sum of each of the observations squared:

$$SST = \sum_{i=1}^{n_f} \sum_{j=1}^{n_o} (y_{ij} - \bar{y})^2 \tag{5}$$

where y_{ij} is an observation on the jth offspring in the ith family. In ANOVA, the total variation (the sum of squares SST) is decomposed into different sources of variation corresponding to different effects in the model.

The model sum of squares (SSA) due to a particular factor (e.g., the family effect) is the sum over all observations of the estimated (family) effect in each observation squared. In balanced data sets this is the difference between the family group mean ($\bar{y}_i$) and the overall mean ($\bar{y}$):

$$SSA = n_o \sum_{i=1}^{n_f} (\bar{y}_i - \bar{y})^2 \tag{6}$$

Notice that the sum of squares for the main effect (SSA) is the sum of all the squared estimates of f_i, because in a balanced data set the estimate of f_i is equal to $(\bar{y}_i - \bar{y})$. In a balanced data set, it is rather simple to determine the expectations for each sum of squares, because the number of observations per class of f is constant (n_o).

The residual sum of squares (SSE) 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):

$$SSE = \sum_{i=1}^{n_f} \sum_{j=1}^{n_o} (y_{ij} - \bar{y}_i)^2 \tag{7}$$

The total sum of squares can be expressed as the sum of the components described above:

$$SST = SSA + SSE \tag{8}$$

In balanced data, it is rather simple to estimate variance components, by setting the mean squares (MS) equal to their expectations E(MS) (estimation by the method of moments). The mean squares are computed as:

$$\begin{aligned} MSA &= SSA/(n_f - 1) \\ MSE &= SSE/(n - n_f) \end{aligned} \tag{9}$$

The expected mean squares (derived from statistical theory on expected value of a sum of squares) are linear functions of the variance components:

$$\begin{aligned} E(MSE) &= \sigma_e^2 \\ E(MSA) &= \sigma_e^2 + \widetilde{n}_o \sigma_f^2 \end{aligned} \tag{10}$$

where $\widetilde{n}_o = [n - (\sum_{i=1}^{n_f} n_{o_i}^2/n)]/(n_f - 1)$ which reduces to n_o with equal family sizes.

Therefore in this simple model we can estimate of the residual variance components ($\hat{\sigma}_e^2$) as:

$$\hat{\sigma}_e^2 = \text{MSE} \tag{11}$$

and the estimate of the family variance ($\hat{\sigma}_f^2$) as:

$$\hat{\sigma}_f^2 = (\text{MSA-MSE})/\widetilde{n}_o \tag{12}$$

2.2.4 Estimation of heritability from variance components

Importantly, the identity Cov(within) = Var(between) (i.e., Cov(within-group observations) = Var(among-group means)) allows us to relate an estimated variance component (e.g., the between-family variance σ_f^2) with the causal underlying variance components (e.g., σ_a^2) that are our focus. To see this consider the phenotype for two full-sibs:

$$\begin{aligned} y_{o1} &= \mu + 0.5a_m + 0.5a_f + 0.5a_{mendelian_1} + e_{o1} \\ y_{o2} &= \mu + 0.5a_m + 0.5a_f + 0.5a_{mendelian_2} + e_{o2} \end{aligned}$$

and therefore phenotypic covariance between the two full-sibs can be expressed as:

$$\begin{aligned} Cov(y_{o1}, y_{o2}) &= Cov(0.5a_m + 0.5a_f + a_{mendelian1} + e_{o1}, 0.5a_m + 0.5a_f + a_{mendelian2} + e_{o2}) \\ &= Cov(0.5a_f, 0.5a_f) + Cov(0.5a_m, 0.5a_m) \\ &= 0.25Cov(a_f, a_f) + 0.25Cov(a_m, a_m) \\ &= 0.25\sigma_a^2 + 0.25\sigma_a^2 \\ &= 0.5\sigma_a^2 \end{aligned}$$

which is the covariance within full-sib families. Here we assume that parents are genetically unrelated ($Cov(a_m, a_f) = 0$).

The phenotypes for the two full-sibs can also be expressed by the ANOVA model (describing groups of individuals), rather than the genetic model (describing genetic effects):

$$\begin{aligned} y_{i1} &= \mu + f_i + e_{i1} \\ y_{i2} &= \mu + f_i + e_{i2} \end{aligned}$$

and therefore the phenotypic covariance can also be expressed as:

$$\begin{aligned} Cov(y_{i1}, y_{i2}) &= Cov(f_i + e_{i1}, f_i + e_{i2}) \\ &= Cov(f_i, f_i) + Cov(e_{i1}, e_{i2}) \\ &= \sigma_f^2 + 0 \\ &= \sigma_f^2 \end{aligned}$$

which is the between-family variance. In general, this corresponds to the identity Cov(within) = Var(between) which for the full-sib family design:

$$0.5\sigma_a^2 = \sigma_f^2$$

Therefore, for half-sib/full-sib families, the heritability can be estimated from between-family variance (σ_f^2) and residual variance (σ_e^2) in ANOVA, as follows:

$$\begin{aligned} \sigma_{hs}^2 &= 0.25\sigma_a^2 \quad \text{(half-sib families)} \\ \sigma_{fs}^2 &= 0.5\sigma_a^2 \quad \text{(half-sib families)} \\ h^2 &= \frac{4\sigma_{hs}^2}{4\sigma_{hs}^2 + \sigma_e^2} \quad \text{(half-sib families)} \\ h^2 &= \frac{2\sigma_{fs}^2}{2\sigma_{fs}^2 + \sigma_e^2} \quad \text{(full-sib families)} \end{aligned}$$

The ANOVA method has several limitations. First, we must assume that parents and families are unrelated. Second, data arising from experimental designs used for estimating genetic parameters are usually not balanced (i.e., number of offspring varies across families). Violations of these assumptions and unbalanced data will lead to biases or errors in the estimation of genetic parameters by the ANOVA method. Accommodations for imbalance (differences in number of offspring per family) are possible in ANOVA, but they are complex. Moreover, imbalance is better accounted for by Restricted Maximum Likelihood approaches, which are also more flexible in genetic analyses.

2.3 Estimating heritability for a general pedigree using Restricted Maximum Likelihood

Genetic parameters are nowadays estimated using restricted maximum likelihood (REML) or Bayesian methods. These methods allow for estimation of genetic parameters using phenotypic information for individuals

from a general pedigree (with arbitrary relationships among them). This method allows for unbalanced data and account for genetic relationships within and between families. REML is based on linear mixed model methodology and uses a likelihood approach to estimate genetic parameters.

2.3.1 Linear mixed model:

The linear mixed model contains the observation vector for the trait(s) of interest (y), the 'fixed effects' that explain systematic differences in y, and the 'random effects' which capture unidentified factors affecting y, e.g., random genetic effects and random residual effects.

A matrix formulation of a general model equation is:

$$y = Xb + a + e$$

where

y : is the vector of observed values of the trait,

b : is a vector of factors, collectively known as fixed effects,

a : is a vector of factors known as random effects,

e : is a vector of residual terms, also random,

X : is a known design matrix that relates the elements of b to their corresponding element in y.

The factors (or 'variables') which describe fixed and random effects, may be either continuous or categorical.

Continuous variables have (theoretically) an infinite range of possible values (e.g., body weight in pigs or grain yield in wheat).

Categorical variables fall in distinct categories (e.g., different herds in animals, or different locations in plants). These variables do not describe a gradient of values along a single axis, like height of individuals (values between 0 and "infinity"). Instead, they have distinct classes (or 'levels'), each of which has its own estimated effect.

In addition to continuous or categorical (factor), it is necessary to distinguish between **fixed** and **random** effects in the linear mixed model.

Fixed effect: If the number of levels of a categorical variables is small or limited to a fixed number, and inferences about that factor are going to be limited to that set of levels, and to no others, then its effects is usually fixed. In other words, if a new sample of observations is made (from a new experiment), and the same levels of that factor are in both samples, then the factor is usually fixed. Continuous variables are usually fixed too (but not always).

Random effect: If the number of levels of a categorical variable is large, then that factor may be random. If the inferences about that factor are going to be made for an entire population of levels, and if the levels of the factor are a sample from an infinitely large population, then that factor is usually random. In other words, if a new sample of observations are made (from a new experiment), and the levels are completely different between the two samples, then the factor is usually random.

2.3.2 Expectation and variance of variables in the linear mixed model:

In the statistical model (specified above) the random effects (a and e) and the phenotypes (y) are considered to be random variables which follow a multivariate normal distribution. In general terms the expectations of these random variables are:

$$\begin{aligned} E(y) &= E(Xb) + E(a) + E(e) \\ &= Xb + 0 + 0 \\ &= Xb \end{aligned}$$

and the variance-covariance matrices are:

$$\begin{aligned} Var(a) &= G \\ &= A\sigma_a^2 \\ Var(e) &= R \\ &= I\sigma_e^2 \\ Var(y) &= G + R = V \\ &= A\sigma_a^2 + I\sigma_e^2 \end{aligned}$$

where G, R and V are square matrices of genetic, residual and phenotypic (co)variances among the individuals, respectively.

2.3.3 Genetic relationships amongs individuals

Estimating heritability using REML (similar to the parent-offspring regression and ANOVA method) requires that the phenotypic covariance between related individuals can be expressed by their genetic relationship and the genetic variance (σ_a^2). Related individuals share more alleles and thus resemble each other (have correlated phenotypes, to an extent that depends on additive genetic relationships).

In general, the genetic covariance with a selection candidate (breeding individual) depends on the additive genetic relationship with the candidate. Examples of different types of additive genetic relationships can be found in the table below.

The additive genetic relationship (A_{ij}) between the various sources (j) and the individual itself, i.e. the candidate to be evaluated (i), can be seen in the table below.

Relative	A_{ij}
Self	1.0
Unrelated	0
Mother	0.5
Father	0.5
Grandparent	0.25
Half-sib	0.25
Full-sib	0.5
Cousin	0.0625
Progeny	0.5
Twin(MZ/DZ)	1/0.5

The A matrix expresses the additive genetic relationship among individuals in a population, and is called the **numerator relationship matrix** A. The matrix A is symmetric and its diagonal elements A_{ii} are equal to $1 + F_i$ where F_i is the **coefficient of inbreeding** of individual i. F_i is defined as the probability that two alleles taken at random from individual i are identical by descent. As such, F_i is also the kinship coefficient of its parents (half their genetic relationship).

Each off-diagonal elements (A_{ij}) is the genetic relationship between individuals i and j. Multiplying the matrix A by the additive genetic variance σ_a^2 leads to the covariance among breeding values. Thus if a_i is the breeding value of individual i then

$$var(a_i) = A_{ii}\sigma_a^2 = (1 + F_i)\sigma_a^2 \tag{13}$$

2.3.4 Algorithm to compute the numerator relationship matrix A

The matrix A can be computed using a recursive method. This method is especially suitable for an implementation by a software program. In what follows the recursive method to compute the components of A is described. Initially, individuals in a pedigree are numbered from 1 to n and ordered such that parents precede their progeny. The following rules are then used to compute the components of A.

- If both parents s and d of individual i are known, then
  - the diagonal element A_{ii} corresponds to: $A_{ii} = 1 + F_i = 1 + \frac{1}{2}A_{sd}$ and
  - the off-diagonal element A_{ji} is computed as: $A_{ji} = \frac{1}{2}(A_{js} + A_{jd})$
  - because A is symmetric $A_{ji} = A_{ij}$
- If only one parent s of individual i is known and assumed unrelated to the mate
  - $A_{ii} = 1$
  - $A_{ij} = A_{ji} = \frac{1}{2}(A_{js})$
- If both parents are unknown
  - $A_{ii} = 1$
  - $A_{ij} = A_{ji} = 0$

2.3.4.1 Numeric Example We are given the following pedigree and we want to compute the matrix A using the recursive method described in 2.3.4.

Table 1: Example Pedigree To Compute Additive Genetic Relationship Matrix

Calf	Sire	Dam
3	1	2
4	1	NA
5	4	3
6	5	2

The first step of the computations of A are the numbering and the ordering of all the individuals. This is already done in the pedigree shown in Table 1. The components of A are computed row-by-row starting with A_{11}.

$$
\begin{aligned}
A_{11} &= 1 + F_1 = 1 + 0 = 1 \\
A_{12} &= 0 = A_{21} \\
A_{13} &= \frac{1}{2}(A_{11} + A_{12}) = 0.5 = A_{31} \\
A_{14} &= \frac{1}{2}A_{11} = 0.5 = A_{41} \\
A_{15} &= \frac{1}{2}(A_{14} + A_{13}) = 0.5 = A_{51} \\
A_{16} &= \frac{1}{2}(A_{15} + A_{12}) = 0.25
\end{aligned}
$$

The same computations are also done for all the other components of the matrix A. The final result for the matrix looks as follows

$$A = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 & 0.5 & 0.25 \\ 0 & 1 & 0.5 & 0 & 0.25 & 0.625 \\ 0.5 & 0.5 & 1 & 0.25 & 0.625 & 0.5625 \\ 0.5 & 0 & 0.25 & 1 & 0.625 & 0.3125 \\ 0.5 & 0.25 & 0.625 & 0.625 & 1.125 & 0.6875 \\ 0.25 & 0.625 & 0.5625 & 0.3125 & 0.6875 & 1.125 \end{bmatrix}$$

As a result, we can see from the components of the above shown matrix A that individuals 1 and 2 are not related to each other. Furthermore from the diagonal elements of A, it follows that individuals 5 and 6 are inbred while individuals 1 to 4 are not inbred. Finally, we can see that different types of relationships were included in this data. In comparison, only two types of relationships could exist in regression and ANOVA analyses: unrelated (e.g., A_{ij}=0 between individuals from different families) or not (e.g., A_{ij}=0.5 between individuals from the same full-sib family).

2.3.5 Likelihood approach for estimating variance components and heritability

Restricted Maximum Likelihood is a method that is used to estimate the parameters (i.e. variance components σ_a^2 and σ_e^2) in the linear mixed model specified above. The general principle used in maximum likelihood methods is to find the set of parameters which maximizes the **likelihood** of the data, i.e., the probability of observations given the model and its parameter estimates: $p(y|\hat{b}, \hat{\sigma}_a^2, \hat{\sigma}_e^2)$.

It is useful to recall that the likelihood $L(\theta|y)$ may be any function of the parameters (θ) that is proportional to $p(y|\theta)$. Maximizing $L(\theta|y)$ leads to obtaining the most likely value of θ ($\hat{\theta}$) given the data y. Usually the likelihood is expressed in terms of its logarithm $l(\theta|\mathbf{y})$ as it makes the algebra easier.

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.

The REML method was developed by Patterson and Thompson [1971] as an improvement of the standard Maximum Likelihood (ML). The ML method was originally proposed by Fisher [1922], and was introduced to variance component estimation by Hartley and Rao [1967]. ML assumes that fixed effects are known without error which is in most cases false and, as consequence, it produces biased estimates of variance components (usually, the residual variance is biased downward). As a solution to this problem, REML estimators maximize only the part of the likelihood which does not depend on the fixed effects, and REML, by itself, does not estimate the fixed effects. This entails that when comparing multiple models by their REML likelihoods, those must contain the same fixed effects, and they might differ only by their random effects (e.g., with or without breeding value).

There are no simple one-step solutions for estimating the variance components based on REML [Lynch and Walsh, 1998]. Instead, we infer the partial derivatives of the likelihoods with respect to the variance components. The solutions to these involve the inverse of the variance-covariance matrix, which themselves includes the variance components, so the variance components estimates are non-linear functions of the variance components. It is therefore necessary to apply iterative methods to obtain the estimates.

From the REML estimate of the variance components, the heritability can easily be computed by

$$\hat{h}^2 = \hat{\sigma}_a^2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2) \tag{14}$$

where the "^" refers to estimators.

2.3.6 Advantages of using REML for estimating genetic parameters

Although REML does not produce unbiased estimates, it is still the method of choice due to the fact that this source of bias is also present but higher in ML estimates [Lynch and Walsh, 1998].

REML 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).

REML can account for selection when the complete mixed model is used with all genetic relationships and all data used for selection is included (Sorensen and Kennedy, 1984; Van der Werf and De Boer, 1990).

There is obviously an advantage in using (RE)ML methods that are more flexible in handling animal and plant breeding data on several (overlapping) generations (and possibly several random effects). However, the use of such methods are "dangerous" in the sense we no longer need to think explicitly about the data structure. For example, to estimate additive genetic variance, we need to have a data set that contains a certain family structure which allows us to separate differences between families from differences within families. Or in other words, we need to differentiate genetic and residual effects, so the structure due to genetic relationships must be different from the structure due to residual effects (i.e., the G and R matrices must be different enough). In comparison 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).

Early REML applications were generally limited to models largely equivalent to those in corresponding ANOVA analyses, considering one random effect only and estimating genetic variances from paternal half sib covariances (so-called sire model in animal breeding). Today, heritability can be estimated based on genetic relationships, inferred from general pedigrees or estimated from genetic markers. Linear mixed models are also used in genetic evaluation, allowing information on all known relationships between individuals to be incorporated simultaneously in the analysis. Linear mixed models can include additional effects to describe the data more accurately: maternal, permanent environmental, cytoplasmic or dominance effects and QTL effects. These effects may be fitted as additional random effects.

3 When to estimate variance components?

In general, the estimation of variance components has to be based on a sufficient amount of data. Depending on the data structure and measurements, estimations can be based on hundreds (in selection experiments) or more than 10,000 observations (in field recorded data). Importantly, in cases where the data set is small, the information from the literature may yield more accurate estimates of variance components. In general, we have to estimate variance without external information if we study a new trait, for which no prior parameter estimates are available, or a different sample: variances and covariances might have changed over time, or due to various evolutionary forces (genetic drift, selection, migration, or mutation).

Generally, it is assumed that variances and covariances, and especially their ratio (like heritability, correlation) do not rapidly change over time. However, it is well known that the genetic variance changes as a consequence of selection or genetic drift. Changes are expected, especially when generation intervals are short, selection intensity is high, or the trait under selection is determined by few causal genes with large effects. Moreover, the circumstances under which measurements are taken can change. If measurement conditions are better controlled, and getting more uniform over time, the environmental variance decreases, and consequently the heritability increases. Finally, the biological basis of a trait may change from one environment to another; for example, feed intake under limited feeding is not really the same as feed intake under ad-lib feeding. In conclusion, there are sufficient reasons for regular estimation of (co-)variance components.

References

R. A. Fisher. On the Mathematical Foundations of Theoretical Statistics. *Philos. Trans. R. Soc. London*, 222(594-604):309–368, 1922. ISSN 1364-503X. doi: 10.1098/rsta.1922.0009. URL http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.1922.0009.

H. O. Hartley and J. N. K. Rao. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1):93–108, 1967. URL http://biomet.oxfordjournals.org/content/54/1-2/93.short.

Michael Lynch and Bruce Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, USA, 1998. ISBN 0-87893-481-2.

H. D. Patterson and R. Thompson. Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58(3):545–554, December 1971. ISSN 00063444. doi: 10.2307/2334389. URL http://www.jstor.org/stable/2334389.