Estimation of Variance Components

Why?

• Better understanding of genetic mechanism
• Needed for prediction of breeding values
  – Selection Index / BLUP
• Needed for optimization of breeding programs
  and prediction of response
## Variance Components

- Add. Genetic
- Residual
- Maternal
- Permanent Environment
  - Litter,
  - Dominance,
  - Herd
- Covariances

## Parameters

- Heritability
- Maternal Heritability
- Repeatability
  - Common full-sib comp’t (“c²”)
- Correlations
  - Phenotypic/ Genetic
When to (re) estimate variance components?

- New trait
- (co)variances change over time due to environmental and/or genetic change
  - Selection
  - Upgrading
  - Trait definition
Variance and Covariance

- **Variance**: measure of differences (extent of)
- **Covariance**: measure of ‘differences in common’
- Between individuals/ between traits

<table>
<thead>
<tr>
<th>sire</th>
<th>Individual Values</th>
<th>Var between Families</th>
<th>Var within Fam</th>
<th>Types of family resemblance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>None</td>
<td>Moderate</td>
<td>Full</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>None</td>
<td>Moderate</td>
<td>None</td>
<td>Large</td>
</tr>
<tr>
<td></td>
<td>Large</td>
<td>Moderate</td>
<td>Large</td>
<td>None</td>
</tr>
</tbody>
</table>

None
Moderate
Large
Relating variance components to underlying effects - give it a meaning!

- Variance between groups = covariance within groups!

- Variance between HS families
  = Covariance among half sibs
  = $\frac{1}{4} V_A$
  They share 25% of their genes!

Variance within HS families
= Residual Variance
= $V_P - \frac{1}{4} V_A$
= $\frac{3}{4} V_A + V_E + V_D$
Relating variance components to underlying effects - give it a meaning!

- Variance between groups = covariance within groups!

- Variance between FS families
  \[ = \text{Covariance among full sibs} = \frac{1}{2} V_A + V_{ec} + \frac{1}{4} V_D \]
  They share 50% of their genes!

Variance within FS families

\[ = \text{Residual Variance} = V_P - \frac{1}{2} V_A - V_{ec} - \frac{1}{4} V_D \]
\[ = \frac{1}{2} V_A + V_{EW} + \frac{3}{4} V_D \]
Analyses of Variance

Principle

- Detect the importance of different sources of effects
- Importance is determined by its contribution to variation
- Variation if derived from sums of squares and df
Analyses of Variance

Example

\[ y_i = \mu + e_i \]

\( \mu = \text{mean (fixed)} \)

\( e_i = \text{residual is random} \)

(cause variation)

\[ \text{Var}(y) = \sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-1) \]

Same as

Calculating sum of squares

\[ \sum_{i=1}^{n} e_i^2 = SSE \]

Equal SS to its expectation

\[ E(SSE) = (n-1) \sigma_e^2 \]
Analyses of Variance

**Example Data**  \( y = [8, 9, 11, 12] \)

**Model:**  \( y_i = \mu + e_i \)

**Sums of squares:**
- **Total:** \( 8^2 + 9^2 + 11^2 + 12^2 = 410 \)
- **Mean:** \( 4 \times 10^2 = 400 \)
- **Residual SS:** \( 10 \)
  \( (= (-2)^2 + (-1)^2 + 1^2 + 2^2) \)
Analyses of Variance

**Example Data**  \( y = [8, 9, 11, 12] \)  \( a: i = 1 \ 1 \ 2 \ 2 \)

Model: \( y_i = \mu + a_i + e_{ij} \)

Estimates: \( \mu = 10 \)  \( a_i = -1.5 \)  \( a_2 = +1.5 \)

<table>
<thead>
<tr>
<th>Sum of squares</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed: 8 ( ) 9 ( ) 11 ( ) 12 ( ) 410 ( ) ( SS_{Total} )</td>
</tr>
<tr>
<td>Mean: 10 ( ) 10 ( ) 10 ( ) 10 ( ) 400 ( ) ( SS_{Mean} )</td>
</tr>
<tr>
<td>a-effect: -1.5 ( ) -1.5 ( ) +1.5 ( ) +1.5 ( ) 9 ( ) ( SSA )</td>
</tr>
<tr>
<td>Residual: -0.5 ( ) +0.5 ( ) -0.5 ( ) +0.5 ( ) 1 ( ) ( SSE )</td>
</tr>
</tbody>
</table>
## ANOVA-Table

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>400</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-effect</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td>$\sigma_e^2 + 2\sigma_a^2$</td>
</tr>
<tr>
<td>Residual</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>$\sigma_e^2$</td>
</tr>
<tr>
<td>Total</td>
<td>410</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: “a-effect” is a classification of data: e.g. according to sires (half sib groups). It relates to variance between groups.

“residual” relates to variance within groups.
Group (e.g. sire) differences relate to variance between groups.

“residual” differences relate to variance within groups.

\[ \sigma_a^2 = 4.25 \quad \sigma_a = 2.1 \]

\[ \sigma_e^2 = 0.5 \quad \sigma_e = 0.7 \]
Summarizing the procedure

Modeling (general)

- Data = fixed effects + random effects
  - E(y) = fixed effects means
  - Var(y) = variance due to random effects

Interpretation

- Statistically:
  - Need sufficient data
  - Need to think about data structure
  - Sampling conditions need to be fulfilled (random?)

- Genetically
  - Translating the components into meaningful parameters
    - (e.g. sire variance = \( \frac{1}{4} V_A \))
$h^2$ estimates from half-sib families
Depend on number in each family (higher number $\rightarrow$ more accuracy)
$h^2$ estimates from half-sib families

Depend on the number of sires (sire families) in the sample (higher number $\Rightarrow$ more accuracy)

- Small sample of sire, bad estimate of sire variance
- Large sample of sire, better estimate of sire variance
### Accuracy: SE of heritability estimate

<table>
<thead>
<tr>
<th>Nr. of records</th>
<th>True heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>0.18</td>
</tr>
<tr>
<td>500</td>
<td>0.08</td>
</tr>
<tr>
<td>1000</td>
<td>0.06</td>
</tr>
<tr>
<td>5000</td>
<td>0.03</td>
</tr>
<tr>
<td>50000</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Effect of data size on SE of heritability
Effect of data size on SE of heritability

![Graph showing the effect of data size on SE of heritability.](chart.png)
Effect of progeny group size on SE of heritability

![Graph showing the relationship between the number of progeny per sire and the standard error (SE) of heritability. The graph includes two lines representing different values of heritability (0.1 and 0.3), with the SE decreasing as the number of progeny increases.]
The following slides are not GENE422 material (reference only)
Methods for variance component estimation

• **ANOVA** - balanced data
• **ANOVA** – unbalanced data
  – Henderson’s methods (SAS etc)

• **Likelihood methods**
  – Maximum Likelihood
  – Restricted maximum Likelihood (REML)

• **Bayesian Methods**
  – Gibbs Sampling
ANOVA-Table for balanced data

<table>
<thead>
<tr>
<th></th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>EMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>400</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-effect</td>
<td>9</td>
<td>1</td>
<td>9</td>
<td>$\sigma_e^2 + n\sigma_a^2$</td>
</tr>
<tr>
<td>Residual</td>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>$\sigma_e^2$</td>
</tr>
<tr>
<td>Total</td>
<td>410</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A-effect refers to differences ‘Between groups’
Residual refers to differences ‘Within groups’
ANOVA in Unbalanced data

Same idea as balanced (previous) but use a weighted number for “n” in: \( \text{EMS}_A = \sigma_e^2 + n\sigma_a^2 \)

Need matrix notation to work out SS and EMS (as in linear models)

Standard method in computer programs such as SAS, Harvey, SPSS etc.

Most general of those is called the “Henderson III method”
Likelihood methods

Each observation has a probability density, determined by its
• distribution
• expected value (e.g. mean) ‘location parameters’
• variance ‘dispersion parameters’

E.g. y with normal distribution, mean $\mu$ and variance $\sigma^2$

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

This is a Probability Density Function (PDF) for the observation
It gives the probability of the observation, given the parameters $\mu$ and $\sigma^2$
But we turn this around and get the likelihood of the parameters given y
Likelihood methods

We can multiply these probability values over the whole data, and include the fact that some of the observations may be related, i.e. we have a *joint distribution*. 

Data vector $y$ with exp. means $E(y) = Xb$ and $\text{var}(y) = V$

The log of the likelihood is:

$$L(b, V \mid X, y) = -\frac{1}{2} N \log(2\pi) - \frac{1}{2} \log(|V|) - \frac{1}{2} (y - Xb)' V^{-1} (y - Xb)$$

The expression gives the likelihood of the parameters $(b, V)$ given data $(X, y)$ in the right-hand side. It is a restricted (or residual) likelihood, after fitting the fixed effects.

*first two terms are expectations* \(\text{the last term is a (residual) sum of squares} \)*
Restricted Maximum Likelihood

- Correct all data first for all fixed effects
- Find the maximum likelihood (solution for variance components) after these corrections
- Usually an iterative procedure is used to solve the problem
- Starting values (for the parameters) are needed to get going
An example of a REML algorithm (EM-algorithm, for illustration only)

1. Solve mixed model equations using a prior value for the variance components (ratio)

\[
\begin{bmatrix}
X'X & XZ \\
Z'X & Z'Z + \lambda A^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{b} \\
\hat{a}
\end{bmatrix} =
\begin{bmatrix}
X'y \\
Z'Y
\end{bmatrix}
\]

2. Solve variance components from the MME-solutions

\[
\sigma_a^2 = \left[ \hat{a}' A^{-1} a + tr(A^{-1} C) \sigma_e^2 \right] / q
\]

\[
\sigma_e^2 = \left[ y'y - \hat{b}'X'y - \hat{a}'Z'y \right] / (N - r(X))
\]

Use a new \( \lambda (= \sigma_e^2 / \sigma_a^2) \) and iterate between 1 and 2
Why is REML better than ANOVA from SAS?

- It is by definition more accurate
- Uses full mixed model equations, so can utilize all animal relationships (animal model)
- Therefore, it has many properties as BLUP, e.g. it accounts for selection
- It allows more complicated mixed models (maternal effects, multiple traits etc) as with BLUP
Further notes on REML procedure

• If using an animal model, heritability is estimated from naturally combining
  – information between families (HS/FS)
  – information from parent-offspring regression

• The method and model are very flexible, but it can be hard to evaluate the estimates based on the data and the data structure
  – e.g. Is there a good family structure?
Evaluating the quality of the parameter estimates

• **Accuracy**
  
  – Look at SE of estimates (although these are approximated!)
  
  – Evaluate effect of number of records, and structure (nr. of groups, usually HS groups, vs nr. per group)

• **Unbiasedness**
  
  – From the data, and the possible effects, evaluate whether there was no bias from selection, or from confounding effects, e.g. sires confounded with herd or management group
Example: Analysis of weaning weight for White Suffolk data on 9700 animals, 15,000 in pedigree

Comparison of including or not including the correlation between direct genetic (A) and maternal (M) effects and the effect of ignoring maternal effects on estimating $h^2$

<table>
<thead>
<tr>
<th></th>
<th>Correlation A-M included</th>
<th>No correlation</th>
<th>No maternal effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>PhenVar</td>
<td>23.45</td>
<td>23.26</td>
<td>23.94</td>
</tr>
<tr>
<td>Heritability</td>
<td>0.25 0.04</td>
<td>0.19 0.03</td>
<td>0.44 0.03</td>
</tr>
<tr>
<td>Maternal Heritab.</td>
<td>0.28 0.04</td>
<td>0.18 0.02</td>
<td></td>
</tr>
<tr>
<td>Correl. direct-matern.</td>
<td>-0.44 0.10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: Analysis of weaning weight for White Suffolk data on 9700 animals, 15,000 in pedigree

The effect of ignoring or including a permanent environmental effect (PE) of dams

<table>
<thead>
<tr>
<th></th>
<th>with PE</th>
<th>without PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenotypic Var.</td>
<td>23.06</td>
<td>23.45</td>
</tr>
<tr>
<td>Heritability (direct)</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>Maternal heritability</td>
<td>0.13</td>
<td>0.28</td>
</tr>
<tr>
<td>Corr Mat-Direct</td>
<td>-0.50</td>
<td>-0.44</td>
</tr>
<tr>
<td>Permanent Env. Ewe</td>
<td>0.12</td>
<td>0.10</td>
</tr>
</tbody>
</table>