Analysis of Half-sib Progeny Test Data of Forest Trees

Fikret Isik

North Carolina State University, Raleigh, USA

You are welcome to use these notes. They are provided as is. Send errors or suggestions/comments to (fisik@ncsu.edu).

If you would like to have more details and solutions using the ASReml software, a textbook I co-authored is available. Isik F, Holland J, Maltecca C. *Genetic data analysis for plant and animal breeding*. Springer International Publishing; 2017 Sep 9.

TABLE of CONTENTS

4.1 Introduction

- **Genetic Materials**: Wind-pollinated parents were tested in field experiments for selection. The progeny from a parent are assumed to be related as half-sibs and constitute as a "family". In another words, we only know one common parent of the progeny, and we assume the other parent is different for each sibling.
- **Experimental Field Designs**: We will give examples for two field designs that are commonly used in agriculture and forestry; Randomized Complete Block Design with single-tree plots and Randomized Complete Block Design with multiple-tree plots (e.g., row plots).
- **Blocking** is an experimental unit that is used to diminish the influence of environmental variation within a test site. Ideally, each block should be homogeneous with no apparent variation within a block. The experimental units (family plots) are randomly assigned within each block.

4.2 Single-Tree-Plot (STP) Design

- A progeny test of *Pinus taeda* was established to predict breeding values of 24 parents (half-sib families) for selection. A randomized complete block design was used in the field. The experiment was replicated at five locations (sites) in North Carolina, USA. There were 15 blocks at each site. Each family had only one tree per block and thus 75 in total. Height of trees was measured after five growing seasons in the field.

4.3 The Statistical Model

[1] $$y_{ijkl} = \mu + S_i + B(S)_{j(i)} + F_k + SF_{ik} + E_{ijkl}$$

where

y_{ijkl} is the *l*th observation of the *j*th block within the *i*th site for the *k*th family;
μ is the overall mean;
S_i is the fixed *i*th site effect (i=1,…,5);
$B(S)_{j(i)}$ is the fixed *j*th block effect within the *i*th site (j=1,…,15);
F_k is the random general combining ability of the *k*th family, normally and independently distributed ~NID (0, $\boldsymbol{\sigma^2_F}$), (k=1,…,24);
SF_{ik} is the random *k*th family by *i*th site interaction effect ~NID (0, $\boldsymbol{\sigma^2_{SF}}$) and
E_{ijkl} is the error term ~NID (0, $\boldsymbol{\sigma^2_E}$).

If we had multiple-tree plot design, then, in addition to all above terms, we would have also a random family by block interaction [FB(S)] term in the model. This is a plot-to-plot error. See Chapter 12 for the details and assumptions behind the linear mixed model.

For this specific example, here is why the factors and their interactions in the model are considered fixed or random:

1. We would like to explain the sources of variation for height in the experiment. How much of the phenotypic variation is due to genetics and how much is due to environment? To answer this question, we test families, **a random sample** of a breeding population. Since families represent a random sample of population (randomly selected), the **family (F), family by site interaction (FS), and within family variation or error (E) are considered random**. In another words, any term in the linear model having a family subscript (k) is considered random.

2. The second reason we define Family effects as random is that, we would like to draw conclusions about the breeding population, not about those families in the experiment. Heritability, additive genetic variance and phenotypic variance are all parameters that refer to a population. If we sample the same population and randomly select another group of families, we would get slightly different results.

3. We have no inherent interest in the blocks or sites used in the experiment. We are not interested in how much total variation is explained by sites or blocks within sites. Instead, we are using sites and blocks to control environmental variation. We can always choose the same sites and set up the same blocks. Since they are not randomly selected, their effects are not random, so that **sites (S) and blocks** [B(S)] **are considered fixed** (see Chapter 12 for more details).

4.4 Implementation with SAS MIXED Procedure

Consider the data set described in section 4.2. The response variable is height (in feet) of trees measured at age five. There are 1632 trees in the experiment. Here, only the first 7 observations, all from the same family at one site, are shown.

Obs	site	block	family	height
1	A	1	F1378	17.8
2	A	2	F1378	20.9
3	A	3	F1378	21.4
4	A	4	F1378	20.5
5	A	6	F1378	18.4
6	A	7	F1378	19.8
7	A	8	F1378	20.0

The linear mixed model given under the heading 4.3 for height data can be analyzed using the analysis of variance (ANOVA). Let's look at an ANOVA results for the data.

BOX 1: Analysis of variance and expected mean squares

Using analysis of variance and expected mean squares (EMS) equations, we can calculate variance components.

SOURCE	DF	SS	MS	Expected Mean Squares (EMS)
site	s-1	SSs	MSs	-
block(site)	s(b-1)	SSb	MSb	-
family	f-1	SSf	MSf	σ^2_E + **bn** σ^2_{SF} + **sbn** σ^2_F
site*family	(s-1)(f-1)	SSsf	MSsf	σ^2_E + **bn** σ^2_{SF}
Residual	remaining	SSe	MSe	σ^2_E
TOTAL	sbfn-1			

Mean squares of family effect in the table is composed of Residual (σ^2_E), site*family (σ^2_{SF}) interaction and family (σ^2_F) effects as shown by the Expected Mean Squares. The coefficients of the EMS are; **b**= number of blocks, **n**= number of trees per family per block and **s**= number of sites. If there were no missing trees in the experiment, then the coefficient of site*family variance would be **bn** =15 (15 blocks*1 tree per family). Similarly, the coefficient of family variance would be **sbn**=75 (5 sites x 15 blocks x 1 tree).

SAS MIXED procedure can be used to obtain the expected mean squares and their coefficients. All you need to do is to add METHOD=TYPE3 to the SAS MIXED ***code-1***.

One of the output tables for height data from the MIXED procedure is given below:

The Mixed Procedure
Type 3 Analysis of Variance

Source	DF	MS	Expected Mean Square
site	4	2103	Var(Res) + 13.3 Var(SF) + Q(S)
block(site)	70	31	Var(Res) + Q(B(S))
family	23	30	Var(Res) + 13.3 Var(SF) + 66.7 Var(F)
site*family	92	4.7	Var(Res) + 13.5 Var(SF)
Residual	1442	4.6	Var(Res)

Notice that the coefficients of EMS are slightly different from fully balanced design. For example, the coefficient for the *family* effect now is 66.7 instead of 75.

What is the family variance component? Family EMS equation is = Var(Res) + 13.3Var(SF) + 66.7Var(F). We need to subtract the Residual MS and site*family MS from family MS and divide the remaining with the coefficient 66.7 to get family variance component. The Var(F) is

σ^2_F = (MSf - MSsf) / 66.7 = (30 – 4.7) / 66.7 = **0.379**

ANOVA method is not the best approach to calculate variance components mainly because the data are ***not*** always balanced. Instead, likelihood methods are standard algorithm in software for mixed models.

Code 1: SAS MIXED code for STP field design

This is a simple SAS MIXED code to analyze the wind-pollinated data described above. In the code below, the UPPERCASE words are SAS options or statements. The lowercase words are the ones we type in the SAS codes.

```
PROC MIXED DATA=op.pine_halfsib METHOD=REML;
   CLASS site block family;
   MODEL height =site block(site);
   RANDOM family site*family;
RUN;
```

Explanation of the code:

1. The name of the SAS data set to be analyzed is **op.pine_halfsib**. The data file name has two parts. The first part before the dot (hbook) is the library name (e.g., a folder) where the data set is located. The second name (.op) is the actual name of the data set that we are referencing.
2. METHOD: This is to define a method for calculation of variance components. If you do not type METHOD=REML, the MIXED procedure will still use REML (restricted maximum likelihood) to calculate variance components because it is the default method in the SAS MIXED procedure. You may specify a different method such as METHOD=TYPE3 to calculate variance components based on ANOVA. TYPE3 is an analysis of variance method which equates the Mean Squares to the Expected Mean Squares Equations and solve for variance components. There are many other statistical methods to calculate variance components.
3. CLASS statement: We list the factors (independent variables) after the CLASS statement. As explained in the section 4.3, **site**, **block** and **famil**y are independent or classification (CLASS) variables in the model.

4. MODEL statement: The response variable height is given after the MODEL statement. You can analyze only one trait at a time with the MIXED procedure. The **fixed effects** terms in the model (**site** and **block(site)**) are listed after the '=' sign. Block effect is nested within sites as shown by putting **site** in the parenthesis right after the **block** term. There is no need to list the intercept. The intercept (μ) is included in the model by default.
5. RANDOM statement: **family** and **family*site are random** terms and they are listed after the RANDOM statement. The **interaction** is shown by putting the star sign (*) in between two or more terms, such as site*family.

Output 1:

Code-1 produces a lot of output (tables) by default. The list of tables produced is given below. We will explain some important tables (1, 2, 7, 9) here marked with bold face fonts. For the rest, you should look at the MIXED procedure syntax in the help system.

1. **Model Information**
2. **Class Level Information**
3. Dimensions
4. Number of Observations
5. Iteration History
6. **Covariance Parameter Estimates**
7. Fit Statistics
8. **Type 3 Tests of Fixed Effects**

Model Information

Data Set	OP.PINE_HALFSIB
Dependent Variable	height
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

- This table (Model Information) is about statistical methods used to analyze data. The name of the data set (OP.PINE_HALFSIB) analyzed, the dependent variable (height) are listed. The method used to calculate the variance components is the default REML (A maximum likelihood-based method).

Class Level Information

Class	Levels	Values
site	5	A B C E F
block	15	1 10 11 12 13 14 15 2 3 4 5 6

```
                                7 8 9
family          24     F1378 F1801 F1805
                       F1853 F1806 F1013
                       F1033 F1051 F1080
                       F1085 F1086 F1070
                       F1806 F1805 F1002
                       F1805 F1003 F1007
                       F1009 F1010 F1003
                       F1005 F1804 F1805
```

- Using this table, you can check your levels of factors (independent variables): Are there 5 sites, 15 blocks per site, and 24 families in the data as expected? If a family is typed as 'f1378' instead of 'F1378', SAS thinks that they are different families. Make sure that there are no such errors in the lists.

Covariance Parameter Estimates

Cov Parm	Estimate
family	0.378
site*family	0.0065
Residual	4.645

- The 'Estimate' column lists the *observational variance components*. The family variance component is $\sigma^2_F = 0.378$. The site by family interaction variance component is $\sigma^2_{SF} = 0.378$. The error or residual variance component is $\sigma^2_E = 4.645$. We can use observed variance components and genetic covariances between relatives to calculate *causal variance components*, such as additive genetic variance (BOX 2).

BOX 2: Observed versus causal variance components and resemblance between half-sibs

Variance components obtained from data are *observational variance components*. We simply breakdown the total phenotypic variance into groups, such as between group component (*family),* and within group component *(Residual*). Using the observational variance components and genetic covariances among relatives, we can calculate the *causal variance components*. Additive genetic variance is the *causal variance component* arises from additive effects of genes that cause resemblance between relatives. Falconer and MacKay (1996) denote observational variance component by the symbol 'σ^2' and the causal components by the symbol 'V'.

We know that when we have half-sibs (one parent is shared, the other parents is different), variance explained by family effect is 1/4 of the additive genetics variance. Where does this relationship come from? We will give an example similar to a work example given by Bruce Walsh in his lecture notes.

Let's say parent P has the overall breeding value of A=(α1, α2). O_1 and O_2 are half-sibs with genetic values of Go1=(α1+α3), and Go2=(α1+α4). They share only one allele (α1) which comes from the mother tree (identical by descent or IBD). What is covariance of genetic values between siblings O1 and O2?

Cov(Go1, Go2) = Cov[(α1+ α3, α1+α4)]
= **Cov(α1, α1)** + Cov(α1, α4) + Cov(α3, α4)

As a rule,
when *x* and *y* are unrelated (*x ≠ y, i.e., not IBD*) then, Cov(x,y) = 0,
When *x* and *y* are related (*x = y, i.e., IBD*) then, Cov(x,y) = Var(A)/2

Cov(α1, α4) = 0 because α1 and α4 are not IDB. Similarly, Cov(α3, α4)=0.
Then, Cov(Go1, Go2) = **Cov(α1, α1)** + 0 + 0

The covariance of α1 with itself is the variance of α1.
Cov(Go1, Go2) = **Cov(α1, α1)**
= Var(α1) = Var(A)/2

This is the covariance of genetic values between two half-sibs.

The degree of resemblance is measured using the coefficient of coancestry (Θxy), which is simply the probability of an allele in offspring being IBD. This probability for half-sibs O1 and O2 is 1/2. When two trees have one allele IBD, the contribution to the genetic covariance is Var(A)/2 as shown above. Thus, the genetic covariance of half-sibs is one quarter of the additive genetic variance.

COV_{HS} = Pr(1 allele IBD) x contribution Var(A)/2) = 1/2 x Var(A)/2 = 1/4 $\boldsymbol{\sigma^2_A}$

Assuming only the additive genetic effects (no dominance or epistatic interactions), family variance component from output 1 is 1/4 of the additive genetic variance:

Var(A) = 4 σ^2_F = 4*0.378 = 1.511

See Lynch and Walsh (1998, Chapter 7), Isik-Holland-Maltecca (217), and Falconer and MacKay (1996, Chapter 9) for more details about the genetic covariances among relatives.

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
site	4	92	452.01	<.0001
block(site)	70	1442	6.72	<.0001

- The F-tests of fixed effects factors (site and blocks within site) are given. Sites are significantly different from each other for height. Similarly, blocks within sites are also significantly different for height.

Code 2: MIXED procedure to obtain covariances of variance components

Code-1 produced many tables but additional output is needed so we can calculate standard errors of variance components as well as standard error of heritability. Since variance components are estimates, we would like to know their precision too. An option called COVTEST tells MIXED procedure to produce standard errors of variance components. Another table called ASYCOV produces the variances and covariances of the variance components. The new options are given in bold fonts in the PROC MIXED code as shown below.

```
PROC MIXED DATA=hbook.Vpine ASYCOV COVTEST;
   CLASS site block family ;
   MODEL Height =site block(site) ;
   RANDOM family site*family ;
  ODS OUTPUT COVPARMS =_varcomp  ASYCOV =_cov ;
RUN;
```

Explanation of the code:

1. The ASYCOV option produces the variances of variance components (diagonal elements) and the covariances (off diagonal elements) between them. If you do not list these options after the PROC MIXED statement, these tables will not be produced as shown in code-1. We need variance of variance components and covariances between variance components to calculate standard error of heritability or standard error of any other function of variance components.
2. COVTEST produces asymptotic standard errors for the variance components. A Z value is calculated for each component by simply dividing the estimate by its standard error. In addition, a Chi-square test of the variance components (Pr Z) is produced.
3. ODS OUTPUT creates SAS output files. The name before the '=' is the table name of the SAS Output Delivery System (ODS) (e.g., COVPARMS). The name after the '=' (e.g., *_cov*) is a name we provide for the table. You may give any name instead of *_varcomp* or

_cov. As a rule for this book, SAS options and statements are given in uppercase letters, where the name we give are lowercase and starts with underscore. You may use a different name for the output table, such as *_varcomp_height*. If you do not use the ODS statement, you will see the tables in the output window but they will not be created as SAS data sets in the work library.

4. COVPARMS=*_ varcomp* requests that the table of the variance components, and their standard errors, approximate Z test values created by the COVTEST option be saved. By default, the SAS MIXED procedure uses the REML (Restricted Expected Maximum Likelihood) method to produce variance components. If you do not use ODS OUTPUT statement and COVPARMS =*_varcomp*, then you can not save the table of variance components, nor their standard errors and Z test scores as a SAS data set.
5. ASYCOV=*_cov* requests that the table of covariances of the variance components be saved in the Work library of SAS. It is the matrix of the variances of variance components (in the diagonal) and the covariances between the variance components (the off-diagonal numbers). We need this table to estimate standard errors of the functions of the variance components (i.e., heritability).

Output 2:

Since we explained most of the MIXED procedure output in ***Code-1***, we are only focusing on the COVPARMS, COVTEST and ASYCOV matrices in this section.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr Z
family	0.3779	0.1328	2.84	0.0022
site*family	0.006511	0.05238	0.12	0.4505
Residual	4.6448	0.1729	26.87	<.0001

- The "Covariance Parameter Estimates" table contains the estimates (variance components), their standard errors, the Z value (Estimate/Stderr) and the Wald test probability value (Pr Z).
- The effects are labeled in the "Cov Parm" column. The Estimates are observed variance components displayed in the Estimate column (e.g family variance components is σ^2_F= 0.378).
- Requesting the COVTEST option in the PROC MIXED statement produced the Standard Error, Z Value, and Pr Z columns. The Standard Error column contains the approximate standard errors of the covariance parameter estimates (variance components). The Z Value column is the ratio of variance component and its approximate standard error. The Pr Z column is the Wald tests of the variance components. Wald tests are Chi-square statistics that

test the null hypothesis that a parameter is 0; in other words, the corresponding variable has no effect given that the other variables are in the model. The Wald are unreliable in small samples.

Asymptotic Covariance Matrix of Estimates

Row	Cov Parm	CovP1	CovP2	CovP3
1	family	**0.01765**	-0.00044	-0.00009
2	site*family	-0.00044	**0.002744**	-0.00215
3	Residual	-0.00009	-0.00215	**0.02988**

- The values in the diagonal of the table are variances of variance components. For example, **0.01765** is the variance of family variance component.
- The covariance [Cov(σ^2_F, σ^2_{SF})] between family variance component and Family x Site interaction variance component is -0.00044.
- We can use these covariances and variances of variance components to calculate standard errors of any function.

BOX 3: Components of variance and their standard errors

The total variance for a given trait is phenotypic variance. Phenotypic variance (V_P) is composed of genetic (V_G) and environmental (V_E) variances.
$V_P = V_G + V_E$

Genetic variance is contributed by the additive (V_A), dominance (V_D) and epistatic (V_I) interactions of genetic effects.

$V_P = V_A + V_D + V_I + V_E$

The main objective of the progeny tests is to partition observed variance into genetics and environmental components. Additive genetic variance, phenotypic variance, heritability and genetic gains are calculated based on variance components.

Here is an example on estimation of additive and phenotypic variances and their standard errors from the output-2.

Causal variances and their standard errors
The family variance (σ^2_F) is an estimate so it has an error (variance) associated with it. Standard error of family variance component is

$SE(\sigma^2_F) = \sqrt{Var(\sigma^2_F)} = \sqrt{0.01765} = 0.133.$

The variance of family variance [$Var(\sigma^2_F)$] comes from the output of SAS MIXED procedure. The table is called 'Asymptotic Covariance Matrix of Estimates'. See an example in Code-2.

We need the variance of family variance component to calculate standard error of additive genetic variance or standard error of heritability.

$\sigma^2_A = 4\sigma^2_F$	Additive genetic variance
$Var(\sigma^2_A) = Var(4\sigma^2_F) = 16Var(\sigma^2_F)$	Variance of additive genetic variance
$SE(\sigma^2_A) = \sqrt{16Var(\sigma^2_F)} = 4\sqrt{Var(\sigma^2_F)}$	Standard error of additive genetic variance

For example, the standard error of additive genetic variance from Output-2 is

$SE(\sigma^2_A) = \sqrt{Var(4\sigma^2_F)} = \sqrt{16Var(0.01765)} = = 0.53$

Phenotypic variance is the sum of the observational components of variance that are included in the Expected Mean Square for the family effect:

$\boldsymbol{\sigma^2_P} = \sigma^2_F + \sigma^2_{SF} + \sigma^2_E$

$= 0.378 + 0.0065 + 4.645 = 5.029$

Variance of phenotypic variance $Var(\sigma^2_P)$:

$\mathbf{Var(\sigma^2_P)} = Var(\sigma^2_F + \sigma^2_{SF} + \sigma^2_E)$

$= Var(\sigma^2_F) + Var(\sigma^2_{SF}) + Var(\sigma^2_E) + 2\,[Cov(\sigma^2_F,\sigma^2_{SF}) + Cov(\sigma^2_F,\sigma^2_E) + Cov(\sigma^2_{SF},\sigma^2_E)]$

The variance of a sum ($\sigma^2_F + \sigma^2_{SF} + \sigma^2_E$) is the variances of each term in the equation, plus 2 times of their covariances.

Using the Asymptotic Covariance Matrix of Estimates table in the Output-2, the variance of phenotypic variance is

= (0.01765+0.002744+0.02988) - 2*(0.00044+0.00009+0.00215) = 0.0449

Standard error of phenotypic variance $SE(\sigma^2_P)$ is simply the square root of the variance.

$\mathbf{SE(\sigma^2_P)} = \sqrt{Var(\sigma^2_P)} = \sqrt{0.0449} = 0.212$

4.5 Using SAS/IML to Estimate Functions of Variance Components

For most of functions of variance components, such as narrow-sense heritability, you may use a spread sheet to do the calculations. However, for more complex calculations or repeated calculations of the same functions, you may consider using software, such as SAS/IML. IML is a product of SAS designed to perform matrix calculations and operations.

Remember, we created a matrix of variance components and named it as *_varcomp* and a matrix of covariances of variance components and named it *_cov* in Code 2. These tables are stored in the WORK library of SAS. We need these tables to calculate heritability and standard error of heritability as shown below.

/* You must have SAS/IML product to run the following code*/

Code 3: Calculation of functions of variance components - 1

We would like to calculate additive, phenotypic variances and heritability.

```
/* Heritability estimate  - 1 */
/* Start IML */
      PROC IML;

      _varcomp={0.378,  0.0065,  4.645 };

      Additive={4 0 0}*_varcomp ;
      Phenotypic={1 1 1}*_varcomp ;
      h2_i=Additive/Phenotypic;

      PRINT _varcomp Additive Phenotypic h2_i [format=6.2];
      QUIT;
```

Explanation of the code:

1. **_varcomp={0.378, 0.0065, 4.645 }**: This is a row vector of variance components. We obtained variance components from the MIXED procedure and created a column vector with 3 rows. In the row vector, the '0.378' is the family variance component, the '00065' is the site by family interaction variance component, and the "4.645" is the error or residual component.

$$\_varcomp = \begin{Bmatrix} 0.378 \\ 0.0065 \\ 4.645 \end{Bmatrix}$$

2. **Additive={4 0 0}*_varcomp**: We would like to calculate additive genetic variance, which is four times of the half-sib family variance (4*0.378). In order to multiply 0.378 with 4, we need to create a Row vector of coefficients {4 0 0}. The product of the row vector of coefficients {4 0 0}and the vector of variance components {_varcomp} will give the additive genetic variance.

$$\text{Additive} = \{4 \quad 0 \quad 0\} * \begin{Bmatrix} 0.378 \\ 0.0065 \\ 4.645 \end{Bmatrix} = 5.12$$

3. **Phenotypic={1 1 1}*_varcomp**: We need the phenotypic variance to calculate heritability. Remember, phenotypic variance is the sum of all variance components that contribute to the Expected Mean Square for the family effect. Multiplying the *_varcomp* vector by the vector of coefficients {1, 1, 1} will give us the phenotypic variance.

$$\text{Phenotypic} = \{1 \quad 1 \quad 1\} * \begin{Bmatrix} 0.378 \\ 0.0065 \\ 4.645 \end{Bmatrix} = 5.0295$$

4. PRINT: In order to see results, we use the PRINT option. Notice that there is no semicolon ';' after the PRINT option.
5. [format= **6.2**]: This is to set the column length to 6 and the number of decimals to 2 for the output.

Output 3:

```
 _VARCOMP   ADDITIVE PHENOTYPIC    H2_I

    0.378      1.512     5.0295    0.30
   0.0065
    4.645
```

- The family variance is 0.378 so the additive genetic variance is 1.512 (4 x 0.378), phenotypic variance is 5.0295, and narrow-sense individual-tree heritability is 0.30. The site x family variance (0.0065) and the error variance (4.645) are also listed in the above table.

Let's add some more calculations to above code. The new terms are bold in the following code.

Code 4: Calculation of functions of variance components – 2

Plant and animal breeders are often interested in percentages of total variance explained by the factors (family, within family etc.) in the experiment. They are also interested in the precision of genetic parameters. In below IML code we added the AsyCov table (covariance table) to calculate standard error of any function of variance components, for example additive genetic variance. The new additions are bold faces in the IML code

```
/* Percent variance, heritability and StdErr – 2 */

PROC IML;

/* Type variance components */
_varcomp={0.378,  0.0065,  4.645};

/* add labels to rows of _varcomp vector using ROWN option*/
ROWN={family site_family Error};

/* Total variance */
Total = SUM (_varcomp) ;

/* Percent Variance Explained by each factor */
VarComp_pct=_varcomp/Total*100;


/* Additive variance */
Additive = {4 0 0} * _varcomp ;

/* Covariance matrix */
_cov={0.01765    -0.00044    -0.00009,
     -0.00044    0.002744    -0.00215,
     -0.00009    -0.00215     0.02988};

/* Variance and StdErr of Additive Variance */
c_n={4, 0, 0};
var_A =c_n` * _cov * c_n ; *<-- variance of Additive var;
SE_A=sqrt(var_A) ;


PRINT
Varcomp_pct [rowname=rown format=5.1]
Additive
var_A
SE_A [format=6.3]
```

```
    h2_i [format=6.3] ;

QUIT;
```

Explanation of the code:

1. ROWN: Names the rows (variance components) in the _varcomp column vector.
2. Using the SUM function of IML, we can easily obtain the total phenotypic variance.
3. **_cov**: We added the covariance matrix *_cov*. Remember, we created this table in the MIXED procedure code 2 (See Output 2). We need variances of variance components and the covariances between them to calculate standard errors of functions of variance components.
4. **c_n**: is a **column vector** of coefficients for additive genetic variance. The commas between elements (4, 0, 0) make it 3 rows and 1 column (a column vector). When we multiply **c_n** with the **_cov** matrix, we actually multiply variance of family variance ($Var\sigma^2_F$) by 4 to obtain variance of additive genetic variance (**var_A**).
5. In matrix algebra in order to multiply a column vector (c_n) with a square matrix (_cov), we must take the transpose of the column vector (c_n`) and multiply it to the square matrix. In the equation below, the bold matrix **[4 0 0]** is the transposed **c_n matrix**. After transpose, the column vector becomes a row vector (1 row, 3 columns).

$$\text{Var\_A} = \mathbf{c\_n}` * \text{\_cov} * \text{c\_n}$$

$$= [4\ 0\ 0] * \text{\_cov} * \begin{bmatrix} 4 \\ 0 \\ 0 \end{bmatrix} = 0.2824$$

Output 4:

SOURCE	VARCOMP_PCT	ADDITIVE	VAR_A	SE_A	H2_I
FAMILY	7.5	1.512	0.2824	0.531	0.30
SITE_FAMILY	0.1				
ERROR	92.4				

- Family factor explained 7.5% of the total phenotypic variance. The site * family interaction term explained only 0.1% of the total phenotypic variance, which is not significantly different from zero.
- Additive genetic variance has a variance of 0.2824 and standard error of 0.531. It is a good idea to present an estimate with its standard error, even if it is an approximation ($\sigma^2_A = 1.51 \pm 0.531$).

Now, let's go further and calculate standard error of heritability using two methods; the Dickerson approximation and the Delta method.

BOX 4: Narrow-sense heritability and its standard error

Narrow-sense heritability is a ratio of additive (σ^2_A) and phenotypic variances (σ^2_P):

$$h_i^2 = \frac{\sigma_A^2}{\sigma_P^2} = \frac{4\sigma_F^2}{\sigma_F^2 + \sigma_{SF}^2 + \sigma_E^2} = 1.511/\ 5.029 = 0.30$$

Variance of heritability Var(h_i^2):

1- Assuming phenotypic variance is a constant, we can use the Dickerson approximation (1969) to calculate variance of heritability:

$$\text{Var}(h_i^2) = \frac{16\text{Var}(\sigma_F^2)}{(\sigma_P^2)^2}$$

2- Delta method: Many genetic parameters are the ratios of variances and covariances, such as heritability. Delta method is a good approximation to obtain the variance of a ratio because it uses all the information of moments. The general formula for the expectation and variance of a ratio (x/y) would be as follows:

$$\text{Var}(x/y) \approx [\ E(x)/E(y)]^2\ [\ \text{Var}(x)/E(x)^2 + \text{Var}(y)/E(y)^2 - 2\text{Cov}(x,y)/E(x)E(y)\ ]$$

Expectation of heritability would be as follows:

$$\text{Var}(h_i^2) = \left(\frac{4\sigma_F^2}{\sigma_P^2}\right)^2 \left[\frac{\text{Var}(4\sigma_F^2)}{(4\sigma_F^2)^2} + \frac{\text{Var}(\sigma_P^2)}{(\sigma_P^2)^2} - \frac{2\text{Cov}(4\sigma_F^2, \sigma_P^2)}{(4\sigma_F^2\sigma_P^2)}\right]$$

Taking the square root, the variance would give standard error. A worked example of Dickerson and Delta method to calculate variance of heritability is given in Code-5. See Appendix 1 in Lynch and Walsh (1998) for the theory of the delta method.

Code 5: Calculation of functions of variance components – 3

```
/* Standard Error of heritability - 3 */

PROC IML;

/* Type variance components */
_varcomp={0.378, 0.0065, 4.645};
```

```
/* Additive genetic variance */
A = {4 0 0}* _varcomp;

/* Phenotypic variance */
P = {1 1 1}* _varcomp;

/* Heritability */
h2_i = A /P;

/* Covariance matrix */
_cov={0.01765    -0.00044    -0.00009,
     -0.00044    0.002744    -0.00215,
     -0.00009    -0.00215     0.02988};

/* Coefficients of numerator (Additive var) */
c_n = {4,0,0};

/* Coefficients of denominator (Pheno. var) */
c_d = {1,1,1};


/* --- You DO NOT need to change the following code ---*/

/* Variance of Additive variance */
var_A   = c_n` * _cov * c_n ;

/* Variance of Phenotypic variance */
var_P   = c_d` * _cov * c_d ;

/* Covariance between Additive and Phenotypic */
cov_A_P = c_n` * _cov * c_d ;

/* Variance of heritability: Dickerson */
var_h2_i_dick = var_A / (P**2);

/* Std Error of heritability: Dickerson */
se_h2_i_dick = SQRT(var_h2_i_dick) ;

/* Variance of heritability: Delta method */
var_h2_i_delta=(h2_i**2)*((var_A/A**2)+(var_P/P**2)-
(2*cov_A_P/(A*P)));

/* Std Error of heritability: Delta method */
SE_h2_i_delta = SQRT(var_h2_i_delta) ;

PRINT
```

```
var_P
var_A
h2_i [format=6.2];
PRINT
var_h2_i_delta
SE_h2_i_delta [format=6.3];
PRINT
var_h2_i_dick
SE_h2_i_dick [format=6.3];
QUIT;
```

Explanation of the code:

1. **A** is a symbol we used for additive genetic variance; **P** is phenotypic variance.
2. **c_d = {1, 1, 1}**: The row vector of coefficients for Phenotypic variance (denominator in heritability). We need these coefficients to calculate variance of P.
3. **var_P**: Variance of phenotypic variance. It is simply adding all the variances of variance components in the equation and subtracting 2 times of covariances between three variance components. See BOX 3 for formula and details.

$$\text{Var\_P} = \mathbf{c\_d`} * \mathbf{\_cov} * \mathbf{c\_d} = \begin{bmatrix}1 & 1 & 1\end{bmatrix} * \text{\_cov} * \begin{bmatrix}1\\1\\1\end{bmatrix} = 0.0449$$

4. **var_h2_i_dick**: Variance of heritability using the Dickerson approximation. Here, we assume that P is a constant, and thus, it does not have a variance and a covariance with A. In order to obtain variance of heritability, all we need to do is divide variance of A with the square of P. Remember; standard error is the square root of variance.
5. **var_h2_i_delta**: Variance of heritability using the Delta method. In contrast to Dickerson assumption, here we use all the info about the variance components. From the _cov matrix, we know that they are not independent, but there are covariances between each pair.

Output 5:

```
           VAR_P       H2_I
           0.0449      0.30

VAR_H2_I_DELTA         SE_H2_I_DELTA
     0.010                 0.098

VAR_H2_I_DICK          SE_H2_I_DICK
    0.011                  0.106
```

- The narrow-sense individual-tree heritability for height is $h^2 = 0.30$. The standard error based on the Delta is 0.098, while it is 0.106 based on the Dickerson method.
- The standard error of heritability based on the Delta method is slightly lower than standard error of heritability based on Dickerson.

BOX 5: Repeatability of family mean and its standard error

In tree improvement programs the main interest could be selection of families not the individual trees within families. We might be interested in ranking the families and making genetic gain predictions based on family selection. For this, we need repeatability or heritability of family mean.

We need to derive family mean formula from the linear model.
Half-sib family means:

$$(\overline{Y}..k.) = F_k + SF._k /s + E.._k./sbn$$

Where, s = number of sites, b = number of blocks per site, n = number of trees per family per block.

The variance (phenotypic) of half-sib family mean:

$$Var(\overline{Y}..k.) = Var\ (F_k + SF._k /s + E.._k./sbn\ )$$

Heritability of half-sib family mean is a ratio of family variance component (σ^2_F) and phenotypic variance of family mean ($\sigma^2_{P\_HS}$):

$$\sigma^2_{P\_HS} = \sigma^2_F + \frac{\sigma^2_{SF}}{s} + \frac{\sigma^2_E}{sbn}$$

If we had multiple-tree plots, then, phenotypic variance half-sib family means would have included the family by block interaction (plot-to-plot error) term:

$$Var(\overline{Y}..k.) = Var\ (F_k + SF._k /s + \mathbf{FB(S).._k} /sb + E.._k./sbl\ )$$

$$\sigma^2_{P\_HS} = \sigma^2_F + \frac{\sigma^2_{SF}}{s} + \frac{\sigma^2_{FB(S)}}{sb} + \frac{\sigma^2_E}{sbl}$$

Where, l = number of trees per plot.

Variance of half-sib family means heritability Var(h^2_{HS}):

1- Assuming phenotypic variance is a constant (Dickerson approximation):

$$\text{Var}(h^2_{HS}) = \frac{\text{Var}(\sigma^2_F)}{(\sigma^2_{P\_HS})^2}$$

2- Delta method:

$$\text{Var}(h^2_{HS}) = \left(\frac{\sigma^2_F}{\sigma^2_{P\_HS}}\right)^2 \left[\frac{\text{Var}(\sigma^2_F)}{(\sigma^2_F)^2} + \frac{\text{Var}(\sigma^2_{P\_HS})}{(\sigma^2_{P\_HS})^2} - \frac{2\text{Cov}(\sigma^2_F, \sigma^2_{P\_HS})}{(\sigma^2_F \sigma^2_{P\_HS})}\right]$$

A worked example from Output-2:

Phenotypic variance of family means:

$$\boldsymbol{\sigma^2_{P\_HS}} = \sigma^2_F + \sigma^2_{SF}/s + \sigma^2_E/sbn$$

where s=number of sites, b=number of blocks, and n=number of trees per family per block.

$\boldsymbol{\sigma^2_{P\_HS}} = 0.378 + 0.0065/5 + 4.645 / 66.7 = 0.441$

Since there are missing trees, sbn=66.6 instead of 75.

Family means heritability:

$$\boldsymbol{h^2_{HS}} = \sigma^2_F / \sigma^2_{P\_HS}$$

$$= 0.378 / 0.441 = 0.86$$

See appendix A1 in Lynch and Walsh (1998) for details of the delta method.

Code 6a: Calculation of functions of variance components – 4

Here, our focus is on the repeatability of half-sib family means (i.e. heritability of half-sib family means) and its standard error. We are interested in heritability of family means because we are usually interested in selecting the best families. Genetic gains from family selection would be based on the heritability of family means, variation among families (phenotypic variance of family means) and the selection intensity.

```
/* Repeatability of family means and SE - 4 */

PROC IML;
   /* variance components */
   _varcomp={0.378, 0.0065, 4.645};
```

```
/* Covariance matrix */
_cov={0.01765    -0.00044    -0.00009,
     -0.00044    0.002744    -0.00215,
     -0.00009    -0.00215     0.02988};

c_n = {1, 0, 0}; *<-Coefficients of numerator;
F=c_n`*_varcomp ; *<-Family variance component;

/* Coefficients of denominator */
site=5 ; numtree=66.7 ;

c_d={1, 1, 1};
c_d[2,1]=1/site ;
c_d[3,1]=1/numtree ;

P_hs = c_d`* _varcomp ;  *<--Phenotypic var of family means;

/* heritability of family means*/
h2_hs = F /P_hs;

/* --- You DO NOT need to change the following code ---*/
var_F    = c_n` * _cov * c_n ; *<--variance of family;
var_P_hs = c_d` * _cov * c_d ; *<--var of Phenotypic var;
cov_F_P  = c_n` * _cov * c_d ; *<--covariance btw F and P_hs;

/* Variance of heritability: Delta method */
var_h2_hs_delta=(h2_hs**2)*((var_F/F**2)+(var_P_hs/P_hs**2)-
(2*cov_F_P/(F*P_hs)));

/* Standard Error of heritability */
SE_h2_hs_delta = SQRT(var_h2_hs_delta) ;

PRINT
c_d ;
PRINT
P_hs [format=6.3]
var_P_hs [format=6.4];
PRINT
h2_hs [format=6.2]
var_h2_hs_delta  [format=6.4]
SE_h2_hs_delta [format=6.3];
QUIT;
```

Explanation of the code:

1. **c_n = {1, 0, 0}** : A row vector of coefficients to get family variance component from the _varcomp row vector. This is done by the subsequent matrix multiplication: **F=c_n`*_varcomp**.
2. Remember that to calculate phenotypic variance of family means, we need to divide the variance components by certain coefficients: $\sigma^2_{P\_HS}$ = 0.378 + 0.0065/5 + 4.645 / 66.7. Here, for simplicity, we assumed that the data are perfectly balanced (no dead trees). The following code obtains these coefficients.

   ```
   /* Coefficients of denominator */
   ```
   site=5; numtree=66.7: We have 5 sites, 15 blocks in each site, 1 tree per family per block. However, because some trees were dead, the actual (average) number of trees per family per site is 66.7 instead of 75.
   c_d={1,1,1}: Create a row vector of coefficients. All 3 elements of the matrix are 1.
   c_d[2,1]=1/site: Make the second element to 1/5.
   c_d[3,1]=1/numtree: Make the third element to 66.7. This is the average number of trees per family across 5 sites.

   Assuming that there are no missing trees, every family has 15 trees in each of 5 sites, the coefficients of denominator (phenotypic variance of family means) would be
   c_d = {1, 1/5, 1/(5*15)} or
   c_d = {1, 0.2, 0.013}

   When the transpose of **c_d** is multiplied to the *_varcomp* (the row vector of variance components), the result would be the phenotypic variance of family means.
   P_hs = c_d`* _varcomp.

3. **h2_hs = F /P_hs**: Heritability of family means. Family variance is divided by the phenotypic variance of family means.

Output 6a:

```
      C_D
        1
      0.2
0.0133333

  P_HS      VAR_P_HS
 0.441      0.0176

 H2_HS      VAR_H2_HS_DELTA       SE_H2_HS_DELTA
  0.86           0.0024              0.049
```

- There is considerable variation among families for height that is due to by genetic factors as suggested by the high heritability of family means (H2_HS=0.86 ± 0.049). The low standard error of 0.049 suggests that the estimate is precise.

- See Box 3 and 4 for the details of formula of the functions of variance components.

What coefficients do we use to calculate heritability of family mean if there is imbalance or if there are missing values?

The above calculation assumes there are no missing trees. In reality, we will always have dead trees in progeny tests. In order to get coefficients for calculation of family mean phenotypic variance, we can run the MIXED Code-1 with adding the **METHOD=TYPE3**as follows:

Code 6b: Obtaining the coefficients for expected mean squares

```
PROC MIXED DATA=op.pine_halfsib METHOD=TYPE3;
   CLASS site block family;
   MODEL height =site block(site);
   RANDOM family site*family;
RUN;
```

Explanation of the code:

1. **METHOD=TYPE3:** Tells MIXED procedure to use the ANOVA method to produce expected mean squares and their coefficients. The default is the REML method (a likelihood-based calculation), which does not produce expected mean squares.

Output 6b:

The Mixed Procedure

Type 3 Analysis of Variance

Source	DF	MS	Expected Mean Square
site	4	2103	Var(Res) + 13.3 Var(SF) + Q(S)
block(site)	70	31	Var(Res) + Q(block(site))
family	23	30	Var(Res) + 13.3 Var(SF) + 66.7 Var(F)
site*family	92	4.7	Var(Res) + 13.5 Var(SF)
Residual	1442	4.6	Var(Res)

- The phenotypic variance of family means would be

c_d = {1, 1/s, 1/(s*n)} or

```
c_d = {1,  1/5,  1/66.7} = {1,  0.2,  0.015}
```

- Thus, on average, a family had **66.7** trees instead of 75 (5 sites x 15 blocks x 1 tree=75). See BOX1 for details.

4.6 Breeding Values

BOX 6: Half-sib family breeding values

In some cases, the primary interest in a progeny test is to make inferences about the random effects, such as breeding values of genotypes (Lynch and Walsh 1998). Breeding values are used to rank parents and select the best ones for future breeding or deployment. Here, we give brief definitions and equations about breeding values for wind-pollinated trees. For more details about breeding values, see related chapters in Falconer and MacKay (1996) and in Lynch and Walsh (1998).

When a parent is crossed with a number of other parents in a breeding population, we measure progeny from all the crosses and estimate a mean performance of that parent. The deviation of the parent mean (X) from the population mean ($\overline{X}$) is **general combining ability (GCA)** (see Falconer and Mackay 1996, page 274).

$GCA = X - \overline{X}$

Breeding value (BV) is twice the expected deviation of its progeny mean (GCA) from the population mean. In another words, it is twice of GCA.

$BV_{HS} = 2GCA$

To express BV, GCA is multiplied by two because a parent can only transmit half of its genes to its progeny. The other half comes from the other parent.

We obtain GCA values of parents by a procedure called **Best Linear Unbiased Predictors (BLUP)**. If we are interested in the effect of a particular site (make inferences about fixed effects) we use a procedure called **Best Linear Unbiased Estimates (BLUE)**. These methods are based on the maximum likelihood theory and are beyond the scope of this handbook.

The solutions from the mixed model for family effect are the GCA values and they are the Best Linear Unbiased Predictions of families. See equation [4] in Chapter 12.

Code 7: Estimation of breeding values

To obtain GCA estimates of families (hence breeding values), we need to add some terms to the SAS MIXED procedure code. The new terms added to the MIXED procedure given in Code 1 are in bold.

```
/* Estimation of breeding values */
   PROC MIXED DATA=op.pine_halfsib ;
      CLASS site block family ;
      MODEL Height=site block(site)/S OUTP=_pd COVB ;
      RANDOM family site*family/S ;
      ODS LISTING EXCLUDE SOLUTIONF SOLUTIONR COVB ;
      ODS OUTPUT SOLUTIONF=s_f SOLUTIONR=s_r COVB=_covb ;
   run;
```

Explanation of the code:

1. The **/S** option after MODEL requests the BLUE of the fixed effects be produced. Such as, the effect of specific sites.
2. OUTP= option after MODEL produces the residuals and predicted values for every tree and saves them in a data file named **_pd**. You may give a different name than _pd. All the raw data are included in file _pd, such as site number, family ID of all the trees. Using the RESIDUALS in the **_pd** data we can calculate individual tree breeding values.
3. COVB option after MODEL is the approximate covariance matrix of fixed-effects parameter estimates. We need this table to estimate standard error of individual tree breeding values.
4. The **/S** options after RANDOM requests the table of the BLUP of random effects be produced.
5. ODS LISTING EXCLUDE statements tell SAS to stop dumping large tables (i.e., SOLUTIONF, SOLUTIONR, COVB) into the output window of SAS. These tables can be very large and fill the output window quickly.
6. ODS OUTPUT creates SAS output files. The name before the '=' is the table name of the SAS Output Delivery System (ODS) (e.g., SOLUTIONR). The name after the '=' (e.g., S_R) is a given table name. You may give any name instead of s_r.

7. SOLUTIONF=**s_f** is the Best Linear Unbiased Estimates (BLUE) of the **fixed effects** (i.e, the intercept, site effects etc.). This table is the solution of the formula 3 in Chapter 12.
8. SOLUTIONR=**s_r** requests that the solution for the random-effects parameters be produced. This is the table of the Best Linear Unbiased Predictors (BLUP) of **random effects** (i.e., General combining ability effects of parents). This table is the solution of the formula 5 in Chapter 12.

Output 7:

- The above MIXED procedure code produces most of the output given for the Code 1. The additional tables we requested (**s_r, s_f, _COVB**) will not be printed because we used the ODS LISTING EXCLUDE statements. But these tables are created and stored in the WORK LIBRARY of SAS. You may click on the EXPLORER tab, and then click on Library and the Work library icon to see those tables. We can also print (not sending to a printer but to see in SAS Output Window) some of those tables using the following codes:

Code 8: Printing BLUE of fixed effects and BLUP of random effects

```
/* BLUE values of sites and blocks */
TITLE 'BLUE of fixed effects ';
PROC PRINT DATA=s_f (OBS=7) NOOBS;
RUN;

/* BLUP values of families */
TITLE 'BLUP of families';
PROC PRINT DATA=s_r (OBS=7) NOOBS;
WHERE EFFECT='family' ;
RUN;
```

Explanation of the code:

1. OBS=7) : The actual files **s_f** or **s_r** can be very large. We would like to see the first seven observations in the SAS output window.
2. NOOBS: SAS by default prints observation number (1,2,3 …) with the data printed in the output window. We did not want to print column of observations.

The BLUE of fixed effects and a partial of BLUP of random effects are printed.

Output 8:

BLUE of fixed effects

Effect	site	block	Estimate	StdErr	DF	tValue	Probt
Intercept			23.1708	0.4874	23	47.54	<.0001
site	A		-3.3602	0.6745	92	-4.98	<.0001
site	B		-4.5161	0.6387	92	-7.07	<.0001
site	C		2.5264	0.6387	92	3.96	0.0002
site	E		0.4126	0.6745	92	0.61	0.5423
site	F		0	.	.	.	.
block(site)	A	1	-1.1481	0.6529	1442	-1.76	0.0789

- The ESTIMATE in the above printout is the **Best Linear Unbiased Estimate (BLUE)** of Sites and Blocks. The degree of freedom (DF), standard errors of BLUEs (STDERR), the t value and probability of t value are given. The t probability values tell us whether the estimate is significantly different from zero or not. Trees in site C had the highest growth (BLUE of C=2.5264), and lowest growth was in site B (BLUE = - 4.5161). These numbers are solutions from the mixed model they are distributed around zero. You may add the grand mean or the use the ESTIMATE statement to get meaningful (interpretable) site estimates.
- The output includes an estimate for the block 1 at site A. The BLUE of that block is -1.1481. There are 15 x 5 estimates for blocks because each of 5 site has 15 blocks but we only presented the estimate for the block 1 at site A.

BLUP of families

family	Estimate	StdErr Pred	DF	tValue	Probt
F1378	-0.8653	0.2745	1442	-3.15	0.0017
F1801	-0.4997	0.2702	1442	-1.85	0.0646
F1805	0.1848	0.2702	1442	0.68	0.4941
F1853	-0.6621	0.2665	1442	-2.48	0.0131
F1806	0.4823	0.2665	1442	1.81	0.0705
F1013	-0.5781	0.2630	1442	-2.20	0.0281
...					

- The ESTIMATE in above printout is the **Best Linear Unbiased Prediction (BLUP)** GCA estimates of Families. The degree of freedom (DF), standard errors of BLUPs (STDERR PRED), the t value and probability of t value are given. The t probability values tell us whether the prediction is significantly different from zero. The BLUPs of families are distributed around zero that is some of the BLUP values are negative and some are positive. Remember that the BLUPs of families are GCA values. In order to calculate breeding values, either we can export the **s_r** table into Microsoft Excel to do calculation or we can continue to use SAS to the job. In below code we used SAS data steps to calculate breeding values of families.

Code 9: Calculation of family breeding values

```
/* Calculation of BLUP family breeding values */
DATA bv_hs ;
  SET s_r (WHERE=(EFFECT='family'));
  BV_HS=2*estimate ;
  FORMAT estimate BV_HS stderrpred 8.3 ;
  KEEP family estimate BV_HS stderrpred;
RUN;
```

Explanation of the code:

The above code is a data step code. The SAS file s_r produced by MIXED procedure is used to calculate family breeding values.

1. The new data file named A includes only the **family effects** (rows) not the **family*site** or any other interaction effects. We used the WHERE clause to select the family effects.
2. We tell SAS to look at the column named EFFECT and keep the rows that have '**family**'.
3. Breeding value is two times of the general combining ability because one parent (i.e. the family term) transmits only half of the genes to its progeny. The other half comes from another parent.
4. The new file A includes FAMILY ids, general combining ability (GCA) and breeding values (BV_HS) of each family and the standard errors of predictions (STDERRPRED).

```
TITLE 'BV of families' ;
PROC PRINT DATA=bv_hs (OBS=6) NOOBS ROUND;
RUN;
```

Output 9:

BV of families

family	Estimate	StdErr Pred	BV_HS
F011378	-0.865	0.275	-1.731
F011801	-0.500	0.270	-0.999
F011805	0.185	0.270	0.370
F011853	-0.662	0.267	-1.324
F021806	0.482	0.266	0.965
F051013	-0.578	0.263	-1.156
F051033	1.322	0.274	2.643

BOX 7: Individual tree breeding values

Individual-tree breeding values (IBV) are obtained by adding parental breeding value to the estimated within-family deviation (**Aw**).

$$IBV = BV_{HS} + \mathbf{Aw}$$

The within family deviation (**Aw**) is the product of residuals from the mixed model and an approximate within family heritability.

$$Aw = \frac{3\sigma_F^2}{\sigma_E^2}\left(y - X\hat{B} - Z\hat{\gamma}\right)$$

Where, σ_F^2 is the family variance component (or variance due to general combining ability of parents), σ_E^2 is the error variance, $X\hat{B}$ is the product of the design matrix **X**, and BLUE of fixed effects, and $Z\hat{\gamma}$ is the product of the design matrix **Z** and BLUP of random effects.

Each measured trait of a tree (y_{ijkl}) is adjusted for fixed and random effects ($y - X\hat{B} - Z\hat{\gamma}$) in the model and then multiplied by approximate within-family heritability ($3\sigma_F^2/\sigma_E^2$) to obtain within family deviation **Aw** (Xiang and Li 2001). Thus, tree breeding values are comparable across blocks or sites.

Code 10: Within family individual tree deviations

```
TITLE 'Print deviation of individual trees ';
PROC PRINT DATA=_pd (OBS=7) NOOBS ROUND;
VAR site block family tree  Pred  StdErrPred DF Resid;
RUN;
```

Explanation of the code:

1. We use the output file ***_pd*** from the MIXED procedure given in Code 7 to calculate tree breeding values. The above code prints out 7 observations from the _pd data.
2. The file is a large one and includes raw data too. Using the VAR option, we limited the output (variables) we want to see.

Output 10:

Breeding values of individual trees

site	block	family	tree	Pred	StdErrPred	DF	RESID
A	1	F1378	1	17.81	0.51	1442	-0.01
A	1	F1801	15	18.14	0.50	1442	-0.34
A	1	F1805	28	18.84	0.50	1442	-2.44

A	1	F1853	41	18.00	0.50	1442	-0.40
A	1	F1806	53	19.15	0.50	1442	0.45
A	1	F1013	67	18.07	0.50	1442	-0.87
A	1	F1033	82	20.00	0.51	1442	-1.20

- The RESID column is the individual tree deviation ($\mathbf{y} - \mathbf{X}\hat{\mathbf{B}} - \mathbf{Z}\hat{\boldsymbol{\gamma}}$). They are adjusted for fixed and random effects. In order to calculate breeding value of a tree, we need to multiply that deviation (RESID) with within-family heritability and add family breeding value to the outcome.

  IBV = BV_{HS} + **Aw**

Example: Let's calculate breeding value of tree 1.

- The parent of tree 1 is F1378 with a breeding value of BV_{HS} = -1.731 (see Output 9).
- The within family heritability is $\frac{3\sigma_F^2}{\sigma_E^2} = \frac{3*0.378}{4.645} = 0.24$
- The within family deviation for tree 1 is -0.01.

Breeding value of tree 1 is;
IBV = -1.731 + 0.24* (-0.01)
IBV = **-1.733**
Similarly, the breeding value of tree 25 of family F1806 would be;
IBV = 0.965 + 0.24*(0.45) = **1.073.**

Progeny receive half of their genes from parents. That's why family breeding values are significant part of individual tree breeding values. In order to add family breeding values to the individual tree deviation, we need to merge two data files as shown below:

Code 11: Calculation of individual tree breeding values

```
   /* Calculation of Individual tree breeding values */
   PROC SORT DATA=bv_hs ; BY family ;
   PROC SORT DATA=_pd ; BY family ;

DATA bv_all ;
     MERGE bv_hs _pd ;
     BY family ;
      h2_w = 3*0.378/4.645; *<--Within family heritability;
      ibv= bv_hs + (RESID*h2_w);
   RUN;
```

Explanation of the code:

1. Using the SORT procedure of SAS, we sort data files **bv_hs** and **_pd** for the common variable **family**. If you do not sort data files, you will not be able correctly merge them.
2. MERGE: In the DATA step we are creating a new data called **bv_all**, by merging two data sets; bv_hs: family breeding values and _pd: Individual tree deviations. MERGE statement in DATA step put two data sets side by side.
3. BY: The by statement make sure that every tree of a family will have new column of family breeding values.
4. Within family heritability (**h2_w**) is calculated as 3/4 of the additive genetic variance (3*0.378) divided by within family variance (error variance component = 4.645).
5. Finally, Individual tree Breeding Values (**ibv**) were calculated by adding family breeding value (**bv_hs**) to the product of h2_w and RESID.

BOX 8: Adjusted breeding values

Breeding values are distributed around zero, being deviations from a mean value. There are negative and positive values. We can simply rank the breeding values and make selections. However, they are not meaningful to interpret or to calculate genetic gains over a checklot (unimproved seed). We may wish to re-express them on a scale relative to a standard "benchmark". We need to estimate the grand mean for the response variable and add it to all the breeding values to obtain interpretable values.

Is it OK to add the arithmetic mean to the breeding values to obtain adjusted breeding values? The answer is it depends. If data are unbalanced, if there are significant growth differences between sites, then arithmetic mean could be unreliable and significantly different from a grand mean based on Best Linear Unbiased Estimates (BLUE) of fixed effects.

For example, for these specific data we used so far, the arithmetic mean with standard error for height was 21.64 ± 0.084 but the grand mean based on BLUE of fixed effect is 21.61 ± 0.137. In the following code we gave an example on how to obtain the grand mean using the ESTIMATE statement in the MIXED Procedure:

Code 12: Calculation of grand mean

```
/* Calculation of grand mean */
   PROC MIXED DATA=op.pine_halfsib ;
```

```
    CLASS site block family;
    MODEL Height = site block(site);
    RANDOM family site*family;
ESTIMATE 'GrandMean' intercept 75 site 15 15 15 15 15 block(site) 1
1 1 1 1  1 1 1 1 1  1 1 1 1 1
            1 1 1 1 1  1 1 1 1 1  1 1 1 1 1
            1 1 1 1 1  1 1 1 1 1  1 1 1 1 1
            1 1 1 1 1  1 1 1 1 1  1 1 1 1 1
            1 1 1 1 1  1 1 1 1 1  1 1 1 1 1/DIVISOR=75;
RUN;
```

Explanation of the code:

1. The ESTIMATE statement produces a Best Linear Unbiased Estimates of the grand mean (overall mean) for the response variable. The ESTIMATE statement should be included after the RANDOM statement.
2. Here, we have **5 sites** each with 15 blocks. The total number of blocks is 75 (5 sites x 15 blocks). Thus, the number of coefficients for blocks must be equal to the total (75). Since the total number of blocks in the experiment is 75, we need to give equal coefficients (75 of 1) to each block. Each block estimate is included in the calculation.
3. DIVISOR: The sum of the estimates for the intercept, sites and blocks would be divided by the total (75) to calculate the grand mean.

Output 12:

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
GrandMean	21.61	0.1367	1442	158.15	<.0001

- Of course, we need to merge family breeding values (Codes 7, 8, 9) with the **_pd** output file (Code 7 and 10) so we can calculate individual tree breeding values. For this task, either we can use SAS data steps or export the files to Excel and make calculations. Here, an example is given about how to merge BV_HS and _pd files using SAS.

Code 12: Adjusted breeding values of families and individual trees

```
/* Code 12: Adjusted breeding values and gain calculation */
   TITLE 'Adjusted breeding values and gains';
   DATA bv_all; SET bv_all ;
```

```
  GrandMean=21.61 ;  checklot=18.59 ;
  Adj_bv_hs = bv_hs + GrandMean ;
  Adj_ibv=ibv + GrandMean ;

  Family_Gain = (Adj_bv_hs - Checklot) / Checklot*100 ;
  Tree_Gain = (Adj_ibv - Checklot) / Checklot*100 ;
  FORMAT ibv adj_ibv Adj_bv_hs 8.2 family_gain tree_gain 8. ;
RUN;

PROC PRINT DATA=bv_all (OBS=14) NOOBS round;
VAR family tree  Adj_bv_hs Adj_ibv family_gain tree_gain ;
RUN;
```

Explanation of the code:

1. SET: Using the set statement, we used the same data (bv_all) to calculate adjusted breeding values for families and individual trees. Genetic gains for selection of families and individual trees were calculated as % deviations over an improved seed source (Checklot). The Checklot is generally field tested together with families.
2. The FORMAT statement is to reduce number of decimals in the variables.

Output 12:

Adjusted breeding values and gains

family	tree	Adj_bv_hs	Adj_ibv	Family_Gain	Tree_Gain
F011378	1	19.88	19.88	7	7
F011378	2	19.88	20.07	7	8
F011378	3	19.88	20.39	7	10
F011378	4	19.88	20.11	7	8
F011378	5	19.88	19.60	7	5
F011378	6	19.88	20.12	7	8
F011378	7	19.88	20.01	7	8
F011378	8	19.88	19.82	7	7
F011378	9	19.88	19.94	7	7
F011378	10	19.88	20.03	7	8
F011378	11	19.88	20.11	7	8
F011378	12	19.88	19.73	7	6
F011378	13	19.88	19.64	7	6
F011378	14	19.88	20.29	7	9

- Genetic gain is typically calculated as selection of a group families and selection of a few of best individuals.

4.1 Using a Macro Code to Predict Breeding Values

You may also use a SAS macro code to estimate family and individual tree breeding values from open-pollinated trials. Macro codes are a way of programming to do a job automatically.

Code 7: A Macro to predict breeding values of families and individual trees

```
/* Call the SAS Macro File */
%INC C:\RESEARCH\Handbook\SAS\CH4OP\Macro_HS_Family_BV_stp.SAS';
```

You need to download the macro code from the handbook web site and save it in a folder on your computer. Use the %INC in SAS editor window to call the macro. Make sure the address (folder) is correct.

```
/* Change UPPERCASE WORDS parameters according to your data */
%HSFamilyBV(dset   = OP.PINE_HALFSIB,
            site   = SITE,
            rep    = BLOCK,
            family = FAMILY,
            msite  = Y,
            y      = Height,
            outpath= C:\RESEARCH\Handbook\Results);
```

Explanation of the code:

1. %INC is a SAS statement to call the Macro Code. The SAS macro file ”Macro_HS_Family_BV_stp.SAS” is saved in a directory. You may save the file in another directory.
2. ***%HSfamilyBV*** is the **name of the macro**. In order to run the macro, you need to change the UPPERCASE variable names according to your data. For example, if the name of your data is PINE05, replace OP.PINE_HALFSIB with PINE05. If column name of the FAMILY is PARENTS, replace ’FAMILY’ with the PARENTS etc.

   After changes, select the code and hit the running man on the tool bar of SAS Editor window.

Outputs:

The Macro code exports the following outputs to a MS Excel file named OP_Y.XLS. If the trait of interest is let say HEIGHT, then the file name will be OP_HEIGHT.XLS.

(1) family breeding values,
(2) individual tree breeding values, which includes additive genetic variance and heritability
(3) variance components,
(4) covariance matrix of the variance components and
(5) BLUE for the fixed effects

4.2 Literature cited

Dickerson, G.E. 1969. Techniques for research in quantitative animal genetics. *In* Techniques and procedures in animal science research. American Society of Animal Sci., Albany, N.Y. pp. 36–79.

Falconer, D.S. and MacKay, T.F.C. 1996. Introduction to Quantitative Genetics, Fourth edition, Longman. 464 p.

Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. SAS System for Mixed Models. Cary, N.C.: SAS Institute Inc. 633 p.

Lynch, M. and B. Walsh. 1998. Genetics and Analysis of Quantitative Traits. Sinauer Associates, Inc. 980 p.

Xiang, B. and Li, B. 2001. A new mixed analytical method for genetic analysis of diallel data. Canadian J. Forest Research. 31: 2252–2259.