Estimates of Variance

By William Beavis, Kendall Lamkey, Katherine Espinosa (ISU)

Introduction

Estimating heritability is a fundamental concept of quantitative genetics. One method for obtaining estimates of heritability is the use of variance and covariance of a known collection of relatives from various types of progeny.

Objectives

- Model components of genetic variances and covariances for purposes of estimating heritability, a fundamental concept of quantitative genetics.
- Explain why estimates of components of genetic variability are limited to the population from which they are estimated.
- Students will derive variance components and recognize the differences amongst components obtained from different progeny used for estimating heritability.
- Students will write out the correct linear models for the correct mean squares and expected mean squares in a ANOVA table, and correctly interpret the ANOVA and algebraically extract the correct values for estimating heritability.
- Leverage of the powerful algebraic equivalence of covariances within groups of relatives to variances among the same groups.

Covariance of Relatives

Recall that $Cov(Y_{ij},Y_{ij'}) = Var(G_i)$, for $j \neq j'$. In the context of genotypic sampling of relatives, this general relationship has a profound and powerful impact on interpretation of ANOVA. It means that the covariance among a sample of relatives can be used to estimate components of genetic variance associated with the genotypic effect.

Table 1 A general ANOVA table for any type of related progeny.

			EMS	
Source	**df**	**MS**	**Variances**	**Covariances**
Reps	$r-1$			
Progeny	$p-1$	MS_3	$\sigma_w^2 + k\sigma^2 + rk\sigma_p^2$	$\sigma_w^2 + k\sigma^2 + rk[Cov(progeny$
Error	$(r-1)(p-1)$	MS_2	$\sigma_w^2 + k\sigma^2$	$\sigma_w^2 + k\sigma^2$
Total	$rp-1$			
Within Progeny	$rp(k-1)$	MS_1	σ_w^2	$\sigma_{we}^2 + [\sigma_T^2 - \text{Cov(progeny)}]$

Note that there are *p* progeny grown in *r* reps. Cov(progeny) refers to the covariance of the progeny, where the progeny can be full-sibs, half-sibs, S_1-progeny, S_2-progeny, testcross progeny, etc. The key is to know the progeny type and take advantage of the general rule that the variance among progeny is equal to the covariance of the progenies.

Note the use of σ^2_T instead of σ^2_G in the within progeny line of the ANOVA table. This is because σ^2_G is usually equal to $\sigma^2_A + \sigma^2_D$ the total variance in a non-inbred random mating population. If the population does not have a random mating structure, then the total variance will be something other than $\sigma^2_A + \sigma^2_D$. For example, the total genetic variance for an F_3 population is

$$\sigma_F^2 = \frac{3}{2}\sigma_A^2 + \frac{3}{2}\sigma_D^2$$

Linear Models for Phenotypic Values

Linear Models for Phenotypic Values

The covariance of relatives is simply that relatives tend to show more phenotypic similarities than with each other than with unrelated individuals. For example let X_{ij} represent an individual from the mating of parent *i* and parent *j*:

Table 2 Descriptions of relationships between individuals X_{ij} and $X_{i'j'}$.

Conditions	Description
$i = i', j = j'$	Full-sibs
$i = j', j = i'$	Recripocal Full-sibs
$i = i', j \neq j'$	Maternal half-sibs
$i \neq i', j = j'$	Paternal half-sibs

Specifying covariance of relatives in terms of genetic variances has the following assumptions:

1. Regular diploid and solely Mendelian inheritance
2. No environmental correlations among relatives
3. No gametic disequilibrium
4. The relatives are not inbred
5. The relatives are considered to be random members of some non-inbred population

With these assumptions, we can specify the covariance of relatives as follows:

$$\text{Cov} = \alpha^2_A + \delta^2_D + \alpha^2 \alpha^2_{AA} + \alpha^2_{AD} + \delta^2 \alpha^2_D + \alpha^3 \alpha^2_{AAA} + \cdots$$

Equation 1

where α is the coefficient of relative relationship, σ^2_A is the additive genetic variance, δ is the dominance relationship coefficient, σ^2_D is the dominance variance, σ^2_{AD}, σ^2_{AAA}, ... are the epistatic variances.

Common Types of Relatives

Using the result of Equation 1 for some common types of relatives it can be shown that:

Covariance of half-sibs with one common parent is:

$$Cov(HS) = \frac{(1+F_A)}{4}\sigma_A^2 + (\frac{(1+F_A)}{4})^2\sigma_{AA}^2 + ...$$

Covariance of full-sibs with parents *A* and *B* is:

$$Cov(FS) = \frac{(2+F_A+F_B)}{4}\sigma_A^2 + \frac{(1+F_A)(1+F_B)}{4}\sigma_D^2 + (\frac{(2+F_A+F_B)}{4})^2\sigma_{AA}^2 + (\frac{(2+F_A+F_B)}{4})(\frac{(1+F_A)(1+F_B)}{4})\sigma_{AD}^2 + (\frac{(1+F_A)(1-F_B)}{4})^2\sigma_{DD}^2 + (\frac{(2+F_A+F_B)}{4})^3\sigma_{AA}^2 + ...$$

F_2 and F_3 progenies

Table 3

Genotype	Freq	GV	F_3 Progeny			F_3 Progeny Mean
			AA	Aa	aa	
AA	1/4	a	1	0	0	a
Aa	1/2	d	1/4	1/2	1/4	1/2d
aa	1/4	-a	0	0	1	-a

F2 and F3 Variances

Total genetic variance among F_2 individuals:

$$\sigma^2_{F_2} = \sigma^2_A + \sigma^2_D = \frac{1}{2}a^2 + \frac{1}{4}d^2$$

Total F_2 phenotypic variation:

$$\sigma^2_{F_2} = \sigma^2_A + \sigma^2_D + \sigma^2_{E1} = \frac{1}{2}a^2 + \frac{1}{4}d^2 + E1$$

where E1 is the non-genetic variation among F_2 plants.

Recall that the F_2 is our reference population for interpretation of genetic results. To estimate the total genetic variation of an F_2, we need the parents and the F_1 (to estimate environmental effects) and the F_2 generation.

F_3 population mean:

$$\frac{1}{4}d$$

Variance among F_3 progeny means:

$$\sigma^2_{\bar{F}_3} = [\frac{1}{4}a^2 + \frac{1}{2}(\frac{1}{2}d)^2 + \frac{1}{4}(-a^2)] - (\frac{1}{4}d)^2)(\frac{1}{2}a^2 + \frac{1}{16}d^2)(\sigma^2_A + \frac{1}{4}\sigma^2_D$$

F3 Variances

Variance within F_3 progeny means:

$$\bar{\sigma}^2_{F_3} = [[\frac{1}{4}a^2 + \frac{1}{2}d^2 + \frac{1}{4}(-a^2)] - (\frac{1}{2}d)^2])(= \frac{1}{4}a^2 + \frac{1}{8}d^2)(\frac{1}{2}\sigma^2_A + \frac{1}{2}\sigma^2_D$$

Total variance among F_3 individuals is then:

$$\sigma^2_{F_3} = \frac{3}{2}\sigma^2_A + \frac{3}{4}\sigma^2_D$$

F_3 progenies can be grown in replicated trials, so a set of equations like the following could be written:

$$\sigma^2_{F_2} = \sigma^2_A + \sigma^2_D + \sigma^2_{E1})(\sigma^2_{\bar{F}_3} = \sigma^2_A + \frac{1}{4}\sigma^2_D + \sigma^2_{E2})(\bar{\sigma}^{\,2}_{F_3} = \frac{1}{2}\sigma^2_A + \frac{1}{2}\sigma^2_D + \sigma^2_{E1})where(\sigma^2_{E2} = \frac{\sigma^2}{r}$$

ANOVA for F3 Progenies

ANOVA for F_3 progenies can be calculated from a replicated experiment.

Table 4 ANOVA for F_3 Progenies.

Source	df	MS	EMS
Reps	$r-1$		
Progeny	$p-1$	M3	$\sigma_e^2 + r\sigma_{F_2}^2$
Error	$(r-1)(p-1)$	M2	σ^2
Total	$rp-1$		
Within Progeny	$rp(k-1)$	M1	$\sigma_{F_a}^2 = \frac{1}{2}\sigma_A^2 + \frac{1}{2}\sigma_D^2 + \sigma_{E1}^2$

Then:

Invalid Equation

Note that the phenotypic variance among F_3 families is:

$$\hat{\sigma}_p^2 = \frac{M3}{rk} = \frac{\sigma_w^2}{rk} + \frac{\sigma^2}{r} + \sigma_c^2$$

Estimate of Heritability

A type of heritability estimate on a progeny mean basis can be calculated as:

$$h^2 = \frac{\sigma_c^2}{\sigma_p^2} = \frac{\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2}{\frac{\sigma_w^2}{rk} + \frac{\sigma^2}{r} + \sigma_c^2}$$

Note that this estimate of heritability contains both additive and dominance variance. Recall that this is an estimate of intra-class correlation, thus it is a type of broad sense heritability.

Limitations of this method (often referred to as Mather's methods)

1. Estimates apply only to specific parents.
2. Estimates for σ^2_{E1} may vary among generations
3. Estimates for a particular set of F_2 plants can be obtained in only one environment
4. Linkage will bias estimates
5. Epistasis is assumed to be absent

Bi-Parental Progenies

Bi-parental progenies are just crosses between individual plants, thus genetically they are full-sib. For example, in a random mating maize population you could cross two individual plants reciprocally and bulk the seed from the two ears. This would produce enough seed to plant FS progeny in 10-20 replications. We could then think about *n* plants and making *n* / 2 full-sib families.

Table 5

Source	df	MS	EMS
Reps	$r-1$		
Among families	$\frac{n}{2}-1$	$M3\ \sigma_w^2+k\sigma^2+rk\sigma_c^2$	$\sigma_w^2+k\sigma^2+rk[Cov(FS)]$
Error	$(r-1)(\frac{n}{2}-1)$	$M2\ \sigma_w^2+k\sigma^2$	$\sigma_w^2+k\sigma^2$
Total	$r\frac{n}{2}-1$		
Within families	$r\frac{n}{2}(k-1)$	$M1\ \sigma_w^2$	$\sigma_w^2+[\sigma_G^2-Cov(FS)]$

Divide Error and Total in table with a horizontal line.

Invalid Equation

Summary

Table 6 Data from Cockerham, 1993.

Progeny Type	**Cov(progeny)**	σ_T^2
Half-sib	$\frac{1}{4}\sigma_A^2$	$\sigma_A^2 + \sigma_D^2$
Full-sib	$\frac{1}{2}\sigma_A^2 + \frac{1}{4}\sigma_D^2$	$\sigma_A^2 + \sigma_D^2$
$S_1(F_{2:3})$	$\sigma_A^2 + \frac{1}{4}\sigma_D^2 + D_1 + \frac{1}{8}D_2$	$\frac{3}{2}\sigma_A^2 + \frac{1}{2}\sigma_D^2 + 2D_1 + \frac{1}{2}D_2 + \frac{1}{4}H^*$
$S_2(F_{3:4})$	$\frac{3}{2}\sigma_A^2 + \frac{1}{8}\sigma_D^2 + 2.5D_1 + \frac{9}{16}D^2 + \frac{1}{16}H^*$	$\frac{7}{4}\sigma_A^2 + \frac{1}{4}\sigma_D^2 + 3D_1 + \frac{3}{4}D_2 + \frac{3}{16}H$
$S_n(F_{4:5})$	$\frac{7}{4}\sigma_A^2 + \frac{1}{16}\sigma_D^2 + 3.25D_1 + \frac{25}{32}D_2 + \frac{3}{64}H^*$	$\frac{15}{8}\sigma_A^2 + \frac{1}{8}\sigma_D^2 + 3.5D_1 + \frac{7}{8}D_2 + \frac{7}{6}$
S_∞	$2\sigma_A^2 + 4D_1 + D_2$	$2\sigma_A^2 + 4D_1 + D_2$

Expected Mean Squares

The AOV tables cannot be interpreted without understanding the expected sources of variability represented by the Mean Squares. In the case of balanced field plot designs with only a few sources of variation the expected mean squares are easily determined. If a particular design involves many sources of random and fixed factors, students have found the approach of Lorenzen and Anderson (1993, Design of Experiments: A No-Name Approach. p 71-72) to be useful.

1. Write the terms of the model with associated subscripts down the left side of the page. Across the top write the single letter subscripts (i,j,k, etc.). Above each subscript place either F or R if the factor associated with that transcript is fixed or random. Above that place the number of levels associated with that subscript (I,J,K, etc.).
2. Enter a 1 in every slot where the subscript at the top is contained within brackets in the term at the left.
3. Enter a 0 in every slot where the subscript at the top is fixed and also contained in the term as the left. Enter a 1 in every slot where the subscript at the top is random and also contained in the terms at the left.
4. Fill in the remaining slots with the number of levels at the top of each column.
5. To compute the Expected Mean Squares (EMS) for a given term having df > 0, start at the bottom and work up. Only consider terms whose indices include all the indices in the term whose EMS you are deriving. Compute the coefficient of this term by covering the columns corresponding to the indices in the term whose EMS you are deriving and multiplying the values in the remaining columns. If there is a 0 column that is not covered, this term need not be written in the EMS. A factor is considered fixed and denoted with a Φ only if all of its indices are fixed. Otherwise it is considered random and denoted by the appropriate σ^2 term.

Using the Algorithm

Notice that this algorithm can be used to compute EMS for all terms in the model, including those that have zero df. A term that has zero df has no expected mean squares. For this reason, we will not compute EMS for terms having zero df even though such terms are in the algorithm to make the EMS of the other terms come out right. Note that this simple algorithm for determining the EMS in an AOV assumes that the data are balanced, i.e., each of the sources of variability (model parameters) have data for all levels, i, j, and k.

To illustrate, let's consider a slightly more complex, but typical RCBD design used by plant breeders to evaluate many genotypes grown in replicates at several environments for purposes of identifying and discarding poor performing genotypes in a cultivar development project.

Step 1

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k}$$

Notice that this algorithm can be used to compute EMS for all terms in the model, including those that have zero df. A term that has zero df has no expected mean squares. For this reason, we will not compute EMS for terms having zero df even though such terms are in the algorithm to make the EMS of the other terms come out right. Note that this simple algorithm for determining the EMS in an AOV assumes that the data are balanced, i.e., each of the sources of variability (model parameters) have data for all levels, i, j, and k.

To illustrate, let's consider a slightly more complex, but typical RCBD design used by plant breeders to evaluate many genotypes grown in replicates at several environments for purposes of identifying and discarding poor performing genotypes in a cultivar development project.

Step 1

Write the terms of the model with associated subscripts down the left side of the page. Across the top write the single letter subscripts (i,j,k, etc.). Above each subscript place either F or R if the factor associated with that transcript is fixed or random. Above that place the number of levels associated with that subscript (I,J,K, etc.).

Factors:

Factor E – Fixed
Factor G – Random
Blocks – Random

Source	E	G	R	EMS

	F	R	R	
	i	j	k	
E_i				
$B(E)_{(i)k}$				
G_j				
GE_{ij}				
$\varepsilon_{(ij)k}$				

Step 2

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k}$$

Step 2

Enter a 1 in every slot where the subscript at the top is contained within brackets in the term at the left.

Factors:

Factor E – Fixed
Factor G – Random
Blocks – Random

Source	**E**	**G**	**R**	**EMS**
	F	**R**	**R**	
	i	**j**	**k**	
E_i				
$B(E)_{(i)k}$			1	
G_j		1		
GE_{ij}	1	1		
$\varepsilon_{(ij)k}$	1	1	1	

Step 3

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k}$$

Step 3

Enter a 0 in every slot where the subscript at the top is fixed and also contained in the term as the left. Enter a 1 in every slot where the subscript at the top is random and also contained in the terms at the left.

Factors:

Factor E – Fixed
Factor G – Random
Blocks – Random

Source	**E**	**G**	**R**	**EMS**
	F	**R**	**R**	
	i	**j**	**k**	
E_i	0			
$B(E)_{(i)k}$	0		1	
G_j		1		
GE_{ij}	1	1		
$\varepsilon_{(ij)k}$	1	1	1	

Step 4

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k}$$

Step 4

Fill in the remaining slots with the number of levels at the top of each column.

Factors:

Factor E – Fixed
Factor G – Random
Blocks – Random

Source	**E**	**G**	**R**	**EMS**
	F	**R**	**R**	
	i	**j**	**k**	
E_i	0	G	R	
$B(E)_{(i)k}$	0	G	1	
G_j	E	1	R	
GE_{ij}	1	1	R	
$\varepsilon_{(ij)k}$	1	1	1	

Step 5

The phenotype Y for this typical field trial will be something like:

$$Y_{ijk} = \mu + E_i + B(E)_{(i)k} + G_j + GE_{ij} + \varepsilon_{(ij)k}$$

Step 5

To compute the EMS for a given term having df > 0, start at the bottom and work up. Only consider terms whose indices include all the indices in the term whose EMS you are deriving. Compute the coefficient of this term by covering the columns corresponding to the indices in the term whose EMS you are deriving and multiplying the values in the remaining columns.

If there is a 0 column that is not covered, this term need not be written in the EMS. A factor is considered fixed and denoted with a Φ only if all of its indices are fixed. Otherwise it is considered random and denoted by the appropriate σ^2 term.

Factors:

Factor E – Fixed

Factor G – Random

Blocks – Random

Source	**E**	**G**	**R**	**EMS**
	F	**R**	**R**	
	I	**j**	**k**	
E_i	0	G	R	$\sigma + G\sigma^2_{B(E)} + \sigma^2_E$
$B(E)_{(i)k}$	0	G	1	$\sigma^2 + G\sigma^2_{B(E)}$
G_j	E	1	R	$\sigma^2 + R\sigma^2_{GE} + RE\sigma^2_G$
GE_{ij}	1	1	R	$\sigma^2 + R\sigma^2_{GE}$
$\varepsilon_{(ij)k}$	1	1	1	σ^2

References

Cockerham, C.C. 1983. Covariances of relatives from self-fertilization. Crop Sci. 23: 1177-1180.

Acknowledgements

This module was developed as part of the Bill & Melinda Gates Foundation Contract No. 24576 for Plant Breeding E-Learning in Africa.

Quantitative Genetics Estimates of Variance Author: William Beavis, Kendall Lamkey, and Katherine Espinosa (ISU)

Multimedia Developers: Gretchen Anderson, Todd Hartnell, and Andy Rohrback (ISU)

How to cite this module: Beavis, W., K. Lamkey, and K. Espinosa. 2016. Estimates of Variance. *In* Quantitative Genetics, interactive e-learning courseware. Plant Breeding E-Learning in Africa. Retrieved from **https://pbea.agron.iastate.edu**.