www.statsoft.com

- Products
- Solutions
- Buy
- Trials
- Support

TextbookVariance Components Mixed Model ANOVA ANCOVA

Looking for info about statistics?

We wrote the book on it.
And you can read it for free!

Variance Components and Mixed Model ANOVA/ANCOVA

- Basic Ideas
  - Properties of Random Effects
- Estimation of Variance Components (Technical Overview)
  - Estimating the Variation of Random Factors
  - Estimating Components of Variation
  - Testing the Significance of Variance Components
  - Estimating the Population Intraclass Correlation

The *Variance Components and Mixed Model ANOVA/ANCOVA* section describes a comprehensive set of techniques for analyzing research designs that include random effects; however, these techniques are also well suited for analyzing large main effect designs (e.g., designs with more than 200 levels per factor), designs with many factors where the higher order interactions are not of interest, and analyses involving case weights.

There are several sections in this electronic textbook that discuss Analysis of Variance for factorial or specialized designs. For a discussion of these sections and the types of designs for which they are best suited, refer to Methods for Analysis of Variance. Note, however, that General Linear Models describes how to analyze designs with any number and type of between effects and compute ANOVA-based variance component estimates for any effect in a mixed-model analysis.

Basic Ideas

Experimentation is sometimes mistakenly thought to involve only the manipulation of levels of the independent variables and the observation of subsequent responses on the dependent variables. Independent variables whose levels are determined or set by the experimenter are said to have *fixed effects*. There is a second class of effects, however, that is often of great interest to the researcher; *random effects* are classification effects where the levels of the effects are assumed to be randomly selected from an infinite population of possible levels.

Many independent variables of research interest are not fully amenable to experimental manipulation, but nevertheless can be studied by considering them to have random effects. For

example, the genetic makeup of individual members of a species cannot at present be (fully) experimentally manipulated, yet it is of great interest to the geneticist to assess the genetic contribution to individual variation on outcomes such as health, behavioral characteristics, and the like. As another example, a manufacturer might want to estimate the components of variation in the characteristics of a product for a random sample of machines operated by a random sample of operators. The statistical analysis of random effects is accomplished by using the *random effect model* if all of the independent variables are assumed to have random effects, or by using the *mixed model* if some of the independent variables are assumed to have random effects and other independent variables are assumed to have fixed effects.

Properties of random effects. To illustrate some of the properties of *random effects*, suppose you collected data on the amount of insect damage to different varieties of wheat. It is impractical to study insect damage for every possible variety of wheat, so to conduct the experiment, you randomly select four varieties of wheat to study. Plant damage is rated for up to a maximum of four plots per variety. Ratings are on a 0 (no damage) to 10 (great damage) scale. The following data for this example are presented in Milliken and Johnson (1992, p. 237).

DATA: wheat.sta 3v

VARIETY	PLOT	DAMAGE
A	1	3.90
A	2	4.05
A	3	4.25
B	4	3.60
B	5	4.20
B	6	4.05
B	7	3.85
C	8	4.15
C	9	4.60
C	10	4.15
C	11	4.40
D	12	3.35
D	13	3.80

To determine the components of variation in resistance to insect damage for *Variety* and *Plot*, an ANOVA can first be performed. Perhaps surprisingly, in the ANOVA, *Variety* can be treated as a fixed or as a random factor without influencing the results (provided that *Type I Sums of squares* are used and that *Variety* is always entered first in the model). The spreadsheet below shows the ANOVA results of a mixed model analysis treating *Variety* as a *fixed effect* and ignoring *Plot*, i.e., treating the plot-to-plot variation as a measure of random error.

ANOVA Results: DAMAGE (wheat.sta)

Effect	Effect (F/R)	df Effect	MS Effect	df Error	MS Error	F	p
{1}VARIETY	Fixed	3	.270053	9	.056435	4.785196	.029275

Another way to perform the same mixed model analysis is to treat *Variety* as a *fixed effect* and *Plot* as a *random effect*. The spreadsheet below shows the ANOVA results for this mixed model analysis.

ANOVA Results for Synthesized Errors: DAMAGE (wheat.sta)

	df error computed using Satterthwaite method						
Effect	Effect (F/R)	df Effect	MS Effect	df Error	MS Error	F	p
{1}VARIETY	Fixed	3	.270053	9	.056435	4.785196	.029275
{2}PLOT	Random	9	.056435	-----	-----	-----	-----

The spreadsheet below shows the ANOVA results for a *random effect* model treating *Plot* as a *random effect* nested within *Variety*, which is also treated as a *random effect*.

ANOVA Results for Synthesized Errors: DAMAGE (wheat.sta)

	df error computed using Satterthwaite method

Effect	Effect (F/R)	df Effect	MS Effect	df Error	MS Error	F	p
{1}VARIETY	Random	3	.270053	9	.056435	4.785196	.029275
{2}PLOT	Random	9	.056435	-----	-----	-----	-----

As can be seen, the tests of significance for the *Variety* effect are identical in all three analyses (and in fact, there are even more ways to produce the same result). When components of variance are estimated, however, the difference between the mixed model (treating *Variety* as fixed) and the random model (treating *Variety* as random) becomes apparent. The spreadsheet below shows the *variance component* estimates for the mixed model treating *Variety* as a *fixed effect*.

Components of Variance (wheat.sta)	
	Mean Squares Type: 1
Source	DAMAGE
{2}PLOT	.056435
Error	0.000000

The spreadsheet below shows the *variance component* estimates for the *random effects* model treating *Variety* and *Plot* as *random effects*.

Components of Variance (wheat.sta)	
	Mean Squares Type: 1
Source	DAMAGE
{1}VARIETY	.067186
{2}PLOT	.056435
Error	0.000000

As can be seen, the difference in the two sets of estimates is that a *variance component* is estimated for *Variety* only when it is considered to be a *random effect*. This reflects the basic distinction between *fixed* and *random effects*. The variation in the levels of random factors is assumed to be representative of the variation of the whole population of possible levels. Thus, variation in the levels of a random factor can be used to estimate the population variation. Even more importantly, covariation between the levels of a random factor and responses on a dependent variable can be used to estimate the population component of variance in the dependent variable attributable to the random factor. The variation in the levels of fixed factors is instead considered to be arbitrarily determined by the experimenter (i.e., the experimenter can make the levels of a fixed factor vary as little or as much as desired). Thus, the variation of a fixed factor cannot be used to estimate its population variance, nor can the population covariance with the dependent variable be meaningfully estimated. With this basic distinction between *fixed effects* and *random effects* in mind, we now can look more closely at the properties of *variance components*.

Estimation of Variance Components (Technical Overview)

The basic goal of *variance component* estimation is to estimate the population covariation between random factors and the dependent variable. Depending on the method used to estimate variance components, the population variances of the random factors can also be estimated, and significance tests can be performed to test whether the population covariation between the random factors and the dependent variable are nonzero.

Estimating the variation of random factors. The ANOVA method provides an integrative approach to estimating variance components, because ANOVA techniques can be used to estimate the variance of random factors, to estimate the components of variance in the dependent variable attributable to the random factors, and to test whether the variance components differ significantly from zero. The ANOVA method for estimating the variance of the random factors begins by constructing the *Sums of squares and cross products* (*SSCP*) matrix for the independent variables. The sums of squares and cross products for the random effects are then residualized on the *fixed effects*, leaving the *random effects* independent of the *fixed effects*, as required in the mixed model (see, for example, Searle, Casella, & McCulloch, 1992). The residualized *Sums of squares and cross products* for each random factor are then divided by their degrees of freedom

to produce the coefficients in the *Expected mean squares* matrix. Nonzero off-diagonal coefficients for the random effects in this matrix indicate confounding, which must be taken into account when estimating the population variance for each factor. For the *wheat.sta* data, treating both *Variety* and *Plot* as random effects, the coefficients in the Expected mean squares matrix show that the two factors are at least somewhat confounded. The *Expected mean squares* spreadsheet is shown below.

Expected Mean Squares (wheat.sta)				
	Mean Squares Type: 1			
Source	Effect (F/R)	VARIETY	PLOT	Error
{1}VARIETY	Random	3.179487	1.000000	1.000000
{2}PLOT	Random		1.000000	1.000000
Error				1.000000

The coefficients in the *Expected mean squares* matrix are used to estimate the population variation of the random effects by equating their variances to their expected mean squares. For example, the estimated population variance for *Variety* using *Type I Sums of squares* would be 3.179487 times the *Mean square* for *Variety* plus 1 times the *Mean square* for *Plot* plus 1 times the *Mean square* for *Error*.

The ANOVA method provides an integrative approach to estimating *variance components*, but it is not without problems (i.e., ANOVA estimates of *variance components* are generally *biased*, and can be *negative*, even though variances, by definition, must be either zero or positive). An alternative to ANOVA estimation is provided by *maximum likelihood* estimation. Maximum likelihood methods for estimating variance components are based on quadratic forms and typically, but not always, require iteration to find a solution. Perhaps the simplest form of maximum likelihood estimation is *MIVQUE(0)* estimation. MIVQUE(0) produces Minimum Variance Quadratic Unbiased Estimators (i.e., MIVQUE). In MIVQUE(0) estimation, there is no weighting of the *random effects* (thus the 0 [zero] after MIVQUE), so an iterative solution for estimating variance components is not required. MIVQUE(0) estimation begins by constructing the *Quadratic sums of squares (SSQ)* matrix. The elements for the random effects in the *SSQ* matrix can most simply be described as the sums of squares of the sums of squares and cross products for each random effect in the model (after residualization on the *fixed effects*). The elements of this matrix provide coefficients, similar to the elements of the *Expected Mean Squares* matrix, which are used to estimate the covariances among the random factors and the dependent variable. The *SSQ* matrix for the *wheat.sta* data is shown below. Note that the nonzero off-diagonal element for *Variety* and *Plot* again shows that the two random factors are at least somewhat confounded.

MIVQUE(0) Variance Component Estimation (wheat.sta)				
	SSQ Matrix			
Source	VARIETY	PLOT	Error	DAMAGE
{1}VARIETY	31.90533	9.53846	9.53846	2.418964
{2}PLOT	9.53846	12.00000	12.00000	1.318077
Error	9.53846	12.00000	12.00000	1.318077

Restricted Maximum Likelihood (REML) and *Maximum Likelihood (ML) variance component* estimation methods are closely related to MIVQUE(0). In fact, in some programs, *REML* and *ML* use *MIVQUE(0)* estimates as start values for an iterative solution for the *variance components*, so the elements of the *SSQ* matrix serve as initial estimates of the covariances among the random factors and the dependent variable for both *REML* and *ML*.

Estimating components of variation. For ANOVA methods for estimating *variance components*, a solution is found for the system of equations relating the estimated population variances and covariances among the random factors to the estimated population covariances between the random factors and the dependent variable. The solution then defines the *variance components*. The spreadsheet below shows the *Type I Sums of squares* estimates of the *variance components* for the *wheat.sta* data.

Components of Variance (wheat.sta)	
	Mean Squares Type: 1
Source	DAMAGE
{1}VARIETY	0.067186
{2}PLOT	0.056435
Error	0.000000

MIVQUE(0) variance components are estimated by inverting the partition of the *SSQ* matrix that does not include the dependent variable (or finding the generalized inverse, for singular matrices), and postmultiplying the inverse by the dependent variable column vector. This amounts to solving the system of equations that relates the dependent variable to the random independent variables, taking into account the covariation among the independent variables. The MIVQUE(0) estimates for the *wheat.sta* data are listed in the spreadsheet shown below.

MIVQUE(0) Variance Component Estimation (wheat.sta)	
	Variance Components
Source	DAMAGE
{1}VARIETY	0.056376
{2}PLOT	0.065028
Error	0.000000

REML and ML variance components are estimated by iteratively optimizing the parameter estimates for the effects in the model. REML differs from ML in that the likelihood of the data is maximized only for the *random effects*, thus REML is a restricted solution. In both REML and ML estimation, an iterative solution is found for the weights for the random effects in the model that maximizes the likelihood of the data. Some programs use MIVQUE(0) estimates as the start values for both REML and ML estimation, so the relation among these three techniques is close indeed. The statistical theory underlying *maximum likelihood* variance component estimation techniques is an advanced topic (Searle, Casella, & McCulloch, 1992, is recommended as an authoritative and comprehensive source). Implementation of maximum likelihood estimation algorithms, furthermore, is difficult (see, for example, Hemmerle & Hartley, 1973, and Jennrich & Sampson, 1976, for descriptions of these *algorithms*), and faulty implementation can lead to variance component estimates that lie outside the parameter space, converge prematurely to nonoptimal solutions, or give nonsensical results. Milliken and Johnson (1992) noted all of these problems with the commercial software packages they used to estimate variance components.

The basic idea behind both REML and ML estimation is to find the set of weights for the *random effects* in the model that minimize the negative of the natural logarithm times the likelihood of the data (the likelihood of the data can vary from zero to one, so minimizing the negative of the natural logarithm times the likelihood of the data amounts to maximizing the probability, or the likelihood, of the data). The logarithm of the REML likelihood and the REML variance component estimates for the *wheat.sta* data are listed in the last row of the *Iteration history* spreadsheet shown below.

Iteration History (wheat.sta)			
	Variable: DAMAGE		
Iter.	Log LL	Error	VARIETY
1	-2.30618	.057430	.068746
2	-2.25253	.057795	.073744
3	-2.25130	.056977	.072244
4	-2.25088	.057005	.073138
5	-2.25081	.057006	.073160
6	-2.25081	.057003	.073155
7	-2.25081	.057003	.073155

The logarithm of the ML likelihood and the ML estimates for the variance components for the *wheat.sta* data are listed in the last row of the *Iteration history* spreadsheet shown below.

Iteration History (wheat.sta)			
	Variable: DAMAGE		
Iter.	Log LL	Error	VARIETY

1	-2.53585	.057454	.048799
2	-2.48382	.057427	.048541
3	-2.48381	.057492	.048639
4	-2.48381	.057491	.048552
5	-2.48381	.057492	.048552
6	-2.48381	.057492	.048552

As can be seen, the estimates of the variance components for the different methods are quite similar. In general, components of variance using different estimation methods tend to agree fairly well (see, for example, Swallow & Monahan, 1984).

Testing the significance of variance components. When maximum likelihood estimation techniques are used, standard linear model significance testing techniques may not be applicable. ANOVA techniques such as decomposing sums of squares and testing the significance of effects by taking ratios of mean squares are appropriate for linear methods of estimation, but generally are not appropriate for quadratic methods of estimation. When ANOVA methods are used for estimation, standard significance testing techniques can be employed, with the exception that any confounding among random effects must be taken into account.

To test the significance of effects in mixed or random models, error terms must be constructed that contain all the same sources of random variation except for the variation of the respective effect of interest. This is done using Satterthwaite's method of denominator synthesis (Satterthwaite, 1946), which finds the linear combinations of sources of random variation that serve as appropriate error terms for testing the significance of the respective effect of interest. The spreadsheet below shows the coefficients used to construct these linear combinations for testing the *Variety* and *Plot* effects.

Denominator Synthesis: Coefficients (MS Type: 1) (wheat.sta)				
	The synthesized MS Errors are linear combinations of the resp. MS effects			
Effect	(F/R)	VARIETY	PLOT	Error
{1}VARIETY	Random		1.000000	
{2}PLOT	Random			1.000000

The coefficients show that the *Mean square* for *Variety* should be tested against the *Mean square* for *Plot*, and that the *Mean square* for *Plot* should be tested against the *Mean square* for *Error*. Referring back to the *Expected mean squares* spreadsheet, it is clear that the denominator synthesis has identified appropriate error terms for testing the *Variety* and *Plot* effects. Although this is a simple example, in more complex analyses with various degrees of confounding among the random effects, the denominator synthesis can identify appropriate error terms for testing the random effects that would not be readily apparent.

To perform the tests of significance of the random effects, ratios of appropriate Mean squares are formed to compute *F* statistics and *p*-values for each effect. Note that in complex analyses, the degrees of freedom for random effects can be fractional rather than integer values, indicating that fractions of sources of variation were used in synthesizing appropriate error terms for testing the random effects. The spreadsheet displaying the results of the ANOVA for the *Variety* and *Plot random effects* is shown below. Note that, for this simple design, the results are identical to the results presented earlier in the spreadsheet for the ANOVA treating *Plot* as a *random effect* nested within *Variety*.

ANOVA Results for Synthesized Errors: DAMAGE (wheat.sta)							
	df error computed using Satterthwaite method						
Effect	Effect (F/R)	df Effect	MS Effect	df Error	MS Error	F	p
{1}VARIETY	Fixed	3	.270053	9	.056435	4.785196	.029275
{2}PLOT	Random	9	.056435	-----	-----	-----	-----

As shown in the spreadsheet, the *Variety* effect is found to be significant at *p* < .05, but as would be expected, the *Plot* effect cannot be tested for significance because plots served as the basic unit of analysis. If data on samples of plants taken within plots were available, a test of the

significance of the *Plot* effect could be constructed.

Appropriate tests of significance for MIVQUE(0) variance component estimates generally cannot be constructed except in special cases (see Searle, Casella, & McCulloch, 1992). Asymptotic (large sample) tests of significance of REML and ML variance component estimates, however, can be constructed for the parameter estimates from the final iteration of the solution. The spreadsheet below shows the asymptotic (large sample) tests of significance for the REML estimates for the *wheat.sta* data.

Restricted Maximum Likelihood Estimates (wheat.sta)				
	Variable: DAMAGE -2*Log(Likelihood)=4.50162399			
Effect	Variance Comp.	Asympt. Std.Err.	Asympt. z	Asympt. p
{1}VARIETY	.073155	.078019	.937656	.348421
Error	.057003	.027132	2.100914	.035648

The spreadsheet below shows the asymptotic (large sample) tests of significance for the ML estimates for the *wheat.sta* data.

Maximum Likelihood Estimates (wheat.sta)				
	Variable: DAMAGE -2*Log(Likelihood)=4.96761616			
Effect	Variance Comp.	Asympt. Std.Err.	Asympt. z	Asympt. p
{1}VARIETY	.048552	.050747	.956748	.338694
Error	.057492	.027598	2.083213	.037232

It should be emphasized that the asymptotic tests of significance for REML and ML variance component estimates are based on large sample sizes, which certainly is not the case for the *wheat.sta* data. For this data set, the tests of significance from both analyses agree in suggesting that the Variety variance component does not differ significantly from zero.

For basic information on ANOVA in linear models, see also *Elementary Concepts*.

Estimating the population intraclass correlation. Note that if the *variance component* estimates for the random effects in the model are divided by the sum of all components (including the error component), the resulting percentages are population intraclass correlation coefficients for the respective effects.