

VARIANCE COMPONENTS: With a bioassay example

Other Names: Components of variation, sources of variance, variance analysis, intraclass correlation, random effects models, analysis of a nested data.

Acronyms: Restricted Maximum Likelihood (REML), Maximum Likelihood (ML), ANOVA, Minimum Variance Quadratic Unbiased Estimator (MIVQUE).

Related topics: ANOVA, assay validation (precision), sampling units, experimental units, maximum likelihood estimation.

Background: When observations are taken from sampled units (for example, runs of an assay), a common question is: how many samples are needed to achieve a certain precision (a specification on how small the standard deviation of the mean must be)? A common and more complex version of the question arises when trying to establish a population mean: how many samples should be taken from a population, and how many times should each sample be tested? There is variation among samples and variation in the measurement process itself (repeated testing of the same sample will show some variation). The goal of variance component estimation is to estimate the variance associated with each of these sources of variation. One can always decrease the standard deviation of the mean by taking more samples, a common practical question becomes, what is the most efficient (or cost-effective) multi-level sampling strategy.

Note: It generally requires more data to estimate a variance than to estimate a mean to the same level of precision. The familiar term "degrees of freedom" is a measure of the number of independent observations used to estimate a variance. When a sample is used to estimate a mean and a standard deviation, the standard deviation is estimated with $n - 1$ degrees of freedom. This is because the mean has been estimated from the data and the mean is used in estimating the standard deviation; hence there are $n - 1$ independent observations used to estimate the standard deviation.

Assumed: The observations are normally distributed (under some conditions this assumption can be relaxed) with each source of variance being constant for all subgroups (this may be true only after a transformation). The data are completely balanced; this means that all similar subgroups have the same numbers of observations (more complex methods allow estimation of variance components from unbalanced data).

Discussion: There are many settings where this type of analysis is appropriate: when constructing a sampling plan, when optimizing a measurement procedure, or when validating an assay procedure. A few examples may illustrate the value of consideration of at least two levels of variance components

When counting insects on corn one may be interested in estimating the mean number of insects/plant. If the insects are spread fairly evenly among the leaves on an infested plant it does not help to sample all the leaves on each plant sampled, a sample of one or two leaves is sufficient. We describe this as a case where the leaf-to-leaf variation in insect numbers is small. On the other hand, if the insects are likely to be clustered on one or a few leaves, it will be cost-effective to sample more leaves from each plant sampled.

For an assay run in 96 well plates, with 3 replicates of each sample on each plate and several plates for each sample assayed, there may be three or more variance sources of variation (variance components) available: variation among replicates on a plate (typically ignored), variation among plates within a day, variation among technicians, variation among days and possibly variation among laboratories.

Consider a simple two-component case. When a bioassay is analyzed we typically can compute a relative potency estimate for each replicate and these can be combined to get an estimate for each plate. We are rarely interested in computing the relative potencies (or the variation in relative potencies) of individual replicates within a plate. It is probably familiar to think about the standard deviation of the estimates of the potency (or log potency; for the rest of this discussion we will work with log potency) from several plates within a day. The first variance component is then the variation in relative potency among plates within a day. When combining assays across several days a common and sensible procedure is to first form day averages (sometimes weighted averages) of the plates within each day. The overall average is then created by combining (sometimes using a weighted average) the estimates of log relative potency from each day. We can also study the variation in the daily estimates of relative potency. The second variance component is the variation in relative potency due to day-to-day variation. Notice that we cannot measure day-to-day variation directly because the variation within day is included in each day's measurements.

Goal: Separating the total variation observed from day to day into the within day component and the between day component is the primary goal in variance components analysis.

Background: Recall that the variance is denoted σ^2 and is estimated with s^2 ,

$$s^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1},$$

where the observed values are the y_i , for $i = 1 \dots n$, for n observations and the mean of the observed values is denoted $\bar{y} = \sum_{i=1}^n y_i / n$. For factors which are fixed we typically estimate the effect associated with each level of the factor, for instance, we would like to know how much far the average result from technician #2 is above the overall average. For other factors we are not interested in a specific level of the factor, such as plate # 47, and it is more useful to describe the typical variation among plates; these factors are called random factors. Some factors are not clearly fixed or random, such as days or labs. For different purposes we may treat a factor as fixed in one analysis and as random in another analysis. Variance components analysis is appropriate for factors which we are interested in studying as random factors.

Measurement: Each observation is now denoted y_{ij} , where i refers to the sample number and j refers to the measurement number for the i th sample.

Calculation: The preferred methods for calculation are the maximum likelihood (ML) method and the REML (restricted maximum likelihood) method. When the data are balanced, the ANOVA method will give estimates which match those given by the ML method. In this module calculations will be shown only for the ANOVA method.

Procedure

(1) Plan a nested experiment to collect variance components. Preserve balance in the design. Let the factor applied to the smallest unit have K levels, the factor applied to the next larger unit have J levels, and the factor applied to the next larger unit have I levels. In a three-level balanced nested design, each of the J mid-sized units will contain K of the smaller units, and each of the I large units will contain J of the mid-sized units. The observations are denoted y_{ijk} where i goes from 1 to I , j runs from 1 to J , and k runs from 1 to K . This example uses three different sizes of experimental units to illustrate what is needed for USP validation; variance components models can be fit with more or fewer levels.

Example

(1) Consider a biological assay which is run with a standard and a test compound on one 96 well plate. The assay outcome is a single measure of relative potency for a lot of test compound compared to the standard. We want to estimate the plate to plate variance (small unit) and the day to day variance (mid-sized unit) for a single technician and the laboratory to laboratory variance (large unit). One technician in each of three labs will run two assays/day on each of four days, obtaining 24 values of the relative potency. The same lot of standard material will be used in each of the 24 assays using standard at 1.5x for the "test" sample. The data: lab, day within lab, plate within lab and day, and log of the relative potency are given in the appendix.

(2) Construct an ANOVA table with one row for each source of variation. The degrees of freedom and mean squares for the rows are given by,

| Source | df | Sums of Squares |
|--------|-------------------|--|
| large | $(I - 1)$ | $\sum_{i=1}^I J * K * (\bar{y}_{i..} - \bar{y}_{...})^2$ |
| mid | $I * (J - 1)$ | $\sum_{i=1}^I \sum_{j=1}^J K * (\bar{y}_{ij.} - \bar{y}_{i..})^2$ |
| small | $I * J * (K - 1)$ | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (y_{ijk} - \bar{y}_{ij.})^2$ |

where the dot in the subscript of \bar{y} indicates which subscripts should be included in the average. For example, $\bar{y}_{ij.} = \sum_{k=1}^K y_{ijk}$, the mean of the K observations in the i, j th group.

(3) Construct the mean squares for each component. To get the mean squares divide the sums of squares in each row by the degrees of freedom in that row.

| Source | Mean Squares |
|--------|--|
| large | $\sum_{i=1}^I J * K * (\bar{y}_{i..} - \bar{y}_{...})^2 / (I - 1)$ |
| mid | $\sum_{i=1}^I \sum_{j=1}^J K * (\bar{y}_{ij.} - \bar{y}_{i..})^2 / I * (J - 1)$ |
| small | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (y_{ijk} - \bar{y}_{ij.})^2 / I * J * (K - 1)$ |

2) The data in the appendix will give the following ANOVA table

| Source | df | Sums of Squares |
|-----------------|----|-----------------|
| labs | 2 | 0.026557434 |
| days(lab) | 9 | 0.025110831 |
| plates(lab day) | 12 | 0.0098736926 |

The plates(lab day) term is sometimes called the error or residual term because it is often left out of the model. Note: the formulas given here are correct, but there are much better algorithms (faster and more accurate) for routine use.

(3) Construct the mean squares for each source of variation.

| Source | Mean Squares |
|-----------------|---------------|
| labs | 0.013278717 |
| days(lab) | 0.0027900923 |
| plates(lab day) | 0.00082280771 |

(4) Write down the expected values of each of the mean squares.

| Source | Expected Mean Squares |
|--------|---|
| large | $\sigma_{\text{small}}^2 + K*\sigma_{\text{mid}}^2 + J*K*\sigma_{\text{large}}^2$ |
| mid | $\sigma_{\text{small}}^2 + K*\sigma_{\text{mid}}^2$ |
| small | σ_{small}^2 |

(5) Set the expected mean squares equal to the observed mean squares and solve to get estimates of the variance components. If an estimate of a variance component is negative, the estimate is set to be zero.

$$\hat{\sigma}_{\text{small}}^2 = \text{MS}_{\text{small}}$$

$$\hat{\sigma}_{\text{mid}}^2 = (\text{MS}_{\text{mid}} - \text{MS}_{\text{small}}) / K$$

$$\hat{\sigma}_{\text{large}}^2 = (\text{MS}_{\text{large}} - \text{MS}_{\text{mid}}) / K * J$$

(4)

| Source | Expected Mean Squares |
|--------|--|
| large | $\sigma_{\text{plates}}^2 + 2*\sigma_{\text{days}}^2 + 8*\sigma_{\text{labs}}^2$ |
| mid | $\sigma_{\text{plates}}^2 + 2*\sigma_{\text{days}}^2$ |
| small | σ_{plates}^2 |

(5)

$$\hat{\sigma}_{\text{plate}}^2 = 0.00082280771$$

$$\hat{\sigma}_{\text{day}}^2 = (0.0027900923 - 0.00082280771) / 2$$

$$= 0.00098364230$$

$$\hat{\sigma}_{\text{labs}}^2 = (0.013278717 - 0.0027900923) / 8$$

$$= 0.0013110781$$

Cautions: The method in this module is appropriate only when the data are balanced. Here balanced means that there are the same number of samples for each group. If there are several levels of grouping, there must be balance at each level. There are methods for unbalanced data, these are discussed in detail in Searle, Casella and McCulloch.

The data used as the example here was generated by simulation. The actual values of the variances used for the simulation are compared to the estimates in the following table. The fact that these estimates are not close to the true variances illustrates the point that it is difficult to estimate variances. This experiment would have given much better estimates if more labs could have been used.

| Source | True | Estimated |
|--------|--------|-----------|
| Labs | 0.0009 | 0.001311 |
| Days | 0.0004 | 0.000984 |
| Plates | 0.0001 | 0.000823 |

Advice: Design a fully balanced study specifically for estimation of variance components. Unless the cost is totally prohibitive, it is wise to have a minimal number of replicates at the low level (for an assay these would be determinations or samples within a run) and as many replicates as possible at the high level (for an assay this would be laboratories). You must have at least two levels for any factor for which you want to estimate a variance component.

In the language
of mathematics:

We begin with a model statement for a two-level nested design,

$$y_{ij} = \mu + A_i + \varepsilon_{ij}$$

where y_{ij} is the observed value, μ is the overall mean, A_i is the deviation from the mean for the i th group ($\mu + A_i$ is the mean of the i th group), and ε_{ij} is the deviation from the group mean for the ij th measurement. We typically assume that the A_i are normally distributed with mean zero and variance σ_g^2 (often denoted $A_i \sim N(0, \sigma_g^2)$), the $\varepsilon_{ij} \sim N(0, \sigma^2)$, and that the A_i are independent of the ε_{ij} (if we are interested in the A_i effects then we use the same model as a fixed effects model, we do not assume that the A_i come from a normal distribution and we are not interested in, and do not estimate σ_g^2).

The example used earlier has three sources of variation, and would have a model statement,

$$y_{ijk} = \mu + A_i + B_{ij} + \varepsilon_{ijk}$$

where y_{ijk} is the observed value, μ is the overall mean, A_i is the deviation from the mean for the i th group ($\mu + A_i$ is the mean of the i th group), B_{ij} is the deviation for the j th subgroup in the i th group, and ε_{ijk} is the deviation from the group mean for the ijk th measurement. We typically assume that the A_i are normally distributed with mean zero and variance σ_g^2 (often denoted $A_i \sim N(0, \sigma_g^2)$), the $B_{ij} \sim N(0, \sigma_h^2)$, and $\varepsilon_{ijk} \sim N(0, \sigma^2)$, and that the A_i , B_{ij} and ε_{ijk} are all independent.

Returning to the two sources of variation model, the group average for group i is \bar{y}_i .

$$\bar{y}_i = \mu + A_i + \bar{\varepsilon}_i.$$

and has variance

$$\begin{aligned} \text{VAR}(\bar{y}_{i.}) &= \text{VAR}(\mu) + \text{VAR}(A_i) + \frac{\text{VAR}(\varepsilon_{ij})}{J} \\ &= \sigma_g^2 + \frac{\sigma^2}{J}. \end{aligned}$$

While the overall mean $\bar{y}_{..}$ which can be written as

$$\bar{y}_{..} = \mu + \bar{A}_{.} + \bar{\varepsilon}_{..}$$

and has variance given by

$$\begin{aligned} \text{VAR}(\bar{y}_{..}) &= \frac{\text{VAR}(A_i)}{I} + \frac{\text{VAR}(\varepsilon_{ij})}{I*J} \\ &= \frac{\sigma_g^2 + \frac{\sigma^2}{J}}{I}. \end{aligned}$$

The expected values of the ANOVA mean squares, given in the following table, are then set equal to the mean squares to solve for estimates of σ^2 and σ_g^2 .

| Source | df | Mean Square | E(Mean Square) |
|--------|------------|-------------|--------------------------|
| groups | $I - 1$ | MSG | $\sigma^2 + J\sigma_g^2$ |
| error | $I(J - 1)$ | MSE | σ^2 |

Our ANOVA or method of moments estimators are then

$$\hat{\sigma}^2 = MSE$$

and

$$\hat{\sigma}_g^2 = (MSG - MSE)/J.$$

Computer programs: Both the ML and REML methods require iterative computations; these computational methods are offered by the SAS[®] procedures VARCOMP, GLM and MIXED, BMDP[®] procedures, the S-PLUS[®] function LME (written by Lindstrom, Bates and Pinheiro and available from statlib: lib.stat.cmu.edu) and are available in other statistics computing packages. The estimation of variance components from unbalanced data requires help from a statistician and high quality software.

Regulatory: A very important part of an assay validation is documentation of the sizes of the variance components for 1) sample to sample variation within a day for one technician, 2) day to day and/or technician to technician variation within a laboratory, and possibly 3) variation from laboratory to laboratory. The United States Pharmacopeia (USP), which is a regulatory document, describes these sources of variation in terms of their variance components

"The precision of an analytical method is usually expressed as the standard deviation or relative standard deviation. Precision may be a measure of either the degree of reproducibility or of repeatability of the analytical method under normal operating conditions. In this context, reproducibility refers to the use of the analytical procedure in different laboratories. Intermediate precision expresses within-laboratory variation, as on different days, or with different analysts or equipment within the same laboratory. Repeatability refers to the use of the analytical procedure within a laboratory over a short period of time using the same analyst with the same equipment. For most purposes, repeatability is the criterion of concern in USP analytical procedures"

References: SAS/STAT Users Guide, Version 6, First Edition, Volume 2. SAS Institute Inc., SAS Circle, Box 8000, Cary, NC 27512-8000.

Snedecor, George W. and Cochran, William G (1980) *Statistical Methods, 7th Edition*, Iowa State University Press.

Searle, S. R.; Casella, George and McCulloch, Charles E. (1992) *Variance Components*, Wiley.

United States pharmacopeia XXIII/National Formulary XVIII, The United States Pharmacopeial Convention, Inc., Rockville, Maryland, 1994.

Appendix: Example Data

| <u>lab</u> | <u>day(lab)</u> | <u>plate(day lab)</u> | <u>log(R)</u> |
|------------|-----------------|-----------------------|---------------|
| 1 | 1 | 1 | 0.21149458 |
| 1 | 1 | 2 | 0.25198847 |
| 1 | 2 | 1 | 0.22796501 |
| 1 | 2 | 2 | 0.15120732 |
| 1 | 3 | 1 | 0.15436127 |
| 1 | 3 | 2 | 0.14955904 |
| 1 | 4 | 1 | 0.13702753 |
| 1 | 4 | 2 | 0.14785860 |
| 2 | 1 | 1 | 0.14981953 |
| 2 | 1 | 2 | 0.12328865 |
| 2 | 2 | 1 | 0.16958164 |
| 2 | 2 | 2 | 0.16421857 |
| 2 | 3 | 1 | 0.16117746 |
| 2 | 3 | 2 | 0.19998647 |
| 2 | 4 | 1 | 0.16050133 |
| 2 | 4 | 2 | 0.18123937 |
| 3 | 1 | 1 | 0.18403708 |
| 3 | 1 | 2 | 0.15972328 |
| 3 | 2 | 1 | 0.19529639 |
| 3 | 2 | 2 | 0.14084238 |
| 3 | 3 | 1 | 0.18880396 |
| 3 | 3 | 2 | 0.15006891 |
| 3 | 4 | 1 | 0.18251738 |
| 3 | 4 | 2 | 0.24847079 |