

BIOL 458 BIOMETRY

Lab 8 - Nested and Repeated Measures ANOVA

PART 1: NESTED ANOVA

Nested designs are used when levels of one factor are not represented within all levels of another factor. Often this is because there is no alternative. For instance, if we were concerned with the effects of acid rain on productivity in British and American lakes, we might select at random 5 lakes in each country and make 10 productivity measurements at the surface. The lakes would constitute a random effect while country would be a fixed effect. However, each lake does not occur in both countries, so lake is, necessarily, nested within country. Such a design confounds the lake by country interaction since to estimate the interaction would require measurements of each lake within both countries, which is impossible. In such a situation, one analyzes the data as if they represent a fully factorial design with all factors completely crossed, but then the interaction term (lake by country in this case) is pooled with the effect of the nested factor (lake in this case), and the country main effect is tested over the effect of the nested factor pooled with the interaction.

Source	df	MS	F
Country	2 - 1	$SS_{\text{Country}}/df_{\text{Country}}$	$MS_{\text{Country}}/MS_{\text{Lake + Lake by Country}}$
Lake + Lake by Country	2(5 - 1)	$SS_{\text{Lake + Lake by Country}}/df_{\text{Lake + Lake by Country}}$	
With Cells (error)	2 x 5 (10 - 1)	$SS_{\text{error}}/df_{\text{error}}$	

The appropriate F ratio is $MS_{\text{country}}/MS_{\text{Lake + Lake by Country}}$. Alternatively, if factor A is the non-nested factor and factor B is the nested factor, the F ratio to test for the effect of the non-nested factor is $F = MS_A/MS_{B(A)}$, where B(A) connotes the effect of factor B nested within factor A and is equal to the sum of main effect of Factor B and the interaction of Factor A and B that one would obtain from treating the data as being generated by a fully crossed ANOVA design. The results of this analysis should be equivalent to an analysis in which the values in each lake were first averaged and a one-way ANOVA was performed on the averages to test for a country effect.

SPSS General Linear Model - Univariate can be used to calculate the sums - of - squares for the crossed design, and then the pooling of sums - of - squares and degrees of freedom, and the appropriate F ratio must then be calculated by hand. Note that for this approach to work, the five lakes in Britain and the five lakes in the USA would each need to be coded 1-5. Using the syntax window in SPSS, one could calculate the nested ANOVA directly. The factor levels for the nested factor should be identified by unique integers (1-10, 1-5 for Britain and 6-10 for USA, in the above

example), and the **SPSS** syntax would be (where dependent, nested-factor, and non-nested factor are the respective variable names)

glm dependent var by non-nested factor, nested factor

/design non-nested factor, nested factor within non-nested factor.

PART 2: REPEATED MEASURES ANOVA

Within subjects designs are used in agricultural and psychological research and have many applications in biology. These designs are called *split-plot*, *repeated measures*, or *generically within subjects*. The primary purpose of these designs is to eliminate uncontrolled variation due to *a priori* differences in primary sampling units from the estimate of experimental error. In this sense we can see that these designs are a way to remove confounding variation by adding classificatory controls or strata.

Repeated Measures ANOVA has been used increasingly in biology for several reasons. The first is that it allows us to better control for inter-subject variability. It allows us to use a subject as its own control. Secondly, it is more economical in use of subjects, which is especially important when subjects (or study sites) are difficult to locate or get to, or are limited in number.

Remember that a repeated measures ANOVA is an extension of the paired *t* - test to more complicated ANOVA designs. As such, one diagnoses the presence of a repeated measures factor in an ANOVA design by the presence of subjects who are observed under all levels of a factor. ANOVA designs can be comprised entirely of repeated measures factors (a full within subjects design), or have a mixture of repeated measures and non-repeated measures factors (a design with both within and between subjects effects).

Univariate repeated measures ANOVA requires, in addition to the normal ANOVA assumptions, an assumption that the correlations between observations within a subject are all the same. This is sometimes referred to as the assumption of compound symmetry or sphericity of the variance-covariance matrix. The *multivariate* approach to repeated measures ANOVA does not require this assumption, but produces multivariate tests of the hypotheses of interest, which may be more difficult for the average reader to comprehend. Under the General Linear Model command in the Statistics Menu in **SPSS**, the procedures GLM - Repeated Measures will generate both the univariate and the multivariate tests. When one violates the assumption of sphericity, the TYPE I error rate is inflated. So to measure the degree of deviation from sphericity, and provide an adjustment to the univariate tests, **SPSS** calculates the Greenhouse-Geisser ϵ (along with other parameters). To adjust for non-sphericity, multiply the numerator and denominator degrees of freedom by the appropriate ϵ value, and evaluate the *F* value reported by **SPSS** for the adjusted degrees of freedom. In **SPSS**, ϵ is calculated and the adjustment of df's is done automatically. See the paper by O'Brien and Kaiser in the

supplemental readings to learn more about these two approaches to repeated measures ANOVA.

In **SPSS**, data for repeated measures factors must all be on the same line for a single subject. If a subject is observed under all three levels of factor A, then all three response values must be on a single line of data. For a two factor design with repeated measures on one factor (with 3 factor levels for the repeated measures factor and 2 for the between subjects factor) data would look like this:

```
1 10.5 20.7 34.7
1  8.9 19.3 27.5
1  9.1 17.5 23.8
2  3.4  7.9 12.4
2  2.5  8.3 13.2
2  3.5  6.9 15.3
```

The integer code represents the level of the between subjects factor in which each individual subject is nested, and the three real numbers in successive columns represent the observation of the random variable of interest in levels 1, 2, and 3 of the within subjects factor, respectively. Entering data on additional repeated measures factors would entail adding the additional observations to each line of the data file since each line represents an individual subject.

Further Instructions for Lab 8

When I ask for a graphical representation of the experimental design, I am looking for a diagram like I have shown in class that shows all the factors in the experiment, their levels, and represents each group of subjects with the letter G appropriately subscripted under each treatment combination. I also ask that you accurately describe the pattern of crossing and nesting of factor levels and subjects.

Nested ANOVA in SPSS

SPSS has no link in its pull down Menus to options that let you fit a Nested ANOVA model. But there are still two ways to do so in **SPSS**. The first is to code your data as if it is a factorial model with the levels of the nested factor crossed with the non-nested factor. In our example, you would code the trees as 1-3 under level 1 of light (shade) and 1-3 under level 2 of light (sun). Then proceed to fit a two factor fixed effects ANOVA model using the **Univariate** procedure that you used in Lab 7. The resulting ANOVA table contains the elements you need to complete the analysis by hand. In a nested ANOVA, one cannot estimate an interaction between the nested factor and the non-nested factor since their levels are not completely crossed. In fact the effect of factor B operating within the levels of A (B(A), nested within non-nested) can be computed by pooling the Sums of Squares of the main effect of factor B with the Sums of Squares of the AB interaction to obtain the Sums of Squares B(A). Similarly one pools the degrees of freedom of the same terms and uses those to compute the Mean Square B(A) as $SS(B(A))/df(B(A))$. So,

$$SS(B(A)) = SS(B) + SS(AB)$$

$$df(B(A)) = df(B) + df(AB)$$

$$MS(B(A)) = SS(B(A))/df(B(A))$$

In our particular example, our nested factor (trees) is a random effects factor since we chose them at random from a large number of possible trees. Therefore, the final test of the effect of factor A will involve computing an F ratio with the MS(B(A)) as the denominator

F test for factor A, $F = MS(A)/MS(B(A))$.

Given that our nested factor is a random effects factor, it also makes little sense to compute an hypothesis test for the B(A) effect, but if one insisted on doing so then the within cell or error mean square would serve as the denominator of the F -ratio.

A second approach is to use a **Syntax** window in **SPSS**. In that **Syntax** window insert and run the following syntax (go to **File Menu, New**, choose **Syntax** file):

```
glm damage by sunshad, tree  
/design sunshad, tree(sunshad).
```

Then in the Syntax window, choose "Run" from the toolbar and select "All."

This syntax assumes that you have coded your trees 1-6 rather than as described above for a factorial model. It also assumes that "tree" is the name of the column in your data file with this coding, "damage" is your response variable, and "sunshad" is the name of the column in your data file that codes for levels of the light (non-nested) factor.

The resulting ANOVA table will have the correct Sums of Squares and DF's for the nested analysis for factor A, B(A), and error. However, it will still have the wrong F -ratio for the test of the effect of factor A, since it uses the MS error in the denominator, not MS(B(A)) in the denominator as it should. Compute this by hand and then determine your significance level from an F -table.

Finally, for our particular problem, an alternative way to analyze the data would be to average the values among the 15 leaves within each tree, and use the tree means to compute an independent groups t -test for differences in leaf damage between trees in the sun versus the shade. The results of this test would be identical to the test of the light effect in the Nested ANOVA (except the t -value would equal the square root of the F value). Hence, it is sometimes possible to turn a nested ANOVA into a simpler problem, particularly if the nesting of factor levels arises because multiple observations are made on the same subject. Here, by using tree averages, we turn a nested random factor (the tree factor) into our subjects, hence removing the nested factor levels and simplifying the analysis.

Repeated Measures ANOVA in SPSS

Repeated measures ANOVA (within-subject designs) are a bit easier to do in **SPSS** than a Nested ANOVA. Go to the **Analyze** menu, **General Linear Model**, and choose **Repeated Measures**.

The **Repeated Measures Define Factors** subwindow will open. In this subwindow you give a name to each repeated measures factor and tell SPSS how many levels each factor has. In our example for lab we have only 1 repeated measures factor (month), and one between subjects factor (food addition). In the subwindow put the name you are giving the repeated measures factor into the “Within Subject Factor Name” box, then in the number of levels box enter “3,” and then click the add button. Now click on the “define” button and SPSS will take you to the “Repeated Measures” subwindow. In this subwindow you first need to associate column names from your data file, with levels of the newly defined within-subjects factor. In our example, you need to tell SPSS which column holds the values of the response variable for June, which for July, and which for August. These variable or column names from your data file should appear in the box on the left. You need to click these names over into the “Within-Subjects Variables” box in the appropriate positions (denjun for level (1), denjul for level (2), and denaug for level (3)). Now you need to click over the name of the variable or column that contains the integer codes for the between subjects factor (food addition) into the Between-Subjects Factor(s) box. This is all you need to do to run the analysis, however, I usually also ask for descriptive statistics and a plot of the treatments means. Click the “Options” button and then click the Descriptive Statistics box (under the blue colored word “Display”). This will give you the means, standard deviations, and sample sizes in each treatment. Click Continue to return to the Repeated Measures subwindow. Now click on the “plot” button to access the Repeated Measures Profile Plots subwindow. The factors in your design should appear in the “Factors” box. Click over the “month” factor to the “Horizontal axis” box and the factor “food addition” to the “Separate lines” box. Click on the “Add” button and then the “Continue” button when you have finished selecting the plots you want. This will return you to the Repeated Measures subwindow. Now you are ready to run the analysis. Click on the “OK” button.

Interpretation of Repeated Measures Results

SPSS produces lots of results for a **Repeated Measures ANOVA**.

Table 1 just tells you how SPSS associates your column/variable names with levels of the within-subjects factors. Just to check to see if you got it right.

Table 2 will be the Descriptive Statistics if you chose that option.

Table 3 will be labeled “Multivariate Tests.” Here you will find Multivariate F - values, degrees of freedom, and significance levels for all your within-subjects tests. The reason for this is that one can conceive of a within-subjects factor as either a single response variable examined under a series of different treatment levels (a univariate approach), or as several different response variables (a multivariate approach). Usually the Univariate and Multivariate approaches give the same answer, although F -

values, dfs, and p-values will not be exactly the same. The reason for using the Multivariate approach is that one only has to meet the assumption of Multivariate Normality and homogeneity of the variance/covariance matrix for this test to work well. However, in the univariate approach one has to meet the more restrictive assumption of sphericity of the variance/covariance matrix.

Table 4 gives you Mauchly's test for Sphericity of your variance/covariance matrix, and three different estimates of Epsilon (a measure of how much your data depart from meeting this assumption). Epsilon is a value between 1 and the lower bound estimate and it is multiplied by the degrees of freedom for both the numerator and denominator mean squares to adjust the test for non-sphericity. Deviations from sphericity lead to too many Type I errors, so by adjusting dfs downward the resulting univariate F -tests have error rates equal to the nominal error rates (if you claim the p value is 0.05 then it will be so if you meet the other assumptions of the test).

Table 5 gives the univariate ANOVA results of your Within-Subjects Tests. SPSS does all possible adjustments for non-sphericity, including not adjusting. I recommend using the Greenhouse-Geisser adjustment. Otherwise this table has all the usual attributes of an ANOVA Table.

Table 6 decomposes the within-subjects effects into a linear, quadratic, and higher order effects. For a within-subjects factor with k treatment levels it can partition the overall effect into $k-1$ single degree of freedom effects. If you are specifically interested in a quadratic alternative hypothesis then this table may be useful.

Table 7 is an ANOVA table that contains the tests of Between-subject effects (food addition in our example). Ignore the intercept line this is just a test of whether or not the mean of your data differs from 0.

Lab 8 - Assignment PART 1: Nested ANOVA

Data on insect damage on oak trees was collected in order to answer questions concerning the effects of shading on damage levels. Six oak trees were selected at random, 3 in the shade and 3 in the sun. The data are included in the file [nest.dat](#). The file has the following form:

Column 1: Light level: 1 = shade; 2 = sun

Column 2: Tree. Trees 1 – 3 (six different trees 3 in sun and 3 in shade).

Column 3: Damage in percent.

Reiterating, this file contains data which describe the response of trees in terms of damage; thus damage is the response variable. The data come from observations taken from six randomly selected oak trees. Three are located in the shade; and three are

located in the sun. Thus this is a nested design, with trees nested within the level of light (shade or sun) and leaves nested within trees.

A sample of 15 leaves was taken at random from each tree.

1.1) What kind of design is this? Draw a graphical representation of the experimental design, where A = light level, B = trees, and G = group of subjects.

1.2) Conduct an appropriate analysis of variance on these data.

a) specify the null hypotheses being tested and turn in the edited output

b) pool the appropriate sums - of - squares and perform the appropriate F test

c) interpret the results of the tests (at $\alpha = 0.05$)

d) How could the data have been treated differently so a nested ANOVA could have been avoided?

PART 2: Repeated Measures ANOVA

In this experiment, ten small-mammal trapping grids were established in order to study the effect of food addition on population densities. Five of the grids received food supplements while the others were not manipulated. The population levels on each grid were monitored 3 times at monthly intervals following the food addition. The data are contained in the file [rep.dat](#). It is in the following form:

Column 1: Food. Where: 1 = no addition; 2 = food added

Column 2: Grid. Grids 1 - 10

Column 3 - 5: Population density in June, July, and August.

Reiterating, here we have 10 grids. 5 have food added; 5 do not. The response variable, population density, was measured on each grid (subject) 3 times (a repeated measures factor).

2.1) Same as 1.1, but this time where A = food level, B = time, and G = subjects.

2.2) Same as 1.2 except no pooling of sums - of - squares is necessary (but apply to data in PART 2)