Biol 206/306 – Advanced Biostatistics
Lab 3 – Analysis of Variance Designs
Fall 2016

By Philip J. Bergmann

0. Laboratory Objectives
1. Refresh your ANOVA knowledge of the single factor ANOVA
2. Learn how to do a single factor ANOVA in R
3. Learn how to do pairwise *post hoc* tests
4. Learn various designs of multi-factor ANOVAs and how to implement them in R
5. Learn how to define linear models in R
6. Learn about accommodating unbalanced designs for ANOVA and in R
7. Learn about nested and repeated-measures ANOVAs

1. Review of ANOVA
Analysis of Variance (ANOVA) is a statistical approach for comparing the group mean for a response variable among two or more groups. When there are two groups, the ANOVA is equivalent to a two-sample t-test. When there are more than two groups, a significant ANOVA F-test tells you that at least one of the group means is significantly different from the others. To determine which means are significantly different *post hoc* tests can be done, such as the Tukey test. A single factor ANOVA model can be represented with the equation:
$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$
where y_{ij} is your observation, μ is the grand mean of the population, α_i is the mean for treatment/group *i*, and ε_{ij} is the residual or error associated with observation *ij*. We will build on this notation throughout the lab.

ANOVA makes four assumptions about the data. Like almost all statistical tests, data are assumed to be randomly sampled and independent of one another. ***What are the other two assumptions made by ANOVA?***

2. Doing a Single Factor ANOVA in R
Data you will be using for the first half of this lab investigate how egg hatching rates for three of the most common species of mosquitos in New England (alb – *Aedes albopictus*, aeg – *A. aegypti*, and tri – *A. triseriatus*) are influenced by the presence of larvae of the same three species in their vicinity. Eggs normally hatch as bacteria grow on their surface, leading to decreased oxygen uptake. Larvae feed on the bacteria, preventing the decrease in oxygen uptake, and this can inhibit hatching. A higher larval density may lead to higher bacterial consumption rates. Data are modified from those of *Edgerly, J.S., Willey, M.S., and Livdahl,*

T.P. 1993. The community ecology of Aedes egg hatching: implications for a mosquito invasion. Ecological Entomology 18: 123-128.

Start by opening the mosquito data in Excel and saving it in tab-delimited format. Import the data into R, assigning it to an object called "mdata". Then examine the data in R by calling the object (typing "mdata") and by using the str() and summary() functions.

A single-factor ANOVA can be done in one of two basic ways in R:

```
> summary(aov(mdata$Hatch~mdata$Density))
> anova(lm(mdata$Hatch~mdata$Density))
```

For now, try the first option only. You get an ANOVA table with degrees of freedom, sums of squares, mean squares, f-statistic, and p-value. However, there is a problem with this output. Notice that there is only one degree of freedom associated with the Density effect, despite there being three densities. How many degrees of freedom would you expect to be associated with the Density effect?

The problem is that because density is represented as numbers (4, 12, 24), R treats the variable not as categorical, but as continuous. What we need for the ANOVA to be correctly done is to get R to treat Density as a factor. One way in which R is flexible is that you can "coerce" objects or variables into different classes.

```
> as.matrix(dataframe_object)
```
Treats a data frame as a matrix. If you assign the above code to a new object, you've converted a data frame to a matrix.

```
> as.factor(vector_object)
```
This converts a vector into a factor, and is what we need for the current situation. Now try this:
```
> summary(aov(mdata$Hatch~as.factor(mdata$Density)))
```

Note that you do not force Hatch into factor form because it is a continuous variable and needs to be treated as such. ***Now also try the second way of doing ANOVA, shown above, but make sure that you coerce Density into factor form again. How many degrees of freedom are associated with Density now?***

How do the results differ between treating Density as a factor versus a vector? Although in this example, the difference isn't massive, in some cases it can be the difference between a non-significant result and a highly significant result, so be careful.

From the above ANOVA (done correctly), what biological hypothesis did we test?

What was the null hypothesis?

Based on the analysis, would you accept or reject the null hypothesis? Explain.

What don't the above analyses tell you?

Looking at the above lines of code for doing ANOVA, both contain nested functions. In the first, **aov** is nested within **summary**, and in the second, **lm** is nested within **anova**. Looking at the first line, you could just as easily do the following:

```
> results <- aov(mdata$Hatch~as.factor(mdata$Density))
> summary(results)
```

Go ahead and do this.

The nested function approach shown first is useful because it is more efficient if you are just interested in the ANOVA summary table. You could also assign the results summary to another object if you wanted to refer to it later. Note that **aov** fits an anova model to the data. This model contains considerable other information that may be very useful, like the residuals.

Explore the aov function as follows:
```
> results
> str(results)
```
As you can see, the **aov** output is very complex in structure, but you need to extract information from it for it to be useful. When you type in "test", it simply giving you degrees of freedom and sums of squares. *How could you use the information you get from typing "test" to create an ANOVA table, as seen in "summary(aov)"?*

You can obtain a vector of the residuals from your ANOVA from the aov object by typing:
```
> residuals(results)
```

This is extremely useful because you can then test the assumptions of the ANOVA residuals. Let's do this now. One way to test for the normality of your data is using the Shapiro-Wilk's test. The Bartlett test is useful for comparing variances among groups. Note that the residuals are in the same order as your original data in the object "mdata", and so you can use a grouping factor from your original data with the residuals to do the Bartlett test. Give it a try:

```
> shapiro.test(x)    # for x, substitute "residuals(test)"
> bartlett.test(x,group) # x are your residuals
                      # group is "Density", which must be a factor
```

The Shapiro-Wilk's tests the null hypothesis that the variable *x* is normally distributed. Hence, if you accept the null hypothesis, the variable is normally distributed. Likewise, the Bartlett test tests the null hypothesis that two or more groups have equal variances. Hence, if you accept the null hypothesis, it means that the groups are homoscedastic.

Conduct these analyses on your residuals. What conclusions do you come to? Are the residuals normal and homoscedastic? Provide the test statistic and p-value for each test. (We will revisit testing for normality in Lab 4).

Notice also in both approaches to doing an ANOVA (the second approach fits a linear model, **lm**, to the data and then produces an ANOVA using the **anova** function), the variables that you want to do the ANOVA on are specified in the notation: **Y~X**. For now, it is sufficient to say that your response goes before the tilde (~), and your grouping/treatment/independent variable goes after it. You will learn more about this shortly, as we move on to more complex designs.

At the moment, there are two unanswered questions about the ANOVA you did above: how big is the effect of density on hatch rate, and which of the density treatments differ from one another? We can address the first question by calculating **effect size**. There are various ways to do this, but a widespread and straightforward on is **partial η^2**, which can be calculated as:

$$partial\ \eta^2 = \frac{SS_A}{SS_A + SS_{resid}}$$

Here, SS is the sums of squares of the main effect (A) and the residuals (resid). As such, partial η^2 gives you a measure of the variation explained by an effect, relative to the variation it explains and unexplained variation. It is termed *partial*, because the denominator is not SS_{total}, so it only considers variation not explained by other effects in the model (see below for ANOVAs that have more than one main effect). One can also calculate η^2, which uses the SS_{total} as the denominator. For a single factor ANOVA, η^2 and partial η^2 are the same, but when we get into more complex models, this will not be the case. In this course, we will use partial η^2 as the measure of effect size when doing an ANOVA. Interpretation of partial η^2 is either relative to other values of η^2, or, as a rule of thumb values below 0.02 are considered small effects, values below 0.13 medium effects, and values of 0.26 or greater are considered large effects. ***What is***

the partial η² for the density effect in the analysis you did above? Would you consider this a small, medium, or large effect?

Once we have a significant factor in an ANOVA, we can determine which treatment levels differ from one another using *post hoc* tests, which are generally some sort of two-sample t-test done in a pair-wise manner, with the p-values adjusted for multiple comparisons in some way. If you get a significant ANOVA, the function to do pair-wise comparisons is:

```
> pairwise.t.test(y, x,
p.adjust.method=c("holm","bonferroni","BH","none"), pool.sd=T,
paired=F, alternative=c("two.sided","less","greater")
```

where y is your response variable, x is the grouping variable, pool.sd specifies whether variances are equal and can be pooled between treatments (T is the abbreviation for TRUE, F for FALSE, all must be capital), paired specifies whether design is repeated measures, and alternative specifies whether tests should be two or one tailed. p.adjust.method determines whether and how you want to correct the p-values. Importantly, you can not correct, use a standard Bonferroni correction, or use the method of Benjamini-Hochberg that we learned in an earlier lab. The advantage of using a correction is that the returned p-value is already corrected for multiple comparisons and can be compared to your desired type I error rate.

Conduct the pair-wise comparisons for your ANOVA using the BH method. What do you conclude?

3. Multi-factor ANOVAs
Two-factor and greater ANOVAs simply build on the single-factor ANOVA, but as things get more complicated, the number of things to keep track of also increases. In this section you will see the model equations for a two and three factor ANOVA, learn about R model notations, learn about taking random versus fixed factors into account in your analysis, and learn what to do when a design is unbalanced. Although most of the information in the previous sections may have been review, the current section may be full of new material.

We can start with the model equations for two and three factor ANOVAs:
Two-Factor ANOVA:
$$y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk}$$
Here notice that we have a second factor, β, with levels *j*. We also have what is called an "interaction term" between the two main effects. With two effects, an individual can be

identified by *ijk*, instead of just *ij*. The interaction is important because it allows you to test if there is an effect of one factor that is dependent on the value of another factor.

Three-Factor ANOVA:
$$y_{ijkl} = \mu + \alpha_i + \beta_j + \delta_k + \alpha\beta_{ij} + \alpha\delta_{ik} + \beta\delta_{jk} + \alpha\beta\delta_{ijk} + \varepsilon_{ijkl}$$
The equation may start to look long and intimidating, but deconstruct it and notice that it simply builds on the two-factor case. Now you have three factors, there are three possible two-way interactions among them, and one possible three-way interaction. You always have that residual/error term that contains any variance that is unexplained by the rest of the model.

The next step is to learn how to represent these model equations in R to design an ANOVA for any set of data. This is actually very simple and follows the equations presented above quite closely. You already know the basic syntax: **Y~A** for a single-factor ANOVA. Note that the grand mean and the residual term are not specified because they are present in all models automatically. Now consider the following notations and their meanings:

Y~A+B	Y is explained by two effects: A and B
Y~A:B	Y is explained by the interaction of A and B
Y~A*B	Equal to Y~A+B+A:B – the full crossed model with A and B
Y~A−B	Exclude term B, so Y~A*B-B is equivalent to Y~A+A:B
Y~B %in% A	A nested model, where B is nested in A, or B(A)
Y~A/B	Equivalent to Y~A+B %in% A, so B is nested in A again
Y~M^n	The fully crossed model of terms in M, including only up to n^{th} order interactions. If M contains three terms and n=2, then the three-way interaction is excluded.

Express the model shown above for a fully crossed three-factor ANOVA using R notation. Do this in two different ways.

The mosquito dataset described at the beginning of section two is designed as a three-factor ANOVA. Revisit the dataset and its description to answer the following questions and do the relevant tasks.

What are the three factors in the mosquito dataset?

What is the biological hypothesis being tested with each one?

What are the null hypotheses being tested by a fully crossed three-factor ANOVA of the dataset?

Do the three-factor ANOVA on the mosquito hatching dataset and complete the ANOVA table below with the values from your analysis. Present everything to two decimal places, except p-value, which should be to 4 decimal places. Please note that there are more rows in this table than you need. Be sure to coerce any effect variable that is a continuous numbered vector into a factor so that you get the right output. Typing saving tip: use the function attach() (i.e., type "attach(mdata)") to make the variables in mdata available as objects. Although they do not appear when you type "ls()", you can now refer to "Density" instead of "mdata$Density". To undo this, type "detach()".

Effect	df	MS	F	P	partial η^2

How could I calculate a sums of squares value from the above table?

3.a. Fixed and Random Effects

When you are doing a single-factor ANOVA, the F-statistic is calculated the same way no matter what the nature of your factor is. However, with higher-factor ANOVAs, whether a factor is a fixed effect or a random effect influences how the F-statistic is calculated. An effect is fixed when the levels of the factor are of particular interest to the researcher and would be the same if someone redid the experiment/study. Examples of this might be sex (male/female) or whether or not a subject receives a drug. Random effects are either randomly sampled from a population of factor levels, and different levels would be chosen if the study were redone, or else involve levels that are not of specific interest and easily could be changed. Examples of this might be families chosen in a population or populations chosen within a species. With a fixed effect, you cannot generalize your conclusion to a larger set of levels, but you can with a random effect.

In the mosquito dataset, you had three factors. At first glance, they may all seem as though they are fixed effects, and an argument could, perhaps be made for this, but one of them most likely is a random effect. ***Which factor in the mosquito dataset could be a random factor? Explain.***

Fortunately, dealing with random effects is straight forward. As a rule of thumb, when calculating the F-statistic for a fixed effect, $F=MS_A/MS_\varepsilon$, where factor A is a fixed factor and MS_ε is the Mean Squares error. This is how R automatically calculates ANOVAs. ***Confirm this by calculating the F-statistic for the egg effect.*** When you have a random effect, the denominator of the F-statistic becomes the MS of the interaction: $F= MS_B/MS_{AB}$, where B is a random factor. This works well with a two-factor ANOVA, but becomes very complicated with a three-or-more-factor ANOVA, where there are multiple interaction terms. The mosquito egg data is an example of this. ***Pretend that your mosquito egg data are a two factor ANOVA, eliminating "egg species" as a factor. Rerun the ANOVA with the two factors "Density" and "larva", and complete the table below (this analysis assumes everything is a fixed factor again):***

Effect	df	MS	F	P	partial η^2
Density					
Larva					
Density:Larva					
Residual					

Now, to the right of the table, calculate a new F-statistic for the effect that could be random. Remember that R can also be used as a calculator to help you. Also to the right of the table, calculate the p-value that goes with your random-effect F-statistic using the following function:
```
> 1-(pf(F, df_n, df_d))
```
Where F is your F-statistic, df_n is the degrees of freedom of the numerator, and df_d is the degrees of freedom of the denominator.

How has the p-value changed?

Two of the three effects in the ANOVA you just did have highly significant p-values. However, what do the effect sizes tell you that allows you to interpret this analysis more fully?

3.b. Unbalanced ANOVA Designs
Another complication that you may run into when analyzing real data using ANOVA is that various datasets may be unbalanced, meaning that they do not have the same sample size for each combination of treatment levels. Again, this is not an issue for a single factor ANOVA, but is for multi-factor ANOVA. Take another look at the mosquito egg hatching dataset and notice that it is perfectly balanced, with the same number of observations in each treatment level.

Let's explore what happens when we unbalance the design. First, let's delete some data arbitrarily, placing the result into a new object:
```
> ub_mdata<-rbind(mdata[3:17,],mdata[19:52,],mdata[56:160,])
```

Confirm that the data are now unbalanced (visually), and do a two-factor ANOVA on Hatch with Density and Larva as factors A and B, respectively. Repeat the ANOVA, but with Density and Larva switched in the model formula (i.e., Hatch~Density*Larva and Hatch~Larva*Density). Again, make sure that you coerce Density to be a factor.

What do you notice about the results?

The aov function calculates the ANOVA statistics using *Type I Sums of Squares*, where the sums of squares for the second effect are calculated after those for the first effect, taking the first effect into account. Also called sequential sums of squares, the order in which the factors are added influences the results. This is acceptable for a balanced design, and is what is taught in most statistics classes because the calculations are straightforward and the effect sums of squares sum to the total sums of squares. This is not the case with type II and III sums of squares, described next.

There are also other ways to calculate sums of squares for an ANOVA. Type II sums of squares calculates each effect sum of squares while adjusting for all terms in the model that do not contain the effect. For example, for the model A*B, SS_A is adjusted for SS_B, but not SS_{AB}, and SS_B is adjusted for SS_A, but not SS_{AB}. In type III sums of squares, each effect SS is adjusted for all other terms, both main effects and interactions. Therefore, SS_A is adjusted for SS_B and SS_{AB}.

When an ANOVA has only one factor or when it is balanced, all SS types give the same results. When an ANOVA is unbalanced, the general recommendation is that type III SS be used. Type II SS may be useful in some rare situations, but this is beyond the scope of this course. You can do a type II or III SS ANOVA using the following function, *which is part of the "car" package, which you need to load before using the function*:
```
> Anova(lm(model), type=c(2,3), test.statistic=c("LR", "F",
"Wald"))
```
Where "model" is your model, specified in the usual way, "type" is the type of sums of squares, and "test.statistic" tells you the type of test you want done (just use the F test for this course).

Note that the function **Anova** is different from the function **anova** – the former gives type II or III SS, the latter gives type I SS!

Repeat your two-factor ANOVA of the mosquito egg data, with the model specified in both ways (Density* Larva and Larva*Density) using the Anova function. Take a look at the ANOVA tables. How do they compare?

How do they compare to the type I SS ANOVA tables?

4. Nested ANOVA
A nested ANOVA is a design that can account for a factor that is nested within another factor. This is important because the nested factor is not independent on account of its nesting, and because the design allows for hierarchically-arranged factors. For example, consider sampling multiple individuals from a number of different families and these are divided among treatments. In this case, individuals are nested within family. Likewise you could be studying multiple ponds and these ponds are distributed within different drainages or located on different islands. In this case, the ponds are nested within drainage or island. The notation for nesting can be counterintuitive: if factor B is nested within factor A, then this is often represented as B(A). In these designs, there is often an assumed lack of interaction between the nested and nesting factors, although there may be interaction between B(A) and another factor, C. Finally, the model equation for a nested ANOVA is:
$$y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$$
which should come as no surprise, given the description above.

For the rest of the exercise, we will move away from the mosquito dataset and instead deal with another stickleback dataset, produced in the Foster/Baker Lab. The response variable is size at first feeding, so the size of a hatchling fish when it feeds for the first time. These data were collected for fish raised at different temperatures covering the natural range (10, 15, 18 and 21°C). Fish were taken from five source populations, which represent lakes that are either shallow and warm or deep and cool. Eight females from each lake were fertilized, and their eggs were raised at the four different temperatures in the lab. The source female is labeled family because each clutch of eggs was split up, with some placed in each aquarium temperature in the lab.

Examine the dataset. For now, let's ignore three of the temperatures and just consider the 18°C treatment. Also ignore the family column for now. This leaves you with size at first

feeding for one temperature, the lake of origin, and whether the lake was warm or cold. What are the factors for this ANOVA design and what are the levels for each factor?

Which factor is nested within the other factor?

Save the dataset as a text file and import it into R, calling it "sdata1". Use the summary and aov functions to run a nested factor ANOVA on the 18°C data with habitat and population as factors. Refer back to section 3 of this lab, if needed, and use the A/B operator to produce your model formula. Fill in the ANOVA table on the next page (again there are more rows in the table than needed).

Effect	MS	Df	F	P	partial η^2

What do the ANOVA results tell you?

Do the ANOVA again, this time with the %in% operator (and anything else you may need). What is the model formula that gives you the same answer as your original A/B operator?

5. Repeated Measures ANOVA
Repeated measures designs are another way of dealing with data non-independence. In this design, a single sampling unit (individual, plot, etc.) is sampled repeatedly, either to evaluate the effects of time, or to evaluate the effects of some treatment. Although at first glance, non-independence of data seems to be a negative, in a repeated measures design, it is actually a strength because you can account for individual-level variation. By accounting for this individual variation, one has more power to detect treatment effects. Consider a situation where you are comparing the time it takes to run a 100m sprint when the track is dry versus wet. You could design this experiment by running a set of sprinters on a dry track and running another set

of sprinters on a wet track. Alternatively, you could run the same group of sprinters on a wet and dry track. This would allow you to account for the fact that some sprinters are better than others and will run slower irrespective of the condition of the track. The linear model for a repeated measures ANOVA is:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

where α is the treatment of interest and β is the individual. Individual is typically a random effect and often includes a single observation per treatment.

Revisit the description of the stickleback size at first feeding dataset and open it in Excel again. Although the dataset is already formatted in a way suitable for analysis as repeated measures using most statistical software, we need to reformat it for R so that temperature is in a new column and appears as any other factor would. Reformat the data so that you have the following columns: habitat, popul, family, temp, and size. Note that you will now have a single "size" column instead of four of them, and there will be four times as many rows as in the original version.

Save this new dataset as a tab-delimited text file (careful not to overwrite your previous dataset), and import it into R as "sdata2". What is the response variable in the dataset you just formatted?

Which factor is the repeated measure?

Which factor contains the sampling units for which you have the repeated measures?

What are the factor(s) that are not part of the repeated measures design (but still potentially important)?

Let's start of with a simple repeated measures ANOVA where there is a single factor whose levels are measured for each sampling unit. Attach the sdata2 object and try the following code to make the model:

```
> rm_aov1 <- aov(size~as.factor(temp)+
  Error(as.factor(family)/as.factor(temp)))
>summary(rm_aov1)
```

Note that the model has familiar components for a single factor ANOVA, plus an error term that is defined as the sampling unit identifier / the repeated factor. This is the typical way of accounting for sampling unit differences. The "/" refers to nesting, just as before, but this time you have family nested in temp because the family is the sampling unit. Also note again that factors whose levels are represented by integers (family and temp) are coerced to be factors.

Examine the output. Find the ANOVA table with the F test in the output. Note that you are not interested in testing for effects of the individual sampling unit, you are just taking those

into account. You are really interested just in the effects of the factor that is repeated ("within" each sampling unit). Reproduce the ANOVA table here. What conclusions can you draw from the analysis?

Effect	df	MS	F	P	partial η^2

Now that you have done a single-factor repeated measures ANOVA, it is time to do something a little more complex. Revisit the description of the stickleback dataset. You already have the data we will use as object "sdata2". You will now do an ANOVA with one factor that is repeated (temperature) and one factor that is not (population). To start, let's rework the data a little so that the model script is cleaner. Instead of constantly coercing vectors into factors, let's make a few new objects that are factors. Make sure that sdata2 is attached and do the following:

```
> fam <- as.factor(family)
```

Do the same thing for temp, calling it tmp.
Why does this not have to be done for the population identifier, popul?

Now do the following and then call your new object:

```
> rm_aov2 <- aov(size~(popul*tmp)+Error(fam/tmp))
> summary(rm_aov2)
```

This line of script deserves some explanation. "size~(popul*tmp) should be familiar: a two-factor ANOVA on size, with popul and tmp as the two factors. The "Error(fam/tmp)" should also be familiar, defining the treatment (tmp) being nested in the individual sampling unit (fam).

Notice that the output to the code above gives you three ANOVA tables. Pay attention to the degrees of freedom in each to get the ones you need. The first uses "fam" as the error term, which is not what you are looking for and the degrees of freedom add up to 8, the number of families. The next uses "fam:tmp" in the error term, which is what you specified in the code. This table gives you results for the test of whether there are effects of "tmp" on the dependent variable (size), and for the test of whether there is an interaction between tmp and population. The third ANOVA table (heading "Error: Within") gives you the result for the effects of popul. So, you need to take care that you retrieve the results that you need because more are given, and some of those do not really test what you want.

Use str() to view the structure of the rm_aov2 object. Notice that there are separate sections for each ANOVA table. This means that if you want to test for normality of residuals, you must do so for the sets of residuals about which you are testing hypotheses. For example, you can access the residuals for the "fam:tmp" table using:

```
> rm_aov2$"fam:tmp"$residuals
```

Assignment (10 points):
Answer the following questions in the context of the analysis rm_aov2.

What are the main biological hypotheses being tested in this analysis?

Looking at the output of rm_aov2, take just the results that are useful (described above), and place them into the ANOVA table below.

Effect	df	MS	F	P	partial η^2
Tmp					
Tmp:Popul					
Residuals (for above)					
Popul					
Residuals (for Popul)					

Are the residuals for the fam:tmp ANOVA table normally distributed? Report the test you use and the results of that test.

Interpret the ANOVA above – what do the tests and effect sizes tell you?

Make a boxplot of size at first feeding for each temperature and insert it below this question. What does this plot tell you about the biology of stickleback?

We have separately done a nested factor ANOVA that takes into account that each population is nested in a habitat (warm or cold) and one that takes into account that family is repeated across temperature treatments. What model formula would you use to account for all of these things in a single ANOVA?