**Lab 4: Anova**

**4.1. One-way ANOVA.**

Let’s first practice performing a simple one-factor (or “one-way”) ANOVA comparing mean weight loss between three different diets. First we need to read in the data “**diet.txt**”. The data is saved in a tab-delimited text file in the lab 4 zip-folder. There are many options to read data into R, so you could have your data saved in many text file formats, or excel format, if you like. Try typing **?read.table** to find out more.

**read.delim("diet.txt")->diet** #names the dataset “diet”

**head(diet)** #look at the first few rows

**tail(diet)** #look at the bottom rows

**str(diet)** #look at the structure of the dataset

**diet** #Look at the whole freakin’ thing #with large datasets

**diet$weight.loss <- diet$pre.weight - diet$weight6weeks** #calc. weight loss

**head(diet)**

#weight.loss has been added as a new variable to the dataset.

#Note that when you use Rstudio, it is often easier just to click on the #dataset in the top-right menu to look at it.

There are many ways to graph group means in R. **boxpot()** plots median +/- 50% (box) and 95% (whiskers) of your data, the latter roughly corresponding to +/- 2 standard deviations. This plot is good if you want to understand your data and its spread. It is NOT good for understanding if you have significant differences, where you instead should plot mean +/- standard errors or confidence intervals (+/- 2SE ~ 95% confidence limits), which both show uncertainty in the prediction of the mean. These estimates depend on the spread in the data (standard deviation), but critically also on sample size. Be sure to revisit the slides of previous lectures if you need to remind yourself about the critical difference between standard deviations and standard errors.

**boxplot(weight.loss ~ Diet, data = diet, col="light gray",**

**ylab = "Weight loss (kg)", xlab = "Diet type")**

**abline(h=0,col="blue")** #Again, type **?abline** or **?boxplot** to learn more

Now let’s perform our model. There are two basic ways to perform ANOVA in R; using the **aov()** and the **lm()** function. The former is specific to ANOVA (aov = ANalysis Of VAriance) and allows you to also specify nested designs with both fixed and random effects. The latter is more general (lm = linear model) and you can use it for most types of designs and analyses (for example, also regression analysis, which will be the focus of the next lecture and lab). However, it is not possible to specify nested designs and random effects with this function, instead you need to call upon its evil stepsister; the lme4 package (that’s complicated and for another course).

**mod1a <- aov(weight.loss ~ Diet,data=diet)**

**mod1b <- lm(weight.loss ~ Diet,data=diet)**

The “~” sign can be thought of as meaning “*depends on*” – weight loss (the dependent response variable) depends on the type of diet (the independent explanatory variable).

**4.1a. Once you have performed the model(s), *you need to check model assumptions (see lectures notes).* For models where the response variable is continuous and assumed to be normally distributed, R has built in diagnostics. Applying plot() to your stored model object gives you *four figures you need to interpret*. Ask your class-mates or your teacher about how to best interpret them if you find it hard. For this lab, we can skip the last panel (bottom right, we will get back to that last panel in lab 5 on regression analysis).**

**par(mfrow=c(2,2))**

**plot(mod1a)**

You can also make a histogram of the model residuals to see if they are normally distributed:

**hist(resid(mod1a))**

Happy? It looks OK, if not to say very good! So let’s look at our models and interpret the output.

**summary(mod1a)**

**summary(mod1b)**

The models are the same but the standard output is different using the lm() and aov() functions; lm() reports P-values for comparisons of treatments to the model intercept (which happens to be Diet A) using the t-distribution, and aov() reports a single P-value for the effect of diet calculated based on the F-distribution.

***Make sure you understand the output.***

**4.1.b What do the model “coefficients” from the lm-table mean? Can you figure it out by plotting the mean weight loss of each group and then comparing it to the coefficients?**

**4.1.c What do the “SS” and “MS” stand for in the anova table? How do these sum up for the different components (total, treatment and residuals)? And how do they relate to the F- and p-value?**

We can also calculate P-values in more ways, which may become important for more complex experimental designs. One way is to download the “**car**” package (you do this in the bottom right panel of Rstudio). Once you have downloaded it, read it into R.

**library(car)**

Typically we test more complex models with type II or type III Sums of Squares. We can do this with the **Anova()** function (notice the capital A) in the car package:

**Anova(mod1a, type =2)**

**Anova(mod1b, type = 2)**

For one-way ANOVA, P-values are the same for type I, II and III sums of squares. They often differ for more complex designs. Type I SS are not recommended to test significance. Choice between type II and III SS depends on your philosophy and the hypothesis tested. In this course we will use type II SS. If you want more info – ask you teacher. If things are a bit much at the moment – be happy that you have heard about the concept, and lets talk more about this some other time.

If the anova is significant, you know that there is a difference between at least two of the groups that you have compared, but you do not know which. To find out which groups differ, we can perform *post hoc* tests. One such test is *Tukey’s Honest Significant Differences (HSD) test* (don’t you just love the name?).

**TukeyHSD(mod1a)**

**plot(TukeyHSD(mod1a))**

**4.1d. Which groups differ, is it what you expected? Try to link the P-values for each contrast to the figure – makes sense? What are the confidence limits depicting here?**

**4.2. Two-way ANOVA.**

Things can get more complicated. Most experiments include more than one explanatory variable, and we often need to control for several effects simultaneously in the same model. The data on dietary effects on weight also include (biological) sex. Let’s see if there is an effect of sex.

**mod2a <- lm(weight.loss ~ Diet + sex, data=diet)**

You have now looked at effect of diet and sex in the same model. Before interpreting the output using the approach from exercise 1, also *remember to check model assumptions by plotting the model* (it will probably be OK this time around as well, but you should always follow this procedure).

**4.2a. Did sex have any effect on weight loss? What was H0 and do you reject or accept it? Did including sex affect your conclusion about the effect of diet?**

In the model above you tested independent effects of diet and sex, but you did not consider whether males and females respond in the same way to the three diets. To do this you have to include an interaction between diet and sex, and you can do this in two ways that are exactly equivalent (the first one writing out each term of the model, the second one faster to write).

**mod2b <- lm(weight.loss ~ Diet + sex + Diet:sex, data=diet)**

**mod2c <- lm(weight.loss ~ Diet \* sex, data=diet)**

**4.2b. After checking model assumptions, answer the same questions as above – did your conclusions change?** To help you interpret the effects, it is always good to plot your results. For plotting interactions, a couple of packages are helpful. The **lattice** package helps you plot graphs similar to the boxplot-function, but with more complicated designs its output is slightly nicer to look at and easier to interpret. Here, for example, we use the “|”-sign to designate that we want to look at the effect of diet per each sex separately.

**library(lattice)**

**bwplot(weight.loss ~ Diet|sex, data = diet)**

Another useful package is the effects package. This allows you to plot the model predictions from any linear model. Both packages have many settings, so you need to read the help-files carefully, but here we just use the defaults.

**library(effects)**

**plot(allEffects(mod2b))**

*Try to link what you see in the plots to the effects reported in the model output/tables.*

**4.3. Generalized linear models (GLMs)**

Not all data are normally distributed, and sometimes data may also be difficult to transform into a normal distribution. In these cases it may be better to model the data using other predefined distributions included in *generalized* linear models (glms). Before continuing, take note that you are not expected to master practical use of glms on this course – this is just to teach you that this route to analyzing data exists and is indeed very common when it comes to analyzing real world datasets. You learned about some common distributions during Lecture 3. We can illustrate a couple of alternatives by creating dummy datasets drawn from a binomial and a Poisson distribution respectively, and then analyzing them with generalized linear models in R using the **glm()** command.

#Poisson data:

A <- rpois(10000, 1) #generates a poisson-variable with mean = variance = 1

B <- rpois(10000, 1.2) #generates a poisson-variable with mean = variance = 1.2

pois.data <- data.frame(c(rep("A", 10000), rep("B",10000)), c(A,B)) #Create a dataset

mod.pois <- glm(pois.data[,2] ~pois.data[,1], family = "poisson") #generalized linear model

#Note that we have specified the distribution assumed for the response using the “family” argument.

summary(mod.pois)

Anova(mod.pois)

**4.3a. Try to interpret the model output.** This is tricky, because when you analyze data with generalized linear models, the response is actually transformed and interpreted on a different scale (often some kind of logarithmic scale). Therefore, the coefficients in the model output are not straightforward to understand, and may need to be back-transformed to make sense. **Try to exponentiate the intercept and the coefficient giving the effect of treatment (A vs. B). In R you do this; exp(value). Recognize something?**

Note also that the P-value can be calculated using the z-value, or tested using a chi-square distribution. The latter is essentially testing how much residual variance there is in your fitted model (including both intercept and treatment effect) relative to a model including only the intercept (which will contain more residual variance – significantly more? – this is what the P-value tells you).

#Binomial model (logistic regression):

A <- rbinom(10000, 1, 0.2) #generates a binomial-variable probability of success = 0.2

B <- rbinom(10000, 1, 0.1) #generates a binomial-variable probability of success = 0.1

binom.data <- data.frame(c(rep("A", 10000), rep("B",10000)), c(A,B)) #Create a dataset

mod.binom <- glm(binom.data[,2] ~binom.data[,1], family = "binomial") #generalized linear model

summary(mod.binom)

Anova(mod.binom)

**4.3b. Go through this model and try to interpret it.** For logistic regression, the logit transformation is used, so the response is again modelled on a log-scale, but as a logged fraction: logit(p) = log(p/(1-p)). By exponentiation and re-arranging we can back-transform the coefficient to get the predicted probability for a given group: Pr = **exp(b1)/[1+exp( b1)]**. **Try to see if you can make sense of the model output for the intercept and the effect of treatment like you did for the Poisson model.**

This is tricky. The take home message is that when you use GLMs your data are no longer modelled on the original scale, and while you can trust the P-values and model statistics (given that the chosen distribution is a good match to your data), you have to think carefully about what your model coefficients mean.

**4.4. Advanced bonus question: Linear Mixed Effects Models (LMEs)**

Read the data in **Jimson.txt**. They consist of the length/width ratio (LenWid) of second seedling leaves of two types of Jimsonweed, called globe (coded as 1) and nominal (coded as 2). Three seeds of each type were planted in 16 pots (the variable named [Pot] gives the pot identification number). Thus, there are 16 pots, but 16\*6 = 96 seeds in the experiment. You might assign G and N as level labels for the types and you also need to make Pot into a factor, because R typically interprets variables with numbers as either continuous or integer variables:

**dat$Type<-factor(dat$Type, levels=c(1, 2), labels=c("G", "N"))**

**dat$Pot <- factor(dat$Pot)**

Let’s first perform an analysis of variance including “pot” as a fixed effect (i.e. we consider the seeds/individual plants as experimental units representing the true level of replication of the effect of weed “type”):

**fm1a <- aov(LenWid ~ Type + Pot + Type:Pot, data=dat)**

**summary(fm1a)**

Let’s also perform an analysis of variance including [Pot] as a random effect (i.e. we consider the pots as randomly selected experimental units representing the true level of replication of the effect of weed “type”). We do this by specifying “Error” terms with aov(). The first error term within the parenthesis below [Pot] estimates how much pots vary in the total growth of seeds. The second error term [Type:Pot] estimates how much pots vary in the difference in performance of the two seed types. Our hypothesis is that the two seed types generally differ, so this second error term seems most relevant for the uncertainty in our estimate of the effect size (the mean difference in growth between type N and G).

**fm1b <- aov(LenWid~Type+Error(Pot+Type:Pot), data=dat)**

**summary(fm1b)**

**4.4a Compare the output from these two models – how do they differ? Do you understand how the P-values are calculated from the sums of squares (SS -> MS -> F -> P-value) for the effect of “type” in each model?**

Now restrict the dataset to only the first two pots

**dat2 <- subset(dat, (Pot=="16533") | (Pot=="16534"))**

*The question is now whether the data from these two pots are sufficient to conclude that globe and nominal differ in length/width ratio*.

**4.4b. Do a two-way analysis of variance on LenWid with [Type] and [Pot] as factors. As you did above for the 16 pots, first treat [Pot] as a fixed effect and then repeat the analysis with [Pot] as a random effect. Which type of analysis do you think is to be preferred?**

*To help you think about this question, it might be of interest to look at the entire dataset, and in particular to check for the presence of a Type × Pot interaction*.

**bwplot(LenWid ~ Type|Pot, data = dat)**

The effect that you are interested in is if one Type has a different Length-width ratio then the other. If the difference between the types are consistent across all the 16 pots (or say in 13 out of the 16) we would feel quite convinced that there is indeed a difference. But with only 2 pots, how sure can you be? – your number of observations could be considered to be the two pots, and not the total number of plants grown from them. To control for this you have calculated the F-ratio for the fixed effect “Type” using the variance (i.e. difference) among the two pots in the effect of Type – but when analyzing pot:type as a random term in your model you only have 1 degree of freedom for the error term! Using the dataset with all 16 pots the sample size for the observation of variance in the effect of Type between pots is increased substantially, and now you see that there is hardly any variance between pots in the effect of type (the effect of Type is more or less the same in all pots), so the main effect of Type is significant in this more powerful test with larger sample size.

*To help you think more about this you can look through pages 519-527 in the R book by Crawley.*