ANOVA (CRD)

- General format of ANOVA in R
- Testing the assumption of homogeneity of variances using Levene's Test
- One-way ANOVA of nested design
- Obtaining and interpreting Components of Variance

Linear Models in R

The primary R function for the analysis of variance of a fixed-effects model is the linear model (**lm()**) function. The **lm()** function has the following basic syntax:

```
lm(model, dataset)
```

The **model** is an explanatory linear model for the response variable. It is a minimalist representation of a general linear effects model, where a response (i.e. dependent variable) is explained by a host of known additive deviations from a base mean:

$$Y_i = \mu + \kappa_1 + \kappa_2 + ... + \kappa_n + \varepsilon_i$$

For a completely randomized design (CRD) with one treatment variable, the independent effect is the treatment and the dependent variable is the response. The general syntax is:

```
dependent variable ~ independent effects
```

For a single-factor CRD, the R script would be:

```
lm(Response ~ Treatment, data = dataname)
```

There can be more complex models, like those below; and the **lm()** function can handle them:

```
y ~ a b              -> main effects of a factorial experiment
y ~ a b a*b          -> including their interaction
```

ANOVAs in R

Now, to view the results of the above linear model analysis in the familiar form of an ANOVA table, one must apply a second R function (**anova()**) to the object produced by the **lm()** function. The general workflow involves the following two lines of code:

```
data.mod <- lm(Response ~ Treatment, data = dataname)
anova(data.mod)
```

Example 1 *[Lab3ex1.R]*

In this experiment, the nitrogen fixation capacities of six different strains of rhizobia on clover are compared. The experiment is arranged as a CRD with 6 treatments (i.e. 6 different strains) and five independent replications (e.g. plots) per treatment.

```
#This script performs a CRD analysis (one-way ANOVA) and tests for
homogeneity of variance

#Load the data, then use the following function to examine the structure of
your data
str(clover_dat, give.attr = F)

#Convert clover_dat to a dataframe and tell R that "Culture" is a
classification variable; re-examine
clover_dat <- as.data.frame(clover_dat)
clover_dat$Culture <- as.factor(clover_dat$Culture)
str(clover_dat, give.attr = F)

#Apply the linear model and look at the results (ANOVA table)
clover_mod<-lm(NLevel ~ Culture, data = clover_dat)
anova(clover_mod)

#Some more results
summary(clover_mod)

#Create box-and-whisker plot of the data
plot(clover_dat)
```

Results of the lm() and anova() functions:

```
Analysis of Variance Table

Response: NLevel
          Df Sum Sq Mean Sq F value    Pr(>F)
Culture    5 847.05 169.409   14.37 1.485e-06 ***
Residuals 24 282.93  11.789
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation Recall that the null hypothesis of an ANOVA is that all means are equal (H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$) while the alternate hypothesis is that at least one mean is not equal (H_1: $\mu_i \neq \mu_j$). With a p-value of less than 0.00001, we soundly reject H_0. A significant component of variation exists in the dataset due to treatments.

Results of the summary() function:

```
Residual standard error: 3.433 on 24 degrees of freedom
Multiple R-squared:  0.7496,  Adjusted R-squared:  0.6975
F-statistic: 14.37 on 5 and 24 DF,  p-value: 1.485e-06
```

Interpretation The "residual standard error" is the "root MSE" (i.e. the square root of the mean square error $\sqrt{11.789} = 3.433$). The R-Square value is a measure of the amount of variation explained by the model:

$$\text{R-Square} = 847.046667 / 1129.974667 = 0.749616$$

Testing the Assumption of Homogeneity of Variances (HOV)

When you perform an ANOVA, you assert that the data under consideration meet the assumptions for which an ANOVA is valid. We briefly discussed normality in Lab 1. Now we will cover a second assumption, that the variances of all compared treatments are homogeneous.

…by an ANOVA of residuals

```
#extract residual values
clover_res<-residuals(clover_mod)

#add these residual values to the original dataframe
clover_dat$res<-clover_res
clover_dat$absres<-abs(clover_res) #Levene's original test
clover_dat$res2<-clover_res^2 #more robust test, using squared residuals
head(clover_dat)

#Test for homogeneity of variances (Levene's Test)
leveneABS_mod<-lm(absres ~ Culture, data = clover_dat)
anova(leveneABS_mod)
leveneRES2_mod<-lm(res2 ~ Culture, data = clover_dat)
anova(leveneRES2_mod)
```

Results:

```
Response: absres
          Df Sum Sq Mean Sq F value  Pr(>F)
Culture    5 48.957  9.7915  3.1451 0.02531 *
Residuals 24 74.718  3.1133

Response: res2
          Df Sum Sq Mean Sq F value Pr(>F)
Culture    5 2563.0  512.59  1.8433 0.1423
Residuals 24 6674.1  278.09
```

Interpretation The original Levene test indicates a violation of HOV, while the more robust test based on squared residuals does not. In 1974, Brown and Forsythe investigated this sort of disparity. They extended the Levene test to use residuals calculated on group medians and trimmed means, in addition to just means, and found *the method based on group medians to be the most generally robust technique*. This is the default technique used by R and the one we will be using:

```
#The Levene Test is available through the "car" package
#install.packages("car")
#library(car)

leveneTest(clover.mod, center = mean) #same as abs(res)
leveneTest(clover.mod, center = median) #the default
```

Results:

```
Levene's Test for Homogeneity of Variance (center = mean)
      Df F value  Pr(>F)
group  5  3.1451 0.02531 *
      24

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  0.9295 0.4794
      24
```

Interpretation H_0 for Levene's Test is that the variances of the treatments are homogeneous, and H_a is that they are not homogeneous. With a p-value of 0.4794, we fail to reject H_0; thus we have no evidence, at the 95% confidence level, that the variances are not homogeneous. *If you would like to see how to perform an ANOVA on the absolute values of median-based residuals, refer to the code on the next page.*

Finally, there are other HOV tests, also available through the "car" package:

```
#alternative HOV tests
bartlett.test(NLevel ~ Culture, data = clover.dat)  #parametric test
fligner.test(NLevel ~ Culture, data = clover.dat)  #non-parametric test
```

```
      Bartlett test of homogeneity of variances

data:  NLevel by Culture
Bartlett's K-squared = 14.2207, df = 5, p-value = 0.01427

      Fligner-Killeen test of homogeneity of variances

data:  NLevel by Culture
Fligner-Killeen:med chi-squared = 3.7264, df = 5, p-value = 0.5894
```

For the interested:
Performing an ANOVA on the absolute values of the median-based residuals (Brown and Forsythe's improved Levene's Test)

```
#Testing for homogeneity of variances using Brown and Forsythe's improved
 Levene's Test,
#based on an ANOVA of the absolute values of median-based residuals

#To find residuals based on MEDIANS, first calculate the medians for the 6
 treatment classes
aggregate(clover_dat$NLevel, list(clover_dat$Culture), median)
#Create a medians column in the original dataset
clover_dat$medians<-
 c(rep(32.1,5),rep(24.8,5),rep(15.8,5),rep(20.5,5),rep(14.2,5),rep(19.1,5))
#Calculate residuals, based on medians
clover_dat$median_resids<-clover_dat$NLevel - clover_dat$medians
#Find the absolute values of those residuals
clover_dat$median_resids_abs<-abs(clover_dat$median_resids)
clover_dat

#Perform the ANOVA
leveneMEDIAN_ABS_mod<-lm(median_resids_abs ~ Culture, data = clover_dat)
anova(leveneMEDIAN_ABS_mod)
```

Results:

```
Analysis of Variance Table

Response: median.resids.abs
          Df  Sum Sq Mean Sq F value Pr(>F)
Culture    5  37.062  7.4123  0.9295 0.4794
Residuals 24 191.392  7.9747
```

The results match those of the default Levene's Test procedure in the "car" package.

One-Way ANOVA of Nested Design

Example 2 *[Lab3ex2.R]*

In this experiment, the effects of different treatments on the growth of mint plants are being measured. It is a purely nested CRD with six levels of treatment, three replications (i.e. pots) per level, and four subsamples (i.e. plants) per replication.

```
#This script performs a one-way ANOVA for a perfectly nested CRD and finds
 variance components

#read in, re-classify, and inspect the data
mint_dat <- as.data.frame(mint_dat)
mint_dat$Trtmt <- as.factor(mint_dat$Trtmt)
mint_dat$Pot <- as.factor(mint_dat$Pot)
mint_dat$Plant <- as.factor(mint_dat$Plant)
str(mint_dat, give.attr = F)
head(mint_dat)
tail(mint_dat)
```

```
'data.frame': 72 obs. of  4 variables:
 $ Trtmt : Factor w/ 6 levels "Cond1","Cond2",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Pot   : Factor w/ 3 levels "Pot1","Pot2",..: 1 1 1 1 2 2 2 2 3 3 ...
 $ Plant : Factor w/ 4 levels "Plant1","Plant2",..: 1 2 3 4 1 2 3 4 1 2 ...
 $ Growth: num  3.5 4 3 4.5 2.5 4.5 5.5 5 3 3 ...
```

```
   Trtmt  Pot   Plant  Growth
1  Cond1  Pot1  Plant1    3.5
2  Cond1  Pot1  Plant2    4.0
3  Cond1  Pot1  Plant3    3.0
4  Cond1  Pot1  Plant4    4.5
5  Cond1  Pot2  Plant1    2.5
...
68 Cond6  Pot2  Plant4      7
69 Cond6  Pot3  Plant1     11
70 Cond6  Pot3  Plant2      7
71 Cond6  Pot3  Plant3      9
72 Cond6  Pot3  Plant4      8
```

```
#Calculating components of variance
#install.packages("lme4")
#library(lme4)
mintCV_mod<-lmer(Growth ~ Trtmt + (1|Trtmt:Pot), data = mint_dat)
summary(mintCV_mod)
```

Output

```
Random effects:
 Groups    Name        Variance Std.Dev.
 Pot:Trtmt (Intercept) 0.3047   0.5520
 Residual              0.9340   0.9665
Number of obs: 72, groups: Pot:Trtmt, 18
```

```
#The ANOVA
mint_mod<-lm(Growth ~ Trtmt/Pot, data = mint_dat)
anova(mint_mod)

#The above model is short-hand for this:
mint2_mod<-lm(Growth ~ Trtmt + Trtmt:Pot, data = mint_dat)
anova(mint2_mod)
```

With the above scripts, R will generate the following ANOVA table, which presents an *incorrect* F test for the effect of treatment:

```
Response: Growth
          Df  Sum Sq Mean Sq F value   Pr(>F)
Trtmt      5 179.642  35.928 38.4662 < 2e-16 ***   WRONG!
Trtmt:Pot 12  25.833   2.153  2.3048 0.01858 *
Residuals 54  50.438   0.934
```

We know from the expected mean squares that the correct error term to use for testing the effect of treatment is **MS$_{Pot}$** (the MSEE), not **MS$_{Plant}$** (MSSE). To conduct the correct test, we type:

```
#Custom F-test for Trtmt
F_Trtmt<-35.928/2.153
F_Trtmt
pf(F_Trtmt, 5, 12, lower.tail = F)
```

This returns the correct F-value (**16.69**) and p-value (**4.88e-05**) for the effect of treatment.

If this seems unnecessarily complicated, it is. Note that the same F value will result if you simply consider the averages of the subsamples, though the MS's will be different. **If you are not concerned with the components of the variance, there is no need to carry out a nested analysis**. Just average the subsamples and treat it as an un-nested experiment, as shown in the following example.

Example 3 *[Lab3ex3.R]*

This is the same experiment as above, but here the measurements of the four plants (subsamples) in each pot have been averaged.

```
#This script performs a one-way ANOVA, where subsamples have been averaged

#read in, re-classify, and inspect the data
mint_dat <- as.data.frame(mint_dat)
mint_dat$Trtmt <- as.factor(mint_dat$Trtmt)
mint_dat$Pot <- as.factor(mint_dat$Pot)
str(mint_dat, give.attr = F)

#The ANOVA
mint_mod<-lm(Growth ~ Trtmt, data = mint_dat)
anova(mint_mod)
```

```
'data.frame': 18 obs. of  3 variables:
 $ Trtmt : Factor w/ 6 levels "Cond1","Cond2",..: 1 1 1 2 2 2 3 3 3 4 ...
 $ Pot   : Factor w/ 3 levels "Pot1","Pot2",..: 1 2 3 1 2 3 1 2 3 1 ...
 $ Growth: num  3.75 4.38 2.88 4.5 3.5 ...
```

```
   Trtmt  Pot  Growth
1  Cond1 Pot1   3.750
2  Cond1 Pot2   4.375
3  Cond1 Pot3   2.875
4  Cond2 Pot1   4.500
5  Cond2 Pot2   3.500
6  Cond2 Pot3   4.375
...
13 Cond5 Pot1   5.500
14 Cond5 Pot2   6.625
15 Cond5 Pot3   7.250
16 Cond6 Pot1   8.250
17 Cond6 Pot2   6.750
18 Cond6 Pot3   8.750
```

```
#The ANOVA
mint_mod<-lm(Growth ~ Trtmt, data = mint_dat)
anova(mint_mod)
```

Output

```
          Df Sum Sq Mean Sq F value    Pr(>F)    
Trtmt      5 44.911  8.9821  16.689 4.881e-05 ***
Residuals 12  6.458  0.5382
```

We arrive at the same result as before, but with much less complication.

For the interested – some sweet code

Wouldn't it be great if you could run an ANOVA on the Pot means without having to first calculate those means in Excel and then re-import a reduced dataset? The following code, which relies on the ddply() function within the "plyr" library, is what you're looking for:

```
pot_means_dat <- ddply(mint_dat, c("Trtmt", "Pot"), summarise,
                   Growth = mean(Growth, na.rm=TRUE))
```

This one-liner creates a reduced dataset for you (I call it "pot_means_dat"), in which the Growth measurements of all the Plants in a given Pot have been averaged. The elements in green are the terms you would need to change, based on your specific dataset. Give it a try! And thanks to Natalie, a previous BIOL933 student, for figuring this one out.