ENV 710 – Applied Data Analysis
Fall 2014
Lab 7: One-way Analysis of Variance

Up to now, we have conducted one- or two-sample tests. Here we extend the statistical principles gained through these tests to slightly more complicated datasets. Recall that the Z-test was used to compare the mean of a single sample to a hypothetical population mean. We used the t-test (and resampling when the data did not meet assumptions of normality) to compare the means of two samples. One-way ANOVA (analysis of variance) is a method used to compare *two or more groups to determine if the means are the same or different*.

For example, we could use ANOVA to ask whether the quantity of fertilizer (none, low, medium and high) results in significantly different levels of plant growth. Note that our treatment, fertilizer, is a *factor*, with four levels. The *dependent variable* (the variable that depends on the level of fertilization) in this case would be plant growth.

The null hypothesis of this experiment is that the mean plant growth is the same for each fertilizer level ($H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$), versus the alternative hypothesis that at least one of means is different from one of the others. Of course, this experiment should be *replicated* and *randomized*. So if plants were being grown in pots in a nursery, for example, we would apply each level of fertilization to 7-10 pots (N = 40 with 10 pots per treatment). The treatment applied to each pot should be randomly chosen to avoid *confounding* effects. The pots with the same treatment should not all be located together in the green house and should not contain potting soil from the same bag, etc.

When using ANOVA, we make a few assumptions: (1) the dependent variable is normally distributed for each level of the treatment (e.g. fertilization); (2) all observations are independent of each other, within and between groups; and (3) variances of each level of treatment are homogeneous. What? Each treatment level has approximately the same variance (the largest standard deviation of a treatment is less than two times the smallest).

The goals of the lab are to:

1. Introduce the concept of ANOVA in an applied manner.
2. Practice one-way ANOVA in **R**, using it to evaluate tree biomass in relation to different forest types.
3. Learn graphing methods to depict the results of ANOVA.

At the end of the lab, there are a few questions to answer. Please type your answers to each problem, including any requested graphs, in a Word document. *Submit your answers and your **R**-code to the class Sakai site under the folder Assignments before 5 pm on Wed., Oct. 15 (Sections 01, 02, and 07).*

More functions in R

In this lab, we introduce a few new **R** commands.

`aov()` - Fits an ANOVA model by a call to `lm` for each stratum.
`edit()`- Invokes the **R** editor for modifying and viewing data frames.
`levels()` - Displays the unique levels of a categorical (factor) variable.
`jitter()` - Adds a small amount of noise to a numeric vector; good for graphing and offsetting potentially overlapping points so that they can all be seen.
`par()`- Sets global graphics parameters. We used it to alter the number of graphs displayed in the graphics window from one to two.
`tapply()` - Stands for table apply. It applies a function (3rd argument) to a variable (1st argument) separately for each group specified by the second argument.
`~` - Tilda, the symbol used in defining expressions for model fitting.

Verifying the assumptions of ANOVA

We are going to use the Africa plots data to test whether the three forest types (logged/hunted forest, logged only forest, pristine forest) differ in biomass. What is the null hypothesis in this case? What is the alternative hypothesis? The "natural experiment" is replicated because there are multiple plots in each forest type. The experiment is randomized in the sense that plot locations were chosen randomly within each forest type. However, this experiment is not perfectly randomized because the plots of each forest type are grouped spatially. This was unavoidable as it was a constraint of the physical environment where the study was conducted, but it does raise questions about the inferences that can be made.

To keep things simple, we will analyze these data for the first census of the plots.

```
adat <- read.csv("Afrplots.csv", header=T)
adat$Site <- as.factor(rep(c(rep(1,10), rep(2, 10),
  rep(3, 10)), 2))
bdat <- adat[adat$CensusNo == 1,]
attach(bdat)
```

To make sure that Site is correctly classed as a factor, we can ask for the levels of the factor.

```
levels(Site)
```

In the above code, we downloaded the data and added a column in the database attributing each forest type a code: 1 = logged/hunted forest; 2 = logged only forest; 3 = pristine forest. Then we created a new database (`bdat`) that only includes the first census and attached the database so that it is in the current **R** search path.

Let's take a look at the data for each sample. The below code changes the number of printing panels that we see on the screen to make it easier to compare plots here and below. The three lines of points add the raw data to the boxplot.

```
par(mfrow=c(2,2))
boxplot(ChaveMoist ~ Site, las=1, ylab = "Biomass",
        xlab = "Sites")
 points(jitter(rep(1, length(Site[Site==1])), f=4),
         ChaveMoist[Site==1], col = "darkgreen")
 points(jitter(rep(2, length(Site[Site==2])), f=4),
         ChaveMoist[Site==2], col = "darkred")
 points(jitter(rep(3, length(Site[Site==3])), f=4),
         ChaveMoist[Site==3], col = "darkblue")
```

There could be some deviation from our assumption of normal distributions.

```
boxplot(log(ChaveMoist) ~ Site, las=1,
        ylab = "log(Biomass)", xlab = "Sites")
points(jitter(rep(1, length(Site[Site==1])), f=4),
       log(ChaveMoist[Site==1]), col = "darkgreen")
points(jitter(rep(2, length(Site[Site==2])), f=4),
       log(ChaveMoist[Site==2]), col = "darkred")
points(jitter(rep(3, length(Site[Site==3])), f=4),
       log(ChaveMoist[Site==3]), col = "darkblue")
```

A log-transformation of the data does not seem to center the median although it might reduce the skew a bit in the tails. What happens if we try qq plots?

```
qqnorm(ChaveMoist[Site==1])
 qqline(ChaveMoist[Site==1])
```

Run the qq plots for the other Sites for the raw data and log-transformed data. Log-transforming the data does not seem to make much of a difference. We can try the Shapiro-Wilk normality test to test whether the distribution of each sample is significantly different from a normal distribution.

```
shapiro.test(ChaveMoist[Site==1])
```

Try the Shapiro-Wilk test for the other two samples.

None of these tests demonstrate that the samples are significantly different from the H_0 of a normal distribution.

If we were really ambitious, we could use resampling method to test our null hypothesis since the data do not appear to fit a normal distribution that well. Before doing the ANOVA, let's test the assumption of homogeneity of variances by evaluating the ratios of the sample standard deviations.

```
sd(ChaveMoist[Site==1])/sd(ChaveMoist[Site==2])
sd(ChaveMoist[Site==2])/sd(ChaveMoist[Site==3])
sd(ChaveMoist[Site==1])/sd(ChaveMoist[Site==3])
```

The ratios range from 0.52 to 1.73. This is less than our criterion of the largest standard deviation being two times bigger than the smallest.

Conducting one-way ANOVA

Now we will conduct the one-way ANOVA. Recall that we are testing for differences among means of the levels of the factor, Site. In other words, do logged/hunted, logged only, and pristine forests all have the same mean biomass?

```
mod1 <- aov(ChaveMoist~factor(Site))
summary(mod1)
```

Note the syntax used to define the model: dependent variable ~ independent variable. We will use the same syntax to define future models, including other linear models like regression and generalized linear models. Here, the `aov()` function expresses the results of the test in the traditional language of analysis of variance, providing the sum-of-squares (`Sum Sq`), mean squares (`Mean Sq`) and *F*-statistic (`F value`).

Interpreting results of our test

Study the results from `summary()` and note the large *F*-statistic and very small *p*-value. These tell us that we should *reject the null hypothesis that the three forest types have the same mean biomass values in favor of the alternative hypothesis that there is a difference in mean biomasses among the forest types*. Note that we do not yet know which forest types are significantly different from each other.

To determine which of the means are significantly different from one another we conduct a *post-hoc* test. We will use Tukey's 'Honest Significant Difference' method.

```
TukeyHSD(mod1)
plot(TukeyHSD(mod1))
```

The output shows the *difference in mean biomass* in pairwise comparisons of the Sites, confidence intervals of the difference, and the probability of the Sites having the same mean biomass. The plot presents the 95% CI's for the *differences* between the pairwise comparisons.

In this example, there is not a statistically significant difference between sites 2 and 1 (logged only and hunted/logged forest) as illustrated by the fact that the CI overlaps 0 and the p-value is greater than 0.05. By contrast, there appear to be significant differences in forest biomass between sites 1 and 3 and sites 2 and 3.

Graphing the results of ANOVA

As you have seen above, a boxplot is a good way to demonstrate results. Barplots are another commonly used method for visualizing results. **R**, inconveniently, does not have a automatic function to do error bars. Therefore, we need to write a function to plot the barplot, and then we have to add the error bars ourselves (+/- 1 standard error). The function is below. Type in the function and then run it to produce a barplot of the biomass data per forest type (logged/hunted, logged only, pristine forest).

```
error.bars <- function(yvalues, se, nm){
    xv <-
     barplot(yvalues, ylim=c(0, (max(yvalues)+max(se))),
      names=nm, ylab=deparse(substitute(yvalues)), las=1)
       g <- (max(xv)-min(xv))/50

  for (i in 1:length(xv)){
      arrows(xv[i], yvalues[i] + se[i], xv[i],
             yvalues[i]-se[i], length=0.1, angle=90,
             code=3)
     }
}
```

The function will not work until we give it y-values (`yvalues`), standard error (`se`), and names (`nm`), which we need to get from the data.

Below `tapply()` is used to find the mean value of biomass by site. After using `tapply()` to find the mean, we use it to find the number of plots per site and the SD of biomass by site, which is used to calculate the SE.

```
site.mean <- tapply(ChaveMoist, list(Site), mean)
site.n <- tapply(ChaveMoist, list(Site), length)
site.sd <- tapply(ChaveMoist, list(Site), sd)
site.se <- site.sd/sqrt(site.n)
site.se[is.na(site.se)==T] <- 0
labls <- as.character(levels(Site))
yvals <- as.vector(site.mean)
```

OK, we now have all the pieces to run the function. Let's run it and see what the plot looks like.

```
error.bars(yvalues = yvals, se = site.se, nm = labls)
```

We could make the plot a lot prettier, and the y-axis label needs work. Personally, I think boxplots or plots of the group means and their 95% confidence intervals represent the results better than the barplot above.

There is a lot more to learn about ANOVA, including how to implement two-way ANOVA's and to account for interactions. We will discuss these in lecture and try them in the future.

Problem #1: Your assignment is to conduct a one-way ANOVA to determine if the average weight of confiscated elephant tusks has decreased over time. Elephants are poached for their ivory, and USFWS authorities confiscate ivory when they find it entering the country. The data in "TuskData.csv" are the average weights of elephant tusks from 20 different seizure sites in 1970, 1990, and 2010. In a one-page lab write-up, please state the H_0 and H_a of your test, describe how you checked the assumptions of your statistical test, and report the results of the ANOVA and a post-

hoc test. Please type your answers in a Word document, and include your R-code in an appendix. Include at least one graph that shows the means and standard errors or confidence intervals of the weight of tusks over the three years. It is not necessary to include a barplot if you prefer to graph your results in a different way.