Lab 7: One-way Analysis of Variance

September 29, 2018

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 problems. 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 analysis of variance (ANOVA) compares the means of two or more groups.

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 (nominal values). The dependent variable (the variable that depends on the level of fertilization) is plant growth. The null hypothesis is that the mean plant growth is the same for each fertilizer level versus the alternative hypothesis that at least one of the means is different from one of the others.

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_a : \text{Not } \mu_1 = \mu_2 = \mu_3 = \mu_4$$

Of course, the 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 10 pots (N = 40 with 10 pots per treatment). The treatment applied to each pot should be randomly chosen to avoid confounding effects. For example, pots undergoing the same treatment should not 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;
3. variances of each level of treatment are homogeneous (that is, 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:

- Introduce the concept of ANOVA in an applied manner.
- Practice one-way ANOVA in R, using it to evaluate differences in mean tree biomass across three forest types.
- Learn graphing methods to depict the results of ANOVA.

Work through the lab, running the example code by typing it into R Studio (do not copy and paste from the pdf) - make sure you know what every line does. At the end of the lab, there are a few problems to answer. Please answer the problems using R Markdown to show your code and any requested graphs. *Submit your answers and your R-code to the class Sakai site under the Assignments folder before 11:55 pm on Wed., Oct., 10 (Sections 01, 02, and 03).*

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

- `par()` - sets global graphics parameters
- `tapply()` - stands for *table apply*, and applies a function (3rd argument) to a variable (1st argument) separately for each group specified by the second argument
- $\sim$ - tilda, the symbol used in defining expressions for model fitting

Verifying the assumptions of ANOVA

We are going to use the Africa plot data to test whether the three forest types (logged & hunted forest, logged only forest, pristine forest) differ in mean levels of aboveground biomass. What is the null hypothesis? What is the alternative hypothesis? This "natural experiment" is replicated in the sense that 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, the experiment is pseudoreplicated because the plots of each forest type are grouped spatially (i.e. treatments were not replicated). This was unavoidable as it was a constraint of the physical environment of the study site, but it does raise questions about the inferences that can be made from the study.

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)
```

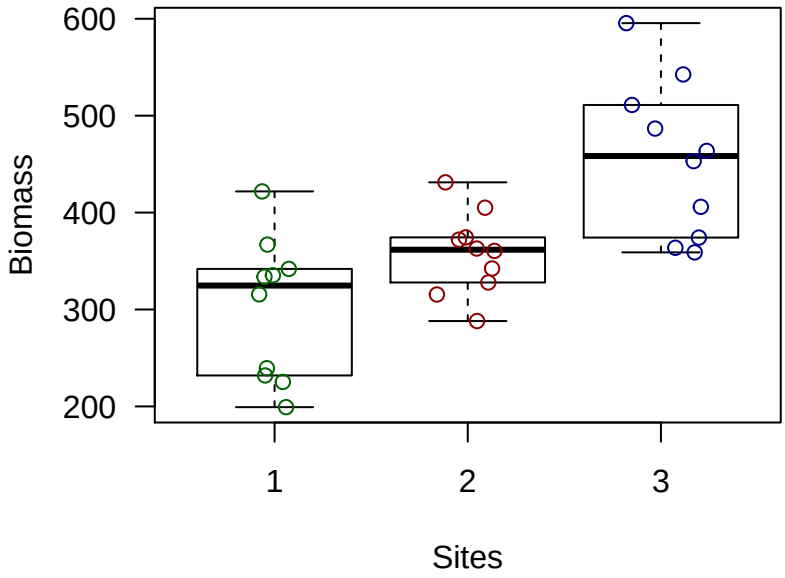
To make sure `Site` is correctly classed as a factor, we evalute the levels of the factor.

```
levels(Site)
```

In the above code, we downloaded the data and added a column, 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.

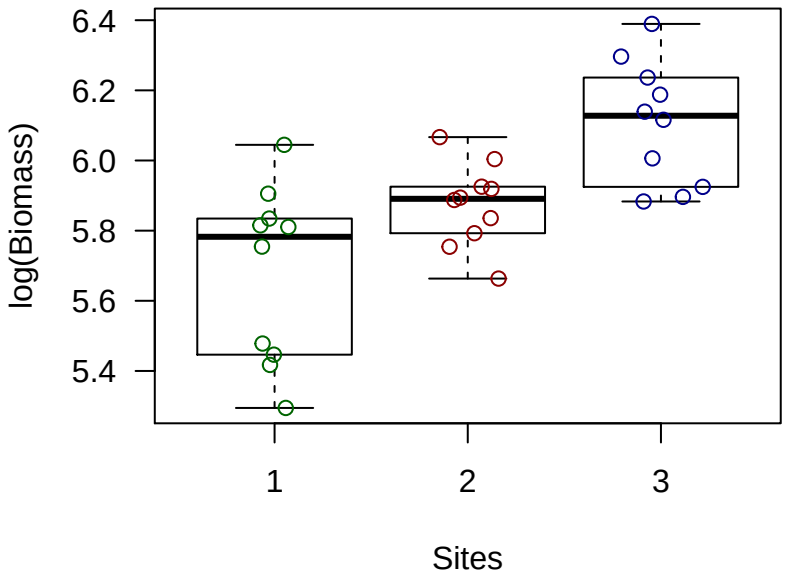
Let's take a look at the data for each sample. The below code changes the number of printing panels to make it easier to compare plots here and below. The three lines of `points` add the raw data to the boxplot. Note that in the `boxplot()` function, we are using the formula format to specify the boxplots. This creates a boxplot of biomass for each levels of `Site`, as long as the independent variable is a factor.

```
par(mfrow=c(1,1))
 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")
```



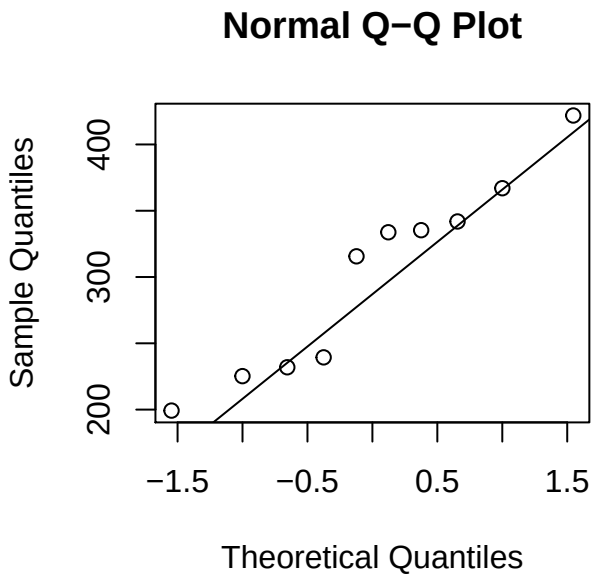
There could be some deviation from our assumption of normally distributed data for each group or level of site. Log transform the data and see if the boxplots look better.

```
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", pch = 21)
 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 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? Here we plot a qqplot for the first site, logged & hunted forest.

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



To Do

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 the null hypothesis that the data for each group comes from a normal distribution.

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

To Do

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

None of these tests demonstrate that the samples are significantly different from the H_0 of a normal distribution. This could be due to a lack of power.

If we were really ambitious, we could use resampling to test our null hypothesis because the data do not appear to fit a normal distribution that well. However, we are going to accept the results of the test and move on.

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 ANOVA language, providing the sum-of-squares (`Sum Sq`), mean squares (`Mean Sq`) and F-statistic (`F value`).

ANOVA's are linear models with with factors as independent variables. Therefore, we could achieve the same results with the linear models function, `lm()`.

```
mod2 <- lm(ChaveMoist~factor(Site))
 summary(mod2)
```

```
##
## Call:
## lm(formula = ChaveMoist ~ factor(Site))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -101.882  -58.691    6.521   39.129  139.966
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     301.13      21.24  14.178 5.00e-14 ***
## factor(Site)2    56.87      30.04   1.893   0.0691 .
## factor(Site)3   154.40      30.04   5.140 2.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 67.16 on 27 degrees of freedom
## Multiple R-squared:  0.5003, Adjusted R-squared:  0.4633
## F-statistic: 13.52 on 2 and 27 DF,  p-value: 8.555e-05
```

The summary of the results of `mod2` provides a lot more information, which we will discuss later, but notice that the last line of the results gives the same `F` statistic and `p` value as `mod`.

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 biomass between at least two of the sites. 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 – Tukey's Honest Significant Difference.

```
TukeyHSD(mod1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = ChaveMoist ~ factor(Site))
##
## $`factor(Site)`
##          diff       lwr      upr     p adj
## 2-1  56.86767 -17.60653 131.3419 0.1599999
## 3-1 154.40412  79.92992 228.8783 0.0000605
## 3-2  97.53645  23.06225 172.0107 0.0084513
```

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

The output shows the difference, `diff`, in mean biomass in pairwise comparisons of the sites, confidence intervals, `lwr` and `upr`, of the difference, and the probability, `p adj` 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 an automatic function to do error bars. Therefore, we need to write a function to create the barplot, and then we have to add the error bars ourselves (± 1 standard error). The function is below. Run it to produce a barplot of the biomass data per forest type.

The function will not work until we give it values for the arguments `yvalues`, the bar heights, `se`, standard error, and `nm`, label names, which we need to get from the data.

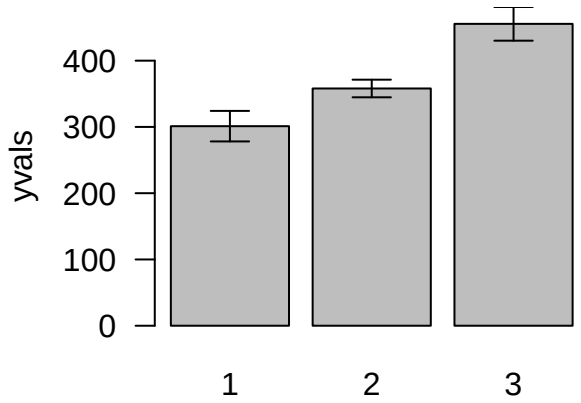
```
error.bars <- function(yvalues, se, nm){
 xv <- barplot(yvalues, ylim=c(0, (max(yvalues)+max(se))),
               names=nm, ylab=deparse(substitute(yvalues)), las=1)

  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)
  }
}
```

Let's use `tapply()` 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 and the SD of biomass by site 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 call `error.bars()` 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. There is a growing concensus that boxplots or plots of the group means and their 95% confidence intervals represent data better than barplots.

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.

Problems

Problem 1

A company is looking for a faster internet service. To decide between two different internet providers (Turbo Net and Speed Web), the company performs an experiment in which it collects data on website loading times from each of the providers (download `Internet.cvs`). Is there a significant difference in the website loading times of the two providers? If so, which provider should the company choose? Make sure to do the following:

(a) specify your hypotheses, (b) check that the data meet the assumptions of the statistical test you plan to use, (c) if your data do not meet the assumptions of normality, test your hypotheses both by transforming the data and by using a nonparametric approach, and (d) write a short paragraph that explains the results of your statistical tests.

Problem 2

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 your 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. 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. Also, please annotate your code to explain what different sections mean.