ANOVA Analysis of Variance

Joseph J. Luczkovich

February 4, 2015

1 Introduction

So far we have looked at comparison of two groups using the t-test. What if there are more than two groups for which you have to compare means? For example, what if there were three fish species in which you wanted to compare growth rates? What if you wanted to compare mean percent cover after one year of trying to grow seagrasses using three different techniques, seed-planting, adult plant transplants, and berm-construction? You need to consider using Analysis of Variance or ANOVA.

2 ANOVA models

ANOVA is the most basic statistical procedure when three or more groups are compared. Like the t-test, it assumes a normally distributed data set, each group forming a bell-shaped curve, with the only differences the location of the true mean for each group and the variance around the group means. The samples from each population are also assumed to be independent, and the basic model assumed a balanced design (e.g., 100 samples per group). The null hypothesis H_o for ANOVA is:

$$\mu_1 = \mu_2 = \mu_3 \tag{1}$$

The alternative hypothesis H_a is:

$$\mu_1 \neq \mu_2 \neq \mu_3 \tag{2}$$

Assumptions: There are three or more groups (k), with independent random samples, and normal errors (normal distribution). One factor or treatment is varied $(1, 2, 3, .k)$ Treatment levels: These are considered factors, or categories. They can be a **fixed** or **random** factor Fixed factor: levels cannot vary (they are unique) Random factor: levels are drawn from a large population.

Here is the ANOVA Model:

$$X_{ij} = \mu + \alpha_i + \sigma_{ij} \tag{3}$$

and this can also be written:

$$X_{ij} = \bar{x}_{\cdot} + (\bar{x}_i - \bar{x}_{\cdot}) + (\bar{x}_{ij} - \bar{x}_i) \tag{4}$$

or in words, the observations are defined by the grand mean + the variation of each group mean around the grand mean + the variation of each observation around its group mean. The total variance is thus partitioned into the grand mean effect, the group mean effect, and the random error of individual samples.

The grand mean is the mean of all the observations no matter what the group membership and is written as $\bar{x}_{\cdot\cdot}$. The group mean is computed separately for each group i through k and is written as $\bar{x}_i$.

3 ANOVA example in R

Here are some plant biomass data taken from a dredge disposal study on a NC marsh. Dredge spoil from a navigation channel was placed on replicate marsh plots (4 per treatment group) at treatment levels of 0, 2, 4, and 10 cm spoil thickness. Then the biomass of marsh plants on each plot was measured in g/m2 after 1 month. Does the mean biomass vary significantly with dredge disposal? Are the means different?

0 cm	2 cm	4 cm	10 cm
37.7	20.4	10.0	0.0
21.4	24.8	3.4	4.0
22.2	32.0	3.4	1.2
14.2	53.3	30.4	0.0
$\bar{x_{0cm}} = 23.875$	$\bar{x_{2cm}} = 32.625$	$\bar{x_{4cm}} = 11.8$	$\bar{x_{10cm}} = 1.3$
$\bar{x}_{\cdot} = 17.4$			

Let's run a simple ANOVA in R. We will enter the data as a vector of biomasses on each plot and another vector of treament levels, then making them factors.

```
> dredgedata <-c(37.7, 21.4, 22.2, 14.2,
+                20.4, 24.8, 32.0, 53.3,
+                10.0,  3.4,  3.4, 30.4,
+                0.0,  4.0,  1.2,  0.0)
> levels<-c(0,0,0,0,
+           2,2,2,2,
+           4,4,4,4,
+           10,10,10,10)
> treat<-factor(levels)
> treat

 [1] 0  0  0  0  2  2  2  2  4  4  4  4  10 10 10 10
Levels: 0 2 4 10
```

Notice how the factor() command turns the levels vector into categories. Now, let's run an ANOVA with the aov() command, using the formula input method.

```
> aov(dredgedata~treat)

Call:
   aov(formula = dredgedata ~ treat)

Terms:
                   treat Residuals
Sum of Squares  2257.185  1433.195
Deg. of Freedom        3        12

Residual standard error: 10.92854
Estimated effects may be unbalanced

> fit1<-aov(dredgedata~treat)
> summary(fit1)

            Df Sum Sq Mean Sq F value  Pr(>F)
treat        3   2257   752.4     6.3 0.00821 **
Residuals   12   1433   119.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The first command runs the ANOVA and prints the output on the screen. The second command places the ANOVA output into a data structure that we called fit1. By typing summary(fit1), we can see what is in the ANOVA output in greater detail. This method of storing the output as a data structire is very useful, and we will use it often later. We can even plot the ANOVA output. Let's interpret the output. Look at the ANOVA table with the degrees of freedom (df), Sum of SqUares (Sum Sq), Mean Squared error (Mean Sq), F-test (F value) and the probability (P). Are the means (at least one mean) different? Which mean is different? We will answer these questiosn in class and discuss ways to compare the means and find which ones are different.