Tutorial 6: Analysis of Variance (ANOVA)

Goal: To provide a more in depth look at the quantitative methods and statistics necessary to examine multi-factor, multi-sample experimental designs.

Note: All text in the Arial font is instruction or explanation. All text in Courier font is input or output from R.

Step-1: One-way analysis of variance model

The simplest place to start out is with a review of the general linear model. Recall:

\[ Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} , \quad \varepsilon_{ij} \approx N(0, \sigma^2) \]

which implies that the null hypothesis sets all groups equal and thus \( \alpha_i=0 \). The \( \varepsilon_{ij} \) terms are assumed to be independent and have the same variance.

And recall that the nice feature of the model is that all of the individual observations can be decomposed in relation to the model:

\[ y_{ij} = \bar{y} + (\bar{y}_i - \bar{y}) + (y_{ij} - \bar{y}_i) \]

The braced terms are what ultimately assist us in determining our variation within groups and the variation between groups, which together sum to the total variation for the model. The F-test is derived from the ratio of the mean squares between relative to the mean squares within. All of this was reviewed in detail in lecture, so will not be repeated here.

Step-2: ANOVA example, reading in the data frame

Let’s jump into an example using R, so you can begin to become familiar with the syntax for ANOVA. Let’s consider an example where we are interested in understanding the relationship between dispersers and seeds in a conservation biology context. The data presented are mean cone size (mass in grams) of lodgepole pine in 16 study sites in three different types of environments in western North America (Edelaar and Benkman 2006). In this example, an “island” refers to a patch of lodgepole pine surrounded by other habitat (e.g., clearcuts) and isolated from other tracts of lodgepole pine. The islands may or may not have squirrels present. The “mainland” refers to an area that is large and extensive in its dominance by lodgepole pine (and always has squirrels present).
How would we approach this problem? Well, it is obvious that we have three groups of data, so ANOVA is appropriate. We also see that given the design of the experiment, and identified as the primary reason for study, is the issue of islands with and without squirrels (because this controls for any effects of forest isolation on mass of lodgepole pine cones. Thus, we are interested in an orthogonal contrast. The data are as follows:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Island, squirrels absent</td>
<td>9.6</td>
</tr>
<tr>
<td>Island, squirrels present</td>
<td>6.8</td>
</tr>
<tr>
<td>Mainland, squirrels present</td>
<td>6.7</td>
</tr>
</tbody>
</table>

To use the linear model (lm) function in R, it is necessary to have the data values in one vector and factor variable(s) describing the division in another. Enter the data into an Excel spreadsheet and save as a *.txt (DOS text) file. Open Notepad (under Accessories) and check to make sure that you have a clean file:

Just as we did in Tutorial-2, we need to now read in this file to R. To reduce the length of the path name, I store my files in a directory of C: called C:\temp, thus:

```r
> squirrels<-read.table("c:\temp\squirrels.txt",header=T)
> attach(squirrels)
```
To confirm the correct read of our data:

```
> squirrels
   cones habitat
 1   9.6  ISA
 2   9.4  ISA
 3   8.9  ISA
 4   8.8  ISA
 5   8.5  ISA
 6   8.2  ISA
 7   6.8  ISP
 8   6.6  ISP
 9   6.0  ISP
10  5.7  ISP
11  5.3  ISP
12  6.7  MSP
13  6.4  MSP
14  6.2  MSP
15  5.7  MSP
16  5.6  MSP
```

**Step-3: Graphing the data and univariate analysis**

As always, the VERY FIRST thing we do is graph the data and do univariate data analysis so we understand what we are working with and alert us to what assumptions we may need to explicitly address. You will likely be tempted to use the summary function:

```
> summary(squirrels)
   cones     habitat
Min.   :5.300   ISA:6
1st Qu.:5.925   ISP:5
 Median :6.650   MSP:5
 Mean   :7.150
3rd Qu.:8.575
 Max.   :9.600
```

Notice how the summary results in an analysis of the quantitative variable and a count summary of the categorical variable. This is useful only in as much as it indicates the number of observations contained in each of the habitat levels; however, the summary stats are for all of the data pooled (i.e., \( n = 16 \)) and not particularly relevant.

A much more informative approach is to use `tapply`:

```
> tapply(cones, habitat, summary)
```
$ISA$

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
</table>

$ISP$

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>5.30</td>
<td>5.70</td>
<td>6.00</td>
<td>6.08</td>
<td>6.60</td>
<td>6.80</td>
</tr>
</tbody>
</table>

$MSP$

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>5.60</td>
<td>5.70</td>
<td>6.20</td>
<td>6.12</td>
<td>6.40</td>
<td>6.70</td>
</tr>
</tbody>
</table>

Tapply is the function by which you can generate tables of summary data. The first argument of \texttt{tapply} is the variable to be summarized, the second is the variable by which the summary is to be classified, and the third argument is the function to be applied (mean, variance, summary, etc.). Often times a good way to screen data early on is the use of a strip-plot. This type of plot permits the visualization of the raw data (which other plots do not). The R-code follows:

```r
> xbar<-tapply(cones,habitat, mean)
> s<-tapply(cones,habitat,sd)
> n<-tapply(cones,habitat,length)
> sem<-s/sqrt(n)
> stripchart(cones~habitat, pch=16,vert=T)
> arrows(1:3,xbar+sem,1:3,xbar-sem,angle=90,code=3,length=.1)
> lines(1:3,xbar,pch=5,type="b")
```

Which yields the following plot:
In this example, to explain the code more fully, we use `pch=16` to request round black plotting dots in the `stripchart` and use `vert=T` to make the strips “vertical”. The error bars have been made using `arrows` (which adds arrows to a plot). Since the arrowhead is fully adjustable, we make the angle between the arm of the arrow and the shaft = 90 degrees; where the length is the length of the arrowhead in the printout (in inches). Finally, `code=3` means that we want a double-sided arrow (head at both ends).

Lastly, `type="b"` indicates that we want both points and lines drawn and we use `pch=5` to get open diamonds for the means (to clearly distinguish from the closed circles for data points).

**Step-4: Testing the assumptions for ANOVA**

After looking at the summary statistics and strip chart plot, we would next test for assumptions. Just like regression, normality testing is performed on the residuals. We can use now very familiar approaches to generate a Q-Q plot:

```r
> qqnorm(residuals(lm(cones~habitat, squirrels)))
> qqline(residuals(lm(cones~habitat, squirrels)))
```

Which yields:

![Normal Q-Q Plot](image)

And an explicit test of the residuals:

```r
> shapiro.test(residuals(lm(cones~habitat, squirrels)))

Shapiro-Wilk normality test
data:  residuals(lm(cones ~ habitat, squirrels))
W = 0.9336, p-value = 0.2781
```
All indications suggest that the data are normally distributed and this assumption is OK.

Testing whether the distribution of a variable has the same variance in all the groups can be done using Bartlett’s test (built into R), but like the Hartley’s F-max test for comparing the ratio of two variances, it is not very robust against departures from normality.

```r
> bartlett.test(cones~habitat)

Bartlett test of homogeneity of variances

data:  cones by habitat
Bartlett's K-squared = 0.307, df = 2, p-value = 0.8577
```

As expected from our inspection of the strip chart, we conclude that there is not difference in variance among the three groups. Suppose our data were not normally distributed (and could not be made to fit that assumption). The Levene test could also be used, but requires a bit more coding (not built in):

```r
> med<-with(squirrels,tapply(cones,habitat,median)) #def med
> ar<-with(squirrels,abs(cones -med[habitat])) #absolute values of med
> anova(lm(ar~habitat,squirrels)) #anova on medians

Analysis of Variance Table

Response: ar

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>habitat</td>
<td>2</td>
<td>0.03750</td>
<td>0.01875</td>
<td>0.2148</td>
<td>0.8095</td>
</tr>
<tr>
<td>Residuals</td>
<td>13</td>
<td>1.13500</td>
<td>0.08731</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Since $P = 0.8095$ we conclude that the variances are homogeneous (i.e., not statistically different). Note: Modified Levene and Bartlett’s will yield the same result when data are normal.

Another interesting approach (which is interactive and similar to what we did in regression analysis) can be had using the `plot` function. Try the following:

```r
> plot(aov(cones~habitat))

Waiting to confirm page change...

Each time you hit the enter key, you will get a new residual plot (4 in all)!

**Step-5: Running the AOVA procedure & MCPs**

Assuming that we have passed the assumptions of normality and homogeneity of variance, it is now safe to proceed to the actual one-way ANOVA. In R:

```r
> anova(lm(cones~habitat))
```
Analysis of Variance Table

Response: cones

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>habitat</td>
<td>2</td>
<td>29.404</td>
<td>14.702</td>
<td>50.085</td>
<td>7.787e-07 ***</td>
</tr>
<tr>
<td>Residuals</td>
<td>13</td>
<td>3.816</td>
<td>0.2935</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Which indicates that there is a highly significant difference ($P = 0.000000078$) of one or more of the groups from each other.

Now what? If no *apriori* reason were built into the design of this experiment, we would proceed with some form of pairwise multiple comparison procedure (MCP) such as Least Significant Difference (LSD), Bonferroni (BON), etc. to help us differentiate amongst the means. The function `p.adj` provides a variety of methods to adjust $P$-values for multiple comparisons:

```r
> pairwise.t.test(cones, habitat, p.adj="bonferroni")

   Pairwise comparisons using t tests with pooled SD

data:  cones and habitat

  ISA     ISP
ISP 3.0e-06 -
MSP 3.6e-06 1

P value adjustment method: bonferroni
```

Shown in the matrix are the $P$-values which clearly indicate that ISA is different from ISP and MSP, but that the latter do not differ from each other.

The R package *multcomp* provides a suite of comparison tests which you can explore at your leisure. An example of a built-in pairwise comparison procedure is, Tukey’s Honestly Significant Difference (HSD), which adjusts for all the pairwise comparisons:

```r
> TukeyHSD(aov(cones~habitat))

Tukey multiple comparisons of means
95% family-wise confidence level


table: 

<table>
<thead>
<tr>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISP-ISA</td>
<td>-2.82</td>
<td>-3.686</td>
<td>1.9537484</td>
</tr>
<tr>
<td>MSP-ISA</td>
<td>-2.78</td>
<td>-3.646</td>
<td>1.9137484</td>
</tr>
<tr>
<td>MSP-ISP</td>
<td>0.04</td>
<td>-0.864</td>
<td>0.9447703</td>
</tr>
</tbody>
</table>
```

The R package *multcomp* provides a suite of comparison tests which you can explore at your leisure. An example of a built-in pairwise comparison procedure is, Tukey’s Honestly Significant Difference (HSD), which adjusts for all the pairwise comparisons:

```r
> TukeyHSD(aov(cones~habitat))

   Tukey multiple comparisons of means
   95% family-wise confidence level

Fit: aov(formula = cones ~ habitat)

   $habitat


%diff lwr upr    p adj
ISP-ISA -2.82 -3.6862516 -1.9537484 0.0000028
MSP-ISA -2.78 -3.6462516 -1.9137484 0.0000033
MSP-ISP  0.04 -0.8647703  0.9447703 0.9925198

Intro to R
Step-6: Contrasts: treatment vs. orthogonal

Clearly, the MCPs show a difference of ISA with ISP and MSP. But, the very design of this experiment begs us to test for the mass of pine cones on islands with squirrels present vs. those islands where the squirrels are absent. This of course is a 1 df comparison or orthogonal contrast.

We first might ask, which factor had the greatest effect on cone mass, and were the three habitats different from each other? To address these issues we need to call `summary.lm` Which yields the following:

```r
> summary(lm(cones~habitat))

Call:
  lm(formula = cones ~ habitat)

Residuals:
   Min     1Q Median     3Q    Max
-0.780 -0.405 -0.040  0.505  0.720

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.9000     0.2212  40.238 4.97e-15 ***
habitatISP  -2.8200     0.3281  -8.596 1.01e-06 ***
habitatMSP  -2.7800     0.3281  -8.474 1.18e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5418 on 13 degrees of freedom
Multiple R-Squared: 0.8851,    Adjusted R-squared: 0.8675
F-statistic: 50.09 on 2 and 13 DF,  p-value: 7.787e-07
```

From this we can see that the mean of the first group is estimated as 8.9 with ISP being 2.82 less than that (8.9000-2.8200=6.0800) and MSP being (8.9000-2.7800=6.1200).

Let's look at the full complement of contrasts:

```r
> contrasts(habitat) #these are “dummy” treatment contrasts
  ISP  MSP
ISA  0  0
ISP  1  0
MSP  0  1
```

You can examine the design matrix to understand the coding:

```r
> model.matrix(lm(cones~habitat, data=squirrels))
```
Now, let’s switch from treatment contrasts (of no real use) to orthogonal contrasts. Since we have three groups there are $K - 1$ or $3 - 1 = 2$ orthogonal contrasts that can be made. The two logical ones are island vs. mainland and island with vs. island without squirrels. In this particular experiment, we had built in a planned contrast (ISA vs. ISP) from the beginning.

> contrasts(habitat) <- cbind(c(-1, -1, 2), c(-1, 1, 0))  # change to orthogonal contrasts in habitat

Note the two properties of orthogonal contrasts: (1) each column of coefficients sums to zero and (2) the products of the two columns sum to zero.

You can now examine the design matrix to understand the coding of the orthogonal contrasts:

> model.matrix(lm(cones~habitat))

(Intercept) habitat1 habitat2
1 1 -1 -1
2 1 -1 -1
3 1 -1 -1
4 1 -1 -1
Now we re-run the summary statement using the orthogonal contrast coefficients:

```r
> summary(lm(cones~habitat))
```

Call:
`lm(formula = cones ~ habitat)`

Residuals:
```
  Min 1Q Median  3Q Max
-0.780 -0.405 -0.040  0.505  0.720
```

Coefficients:
```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.03333    0.13595  51.735  < 2e-16 ***
habitat1    -0.45667    0.09753  -4.682 0.000429 ***
habitat2    -1.41000    0.16404  -8.596 1.01e-06 ***
```

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5418 on 13 degrees of freedom
Multiple R-Squared: 0.8851,    Adjusted R-squared: 0.8675
F-statistic: 50.09 on 2 and 13 DF,  p-value: 7.787e-07

Which tells us that there is a significant difference for both of our contrasts.
Step-7: Kruskal-Wallis: the nonparametric alternative

And, as expected, there is a simple nonparametric alternative to the one-way ANOVA when assumptions can not be met. Following our current example:

```r
> kruskal.test(cones~habitat)

   Kruskal-Wallis rank sum test

data:  cones by habitat
Kruskal-Wallis chi-squared = 10.6038, df = 2, p-value = 0.004982

```

Note the much larger P-value with the K-W test, which points to its decreased efficiency when assumptions have been met and a parametric alternative is possible.

Problem: Example 10.11 (p.199) [Modified].

Using the limnology data from example 10.11 (Zar, 1999), do a complete work up of the data. Read the data into a dataframe, summarize the data, test for all assumptions, perform the appropriate test (either ANOVA or K-W test), draw a conclusion. Lastly, provide a presentation grade graphic that you would turn in to a journal. This should be a bar graph, with 4 bars (labeled pond-1 to pond-4), and include ± SEs around each mean. Insert letters above the bars to indicate which ponds are significantly different from the others.