MSDS650 Week 2 ANOVA Assignment - Nathan Worsham

Following the ANOVA exercises from http://www.r-tutor.com/category/statistical-concept/anova, I created the text files the exercises requested.

0.1 Completely Randomized Design

```
> setwd("/Users/worshamn/Dropbox/Documents/Regis/MSDS650/week2/")
```

Here I copied the text required to a csv file in VI and named it fastfoot-1.txt.

```
> df1 = read.table("fastfood-1.txt", header=TRUE); df1
  Item1 Item2 Item3
1    22    52    16
2    42    33    24
3    44     8    19
4    52    47    18
5    45    43    34
6    37    32    39
```

Here all of the rows are smashed together into a single vector:

```
> r = c(t(df1)); r
 [1] 22 52 16 42 33 24 44  8 19 52 47 18 45 43 34 37 32 39
```

The menu items are designated as "treatment levels":

```
> f = c("Item1", "Item2", "Item3")
```

Nothing fancy here, just setting the number of treatment levels and observations per treatment:

```
> k = 3
> n = 6
```

Similar to when we turned the data frame into a single vector, the same is done for the treatment factors using the `gl` function:

```
> tm = gl(k, 1, n*k, factor(f)); tm
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
  Item1 Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
```

Now pit the two against each other using ANOVA:

```
> av = aov(r ~ tm)
```

Finally the way to see the ANOVA results is to use the `summary` command:

```
> summary(av)
            Df Sum Sq Mean Sq F value Pr(>F)
tm           2  745.4   372.7   2.541  0.112
Residuals   15 2200.2   146.7
```

The recommended exercise to use vertical columns rather than horizontal rows:

```
> #vertical
> rAsC = cbind(c(t(df1))); rAsC
      [,1]
 [1,]   22
 [2,]   52
 [3,]   16
 [4,]   42
 [5,]   33
 [6,]   24
 [7,]   44
 [8,]    8
 [9,]   19
[10,]   52
[11,]   47
[12,]   18
[13,]   45
[14,]   43
[15,]   34
[16,]   37
[17,]   32
[18,]   39
> f = c("Item1", "Item2", "Item3")
> k = 3
> n = 6
> tm = gl(k, 1, n*k, factor(f)); tm
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
  Item1 Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
> av = aov(rAsC ~ tm)
> summary(av)
            Df Sum Sq Mean Sq F value Pr(>F)
tm           2  745.4   372.7   2.541  0.112
Residuals   15 2200.2   146.7
```

In this instance the P value is 0.112 which is well above 0.05, so the null hypothesis that the 3 menu items have the same mean sales is not rejected.

0.2 Randomized Block Design

```
> df2 = read.table("fastfood-2.txt", header=TRUE); df2
  Item1 Item2 Item3
1    31    27    24
```

```
2    31    28    31
3    45    29    46
4    21    18    48
5    42    36    46
6    32    17    40
> r = c(t(df2)); r
 [1] 31 27 24 31 28 31 45 29 46 21 18 48 42 36 46 32 17 40
> f = c("Item1", "Item2", "Item3")
> k = 3
> n = 6
> tm = gl(k, 1, n*k, factor(f)); tm
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
  Item1 Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
> blk = gl(n, k, k*n); blk
 [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6
Levels: 1 2 3 4 5 6
> av = aov(r ~ tm + blk)
> summary(av)
            Df Sum Sq Mean Sq F value Pr(>F)
tm           2  538.8  269.39   4.959 0.0319 *
blk          5  559.8  111.96   2.061 0.1547
Residuals   10  543.2   54.32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #vertical
> rAsC = cbind(c(t(df2))); rAsC
      [,1]
 [1,]   31
 [2,]   27
 [3,]   24
 [4,]   31
 [5,]   28
 [6,]   31
 [7,]   45
 [8,]   29
 [9,]   46
[10,]   21
[11,]   18
[12,]   48
[13,]   42
[14,]   36
[15,]   46
[16,]   32
[17,]   17
[18,]   40
> f = c("Item1", "Item2", "Item3")
> k = 3
```

```
> n = 6
> tm = gl(k, 1, n*k, factor(f)); tm
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
  Item1 Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
> blk = gl(n, k, k*n); blk
 [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6
Levels: 1 2 3 4 5 6
> av = aov(rAsC ~ tm + blk)
> summary(av)
            Df Sum Sq Mean Sq F value Pr(>F)
tm           2  538.8  269.39   4.959 0.0319 *
blk          5  559.8  111.96   2.061 0.1547
Residuals   10  543.2   54.32
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this example each tester is given a chance at each factor. The p value for the experiment is less than 0.05 so the null hypothosis is rejected meaning that the mean sales volume of the new menu items are not all equal.

0.3 Factorial Design

```
> df3 = read.csv("fastfood-3.csv")
> r = c(t(df3))
> r
 [1] 25 39 36 36 42 24 31 39 28 26 35 29 51 43 42 47 39 36 47 53 32 52 46 33
> f1 = c("Item1", "Item2", "Item3")
> f2 = c("East", "West")
> k1 = length(f1)
> k2 = length(f2)
> n = 4
> tm1 = gl(k1, 1, n*k1*k2, factor(f1))
> tm1
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
  Item1 Item2 Item3 Item1 Item2 Item3 Item1
[20] Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
> tm2 = gl(k2, n*k1, n*k1*k2, factor(f2))
> tm2
 [1] East East East East East East East East East East East East West West
  West West West West West West West West
[23] West West
Levels: East West
> av = aov(r ~ tm1 * tm2)
> summary(av)
            Df Sum Sq Mean Sq F value   Pr(>F)
tm1          2  385.1   192.5   9.554  0.00149 **
```

```
tm2          1  715.0   715.0  35.481 1.23e-05 ***
tm1:tm2      2  234.1   117.0   5.808  0.01132 *
Residuals   18  362.8    20.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> #column excercise
> rAsC = cbind(c(t(df3)))
> rAsC
      [,1]
 [1,]   25
 [2,]   39
 [3,]   36
 [4,]   36
 [5,]   42
 [6,]   24
 [7,]   31
 [8,]   39
 [9,]   28
[10,]   26
[11,]   35
[12,]   29
[13,]   51
[14,]   43
[15,]   42
[16,]   47
[17,]   39
[18,]   36
[19,]   47
[20,]   53
[21,]   32
[22,]   52
[23,]   46
[24,]   33
> f1 = c("Item1", "Item2", "Item3")
> f2 = c("East", "West")
> k1 = length(f1)
> k2 = length(f2)
> n = 4
> tm1 = gl(k1, 1, n*k1*k2, factor(f1))
> tm1
 [1] Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3 Item1 Item2 Item3
  Item1 Item2 Item3 Item1 Item2 Item3 Item1
[20] Item2 Item3 Item1 Item2 Item3
Levels: Item1 Item2 Item3
> tm2 = gl(k2, n*k1, n*k1*k2, factor(f2))
> tm2
 [1] East East East East East East East East East East East East West West
  West West West West West West West West
```

```
[23] West West
Levels: East West
> av = aov(rAsC ~ tm1 * tm2)
> summary(av)
            Df Sum Sq Mean Sq F value   Pr(>F)
tm1          2  385.1   192.5   9.554  0.00149 **
tm2          1  715.0   715.0  35.481 1.23e-05 ***
tm1:tm2      2  234.1   117.0   5.808  0.01132 *
Residuals   18  362.8    20.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this example, there are 2 factors instead of one–menu items and marketing regions. Both factors are deemed different as both p values are well below the 0.05 significance level. It goes on to conclude that there is a "possible" interaction between the factors because the final p value is also below the 0.05 mark.

In all three cases the extra "exercise" of changing from horizontal rows to vertical columns did not make much sense to me as all I had to do was used the `cbind` command to change it vertical but after that the code remained the exact same.

After running through all of the examples, I was still confused on how to use ANOVA but it is at least clear in these examples that the p value was the line that divides. David Lane from onlinestatbook.com (n.d.) gives some helpful plain english on what ANOVA does which is that it "is used to test general rather than specific differences among means" and then he goes on to say that it tests the null hypothesis–which the examples spoke often about–that all the means are equal. When the null hypothesis is rejected, the conclusion is that at least one mean is different from at least one other mean but does not reveal which means are different from which (Lane, n.d.), this I believe is why he used the term general versus specific. In these examples when the p-value is less than 0.05 they are recommending to reject the null hypothesis. So it would seem that ANOVA is used as a generic test for comparing the difference or variation between the means of some variable between two or more groups (Boslaugh, 2014). David Lane points out–at first it seems odd that it is not called "Analysis of Means" but variance is what is analyzed to make inferences (Lane, n.d.)

Since it is used to analyze many types of experimental design, it would seem then that the types of questions it answers are

- In the experiment, can variation (assuming there is any) be explained by the "grouping introduced by the classification factor(s)? (r-bloggers.com, 2010)
- Or on average do any statistically significant (not due to chance) differences exist between the treatment and control groups?

So wanting the experiment to "determine the effect and interplay of factors" on the treatment group, the design of the experiment is important to limit "the impact of variability" (Krzywinski, 2014). Put another way, experimental design–when done properly–allows us to see both if the treatment causes the outcome and if the lack of treatment does not cause the outcome (Trochim, 2006). The design chosen can reduce the amount of trials participants have to take but possibly need a larger sample or allow a smaller number of participants with likely more trials required (Boslaugh, 2014). Either of these could translate to costs associated with the experiment or even different results, this is one reason why much care

should go into the experiment design. However depending on the circumstances, the design that is best may not be feasible or ethical, likely there will always be a compromise between what is ideal and what is feasible, but the design should be directed by what is most important to the question at hand (Boslaugh, 2014).

References

Lane, David. n.d. Retrieved from http://onlinestatbook.com/2/analysis_of_variance/intro.html, http://onlinestatbook.com/2/logic_of_hypothesis_testing/significance.html R-bloggers.com, 2010. Retrieved from http://www.r-bloggers.com/one-way-analysis-of-variance-anova/ Krzywinski, Martin. 2014. Retrieved from http://www.nature.com/nmeth/journal/v11/n6/full/nmeth.2974.html Trochim, William. 2006. Retrieved from http://www.socialresearchmethods.net/kb/desexper.php Boslaugh, Sarah. 2014. Statistics in a Nutshell. Chapter 8 - Introduction to Regression and ANOVA. Chapter 18 - Research Design.