One-way ANOVA

We are often interested in determining whether the means from more than two populations or groups are equal or not. To test whether the difference in means is statistically significant we can perform analysis of variance (ANOVA) using the R function aov(). If the ANOVA F-test shows there is a significant difference in means between the groups we may want to perform multiple comparisons between all pair-wise means to determine how they differ.

A. Analysis of Variance

The first step in our analysis is to graphically compare the means of the variable of interest across groups. It is possible to create side-by-side boxplots of measurements organized in groups using the function plot(). Simply type

```
plot(response  ~ factor, data=data_name)
```

where *response* is the name of the response variable and *factor* the variable that separates the data into groups. Both variables should be contained in a data frame called *data_name*.

Ex. A drug company tested three formulations of a pain relief medicine for migraine headache sufferers. For the experiment 27 volunteers were selected and 9 were randomly assigned to one of three drug formulations. The subjects were instructed to take the drug during their next migraine headache episode and to report their pain on a scale of 1 to 10 (10 being most pain).

Drug A	4	5	4	3	2	4	3	4	4
Drug B	6	8	4	5	4	6	5	8	6
Drug C	6	7	6	6	7	5	6	5	5

To make side-by-side boxplots of the variable *pain* grouped by the variable *drug* we must first read in the data into the appropriate format.

```
> pain = c(4, 5, 4, 3, 2, 4, 3, 4, 4, 6, 8, 4, 5, 4, 6, 5, 8, 6, 6, 7, 6, 6, 7, 5, 6, 5, 5)
> drug = c(rep("A",9), rep("B",9), rep("C",9))
> migraine = data.frame(pain,drug)
```

Note the command rep("A",9) constructs a list of nine A's in a row. The variable drug is therefore a list of length 27 consisting of nine A's followed by nine B's followed by nine C's.

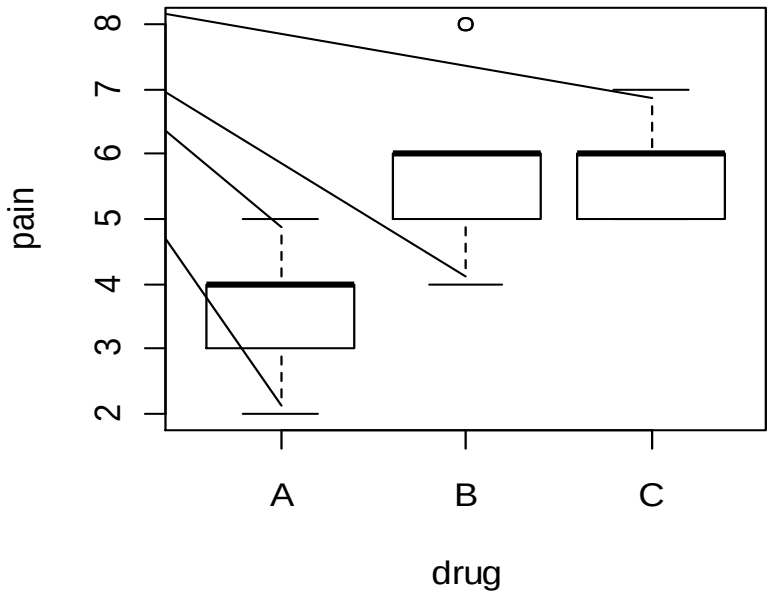
If we print the data frame migraine we can see the format the data should be on in order to make side-by-side boxplots and perform ANOVA (note the output is cut-off between observations 6-25 for space purposes).

```
> migraine
   pain drug
1     4    A
2     5    A
3     4    A
4     3    A
5     2    A
6     4    A
…
25    6    C
26    5    C
27    5    C
```

We can now make the boxplots by typing:

```
> plot(pain ~ drug, data=migraine)
```

The output of this program is shown below:



From the boxplots it appears that the mean pain for drug A is lower than that for drugs B and C.

Next, the R function aov() can be used for fitting ANOVA models. The general form is

```
aov(response  ~ factor, data=data_name)
```

where *response* represents the response variable and *factor* the variable that separates the data into groups. Both variables should be contained in the data

frame called *data_name*. Once the ANOVA model is fit, one can look at the results using the summary() function. This produces the standard ANOVA table.

Ex. Drug company example continued.

```
> results = aov(pain ~ drug, data=migraine)
> summary(results)
             Df  Sum Sq  Mean Sq  F value    Pr(>F)
drug          2  28.222  14.1111    11.906 0.0002559 ***
Residuals  24  28.444   1.1852
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Studying the output of the ANOVA table above we see that the F-statistic is 11.91 with a p-value equal to 0.0003. We clearly reject the null hypothesis of equal means for all three drug groups.

B. Multiple comparisons

The ANOVA F-test answers the question whether there are significant differences in the K population means. However, it does not provide us with any information about how they differ. Therefore when you reject H_0 in ANOVA, additional analyses are required to determine what is driving the difference in means. The function pairwise.t.test computes the pair-wise comparisons between group means with corrections for multiple testing. The general form is

```
pairwise.t.test(reponse, factor, p.adjust = method, alternative = c("two.sided",
"less", "greater"))
```

Here *response* is a vector of observations (the response variable), *factor* a list of factors and p.adjust is the correction method (e.g., “Bonferroni”).

Ex. Drug company example continued.

```
> pairwise.t.test(pain, drug, p.adjust="bonferroni")

        Pairwise comparisons using t tests with pooled SD

        data:  pain and drug

          A       B
        B 0.00119 -
        C 0.00068 1.00000

        P value adjustment method: bonferroni
```

The results state that the difference in means is not significantly different between drugs B and C (p-value = 1.00), but both are significantly different from drug A (p-values = 0.00119 and 0.00068, respectively). Hence, we can conclude that the mean pain is significantly different for drug A.

Another multiple comparisons procedure is Tukey's method (a.k.a. Tukey's Honest Significance Test). The function TukeyHSD() creates a set of confidence intervals on the differences between means with the specified family-wise probability of coverage. The general form is

```
TukeyHSD(x, conf.level = 0.95)
```

Here x is a fitted model object (e.g., an aov fit) and conf.level is the confidence level.

Ex. Drug company example continued.

```
> results = aov(pain ~ drug, data=migraine)
> TukeyHSD(results, conf.level = 0.95)

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

Fit: aov(formula = pain ~ drug, data = migraine)

$drug
        diff       lwr       upr     p adj
B-A 2.1111111 0.8295028  3.392719 0.0011107
C-A 2.2222222 0.9406139  3.503831 0.0006453
C-B 0.1111111 -1.1704972 1.392719 0.9745173
```

These results show that the B-A and C-A differences are significant (p=0.0011 and p=0.00065, respectively), while the C-B difference is not (p=0.97). This confirms the results obtained using Bonferroni correction.