Group Project R

Matthew Farr

November 26, 2018

Parametric and Non-parametric ANOVA analysis in R

First we want to load the dataset into R and omit any missing values using `read.csv()` and `na.omit()`.

```
diet.R = read.csv("./diet.csv", header = TRUE)
diet.R = na.omit(diet.R)
```

Then we need to tell R that Diet and gender are factors by using `as.factor()` and `factor()`.

```
diet.R$Diet = as.factor(diet.R$Diet)
diet.R$gender = factor(diet.R$gender,c(0,1))
```

Then we need to create the variable of interest 'weightlost'.

```
diet.R$weightlost = diet.R$pre.weight - diet.R$weight6weeks
```

We then split the dataset into subsets by 'gender' using `subset()` so that we now have a dataset for females and a dataset for males.

```
Diet.female = subset(diet.R, gender==0)
Diet.male = subset(diet.R, gender==1)
```

After all of this we can start our analysis.

Parametric ANOVA

We use 'Diet' as the grouping variable and use `aov()` to create our parametric one-way ANOVA model for the female dataset. We then use `summary()` to analyze our model.

```
anovaFemale = aov(weightlost~Diet, data = Diet.female)
summary(anovaFemale)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Diet         2  92.32   46.16   10.64 0.000197 ***
## Residuals   40 173.53    4.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above output of the parametric one-way ANOVA for the female dataset, we can see that the p-value is much smaller than 0.05 so we can conclude that there is at least one group that is statistically different from the other groups in the female dataset.

In order to check which group is statistically different, we run the Tukey post hoc test for pairwise comparison following a one-way ANOVA using `TukeyHSD()`.

```r
TukeyHSD(anovaFemale)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weightlost ~ Diet, data = Diet.female)
##
## $Diet
##           diff        lwr      upr     p adj
## 2-1 -0.4428571 -2.3589312 1.473217 0.8406368
## 3-1  2.8300000  0.9461311 4.713869 0.0020846
## 3-2  3.2728571  1.3889883 5.156726 0.0003833
```

From the above result we can see that there is a statistically significant difference in weight loss between the 'Diet 3' group and the 'Diet 1' group and between the 'Diet 3' group and the 'Diet 2' group. Because of this we can determine that the 'Diet 3' group is statistically different from the other two groups.

We then do the same thing for the male dataset.

```r
anovaMale = aov(weightlost~Diet, data = Diet.male)
summary(anovaMale)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Diet         2    2.0   1.001   0.148  0.863
## Residuals   30  202.8   6.760
```

From the above output of the parametric one-way ANOVA for the male dataset, we can see that the p-value is larger than 0.05 so we fail to reject the null hypothesis.

Let's confirm this by looking at the Tukey post hoc test.

```r
TukeyHSD(anovaMale)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weightlost ~ Diet, data = Diet.male)
##
## $Diet
##          diff       lwr      upr     p adj
## 2-1 0.4590909 -2.341516 3.259698 0.9141710
## 3-1 0.5833333 -2.161144 3.327810 0.8602563
## 3-2 0.1242424 -2.551325 2.799809 0.9928028
```

As you can see there is no statistical significane between the three groups in the male dataset.

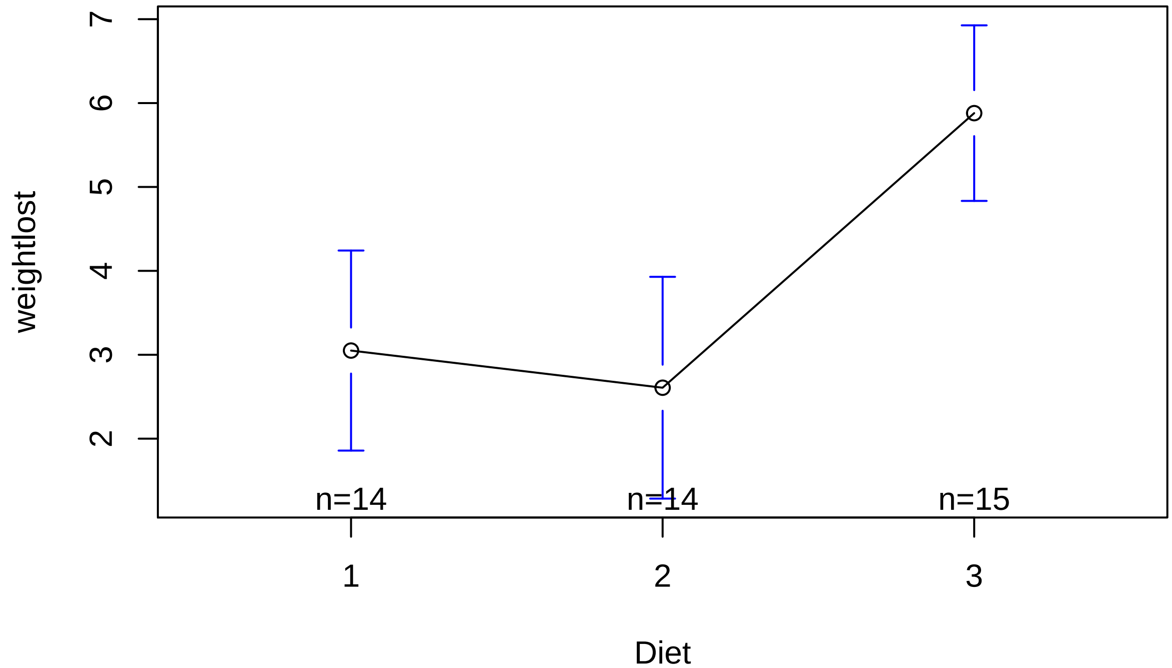
These plots help visualize these two Tukey post hoc tests. Here is the plot for the female data.

```r
library(gplots)
```

```
##
## Attaching package: 'gplots'
```

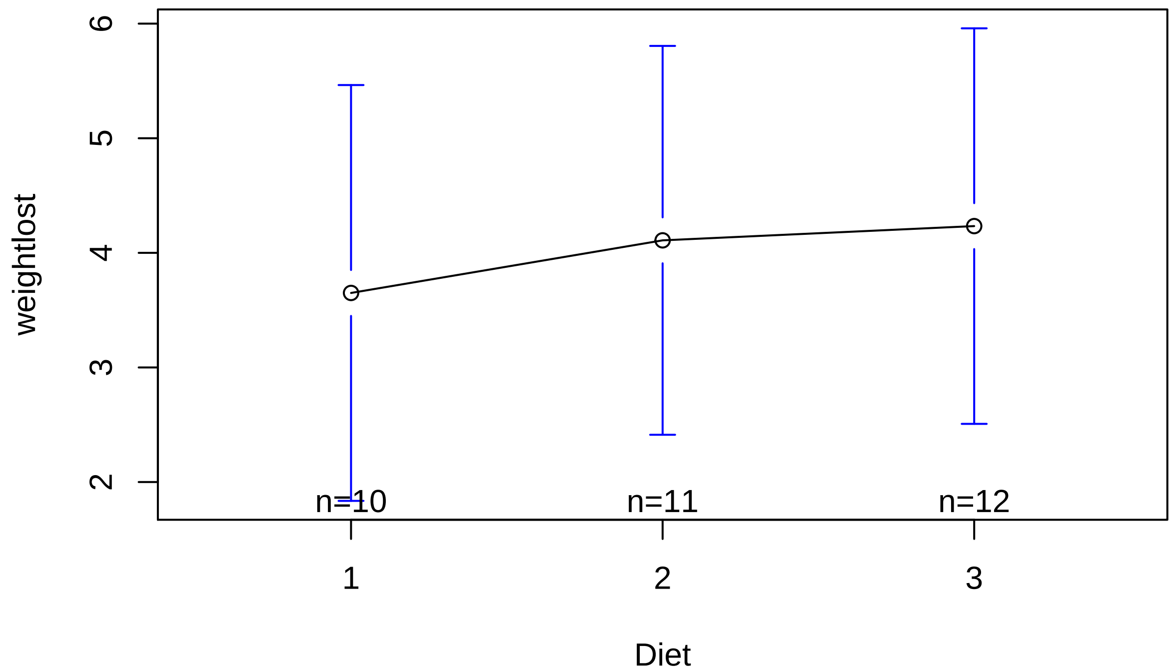
```
## The following object is masked from 'package:stats':
##
##     lowess
```

```r
plotmeans(weightlost~Diet, data = Diet.female)
```



And here is the plot for the male data.

```r
plotmeans(weightlost~Diet, data = Diet.male)
```



We could also do parametric two-way ANOVA for the whole dataset. For this we still use `aov()` but now we regard 'gender' and 'Diet as two grouping variables.

```
anova2 = aov(weightlost~gender*Diet,data=diet.R)
summary(anova2)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## gender       1    0.3   0.278   0.052 0.82062
## Diet         2   60.4  30.209   5.619 0.00546 **
## gender:Diet  2   33.9  16.952   3.153 0.04884 *
## Residuals   70  376.3   5.376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The gender:Diet interaction is statistically significant at the $p = 0.04884$ level. There was no statistically significant difference in weight loss between gender ($p = 0.82062$), but there were statistically significant differences between Diet groups ($p = 0.00546$).

Nonparametric ANOVA: Kruskal-Wallis Test

We regard 'Diet' as the grouping variable and use `kruskal.test()` to do nonparametric one-way ANOVA, i.e. Kruskal-Wallis test for the female data.

```r
kruskal.test(weightlost~Diet, data = Diet.female)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weightlost by Diet
## Kruskal-Wallis chi-squared = 14.545, df = 2, p-value = 0.0006945
```

We get a p-value that is much smaller than 0.05 so we can reject the null hypothesis and conclude that there is at least one group statistically different from the other groups in the female dataset. This is the same conclusion we got in the parametric one-way ANOVA for the female data.

Now we do the same thing for the male dataset.

```r
kruskal.test(weightlost~Diet, data = Diet.male)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weightlost by Diet
## Kruskal-Wallis chi-squared = 0.62218, df = 2, p-value = 0.7326
```

We get a p-value that is larger than 0.05 so there is no statistically significant difference in weight loss between the three groups in the male dataset. This is the same result we got from the parametric one-way ANOVA for the male data.