



The following resources are associated:
Statistical Hypothesis testing, Checking normality in R and the Excel dataset 'Diet.csv'

One-way (between-groups) ANOVA in R

Dependent variable: Continuous (scale/interval/ratio),

Independent variable: Categorical (at least 3 unrelated/ independent groups)

Common Applications: Used to detect a difference in means of 3 or more independent groups. It can be thought of as an extension of the independent t-test for and can be referred to as 'between-subjects' ANOVA.

Data: The data set Diet.csv contains information on 78 people who undertook one of three diets. There is background information such as age, gender (Female=0, Male=1) and height. The aim of the study was to see which diet was best for losing weight so the independent variable (group) is diet. To open the file use the `read.csv()` command.

	gender	Age	Height	pre.weight	Diet	weight6weeks
1	NA	41	171	60	2	60.0
2	NA	32	174	103	2	103.0
3	0	44	174	58	2	60.1
4	0	37	172	58	2	56.0
5	0	22	159	58	2	54.2

Female = 0

Diet 1, 2 or 3

You will need to change the command depending on where you have saved the file.

```
dietR<-read.csv("D:\\diet.csv",header=T, sep=", ")
```

Tell R to use the diet dataset until further notice using `attach(dataset)` so 'Height' can be used instead of `dietR$Height`. Tell R that 'Diet' is a factor using `as.factor(variable)`.

```
attach(dietR)
```

```
Diet<-as.factor(Diet)
```

Calculate the weight lost by person (difference in weight before and after the diet) and add the variable to the dataset. Then attach the data again.

```
dietR$weightlost<-pre.weight-weight6weeks
```

```
attach(dietR)
```

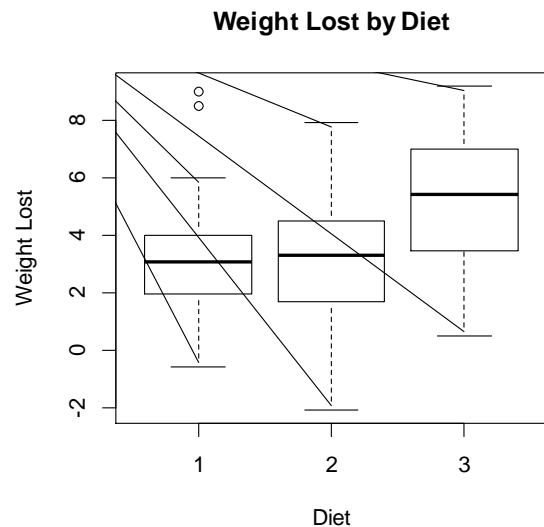


Before carrying any analysis, summarise weight lost by diet using a box-plot or interval plot and some summary statistics. Do the group means and standard deviations look similar or very different? Calculate means and standard deviations for weight lost by diet using the `tapply(dependent, independent, summary statistic required, na.rm=T)` command e.g. `tapply(weightlost, Diet, mean, na.rm=T)`. `na.rm=T` removes rows where missing values exist.

```
> mean<-tapply(weightlost,Diet,mean,na.rm=T)
> sd<-tapply(weightlost,Diet,sd,na.rm=T)
> #Combine in one table and give rownames
> results1<-cbind(mean,sd)
> rownames(results1)<-paste("Diet",1:3,sep=" ")
> #Round all the summary statistics to 2 decimal places.
> round(results1,2)
      mean  sd
Diet 1 3.30 2.24
Diet 2 3.03 2.52
Diet 3 5.15 2.40
```

To produce a boxplot of weight lost by diet:

```
boxplot(weightlost~Diet,main='Weight
Lost by Diet',xlab='Diet',
ylab='Weight Lost')
```



Diet 3 seems better than the other diets as the mean weight lost is greater. The standard deviations are similar so weight lost within each group is equally spread out.

ANOVA stands for 'Analysis of variance' as it uses the ratio of between group variation to within group variation, when deciding if there is a statistically significant difference between the groups. **Within group variation** measures how much the individuals vary from their group mean. Each difference between an individual and their group mean is called a **residual**. These residuals are squared and added together to give the sum of the squared residuals or the within group sum of squares (SS_{within}). **Between group variation** measures how much the group means vary from the overall mean (SS_{between}).

Steps in R and output

To carry out a one way ANOVA use `aov(dependent~independent, give the ANOVA model a name e.g. anovaD and use summary() to see the output.`

```
anovaD<-aov(weightlost~Diet)
summary(anovaD)
```

$$F = \text{Test statistic} \\ \frac{MS_{\text{Diet}}}{MS_{\text{error}}} = \frac{35.55}{5.74} = 6.197$$

```
> summary(anovaD)
      Df Sum Sq Mean Sq F value Pr(>F)
Diet    2   71.1   35.55   6.197 0.00323 **
Residuals 75  430.2    5.74
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P = p-value = sig
= $P(F > 6.197)$
p = 0.00323



When writing up the results, it is common to report certain figures from the ANOVA table.

$F(df_{\text{between}}, df_{\text{within}}) = \text{Test Statistic}, p = \rightarrow F(2, 75) = 6.197, p = 0.003$

There was a significant difference in mean weight lost [$F(2,75)=6.197, p = 0.003$] between the diets.

Post Hoc Tests

ANOVA tests the null hypothesis 'all group means are the same' so the resulting p-value only concludes whether or not there is a difference between one or more pairs of groups. If the ANOVA is significant, further 'post hoc' tests have to be carried out to confirm where those differences are. The post hoc tests are mostly t-tests with an adjustment to account for the multiple testing. *Tukey's* is the most commonly used post hoc test but check if your discipline uses something else. Use the command `TukeyHSD(anovaD)`.

```
> TukeyHSD(anovaD)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weightlost ~ Diet)

$Diet
      diff      lwr      upr    p adj
2-1 -0.2740741 -1.8806155  1.332467 0.9124737
3-1  1.8481481  0.2416067  3.454690 0.0201413
3-2  2.1222222  0.5636481  3.680796 0.0047819
```

Report each of the three pairwise comparisons e.g. there was a significant difference between diet 3 and diet 1 ($p = 0.02$). Use the mean difference between each pair e.g. people on diet 3 lost on average 1.85 kg more than those on diet 1 or use individual group means to conclude which diet is best.

Checking the assumptions for one-way ANOVA

Assumptions	How to check	What to do if the assumptions is not met
Residuals should be normally distributed	Use histogram, QQ plots and normality tests as diagnostic tools (see the Checking normality in R resource for more details)	If the residuals are very skewed, the results of the ANOVA are less reliable so the Kruskal- Wallis test should be used instead (see the Kruskal-Wallis in R resource)
Homogeneity (equality) of variance: The variances (SD squared) should be similar for all the groups	Use the Levene's test of equality of variances through the package <code>car</code> <code>library(car)</code> <code>leveneTest(weightlost~Diet)</code> If p - value > 0.05, equal variances can be assumed and the ANOVA results are valid	If p - value < 0.05, the results of the ANOVA are less reliable. The Welch test is more appropriate and can be accessed via <code>library(car)</code> <code>oneway.test(weightlost~Diet)</code> The Games Howell post hoc test should also be used instead of Tukeys but does not exist in R

Checking the assumptions for this data

Ask for the standardised residuals (difference between each individual and their group mean) and give them a name (res).

```
res<-anovaD$residuals
```



Produce a histogram of the residuals.

```
hist(res, main="Histogram of standardised residuals", xlab="Standardised residuals")
```

The Levene's test for equality of variances is in the additional 'car' package.

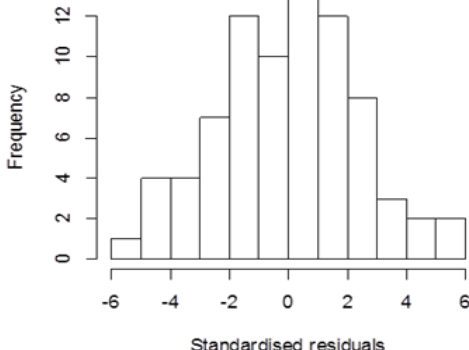
```
library(car)
```

If this command does not work, you will need to go to the Packages --> Install package(s) and select the UK (London)CRAN mirror. Then look for the package 'car' and click. A lot of extra menus will download as well so you shouldn't need to do this again. Then try

```
library(car) again.
```

Once loaded, carry out Levene's test.

```
leveneTest(weightlost~Diet)
```

Homogeneity Assumption	Normality Assumption
<pre>> library(car) > leveneTest(weightlost~Diet) Levene's Test for Homogeneity of Variance (center = median) Df F value Pr(>F) group 2 0.6257 0.5377 75 . .</pre> <p>As p - value (0.5377) > 0.05, equal variances can be assumed</p>	<p>Histogram of standardised residuals</p>  <p>The residuals are normally distributed</p>

Reporting ANOVA

A one-way ANOVA was conducted to compare the effectiveness of three diets. Normality checks and Levene's test were carried out and the assumptions met.

There was a significant difference in mean weight lost [$F(2,75)=6.197$, $p = 0.003$] between the diets. Post hoc comparisons using the Tukey test were carried out. There was a significant difference between diets 1 and 3 ($p = 0.02$) with people on diet 3 lost on average 1.85 kg more than those on diet 3. There was also a significant difference between diets 2 and 3 difference ($p = 0.005$) with people on diet 3 lost on average 2.12 kg more than those on diet 2.

