Code

ANOVA with R: analysis of the *diet* dataset

D.-L. Couturier / R. Nicholls / M. Fernandes

Last modified: 09 Mar 2020

A full version of the dataset *diet* may be found online on the U. of Sheffield website https://www.sheffield.ac.uk/polopoly_fs/1.570199!/file/stcp-Rdataset-Diet.csv (https://www.sheffield.ac.uk/polopoly_fs/1.570199!/file/stcp-Rdataset-Diet.csv).

A slightly modified version is available in the data file is stored under data/diet.csv. The data set contains information on 76 people who undertook one of three diets (referred to as diet *A*, *B* and *C*). There is background information such as age, gender, and height. The aim of the study was to see which diet was best for losing weight.

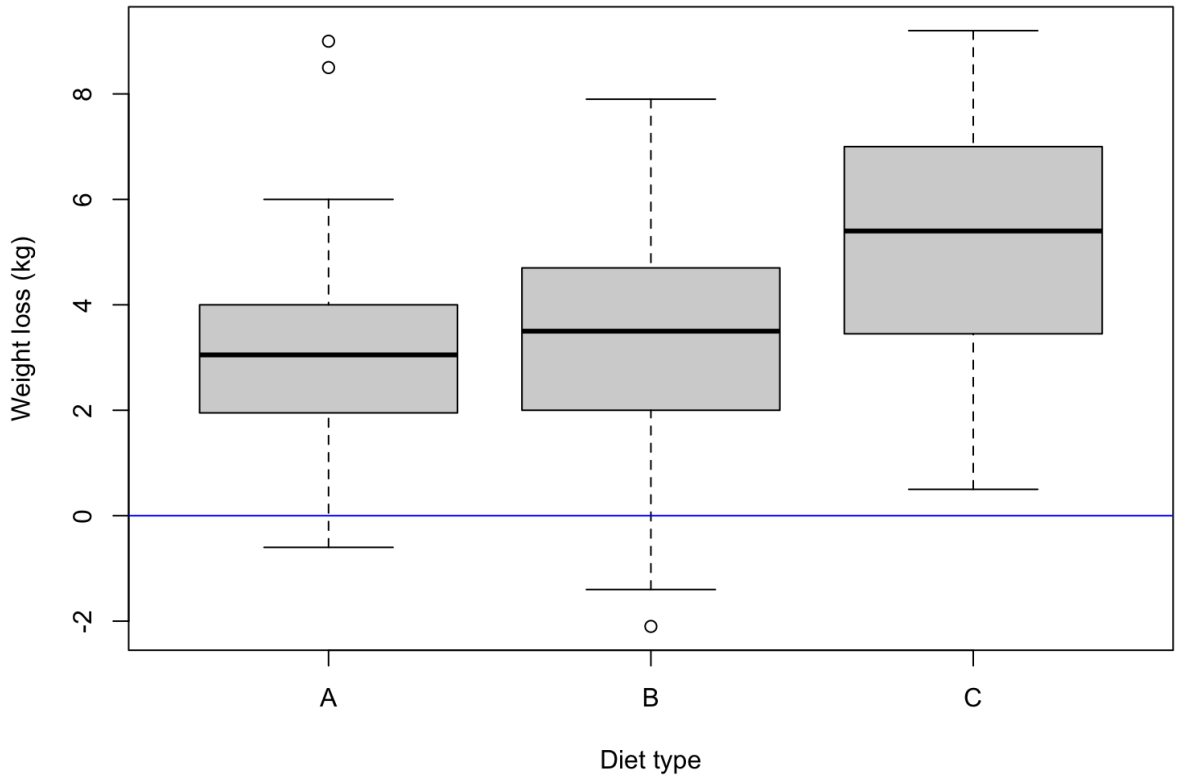
Section 1: importation and descriptive analysis

Lets starts by

- importing the data set *diet* with the function `read.csv()`
- defining a new column *weight.loss*, corresponding to the difference between the initial and final weights (respectively the corresponding to the columns `initial.weight` and `final.weight` of the dataset)
- displaying *weight loss* per *diet type* (column `diet.type`) by means of a boxplot.

Hide

```
diet = read.csv("data/diet.csv",row.names=1)
diet$weight.loss = diet$initial.weight - diet$final.weight
boxplot(weight.loss~diet.type,data=diet,col="light gray",
        ylab = "Weight loss (kg)", xlab = "Diet type")
abline(h=0,col="blue")
```



Section 2: ANOVA

Lets

- perform a Fisher's, Welch's and Kruskal-Wallis one-way ANOVA, respectively by means of the functions `aov()`, `oneway.test()` and `kruskal.test`,
- display and analyse the results: Use the function `summary()` to display the results of an R object of class `aov` and the function `print()` otherwise.

Hide

```
diet.fisher  = aov(weight.loss~diet.type,data=diet)
diet.welch   = oneway.test(weight.loss~diet.type,data=diet)
diet.kruskal = kruskal.test(weight.loss~diet.type,data=diet)

summary(diet.fisher)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## diet.type    2   60.5  30.264   5.383 0.0066 **
## Residuals   73  410.4   5.622
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
print(diet.welch)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  weight.loss and diet.type
## F = 5.2693, num df = 2.00, denom df = 48.48, p-value = 0.008497
```

Hide

```
print(diet.kruskal)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  weight.loss by diet.type
## Kruskal-Wallis chi-squared = 9.4159, df = 2, p-value = 0.009023
```

Note that, when the interest lies in the difference between two means, the Fisher's ANOVA (fonction `aov()`) and the Student's t-test (function `t.test()` with argument `var.equal` set to `TRUE`) leads to the same results. Let check this by comparing the mean weight losses of *Diet A* and *Diet C*.

Hide

```
summary(aov(weight.loss~diet.type,data=diet[diet$diet.type!="B",]))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## diet.type    1   43.4    43.4   8.036 0.00664 **
## Residuals   49  264.6     5.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hide

```
t.test(weight.loss~diet.type,data=diet[diet$diet.type!="B",],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  weight.loss by diet.type
## t = -2.8348, df = 49, p-value = 0.006644
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.1582988 -0.5379975
## sample estimates:
## mean in group A mean in group C
##        3.300000        5.148148
```

Section 3: Model check

Lets first

- define the Fisher's and Welch's residuals by subtracting the mean of each group to the weight loss of the corresponding participants
- define the Kruskal's residual's by subtraction the median of each group to the weight loss of the corresponding participants

The mean or median of each group may be obtained by means of the function `tapply()` which allows a apply a function (like `mean` or `median`) to and by

Hide

```
# mean and median weight loss per group:
mean_group   = tapply(diet$weight.loss,diet$diet.type,mean)
median_group = tapply(diet$weight.loss,diet$diet.type,median)
mean_group
```

```
##        A        B        C
## 3.300000 3.268000 5.148148
```

Hide

```
median_group
```

```
##    A    B    C
## 3.05 3.50 5.40
```

Hide

```
# residuals:
diet$resid.mean   = (diet$weight.loss - mean_group[as.numeric(diet$diet.type)])
diet$resid.median = (diet$weight.loss - median_group[as.numeric(diet$diet.type)])
diet[1:10,]
```

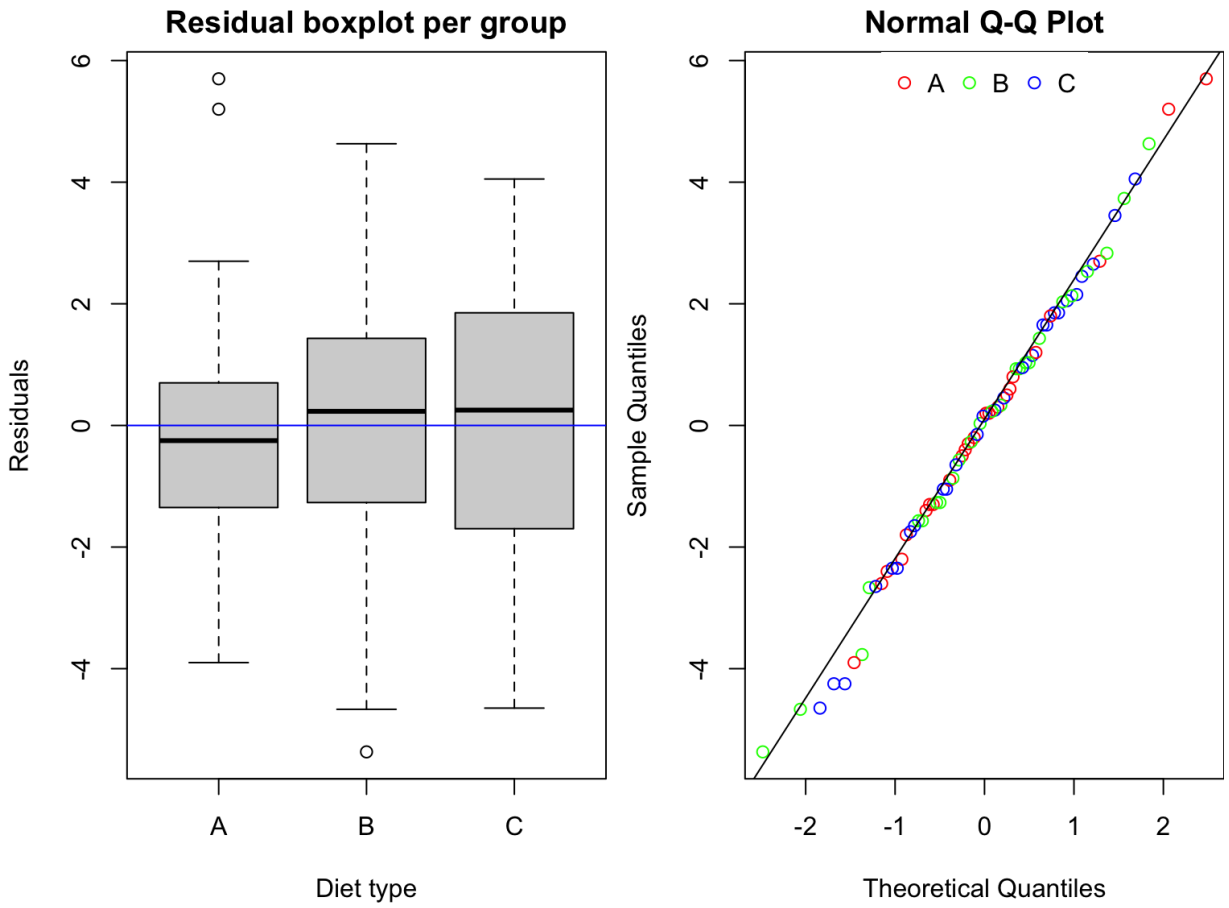
```
##    gender age height diet.type initial.weight final.weight weight.loss resid.mea
n
## 1  Female  22    159         A             58         54.2         3.8        0.
5
## 2  Female  46    192         A             60         54.0         6.0        2.
7
## 3  Female  55    170         A             64         63.3         0.7       -2.
6
## 4  Female  33    171         A             64         61.1         2.9       -0.
4
## 5  Female  50    170         A             65         62.2         2.8       -0.
5
## 6  Female  50    201         A             66         64.0         2.0       -1.
3
## 7  Female  37    174         A             67         65.0         2.0       -1.
3
## 8  Female  28    176         A             69         60.5         8.5        5.
2
## 9  Female  28    165         A             70         68.1         1.9       -1.
4
## 10 Female  45    165         A             70         66.9         3.1       -0.
2
##    resid.median
## 1          0.75
## 2          2.95
## 3         -2.35
## 4         -0.15
## 5         -0.25
## 6         -1.05
## 7         -1.05
## 8          5.45
## 9         -1.15
## 10         0.05
```

Then, lets

- display a boxplot of the residuals per group to assess if (i) the variance per groups are similar (ii) normality of the residuals per group seems credible
- display a QQ-plot of the residuals of the mean model to assess if normality of the residuals seems credible

Hide

```
par(mfrow=c(1,2),mar=c(4.5,4.5,2,0))
#
boxplot(resid.mean~diet.type,data=diet,main="Residual boxplot per group",col="light
        gray",xlab="Diet type",ylab="Residuals")
abline(h=0,col="blue")
#
col_group = rainbow(nlevels(diet$diet.type))
qqnorm(diet$resid.mean,col=col_group[as.numeric(diet$diet.type)])
qqline(diet$resid.mean)
legend("top",legend=levels(diet$diet.type),col=col_group,pch=21,ncol=3,box.lwd=NA)
```



Finally, lets

- perform a Shapiro's test to assess is there is enough evidence that the residuals are not normally distributed (by means of the function `shapiro.test()`)
- perform a Bartlett's test to assess is there is enough evidence that the residuals per group do not have different variance (by means of the function `bartlett.test()`.)

Hide

```
shapiro.test(diet$resid.mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  diet$resid.mean
## W = 0.99175, p-value = 0.9088
```

Hide

```
bartlett.test(diet$resid.mean~as.numeric(diet$diet.type))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  diet$resid.mean by as.numeric(diet$diet.type)
## Bartlett's K-squared = 0.21811, df = 2, p-value = 0.8967
```

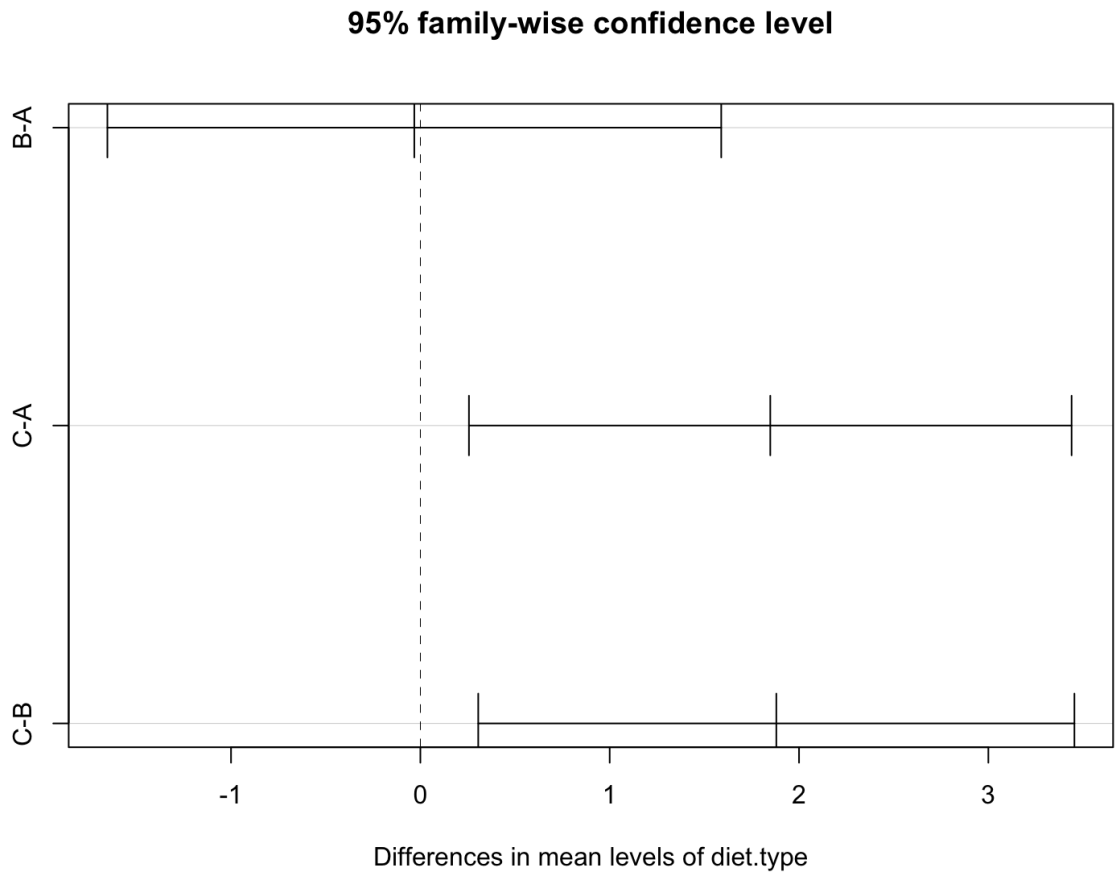
Section 4: Mutiple comparisons

Lets

- perform a Tukey HSD test to define which group pair(s) have different means (by means of the function `TukeyHSD()`)
- compare the Tukey HSD confidence interval size for the difference of means between the weight losses of *Diet A* and *Diet B* with the one obtained by means of a Student's t-test (function `t.test()` with argument `var.equal` set to `TRUE`)

Hide

```
plot(TukeyHSD(diet.fisher))
```



Hide

```
t.test(weight.loss~diet.type,data=diet[diet$diet.type!="C",],var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  weight.loss by diet.type
## t = 0.0475, df = 47, p-value = 0.9623
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.323275  1.387275
## sample estimates:
## mean in group A mean in group B
##           3.300           3.268
```

Section 5: Two-way ANOVA

Lets

- perform a two-way ANOVA to assess if the weight loss means are different per levels of the factors *Diet* and/or *Age*.
- compare the output of the function `aov()` to the one of the function `lm()`.

Hide

```
diet.fisher = aov(weight.loss~diet.type*gender,data=diet)
summary(diet.fisher)
```

```
##                  Df Sum Sq Mean Sq F value  Pr(>F)
## diet.type         2   60.5  30.264   5.629 0.00541 **
## gender            1    0.2   0.169   0.031 0.85991
## diet.type:gender  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
```

Hide

```
anova(lm(weight.loss~diet.type*gender,data=diet))
```

```
## Analysis of Variance Table
##
## Response: weight.loss
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## diet.type         2  60.53 30.2635  5.6292 0.005408 **
## gender            1   0.17  0.1687  0.0314 0.859910
## diet.type:gender  2  33.90 16.9520  3.1532 0.048842 *
## Residuals        70 376.33  5.3761
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Section 5: Practicals

Analyse the two following datasets with the suitable analysis:

(i) *amess.csv*

The data for this exercise are to be found in *amess.csv*. The data are the red cell folate levels in three groups of cardiac bypass patients given different levels of nitrous oxide (N2O) and oxygen (O2) ventilation. (There is a reference to the source of this data in Altman, Practical Statistics for Medical Research, p. 208.) The treatments are

- 50% N2O and 50% O2 continuously for 24 hours
- 50% N2O and 50% O2 during the operation
- No N2O but 35-50% O2 continuously for 24 hours

(ii) *globalBreastCancerRisk.csv*

The file *globalBreastCancerRisk.csv* gives the number of new cases of Breast Cancer (per population of 10,000) in various countries around the world, along with various health and lifestyle risk factors.

Let's suppose we are initially interested in whether the number of breast cancer cases is significantly different in different regions of the world.

Visualise the distribution of breast cancer incidence in each continent. Check how many observations belong to each group (continent). Are there any groups that you would consider removing/grouping before performing the analysis ?