Hands-on Biostatistics

T-test and ANOVA in R

18/7/2020

Install necessary packages

```
#install("ggplots", "PairedData", "car", "dplyr")
```

One sample t-test

The one-sample t test used to compare an observed mean with a theoretical mean.

In our example we want to know if the HbA1c mean level in 5 diabetic patients is equal to 6, the mean assumed for the normal population.

```
#Read data (HbA1c test for 5 diabetic patients)

x <- c(8.5, 9.3, 7.9, 9.2, 10.3)
```

Compute the t-test

```
#One sample t-test

t.test(x, mu = 6) #(two-sided)
```

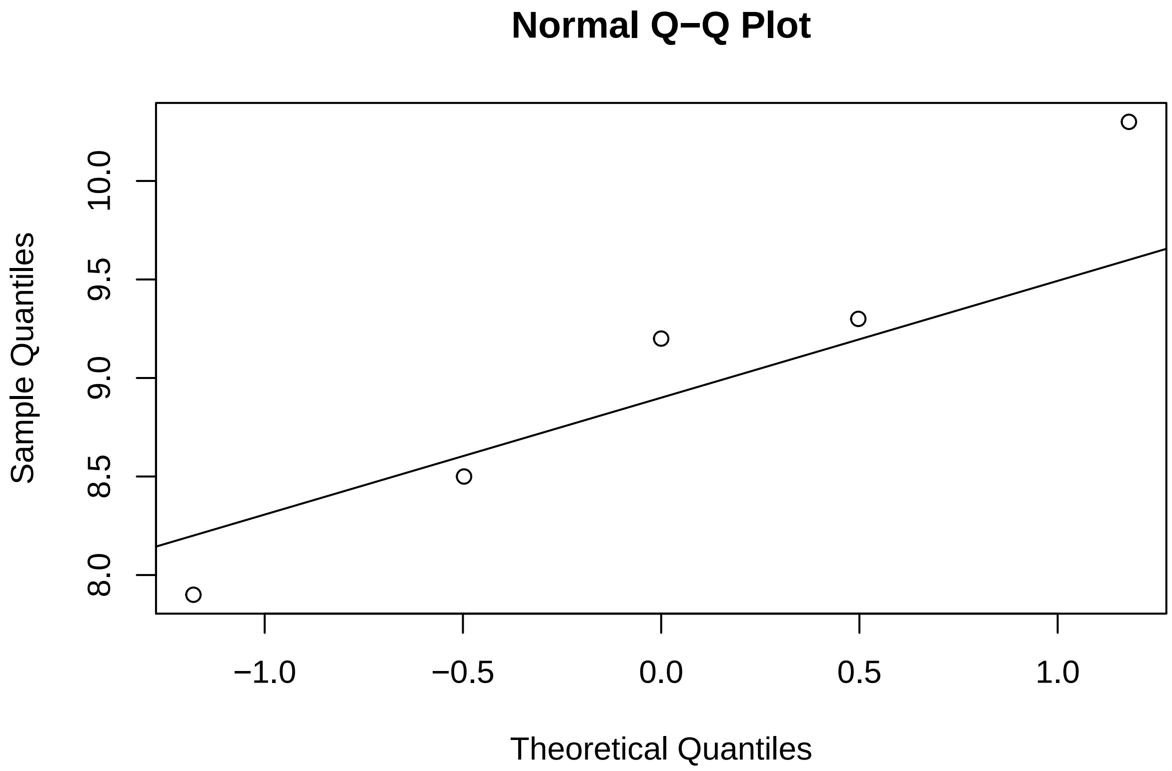
```
##
##  One Sample t-test
##
## data:  x
## t = 7.5159, df = 4, p-value = 0.001677
## alternative hypothesis: true mean is not equal to 6
## 95 percent confidence interval:
##   7.916997 10.163003
## sample estimates:
## mean of x
##      9.04
```

```
# Help on t.test function
?t.test
```

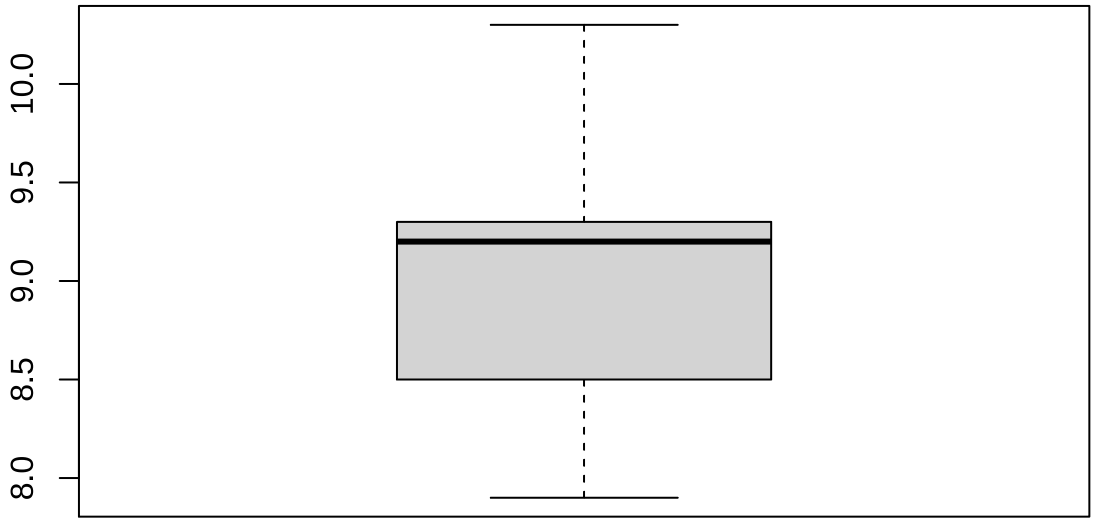
```
## starting httpd help server ... done
```

Check assumptions: Normality of the data

```
#Graphs
qqnorm(x)
qqline(x)
```


Normal Q−Q Plot
Sample Quantiles
8.0
8.5
9.0
9.5
10.0
−1.0
−0.5
0.0
0.5
1.0
Theoretical Quantiles

```r
boxplot(x)
```



```
#Formal test
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 0.9773, p-value = 0.9197
```

###Documentation:

A one-sample t-test was run to determine whether HbA1c level in diabetic subjects was different to normal, defined as a HbA1c of 6.0. HbA1c level was normally distributed, as assessed by Shapiro-Wilk's test (p > 0 .05) and there were no outliers in the data, as assessed by inspection of a boxplot. Mean HbA1c (9.04, 95% CI, 7.92 to 10.16) was higher than the normal HbA1c of 6.0, a statistically significant difference (t(4)= 7.5159; p=0.0017).

The independent t test

We will use an independent t-test to understand whether weight differed based on gender (i.e., our dependent variable is "weight" and our independent variable is "gender", which has two groups: "man" and "woman").

```
# Data in two numeric vectors
women_weight <- c(38.9, 61.2, 73.3, 21.8, 63.4, 64.6, 48.4, 48.8, 48.5)
men_weight <- c(67.8, 60, 63.4, 76, 89.4, 73.3, 67.3, 61.3, 62.4)

# Create a data frame
my_data <- data.frame(
  group = rep(c("Woman", "Man"), each = 9),
  weight = c(women_weight,  men_weight)
)
```

Compute summary statistics

```
library("dplyr")
```

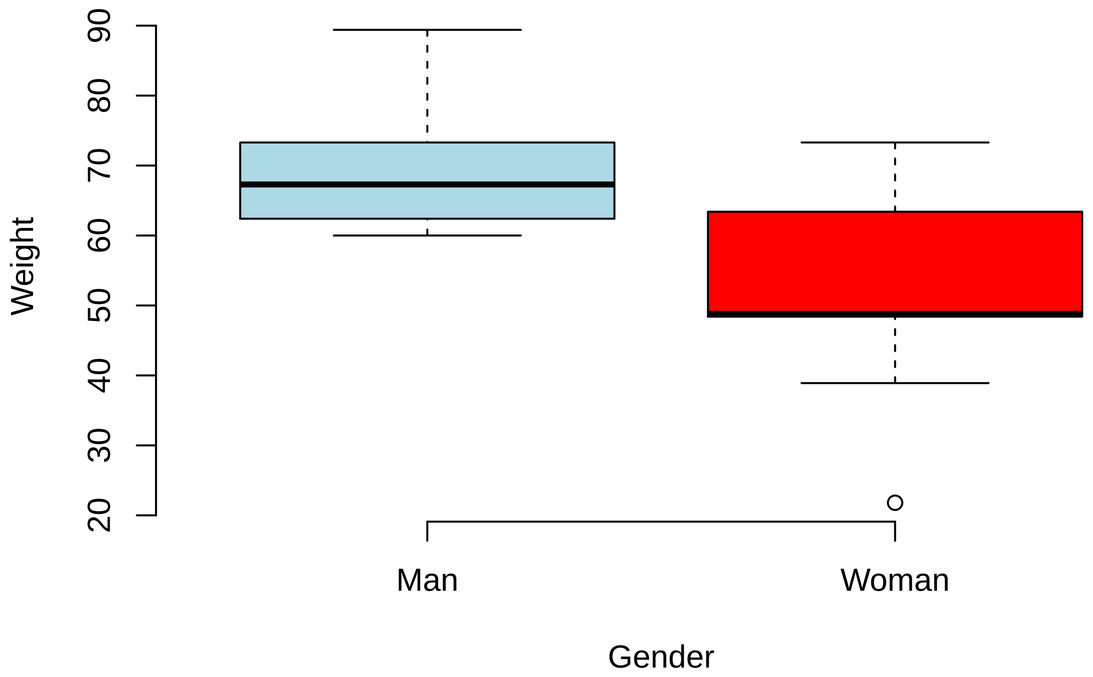
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
group_by(my_data, group) %>%
  summarise(
    count = n(),
    mean = mean(weight, na.rm = TRUE),
    sd = sd(weight, na.rm = TRUE)
  )
```

```
## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 2 x 4
##   group count  mean    sd
##   <chr> <int> <dbl> <dbl>
## 1 Man       9  69.0  9.38
## 2 Woman     9  52.1 15.6
```

Visualize the data

```
# Plot weight by group and color by group

# Box plot
boxplot(weight ~ group, data = my_data,
        xlab = "Gender", ylab = "Weight",
        frame = FALSE, col = c("lightblue", "red"))
```



Test homogeneity of variances

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
leveneTest(weight ~ group, data = my_data)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  1.5888 0.2256
##       16
```

Equal variances assumed p>0.05.

Compute t-test

```
# Compute t-test using t.test function
 t2<- t.test(weight ~ group, data = my_data, var.equal = TRUE) #assume equal variances
 t2
```

```
##
##  Two Sample t-test
##
## data:  weight by group
## t = 2.7842, df = 16, p-value = 0.01327
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   4.029759 29.748019
## sample estimates:
##   mean in group Man mean in group Woman
##            68.98889            52.10000
```

```
#t2$stderr
#t2$statistic
#t2$estimate
```

Test for normality

```
# Shapiro-Wilk normality test for Men's weights
with(my_data, shapiro.test(weight[group == "Man"]))# p = 0.1
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight[group == "Man"]
## W = 0.86425, p-value = 0.1066
```

```
# Shapiro-Wilk normality test for Women's weights
with(my_data, shapiro.test(weight[group == "Woman"])) # p = 0.6
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight[group == "Woman"]
## W = 0.94266, p-value = 0.6101
```

Documentation:

An independent-samples t-test was run to determine if there were differences in weight between males and females. There were no outliers in the data, as assessed by inspection of a boxplot. Weight for each level of gender was normally distributed, as assessed by Shapiro-Wilk test ($p > 0.05$), and there was homogeneity of variances, as assessed by Levene's test for equality of variances ($p = 0.225$). The weight was larger for males (68.98) than female (52.1), a statistically significant difference, $t(16) = 2.784$, $p = 0.013$.

The paired t test

The paired t test used to compare two sets of paired samples.

In this example we consider the weight of the mice before and after a specific treatment. We want to know if the difference in weight before and after the treatment is different from zero.

```
# Data in two numeric vectors
# Weight of the mice before treatment
before <-c(200.1, 190.9, 192.7, 213, 241.4, 196.9, 172.2, 185.5, 205.2, 193.7)
# Weight of the mice after treatment
after <-c(392.9, 393.2, 345.1, 393, 434, 427.9, 422, 383.9, 392.3, 352.2)
# Create a data frame
my_data2 <- data.frame(
                group = rep(c("before", "after"), each = 10),
                weight = c(before,  after)
                )

# Print all data
print(my_data2)
```

```
##     group weight
## 1  before  200.1
## 2  before  190.9
## 3  before  192.7
## 4  before  213.0
## 5  before  241.4
## 6  before  196.9
## 7  before  172.2
## 8  before  185.5
## 9  before  205.2
## 10 before  193.7
## 11  after  392.9
## 12  after  393.2
## 13  after  345.1
## 14  after  393.0
## 15  after  434.0
## 16  after  427.9
## 17  after  422.0
## 18  after  383.9
## 19  after  392.3
## 20  after  352.2
```

Compute summary statistics

```
library("dplyr")
group_by(my_data2, group) %>%
  summarise(
    count = n(),
    mean = mean(weight, na.rm = TRUE),
    sd = sd(weight, na.rm = TRUE)
  )
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

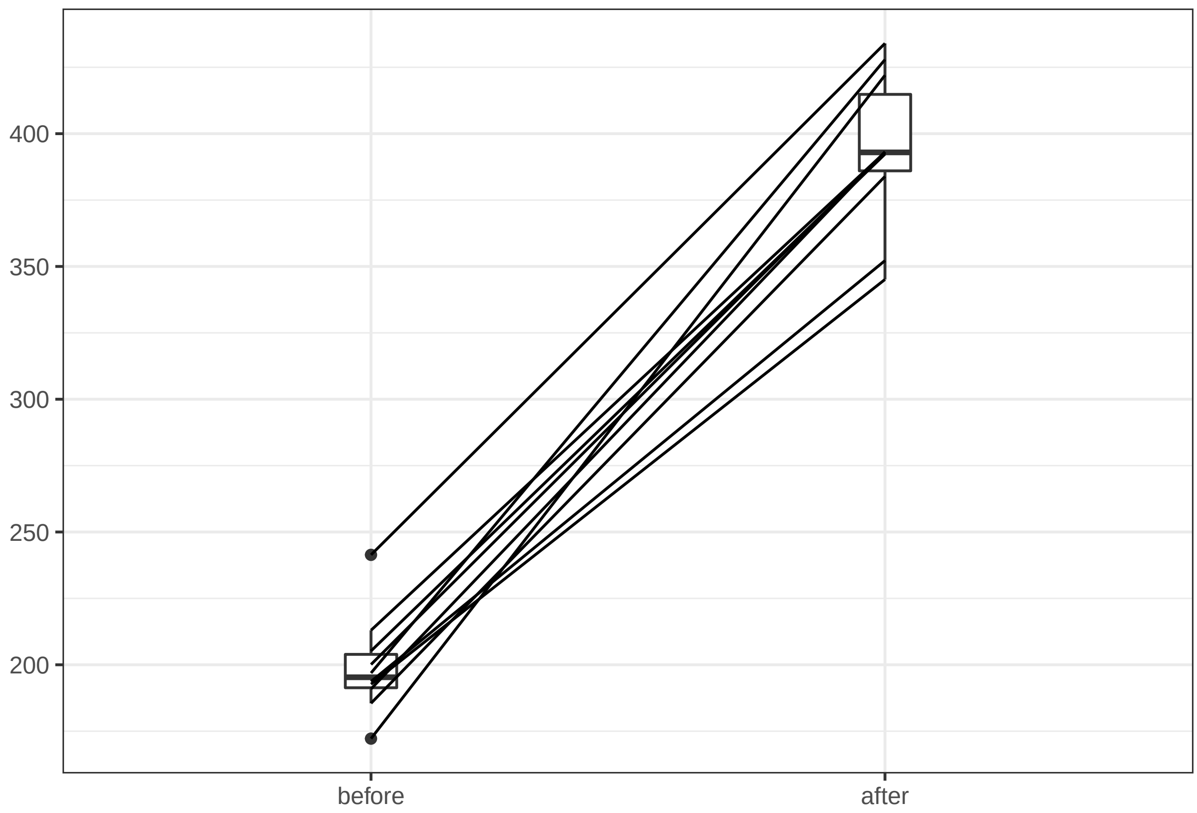
```
## # A tibble: 2 x 4
##   group  count  mean    sd
##   <chr>  <int> <dbl> <dbl>
## 1 after     10  394.  29.4
## 2 before    10  199.  18.5
```

Plot paired data

```r
# Subset weight data before treatment
before <- subset(my_data2,  group == "before", weight,
                 drop = TRUE)
# subset weight data after treatment
after <- subset(my_data2,  group == "after", weight,
                 drop = TRUE)

# Plot paired data
library(PairedData)
```

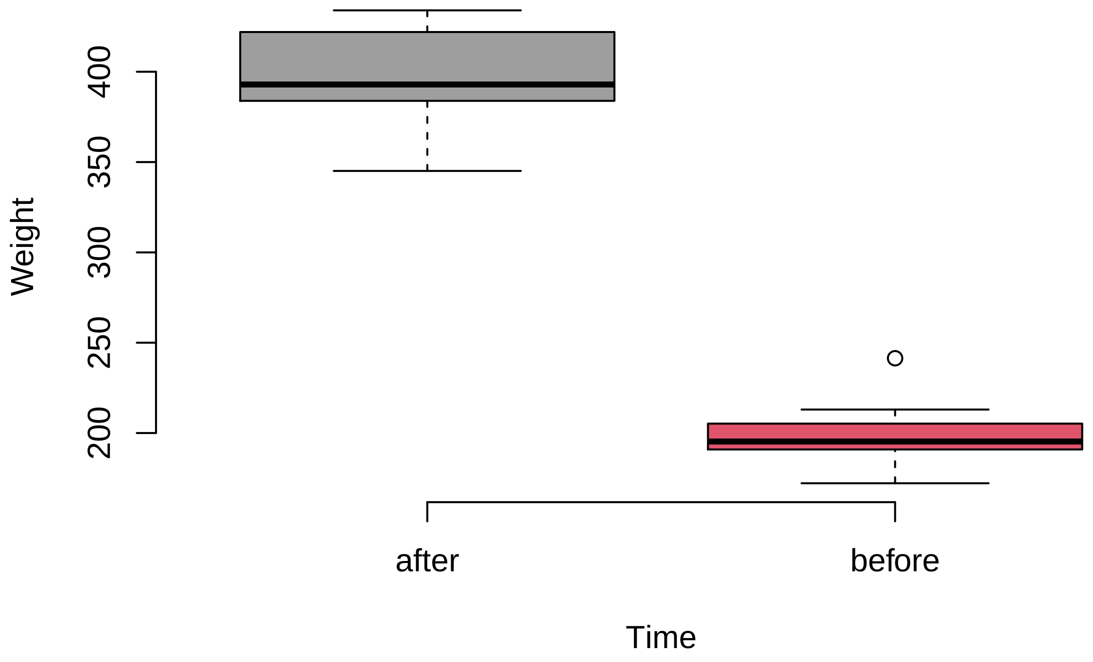
```
## Loading required package: MASS

##
## Attaching package: ’MASS’

## The following object is masked from ’package:dplyr’:
##
##     select

## Loading required package: gld

## Loading required package: mvtnorm

## Loading required package: lattice

## Loading required package: ggplot2

##
## Attaching package: ’PairedData’

## The following object is masked from ’package:base’:
##
##     summary
```

```r
pd <- paired(before, after)
plot(pd, type = "profile") + theme_bw()
```



Visualize data

```
# Box plot
boxplot(weight ~ group, data = my_data2,
        xlab = "Time", ylab = "Weight",
        frame = FALSE, col = c("8", "2")) #color with numbers is an option
```



```
# Compute t-test
res <- t.test(before, after, paired = TRUE)
res
```

```
##
##  Paired t-test
##
## data:  before and after
## t = -20.883, df = 9, p-value = 6.2e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -215.5581 -173.4219
## sample estimates:
## mean of the differences
##                 -194.49
```

Check normality of the differences

```
d=before-after #Calculate differences

shapiro.test(d) #test for normality
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d
## W = 0.94536, p-value = 0.6141
```

Documentation:

A paired-samples t-test was run to determine if there were differences in weight before and after treatment in mice. There were no extreme outliers in the data, as assessed by inspection of a boxplot. Weight differences were normally distributed, as assessed by Shapiro-Wilk test ($p > 0.05$). The weight was larger after the treatment compared to before, a statistically significant difference, $t(16) = -20.88$, $p < 0.0001$.

One-way ANOVA

The one-way analysis of variance (ANOVA), also known as one-factor ANOVA, is used to compare means in a cases where there are two or more groups. The data is organized into several groups base on the so called factor variable.

In this example we run an ANOVA test to see if physical activity has an effect in the ability to cope with stress. Our factor is physical activity with 4 levels: "Sedentary", "Low", "Moderate", "High", and our dependent/response variable is the ability to cope with workplace-related stress (CWWS score).

Read the data

```r
# Read and prepare the data

library(readxl)
data_anova <- read_excel("one-way-anova.xlsx", col_types = c("numeric", "numeric"))

# Create the factor variable or better trasform group to factor.

data_anova <- within(data_anova, {

  group <- factor(group, levels = 1:4, labels = c("Sedentary", "Low", "Moderate", "High"))

})

head(data_anova)
```

```
## # A tibble: 6 x 2
##   group     coping_stress
##   <fct>             <dbl>
## 1 Sedentary          3.18
## 2 Sedentary          3.28
## 3 Sedentary          3.82
## 4 Low                4.11
## 5 Sedentary          4.12
## 6 Low                4.37
```

```r
# Now the data are ready to analyze.
```

Explore the data

Compute summary statistics

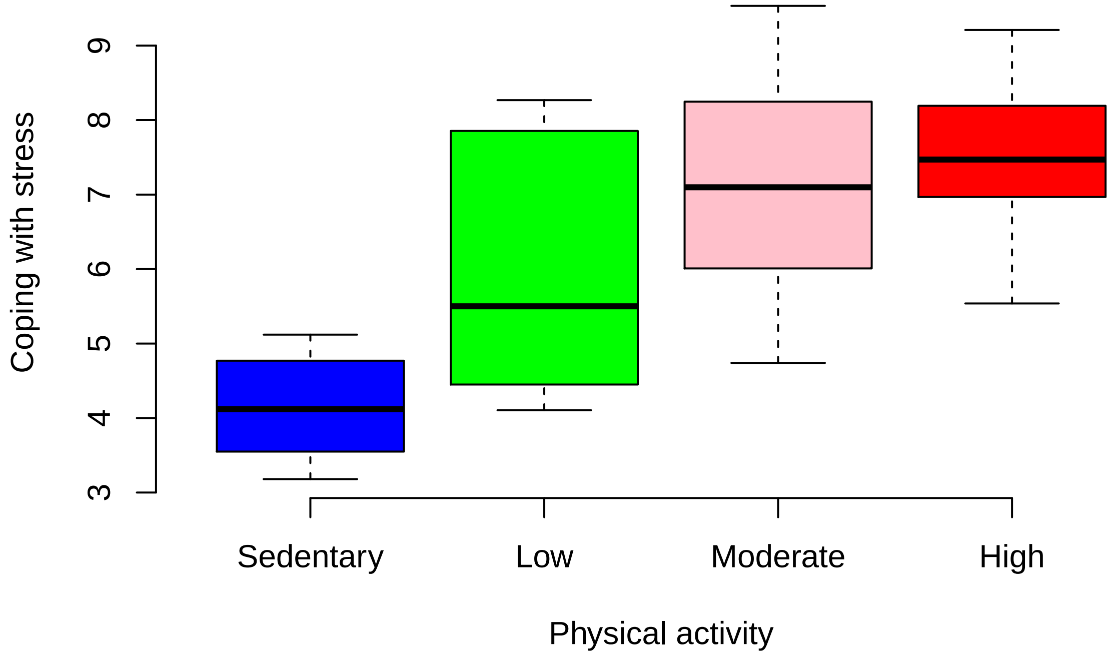
```r
library(dplyr)

group_by(data_anova, group) %>%
  summarise(
    count = n(),
    mean = mean(coping_stress, na.rm = TRUE),
    sd = sd(coping_stress, na.rm = TRUE)
  )
```

```
## ‘summarise()‘ ungrouping output (override with ‘.groups‘ argument)
```

```
## # A tibble: 4 x 4
##   group     count  mean    sd
##   <fct>     <int> <dbl> <dbl>
## 1 Sedentary     7  4.15 0.771
## 2 Low           9  5.88 1.69
## 3 Moderate      8  7.12 1.57
## 4 High          7  7.51 1.24
```

Visualize your data

```
# Box plot
boxplot(coping_stress ~ group, data = data_anova,
        xlab = "Physical activity", ylab = "Coping with stress",
        frame = FALSE, col = c("blue", "green", "pink", "red"))
```



```
# plotmeans
library("gplots")
```

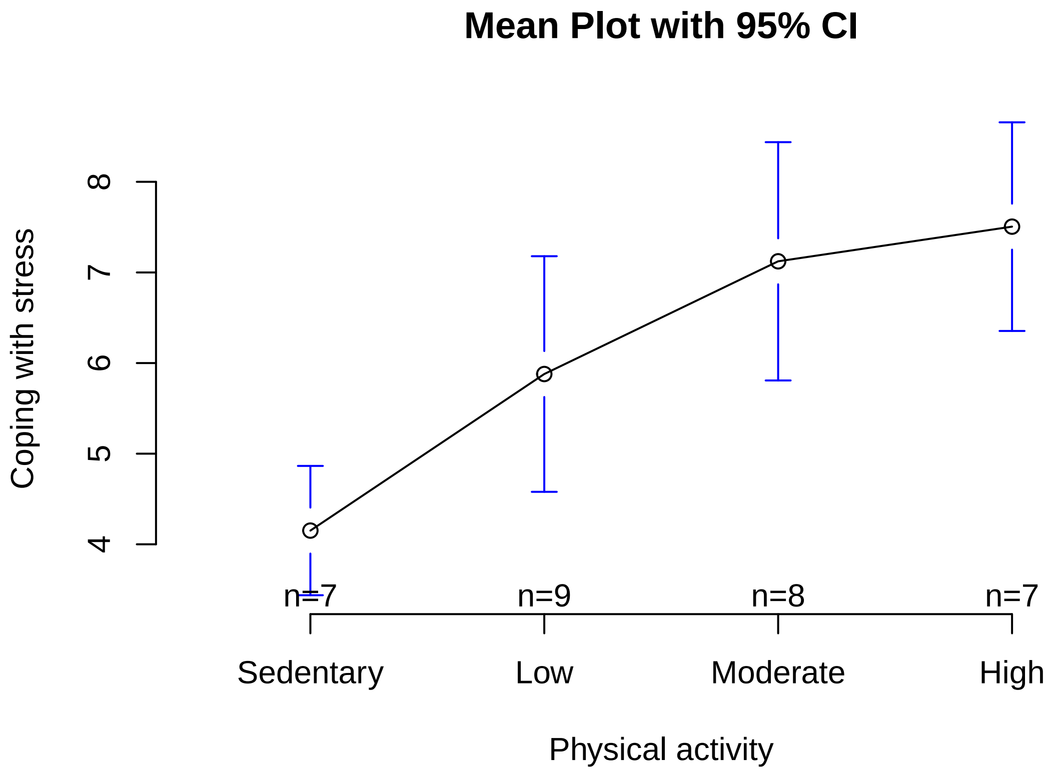
```
##
## Attaching package: ’gplots’
```

```
## The following object is masked from ’package:stats’:
##
```

```
##     lowess
```

```
plotmeans(coping_stress ~ group, data = data_anova, frame = FALSE,
          xlab = "Physical activity", ylab = "Coping with stress",
          main="Mean Plot with 95% CI")
```

```
## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter

## Warning in axis(1, at = 1:length(means), labels = legends, ...): "frame" is not
## a graphical parameter

## Warning in plot.xy(xy.coords(x, y), type = type, ...): "frame" is not a
## graphical parameter
```



Compute one-way ANOVA test

The R function aov() can be used to answer to our research question. The function summary.aov() is used to summarize the analysis of the variance model.

```
# Compute the analysis of variance
res.aov <- aov(coping_stress ~ group, data = data_anova)
# Summary of the analysis
summary(res.aov)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## group        3  49.03  16.344   8.316 0.000445 ***
## Residuals   27  53.07   1.965                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output includes the columns F value and Pr(> F) corresponding to the p-value of the test.

Interpret the result of one-way ANOVA tests

The ability to cope with workplace-related stress (CWWS score) was statistically significantly different for different levels of physical activity group, F(3, 27) = 8.316, p = 0.000445.

Multiple pairwise-comparison between the means of groups

In one-way ANOVA test, a significant p-value indicates that some of the group means are different, but we don't know which pairs of groups are different. It's possible to perform multiple pairwise-comparison, to determine if the mean difference between specific pairs of group are statistically significant.

As the ANOVA test is significant, we can compute Tukey HSD (Tukey Honest Significant Differences, R function: TukeyHSD()) for performing multiple pairwise-comparison between the means of groups. The function TukeyHD() takes the fitted ANOVA as an argument.

```
TukeyHSD(res.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = coping_stress ~ group, data = data_anova)
##
## $group
##                         diff        lwr      upr     p adj
## Low-Sedentary      1.7276175 -0.2057757 3.661011 0.0923527
## Moderate-Sedentary 2.9715262  0.9859704 4.957082 0.0018413
## High-Sedentary     3.3540854  1.3034122 5.404759 0.0006806
## Moderate-Low       1.2439086 -0.6202750 3.108092 0.2835038
## High-Low           1.6264679 -0.3069254 3.559861 0.1226045
## High-Moderate      0.3825593 -1.6029965 2.368115 0.9517285
```

Interpretation:

The result of the multiple pairwise-comparison between the means of groups show a significant difference (p<0.05) between the sedentary group and the moderate and high groups.

Check ANOVA assumptions: test validity?

The ANOVA test assumes that, the data are normally distributed and the variance across groups are homogeneous. We can check that with some diagnostic tests and plots.

1. Check the homogeneity of variance assumption

You can use formal Levene's test, which is less sensitive to departures from normal distribution. The function leveneTest() [in car package] will be used:

```
library(car)
leveneTest(coping_stress ~ group, data = data_anova)
```

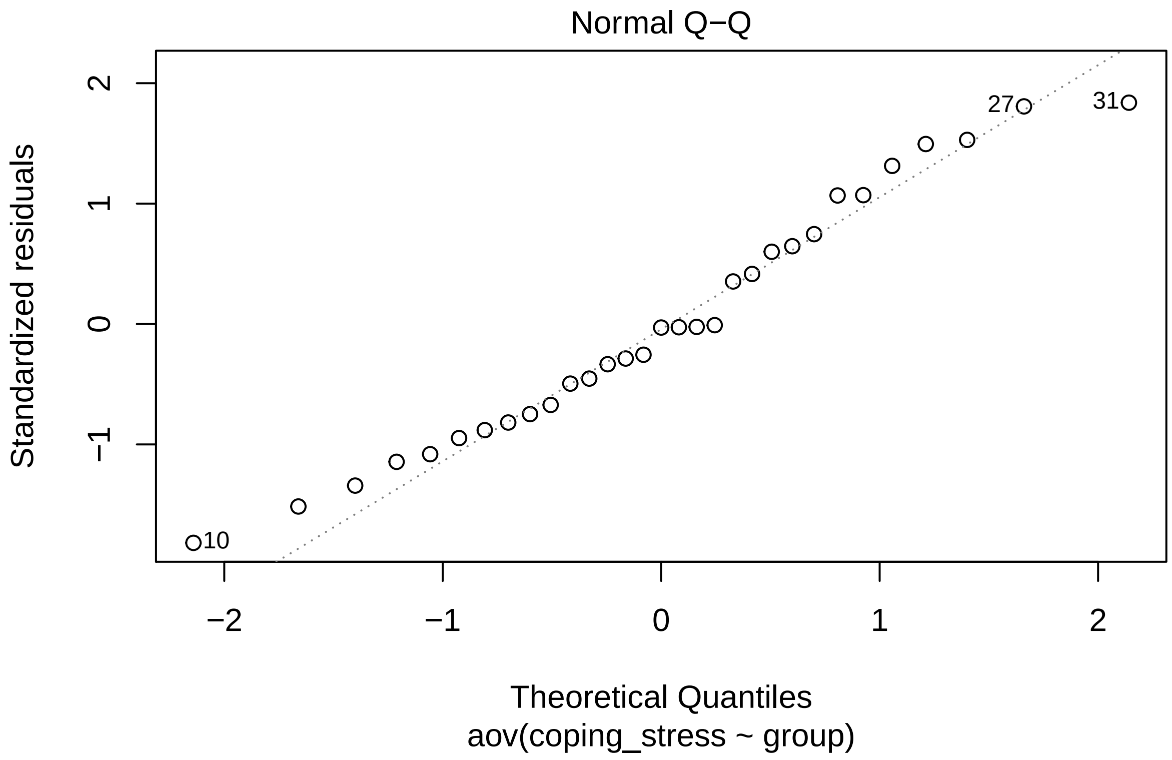
```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  3  1.6308 0.2055
##       27
```

From the output above we can see that the p-value > 0.05, therefore, we can assume the homogeneity of variances in the different groups.

2. Check the normality assumption

Normality plot of residuals. In the plot below, the quantiles of the residuals are plotted against the quantiles of the normal distribution. A 45-degree reference line is also plotted. The normal probability plot of residuals is used to check the assumption that the residuals are normally distributed. It should approximately follow a straight line.

```
# 2. Normality
plot(res.aov, 2)
```



As all the points fall approximately along this reference line, we can assume normality.

The conclusion above, is supported by the Shapiro-Wilk test on the ANOVA residuals (W = 0.96, p = 0.450) which finds no indication that normality is violated.

```
# Extract the residuals
aov_residuals <- residuals(object = res.aov )
# Run Shapiro-Wilk test
shapiro.test(x = aov_residuals )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aov_residuals
## W = 0.9674, p-value = 0.4506
```

References

http://www.sthda.com/english/wiki/one-way-anova-test-in-r

https://statistics.laerd.com/r-tutorials/independent-samples-t-test-using-r-excel-and-rstudio-3.php

http://www.sthda.com/english/wiki/wiki.php?title=paired-samples-t-test-in-r

http://www.stat.auckland.ac.nz/~paul/RGraphics/rgraphics.html

http://addictedtor.free.fr/graphiques/

http://addictedtor.free.fr/graphiques/thumbs.php?sort=votes

http://www.statmethods.net/advgraphs/layout.html

http://socserv.mcmaster.ca/jfox/

Quick R http://www.statmethods.net/

https://statistics.laerd.com/