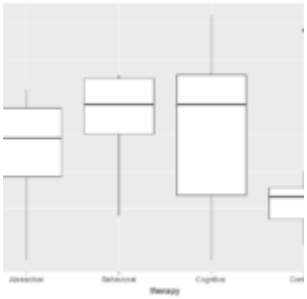
One way MANOVA exercises



In ANOVA our interest lies in knowing if one continuous dependent variable is affected by one or more categorical independent variables. MANOVA is an extension of ANOVA where we are now able to understand how several dependent variables are affected by independent variables. For example consider an investigation where a medical investigator has developed 3 back pain therapies. Patients are enrolled for a 10 week trial and at the end the investigator interviews them on reduction of physiological, emotional and cognitive pain. Interest is in knowing which therapy is best at reducing pain.

Just like in ANOVA we can have one way or two way MANOVA depending on number of independent variables.

When conducting MANOVA it is important to understand the assumptions that need to be satisfied so that the results are valid. The assumptions are explained below.

- The observations are independent. Observations that are collected over time, over space and in any groupings violate the assumption of independence.
- The data follows a multivariate normal distribution. When observations are many we can rely on the central limit theorem (CLT) to achieve normality. It has been generally accepted any distribution with more than that observations will follow a normal distribution. MANOVA is robust to any non-normality that arises from skewness but it is not robust to non-normality resulting from

outliers. Outliers should be checked and appropriate action taken. Analysis can be done with and without the outliers to check sensitivity.

- The variance in all the groups is homogeneous. A Bartlett test is useful for assessing the homogeneity of variance. MANOVA is not robust to deviations from the assumption of normality therefore a transformation is required to stabilize variance.

MANOVA can be used to understand the interactions and main effects of independent variables. The four test statistics that can be used are Wilk's lambda, Pillai trace, Hotelling-Lawley trace and Roy's maximum root. Among the four test statistics Pillai is least affected by any violations in assumptions but Wilk's is the most commonly used.

In this first part of MANOVA exercises we will use data from a study investigating a control and three therapies aimed at reducing symptoms of koro. Forty patients were selected for inclusion in the study and 10 patients were assigned to each of the four groups. Interest is in understanding which therapy is best in reducing symptoms. We will create three variables that hold change in indices before and after treatment. Here we have one independent variable and three dependent variables resulting in a one way MANOVA.

Solutions to these exercises can be found here

Exercise 1

Import data into R

Exercise 2

Check the number of observations in each group

Exercise 3

Create the variables that hold the change in indices

Exercise 4

Summarize the change variables

Exercise 5

Get descriptive statistics for each therapy

Exercise 6

Obtain the correlation matrix

Exercise 7

Check for outliers

Exercise 8

Check for homogeneity of variance

Exercise 9

Run MANOVA test with outliers

Exercise 10

Interpret results

One way MANOVA in R solutions

These are solutions to exercises found here

```
####################
#                  #
#    Exercise 1    #
#                  #
####################
# Read in SPSS data
```

```
library(foreign)
koro.data                                                    =
read.spss("C:/Users/INVESTS/Downloads/koro.sav",to.data.frame
= TRUE)
View(koro.data)
attach(koro.data)
####################
#                  #
#    Exercise 2    #
#                  #
####################
#Check number of observations in each group
table(therapy)
```

```
## therapy
## Abreaction Behavioral Cognitive  Control
##         10         10         10         10
```

```
####################
#                  #
#    Exercise 3    #
#                  #
####################
#Create the variables that hold the change in indices
koro.data$si.diff = si_post - si_pre
koro.data$sf.diff = sf_post - sf_pre
koro.data$oa.diff = oa_post - oa_pre
####################
#                  #
#    Exercise 4    #
#                  #
####################
#Summarize the change variables
library(pastecs)
indices = koro.data[,c(3,10,11,12)]
stat.desc(indices)
```

```
##          therapy    si.diff     sf.diff    oa.diff
## nbr.val       NA  40.000000  40.0000000  40.000000
## nbr.null      NA   1.000000   0.0000000   1.000000
## nbr.na        NA   0.000000   0.0000000   0.000000
## min           NA  -7.000000 -16.0000000 -13.000000
## max           NA  26.000000  38.0000000  20.000000
```

```
## range         NA  33.000000  54.0000000  33.000000
## sum           NA 328.000000 484.0000000 236.000000
## median        NA   9.500000  13.5000000   6.000000
## mean          NA   8.200000  12.1000000   5.900000
## SE.mean       NA   1.392839   1.7646093   1.189592
## CI.mean       NA   2.817282   3.5692593   2.406176
## var           NA  77.600000 124.5538462  56.605128
## std.dev       NA   8.809086  11.1603694   7.523638
## coef.var      NA   1.074279   0.9223446   1.275193
```

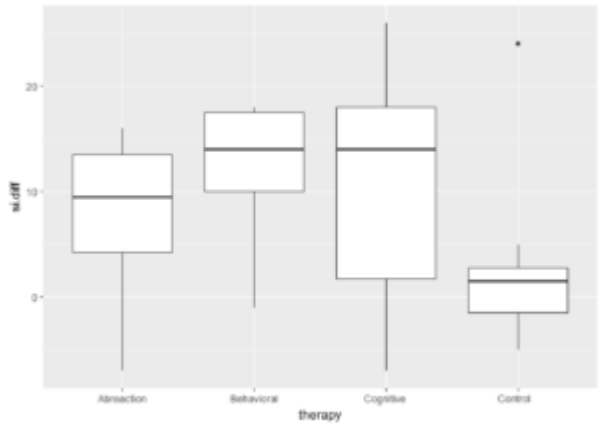
```
####################
#                  #
#    Exercise 5    #
#                  #
####################
#Get descriptive statistics for each therapy
library(psych)
describeBy(indices[-1],therapy)
```

```
## group: Abreaction
##          vars  n mean    sd median trimmed   mad min max
range  skew
## si.diff    1 10  7.8  7.60    9.5    8.62  7.41  -7  16
23 -0.69
## sf.diff    2 10 19.1 12.38   18.5   18.75 17.79   3  38
35  0.17
## oa.diff    3 10 10.1  7.16   11.5   10.12  8.15   0  20
20 -0.18
##         kurtosis   se
## si.diff    -1.00 2.40
## sf.diff    -1.58 3.91
## oa.diff    -1.63 2.26
## ------------------------------------------------------------
## group: Behavioral
##          vars  n mean   sd median trimmed  mad min max range
skew kurtosis
## si.diff    1 10 12.1 6.45   14.0   13.00 5.93  -1  18    19
-0.78    -0.83
## sf.diff    2 10 16.0 6.11   17.5   16.88 5.19   3  22    19
-0.88    -0.55
## oa.diff    3 10  6.3 7.67    7.5    6.38 4.45  -8  20    28
-0.17    -0.60
```

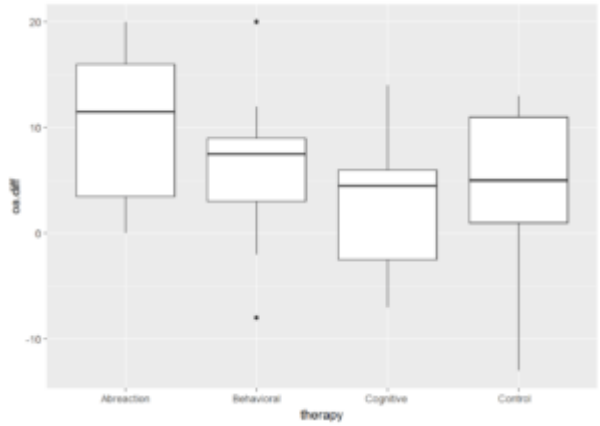
```
##           se
## si.diff 2.04
## sf.diff 1.93
## oa.diff 2.43
## ------------------------------------------------------------
## group: Cognitive
##          vars  n mean    sd median trimmed   mad min max
range  skew
## si.diff    1 10 10.1 10.94   14.0   10.25 11.12  -7  26
33 -0.24
## sf.diff    2 10 12.3  8.60   14.5   12.88  7.41  -3  23
26 -0.52
## oa.diff    3 10  2.9  6.35    4.5    2.75  3.71  -7  14
21 -0.04
##         kurtosis   se
## si.diff    -1.53 3.46
## sf.diff    -1.30 2.72
## oa.diff    -1.16 2.01
## ------------------------------------------------------------
## group: Control
##          vars  n mean   sd median trimmed  mad min max range
skew kurtosis
## si.diff    1 10  2.8 7.98    1.5    1.12 3.71  -5  24    29
1.73     2.09
## sf.diff    2 10  1.0 8.18    2.0    1.38 5.93 -16  15    31
-0.41    -0.20
## oa.diff    3 10  4.3 7.89    5.0    5.38 7.41 -13  13    26
-0.77    -0.33
##           se
## si.diff 2.52
## sf.diff 2.59
## oa.diff 2.49

####################
#                  #
#    Exercise 6    #
#                  #
####################
#Obtain the correlation matrix
library(Hmisc)
rcorr(as.matrix(indices[-1]),type = "pearson")
```

```
##         si.diff sf.diff oa.diff
## si.diff    1.00    0.56    0.41
## sf.diff    0.56    1.00    0.41
## oa.diff    0.41    0.41    1.00
##
## n= 40
##
##
## P
##         si.diff sf.diff oa.diff
## si.diff         0.0002  0.0079
## sf.diff 0.0002          0.0089
## oa.diff 0.0079  0.0089
```

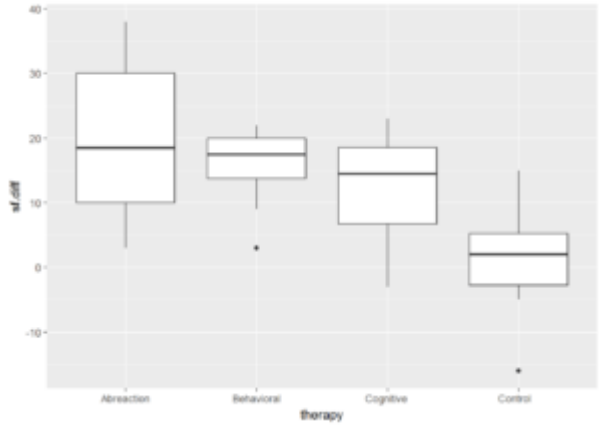
```
#Our variables are moderately correlated. When variables are
highly correlated some need to be dropped
####################
#                  #
#    Exercise 7    #
#                  #
####################
#Check for univariate and multivariate outliers
library(ggplot2)
#Check univariate outliers
ggplot(indices,aes(x=therapy,y=si.diff)) + geom_boxplot()
```




```
ggplot(indices,aes(x=therapy,y=oa.diff)) + geom_boxplot()
```

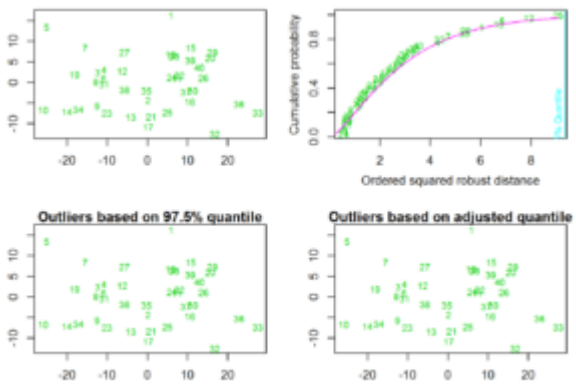



```
ggplot(indices,aes(x=therapy,y=sf.diff)) + geom_boxplot()
```




```
#Box plots show some observations are outliers
#Check multivariate outliers
library(mvoutlier)
aq.plot(indices[-1])
```

```
## Projection to the first and second robust principal components.
## Proportion of total variation (explained variance): 0.7961158
```



```
## $outliers
##  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE
## [12] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE
## [23] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE
## [34] FALSE FALSE FALSE FALSE FALSE FALSE FALSE

#No observations were identified as multivariate outliers
####################
#                  #
#    Exercise 8    #
#                  #
####################
#Check for  homogeneity of variance
bartlett.test(sf.diff~therapy, data = indices)

##
##  Bartlett test of homogeneity of variances
##
## data:  sf.diff by therapy
## Bartlett's K-squared = 4.3844, df = 3, p-value = 0.2228

bartlett.test(si.diff~therapy,data = indices)

##
##  Bartlett test of homogeneity of variances
##
## data:  si.diff by therapy
## Bartlett's K-squared = 2.6569, df = 3, p-value = 0.4476

bartlett.test(oa.diff~therapy,data = indices)

##
##  Bartlett test of homogeneity of variances
##
## data:  oa.diff by therapy
## Bartlett's K-squared = 0.46566, df = 3, p-value = 0.9264

#There was no evidence of departure from homogeneity of
variance
####################
#                  #
```

```
#    Exercise 9    #
#                  #
####################
#Run MANOVA with outliers
manova.analysis = manova(as.matrix(indices[-1])~therapy)
summary(manova.analysis)
```

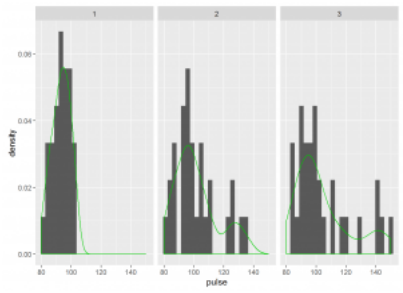
```
##           Df  Pillai approx F num Df den Df   Pr(>F)
## therapy    3 0.63123   3.1978      9    108 0.001812 **
## Residuals 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

```
summary(manova.analysis,test = "Wilks")
```

```
##           Df   Wilks approx F num Df den Df   Pr(>F)
## therapy    3 0.46438   3.4126      9 82.898 0.001301 **
## Residuals 36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

```
#Both Pillai and Wilk's lambda showed significance
####################
#                  #
#    Exercise 10   #
#                  #
####################
#We did not find violation of any of assumptions required for
MANOVA analysis
#However we identified some observations that were outliers
#In part 2 of MANOVA exercises we will drop the outliers and
repeat the analysis
#In part 2 we will explain what to do when we have a
significant result
```

Repeated measures ANOVA in R Exercises



One way, two way and n way ANOVA are used to test difference in means when we have one, two and n factor variables. A key assumption when performing these ANOVAs is that the measurements are independent. When we have repeated measures this assumption is violated, so we have to use repeated measures ANOVA. Repeated measures designs occur often in longitudinal studies where we are interested in understanding change over time. For example a medical researcher would be interested in assessing the level of depression before and after a surgery procedure. Repeated measures designs are not limited to longitudinal studies, they can also be used when you have an important variable you would like to repeat measures. For example in a fitness experiment you can repeat your measures at different intensity levels. Repeated measures ANOVA can be considered an extension of the paired t test.

Before diving deeper into repeated measures ANOVA you need to understand terminology used. A ***subject*** is a member of the sample under consideration. In our medical study introduced earlier an individual patient is a subject. The ***within-subjects factor*** is the variable that identifies how the dependent variable has been repeatedly measured. In our medical study we would measure depression 4 weeks before surgery, 4 weeks after surgery and 8 weeks after surgery. The different conditions when repeated measurements are made are referred to as trials. A ***between-subjects factor*** identifies

independent groups in the study. For example if we had two different procedures this would be the between subjects factor. These conditions are referred to as groups. Repeated measures analysis requires balance in between-subjects factor. For example subjects in each of surgery procedures need to be equal.

With a repeated measures design we are able to test the following hypotheses.

1. There is no within-subjects main effect
2. There is no between-subjects main effect
3. There is no between subjects interaction effect
4. There is no within subject by between subject interaction effect

There are two assumptions that need to be satisfied when using repeated measures.

1. The dependent variable is normally distributed in each level of the within-subjects factor. Repeated measures analysis is robust to violations of normality with a large sample size which is considered at least 30 subjects. However the accuracy of p values is questionable when the distribution is heavily skewed or thick tailed.
2. The variance across the within subject factor is equal. This is the sphericity assumption. Repeated measures analysis is not robust to this assumption so when there is a violation power decreases and a corresponding increase in probability of a type II error occurs. A Mauchly’s test assesses the null hypothesis variance is equal. The sphericity assumption is only relevant when there are more than 2 levels of the within subjects factor.

When the sphericity assumption is violated we make corrections by adjusting the degrees of freedom. Corrections available are

Greenhouse-Geisser, Huynh-Feldt and Lower bound. To make a decision on appropriate correction we use a Greenhouse-Geisser estimate of sphericity (ξ). When $\xi < 0.75$ or we do not know anything about sphericity the Greenhouse-Geisser is the appropriate correction. When $\xi > 0.75$ Huynh-Feldt is the appropriate correction.

For this exercise we will use data on pulse rate exer. People were randomized to two diets, three exercise types and pulse was measured at three different time points. For this data time points is the within-subjects factor. The between-subjects factors are diet and exercise type

The solutions to the exercises below can be found here

Exercise 1

Load the data and inspect its structure

Exercise 2

Check for missing values

Exercise 3

Check for balance in between-subjects factor

Exercise 4

Generate descriptive statistics for the sex variable which is a between subjects factor

Exercise 5

Generate descriptive statistics for the treatment level variable which is a between subjects factor

Exercise 6

Generate descriptive statistics for the weeks variable which is the within subjects factor

Exercise 7

Use histograms to assess distribution across within subjects factor.

Exercise 8

Perform a repeated measures analysis with only the within subjects factor

Exercise 9

Perform a repeated measures analysis with the within subjects factor and one between subjects factor

Exercise 10

Perform a repeated measures analysis with the within subjects factor and two between subjects factors

---

Repeated measures ANOVA in R Solutions

Solutions to exercises found here

```
####################
#                  #
#    Exercise 1    #
#                  #
####################
#load the data
setwd("H:/data analysis")
library(foreign)
exercise = read.csv("exer.csv")
attach(exercise)
```

```
head(exercise)
```

```
##   id diet exertype pulse time
## 1  1    1        1    85    1
## 2  1    1        1    85    2
## 3  1    1        1    88    3
## 4  2    1        1    90    1
## 5  2    1        1    92    2
## 6  2    1        1    93    3
```

```
####################
#                  #
#    Exercise 2    #
#                  #
####################
#check for missing values
sapply(exercise, function(x) sum(is.na(x)))
```

```
##       id     diet exertype    pulse     time
##        0        0        0        0        0
```

```
#check structure of data and perform necessary conversion to
factors
str(exercise)
```

```
## 'data.frame':    90 obs. of  5 variables:
##  $ id      : int  1 1 1 2 2 2 3 3 3 4 ...
##  $ diet    : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ exertype: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ pulse   : int  85 85 88 90 92 93 97 97 94 80 ...
##  $ time    : int  1 2 3 1 2 3 1 2 3 1 ...
```

```
exercise$time = as.factor(time)
exercise$exertype = as.factor(exertype)
exercise$diet = as.factor(diet)
exercise$id = as.factor(id)

####################
#                  #
#    Exercise 3    #
#                  #
####################
#check balance of between subjects factors
table(diet)
```

```
## diet
##  1  2
## 45 45

table(exertype)

## exertype
##  1  2  3
## 30 30 30

table(diet,exertype)

##     exertype
## diet  1  2  3
##    1 15 15 15
##    2 15 15 15

#The design is balanced
####################
#                  #
#    Exercise 4    #
#                  #
####################
#get descriptive statistics across the diet variable
library(psych)

## Warning: package 'psych' was built under R version 3.3.1

describeBy(pulse,diet)

## group: 1
##    vars  n  mean    sd median trimmed  mad min max range
skew kurtosis
## X1    1 45 95.96 10.48     94   94.84 7.41  80 132    52
1.21     1.72
##      se
## X1 1.56
## ------------------------------------------------------------
## group: 2
##    vars  n   mean    sd median trimmed   mad min max range
skew kurtosis
## X1    1 45 103.44 17.55     99   101.3 10.38  83 150    67
1.16     0.27
##      se
```

```
## X1 2.62
```

```
####################
#                  #
#    Exercise 5    #
#                  #
####################
#get descriptive statistics across exercise type
describeBy(pulse,exertype)
```

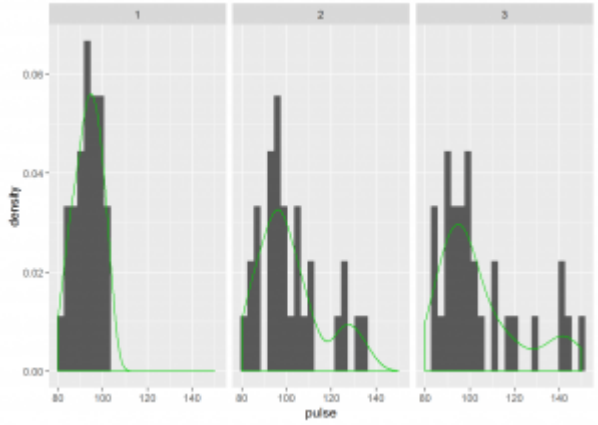
```
## group: 1
##    vars  n  mean   sd median trimmed  mad min max range
skew kurtosis
## X1    1 30 90.83 5.83   91.5   90.88 7.41  80 100    20
-0.15    -1.22
##      se
## X1 1.06
## ------------------------------------------------------
## group: 2
##    vars  n mean   sd median trimmed  mad min max range skew
kurtosis   se
## X1    1 30 95.2 6.78   95.5   95.21 8.15  84 109    25 0.02
-1.08 1.24
## ------------------------------------------------------
## group: 3
##    vars  n   mean    sd median trimmed   mad min max range
skew kurtosis
## X1    1 30 113.07 17.62    110  111.88 19.27  87 150    63
0.48     -1.1
##      se
## X1 3.22
```

```
####################
#                  #
#    Exercise 6    #
#                  #
####################
#get descriptive statistics across time points
describeBy(pulse,time)
```

```
## group: 1
##    vars  n  mean   sd median trimmed  mad min max range
skew kurtosis
```

```
## X1    1 30 93.13 6.15   93.5   93.29 6.67  80 103    23
-0.24    -0.9
##      se
## X1 1.12
## ------------------------------------------------------------
## group: 2
##    vars  n   mean    sd median trimmed   mad min max range
skew kurtosis
## X1    1 30 101.53 14.56   97.5  100.12 10.38  82 135    53
0.82    -0.32
##      se
## X1 2.66
## ------------------------------------------------------------
## group: 3
##    vars  n   mean    sd median trimmed   mad min max range
skew kurtosis
## X1    1 30 104.43 18.88   99.5  102.04 14.08  83 150    67
1.06    -0.08
##      se
## X1 3.45

####################
#                  #
#    Exercise 7    #
#                  #
####################
#use histograms to assess the distribution at at each time
point
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.1

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

ggplot(exercise,aes(x=pulse)) + geom_histogram(binwidth =
3,aes(y=..density..)) + facet_grid(. ~ time) +
geom_density(col=3)
```

```
#observations at time points 2 and 3 seem to depart from
normality
```




```
####################
#                  #
#    Exercise 8    #
#                  #
####################
#function ezAnova is one of the ways we can do repeated
measures
#we will use library ez, so you need to install it
library(ez)

## Warning: package 'ez' was built under R version 3.3.1

#perform analysis with only the within subject factor
ex1 = ezANOVA(exercise,dv = .(pulse), wid = .(id), within =
.(time),detailed = TRUE)
ex1

## $ANOVA
##         Effect DFn DFd      SSn     SSd           F
p p<.05
## 1 (Intercept)    1   29 894608.1 12488.9 2077.33547
1.525579e-28      *
## 2         time   2   58   2066.6  5093.4   11.76648
5.137456e-05     *
##         ges
## 1 0.9807252
## 2 0.1051764
##
## $`Mauchly's Test for Sphericity`
##   Effect         W           p p<.05
## 2   time 0.6602812 0.002993676     *
##
```

```
## $`Sphericity Corrections`
##   Effect       GGe        p[GG] p[GG]<.05       HFe
p[HF] p[HF]<.05
## 2   time 0.7464253 0.0003118343          * 0.7777163
0.0002493935         *
```

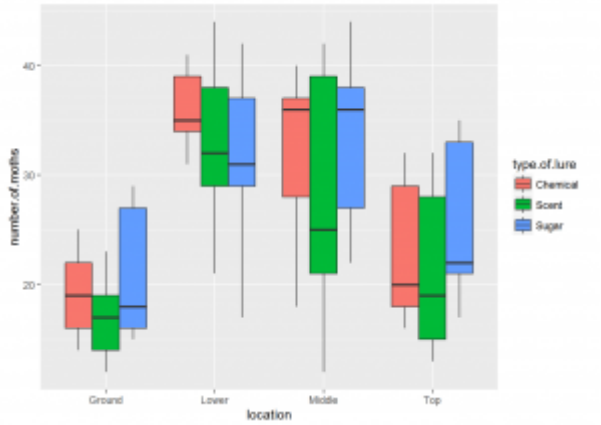
```
####################
#                  #
#    Exercise 9    #
#                  #
####################
#Perform a repeated measures analysis with the within subjects
factor and one between subjects factor
ex2 = ezANOVA(exercise,dv = .(pulse), wid = .(id), within =
.(time),between = .(diet), detailed = TRUE)
ex2
```

```
## $ANOVA
##        Effect DFn DFd         SSn       SSd           F
p p<.05
## 1 (Intercept)   1  28 894608.1000 11227.022 2231.137189
3.034079e-28     *
## 2        diet   1  28   1261.8778 11227.022    3.147101
8.693862e-02
## 3        time   2  56   2066.6000  4900.578   11.807751
5.264184e-05     *
## 4   diet:time   2  56    192.8222  4900.578    1.101711
3.393955e-01
##         ges
## 1 0.98229168
## 2 0.07256559
## 3 0.11358565
## 4 0.01181478
##
## $`Mauchly's Test for Sphericity`
##      Effect        W           p p<.05
## 3      time 0.673336 0.004798766     *
## 4 diet:time 0.673336 0.004798766     *
##
## $`Sphericity Corrections`
##      Effect       GGe        p[GG] p[GG]<.05      HFe
p[HF]
```

```
## 3      time 0.7537703 0.0003019441          * 0.787369
0.0002376889
## 4 diet:time 0.7537703 0.3264339487            0.787369
0.3286130371
##   p[HF]<.05
## 3         *
## 4

####################
#                  #
#    Exercise 10   #
#                  #
####################
#Perform a repeated measures analysis with the within subjects
factor and two between subjects factors
ex3 = ezANOVA(exercise,dv = .(pulse), wid = .(id), within =
.(time),between = .(diet,exertype), detailed = TRUE)
ex3

## $ANOVA
##                 Effect DFn DFd          SSn     SSd
F            p
## 1          (Intercept)   1  24 894608.1000 2085.2
10296.659505 4.034818e-33
## 2                 diet   1  24    1261.8778 2085.2
14.523819 8.482600e-04
## 3             exertype   2  24    8326.0667 2085.2
47.915212 4.166101e-09
## 5                 time   2  48    2066.6000 1563.6
31.720645 1.662197e-09
## 4       diet:exertype   2  24     815.7556 2085.2
4.694546 1.902303e-02
## 6            diet:time   2  48     192.8222 1563.6
2.959666 6.136514e-02
## 7       exertype:time   4  48    2723.3333 1563.6
20.900486 4.991713e-10
## 8 diet:exertype:time   4  48     613.6444 1563.6
4.709474 2.750071e-03
##   p<.05       ges
## 1     * 0.99593791
## 2     * 0.25696611
## 3     * 0.69529515
```

```
## 5     * 0.36158449
## 4     * 0.18271820
## 6       0.05019292
## 7     * 0.42738172
## 8     * 0.14396538
##
## $`Mauchly's Test for Sphericity`
##               Effect         W         p p<.05
## 5               time 0.9241579 0.4037205
## 6          diet:time 0.9241579 0.4037205
## 7      exertype:time 0.9241579 0.4037205
## 8 diet:exertype:time 0.9241579 0.4037205
##
## $`Sphericity Corrections`
##               Effect       GGe        p[GG] p[GG]<.05
HFe
## 5               time 0.9295044 5.503833e-09         *
1.004364
##  6               diet:time  0.9295044  6.568963e-02
1.004364
## 7      exertype:time 0.9295044 1.840672e-09         *
1.004364
## 8 diet:exertype:time 0.9295044 3.590363e-03         *
1.004364
##          p[HF] p[HF]<.05
## 5 1.662197e-09        *
## 6 6.136514e-02
## 7 4.991713e-10        *
## 8 2.750071e-03        *
```

Two Way ANOVA in R Exercises



One way analysis of variance helps us understand the relationship between one continuous dependent variable and one categorical independent variable. When we have one continuous dependent variable and more than one independent categorical variable we cannot use one way ANOVA. When we have two independent categorical variable we need to use two way ANOVA. When we have more than two categorical independent variables we need to use N way ANOVA.

In two way ANOVA there are three hypotheses of interest as listed below

1. H: There is an effect of the first factor on the dependent continuous variable (main effect)
2. H: There is an effect of the second factor variable on the dependent continuous variable (main effect)
3. H: There is a combined effect of the first and second factor variable on the continuous dependent variable (interaction)

The above hypotheses can be extended from two factor variables to N factor variables.

For results of two way ANOVA to be valid there are several assumptions that need to be satisfied. They are listed below.

1. Observations must be independent within and across groups
2. Observations are approximately normally distributed.

3. There is equal variance in the observations
4. We should not have any outliers especially when our design is unbalanced
5. The errors are independent

When the normality and equal variance assumptions are violated you need to transform your data.

In this exercise we will use data on a moth experiment which is available here here. The data is not well formatted in that link so use this csv file moth-trap-experiment.

The dependent variable is the number of moths in a trap. The independent variables are location and type of lure. There were four locations (top, middle, lower and ground). There were three types of lure (scent, sugar and chemical).

Solutions to these exercises are found here

Exercise 1

Read in the data and inspect its structure

Exercise 2

Create summary statistics for location

Exercise 3

Create summary statistics for type of lure

Exercise 4

Create boxplots for each category

Exercise 5

Check for normality

Exercise 6

Check for equality of variance

Exercise 7

Take a log transformation of our data

Exercise 8

Perform a power analysis

Exercise 9

Perform anova

Exercise 10

Check homogeneity of variance

Two Way ANOVA in R Solutions

Solutions to exercises found here

```
####################
#                  #
#    Exercise 1    #
#                  #
####################
#Read in the moth experiment data
setwd("H:/datasets")
moth.experiment = read.csv("moth trap experiment.csv", header
= TRUE)

#Inspect structure of the data
head(moth.experiment)
```

```
##   number.of.moths location type.of.lure
## 1              32      Top     Chemical
## 2              29      Top     Chemical
## 3              16      Top     Chemical
## 4              18      Top     Chemical
## 5              20      Top     Chemical
## 6              37   Middle     Chemical

#check if our design is balanced
table(moth.experiment$location,moth.experiment$type.of.lure)

##
##          Chemical Scent Sugar
##   Ground        5     5     5
##   Lower         5     5     5
##   Middle        5     5     5
##   Top           5     5     5

#our design is balanced because we have equal observations in
each cell
####################
#                  #
#    Exercise 2    #
#                  #
####################
#get summary statistics for location group
library(psych)

## Warning: package 'psych' was built under R version 3.3.1

describeBy(moth.experiment$number.of.moths,moth.experiment$loc
ation)

## group: Ground
##    vars  n  mean   sd median trimmed  mad min max range
skew kurtosis   se
## X1    1 15 19.07 5.09     18   18.85 5.93  12  29    17
0.52    -1.06 1.31
## ------------------------------------------------------------
## group: Lower
##    vars  n  mean  sd median trimmed  mad min max range skew
kurtosis   se
## X1    1 15 33.33 7.5     34   33.77 7.41  17  44    27 -0.6
-0.5 1.94
```

```
## ------------------------------------------------------------
## group: Middle
##    vars  n mean   sd median trimmed   mad min max range
skew kurtosis
## X1    1 15   31 9.79     36   31.46 11.86  12  44    32
-0.39    -1.29
##      se
## X1 2.53
## ------------------------------------------------------------
## group: Top
##    vars  n  mean   sd median trimmed mad min max range skew
kurtosis   se
## X1    1 15 23.33 7.41     21   23.23 8.9  13  35    22 0.24
-1.63 1.91

####################
#                  #
#    Exercise 3    #
#                  #
####################
#get summary statistics for type of lure group
describeBy(moth.experiment$number.of.moths,moth.experiment$typ
e.of.lure)

## group: Chemical
##    vars  n mean   sd median trimmed  mad min max range
skew kurtosis   se
## X1    1 20 27.5 9.06   28.5   27.44 12.6  14  41    27
-0.01    -1.61 2.03
## ------------------------------------------------------------
## group: Scent
##    vars  n  mean    sd median trimmed   mad min max range
skew kurtosis
## X1    1 20 24.75 10.29     22   24.06 11.12  12  44    32
0.43     -1.2
##     se
## X1 2.3
## ------------------------------------------------------------
## group: Sugar
##    vars  n mean   sd median trimmed   mad min max range
skew kurtosis   se
## X1    1 20 27.8 9.06     28   27.44 11.12  15  44    29
```

```
0.14    -1.35 2.03
```

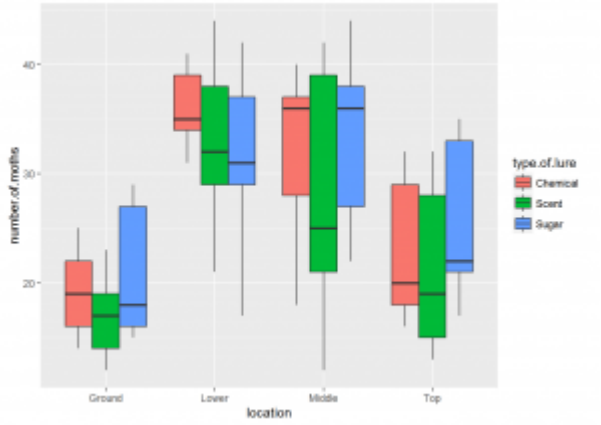
```
####################
#                  #
#    Exercise 4    #
#                  #
####################
#Create boxplots using the two factor variables
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.1
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
ggplot(moth.experiment, aes(x=location,y=number.of.moths, fill
= type.of.lure)) + geom_boxplot()
```




```
####################
#                  #
#    Exercise 5    #
#                  #
####################
#Check for normality of observations

shapiro.test(moth.experiment$number.of.moths)
```

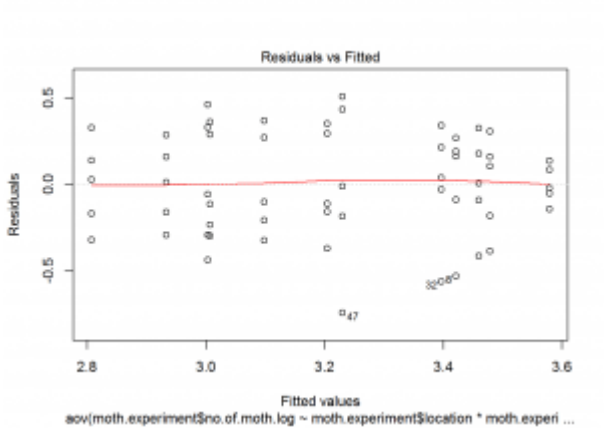
```
##
##  Shapiro-Wilk normality test
##
```

```
## data:  moth.experiment$number.of.moths
## W = 0.94533, p-value = 0.009448

#shapiro test shows our data is not normally distributed
####################
#                  #
#    Exercise 6    #
#                  #
####################
#Check for equality of variance across the two groups so we
will log transform our data
library(car)

## Warning: package 'car' was built under R version 3.3.1

##
## Attaching package: 'car'

## The following object is masked from 'package:psych':
##
##     logit

leveneTest(moth.experiment$number.of.moths~moth.experiment$loc
ation*moth.experiment$type.of.lure)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 11  0.6377 0.7875
##       48

#the levene test shows our data is normally distributed
####################
#                  #
#    Exercise 7    #
#                  #
####################
#take a log transformation of number of moths and check
normality and equal variance
no.of.moth.log = log(moth.experiment$number.of.moths)
moth.experiment$no.of.moth.log  = no.of.moth.log
shapiro.test(moth.experiment$no.of.moth.log)

##
##  Shapiro-Wilk normality test
```

```
##
## data:  moth.experiment$no.of.moth.log
## W = 0.94746, p-value = 0.01185

#the log transformation is not very effective in normalizing
the data
#the appropriate transformation is left as an exercise to the
reader
#this will help the reader appreciate challenges of analyzing
data
leveneTest(moth.experiment$no.of.moth.log~moth.experiment$loca
tion*moth.experiment$type.of.lure)

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group 11  0.5978 0.8211
##       48

####################
#                  #
#    Exercise 8    #
#                  #
####################
#perform a power analysis
#our design has 2 factors with 3 and 4 levels, we have 5
observations in each group
# our df for the mean squared term is 4*3(5-1)=48
#We choose a medium effect size of 0.25
library(pwr)

## Warning: package 'pwr' was built under R version 3.3.1

pwr.f2.test(u=2,v=48,f2=(0.25*0.25))

##
##      Multiple regression power calculation
##
##               u = 2
##               v = 48
##              f2 = 0.0625
##       sig.level = 0.05
##           power = 0.3210203

####################
```

```
#                  #
#    Exercise 9    #
#                  #
####################
#perform anova
moth.anova                                                    =
aov(moth.experiment$no.of.moth.log~moth.experiment$location*mo
th.experiment$type.of.lure)
#location has an effect on number of moths
#type of lure does not have an effect on number of moths
#the combined effect of location and type of lure does not
have an effect on number of moths

#when you have an unbalanced design R does not issue any
warnings
#to correctly analyze an unbalanced design we can use the
Anova function in car library
#we pass results of aov function and specify we would like to
use Type III sums of squares
library(car)
Anova(moth.anova,type = "III")

## Anova Table (Type III tests)
##
## Response: moth.experiment$no.of.moth.log
##                                                          Sum
Sq Df  F value
##                                              (Intercept)
43.018   1 427.6842
##                                 moth.experiment$location
1.302   3    4.3144
##                            moth.experiment$type.of.lure
0.102   2    0.5054
##    moth.experiment$location:moth.experiment$type.of.lure
0.196   6    0.3245
##                                               Residuals
4.828 48
##
Pr(>F)
## (Intercept)                                             <
2.2e-16 ***
##                                 moth.experiment$location
```

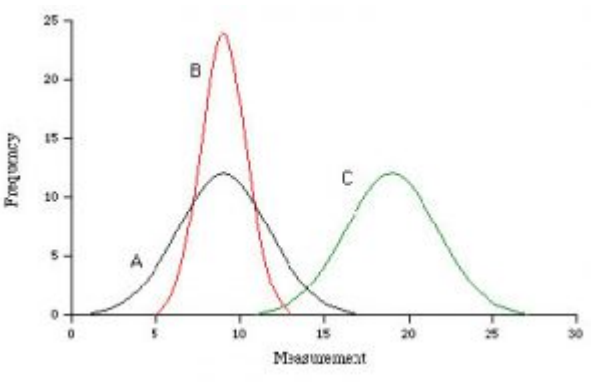
```
0.008988 **
##                                 moth.experiment$type.of.lure
0.606429
##    moth.experiment$location:moth.experiment$type.of.lure
0.920916
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1

####################
#                  #
#    Exercise 10   #
#                  #
####################
#check for homogeneity of residuals
plot(moth.anova,1)
```




```
#homogeneity assumption is not violated but points 47 and 32
are marked as outliers.
#Remember our data still had some non normality
```

One Way Analysis of Variance

Exercises



When we are interested in finding if there is a statistical difference in the mean of two groups we use the t test. When we have more than two groups we cannot use the t test, instead we have to use analysis of variance (ANOVA). In one way ANOVA we have one continuous dependent variable and one independent grouping variable or factor. When we have two groups the t test and one way ANOVA are equivalent.

For our one way ANOVA results to be valid there are several assumptions that need to be satisfied. These assumptions are listed below.

1. The dependent variable is required to be continuous
2. The independent variable is required to be categorical with or more categories.
3. The dependent and independent variables have values for each row of data.
4. Observations in each group are independent.
5. The dependent variable is approximately normally distributed in each group.
6. There is approximate equality of variance in all the groups.
7. We should not have any outliers

When our data shows non-normality, unequal variance or presence of outliers you can transform your data or use a non-parametric test like Kruskal-Wallis. It is good to note Kruskal-Wallis does not require normality of data but still requires equal variance in your groups.

For this exercise we will use data on patients having stomach,

colon, ovary, brochus, or breast cancer. The objective of the study was to identify if the number of days a patient survived was influenced by the organ affected. Our dependent variable is Survival measured in days. Our independent variable is Organ. The data is available here http://lib.stat.cmu.edu/DASL/Datafiles/CancerSurvival.html and a cancer-survival file has been uploaded

Solutions to these exercises can be found here

Exercise 1

Load the data into R

Exercise 2

Create summary statistics for each organ

Exercise 3

Check if we have any outliers using boxplot

Exercise 4

Check for normality using Shapiro.wilk test

Exercise 5

Check for equality of variance

Exercise 6

Transform your data and check for normality and equality of variance.

Exercise 7

Run one way ANOVA test

Exercise 8

Perform a Tukey HSD post hoc test

Exercise 9

Interpret results

Exercise 10

Use a Kruskal-Wallis test

One Way Analysis of Variance Solutions

These are the solutions to the exercises One Way Analysis of Variance here

```
####################
#                  #
#    Exercise 1    #
#                  #
####################

#Read in the cancer survival.csv data
setwd("H:/datasets")
cancer.survival = read.csv("cancer survival.csv", header = TRUE)

#Inspect structure of the data
head(cancer.survival)
```

```
##   Survival   Organ
## 1      124 Stomach
## 2       42 Stomach
## 3       25 Stomach
## 4       45 Stomach
## 5      412 Stomach
```

```
## 6       51 Stomach
```

```
####################
#                  #
#    Exercise 2    #
#                  #
####################
```

```
#Get summary statistics for each organ
#You need to install library psych
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.3.1
```

```
describeBy(cancer.survival$Survival,cancer.survival$Organ)
```

```
## group: Breast
##    vars  n    mean      sd median trimmed    mad min  max
range skew
## X1    1 11 1395.91 1238.97   1166 1280.33 662.72  24 3808
3784 0.81
##    kurtosis     se
## X1     -0.7 373.56
## ------------------------------------------------------------
## group: Bronchus
##    vars  n   mean     sd median trimmed    mad min max
range skew kurtosis
## X1    1 17 211.59 209.86    155   181.2 133.43  20 859
839 1.75     2.66
##      se
## X1 50.9
## ------------------------------------------------------------
## group: Colon
##    vars  n   mean     sd median trimmed    mad min  max
range skew
## X1    1 17 457.41 427.17    372   394.2 244.63  20 1843
1823 1.96
##    kurtosis    se
## X1     3.76 103.6
## ------------------------------------------------------------
## group: Ovary
##    vars n   mean      sd median trimmed    mad min  max
range skew
```

```
## X1    1 6 884.33 1098.58    406  884.33 386.96  89 2970
2881 1.01
##    kurtosis     se
## X1    -0.75 448.49
## ------------------------------------------------------------
## group: Stomach
##    vars  n mean     sd median trimmed    mad min  max range
skew kurtosis
## X1    1 13  286 346.31    124  234.64 121.57  25 1112  1087
1.27     0.25
##       se
## X1 96.05

####################
#                  #
#    Exercise 3    #
#                  #
####################
#Create boxplots to identify any outliers
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.1

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

ggplot(cancer.survival,aes(x = Organ,y=Survival, color =
Organ)) + geom_boxplot() + stat_summary(fun.y=mean,
geom="point", shape=23, size=4) + ggtitle("Survival time of
patients affected by different cancers")
```

```
####################
#                  #
#    Exercise 4    #
#                  #
####################
#Check for normality in each group
 with(cancer.survival,tapply(Survival,Organ,shapiro.test))
```

```
## $Breast
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.86857, p-value = 0.07431
##
##
## $Bronchus
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.76596, p-value = 0.0007186
##
##
## $Colon
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.76056, p-value = 0.0006134
##
##
## $Ovary
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.76688, p-value = 0.029
##
##
## $Stomach
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.75473, p-value = 0.002075
```

```
####################
#                  #
#    Exercise 5    #
```

```
#                   #
#####################
#Check for equality of variance
 library(car)
```

```
## Warning: package 'car' was built under R version 3.3.1
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
leveneTest(Survival~Organ, data = cancer.survival)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value   Pr(>F)
## group  4  4.4524 0.003271 **
##       59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

```
#####################
#                   #
#    Exercise 6     #
#                   #
#####################
#Apply a log transformation to survival time and check for
normality and equality of variance.
cancer.survival$log.survival = log(cancer.survival$Survival)
with(cancer.survival,tapply(log.survival,Organ,shapiro.test))
```

```
## $Breast
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.802, p-value = 0.009995
##
##
## $Bronchus
##
```

```
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.98047, p-value = 0.9613
##
##
## $Colon
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.92636, p-value = 0.1891
##
##
## $Ovary
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.983, p-value = 0.9655
##
##
## $Stomach
##
##  Shapiro-Wilk normality test
##
## data:  X[[i]]
## W = 0.92837, p-value = 0.3245
```

```
leveneTest(log.survival~Organ, data = cancer.survival)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4  0.6685 0.6164
##       59
```

```
####################
#                  #
#    Exercise 7    #
#                  #
####################
#Perform one way anova
aov1 = aov(log.survival~Organ,cancer.survival)
```

```
summary(aov1)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Organ        4  24.49   6.122   4.286 0.00412 **
## Residuals   59  84.27   1.428
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
```

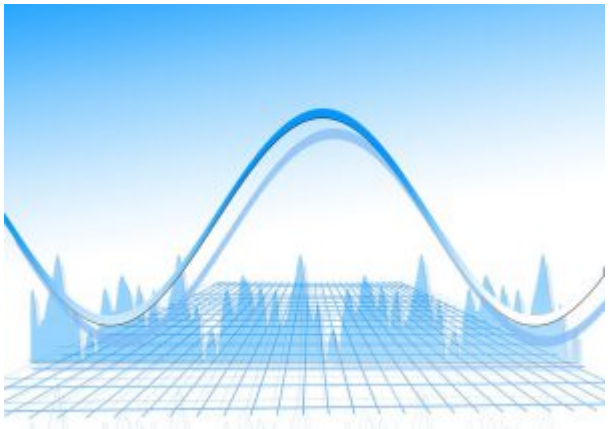
```
####################
#                  #
#    Exercise 8    #
#                  #
####################
#Perform a Tukey HSD comparison
TukeyHSD(aov1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log.survival ~ Organ, data =
cancer.survival)
##
## $Organ
##                          diff       lwr        upr     p adj
## Bronchus-Breast  -1.60543320 -2.906741 -0.3041254 0.0083352
## Colon-Breast     -0.80948110 -2.110789  0.4918267 0.4119156
## Ovary-Breast     -0.40798703 -2.114754  1.2987803 0.9615409
## Stomach-Breast   -1.59068365 -2.968399 -0.2129685 0.0158132
## Colon-Bronchus    0.79595210 -0.357534  1.9494382 0.3072938
## Ovary-Bronchus    1.19744617 -0.399483  2.7943753 0.2296079
## Stomach-Bronchus  0.01474955 -1.224293  1.2537924 0.9999997
## Ovary-Colon       0.40149407 -1.195435  1.9984232 0.9540004
## Stomach-Colon    -0.78120255 -2.020245  0.4578403 0.3981146
## Stomach-Ovary    -1.18269662 -2.842480  0.4770864 0.2763506
```

```
####################
#                  #
#    Exercise 9    #
#                  #
####################
#Interpret results
#our data showed departure from normality and equality of
```

```
variance. Perhaps unequal variance was due to our unbalanced
design (we had unequal samples in our groups)
#kruskal-wallis test would still not be appropriate because it
relies on equal variance
#a log transformation was useful in stabilizing variance.
#normality was violated in the breast group even after
transformation. Anova is robust to slight deviations from
normality
#differences between groups were statistically significant
#kruskal-wallis leads to same conclusion.

####################
#                  #
#    Exercise 10   #
#                  #
####################
#use a kruskal-wallis test
kruskal.test(log.survival~Organ,cancer.survival)

##
##  Kruskal-Wallis rank sum test
##
## data:  log.survival by Organ
## Kruskal-Wallis chi-squared = 14.954, df = 4, p-value =
0.004798
```

Paired t-test in R Exercises



The paired samples t test is used to check if there are any differences in the mean of the same sample at two different time points. For example a medical researcher collects data on the same patients before and after a therapy. A paired t test

will show if the therapy improves patient outcomes.

There are several assumptions that need to be satisfied so that results of a paired t test are valid. They are listed below

- The measured variable is continuous
- The differences between the two groups are approximately normally distributed
- We should not have any outliers in our data
- An adequate sample size is required

For this exercise we will use the anorexia data set available in package MASS. The data set contains weights of girls before and after anorexia treatment. Our interest is to know if the treatment caused any change in weight.

Solutions to these exercises can be found here

Exercise 1

Load the data and inspect its structure

Exercise 2

Generate descriptive statistics on weight before treatment

Exercise 3

Generate descriptive statistics on weight after treatment

Exercise 4

Create a new variable that contains the differences in weight before and after treatment

Exercise 5

Create a boxplot to identify any outliers

Exercise 6

Create a histogram with a normal curve to visually inspect normality

Exercise 7

Perform a normality test on the differences

Exercise 8

Perform a power analysis to assess sample adequacy

Exercise 9

Perform a paired t test

Exercise 10

Interpret the results

Paired t test in R Solutions

Solutions to exercises on paired t-test found here

```
####################
#                  #
#    Exercise 1    #
#                  #
####################

#Load package MASS
library(MASS)
#attach anorexia data so that variables are easily accessible
attach(anorexia)
#Inspect structure of the data
head(anorexia)

##   Treat Prewt Postwt
## 1  Cont  80.7   80.2
```

```
## 2  Cont  89.4   80.1
## 3  Cont  91.8   86.4
## 4  Cont  74.0   86.3
## 5  Cont  78.1   76.1
## 6  Cont  88.3   78.1
```

```
####################
#                  #
#    Exercise 2    #
#                  #
####################
```

```
#descriptive statistics on weight before treatment
#You need to install package psych

library(psych)
```

```
## Warning: package 'psych' was built under R version 3.3.1
```

```
describe(anorexia$Prewt)
```

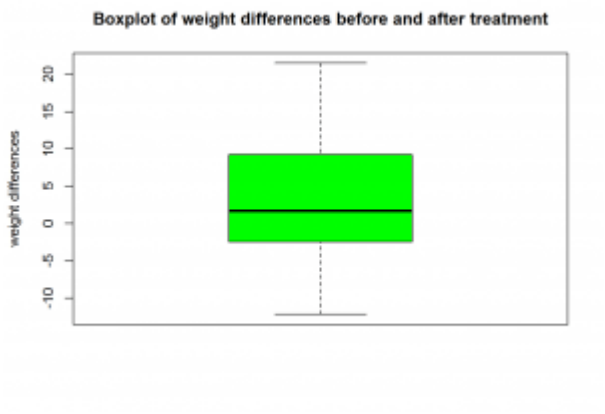
```
##    vars  n  mean   sd median trimmed  mad min  max range
skew kurtosis
## X1    1 72 82.41 5.18   82.3   82.47 5.49  70 94.9  24.9
-0.05    -0.16
##      se
## X1 0.61
```

```
####################
#                  #
#    Exercise 3    #
#                  #
####################
```

```
#descriptive statistics on weight after treatment

describe(anorexia$Postwt)
```

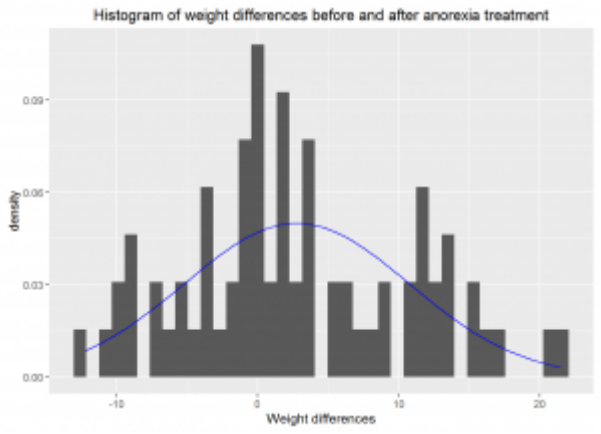
```
##    vars  n  mean   sd median trimmed  mad  min   max range
skew kurtosis
## X1    1 72 85.17 8.04  84.05   84.82 9.56 71.3 103.6  32.3
0.36    -0.81
##      se
## X1 0.95
```

```
####################
#                  #
#    Exercise 4    #
#                  #
####################


#create a new variable containing differences
weight.differences = Postwt - Prewt

####################
#                  #
#    Exercise 5    #
#                  #
####################

#create a boxplot to identify any outliers in our data

boxplot(weight.differences,main = "Boxplot of weight
differences before and after treatment",ylab = "weight
differences",col = "green")
```




```
####################
#                  #
#    Exercise 6    #
#                  #
####################
#Create a histogram to visually assess normality
#In exercise 5 we used base graphics to produce a boxplot
#A more flexible way of data visualization is using package
ggplot
#Install package ggplot if you have not
```

```
#load ggplot
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.1

##
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

#attach weight.differences to anorexia data frame
anorexia$weight.differences = weight.differences

#Create a histogram with a density curve to visually inspect
normality

ggplot(anorexia,aes(x=weight.differences))         +
geom_histogram(aes(y=..density..),binwidth  =  0.9)  +
stat_function(fun = dnorm, colour = "blue",args = list(mean =
mean(anorexia$weight.differences),        sd        =
sd(anorexia$weight.differences)))                    +
scale_x_continuous(name="Weight   differences")    +
ggtitle("Histogram of weight differences before and after
anorexia treatment")
```




```
####################
#                  #
#    Exercise 7    #
#                  #
####################
#Test if the weight differences are normally distributed
```

```
shapiro.test(weight.differences)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  weight.differences
## W = 0.97466, p-value = 0.1544
```

```
####################
#                  #
#    Exercise 8    #
#                  #
####################
#Perform a power analysis to check the sample size has adequate power to detect a difference if it exists
#install package pwr and load it
library(pwr)
```

```
## Warning: package 'pwr' was built under R version 3.3.1
```

```
pwr.t.test(n=72,d=0.5,sig.level = 0.05,type = c("paired"))
```

```
##
##      Paired t test power calculation
##
##               n = 72
##               d = 0.5
##       sig.level = 0.05
##           power = 0.9869471
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```
####################
#                  #
#    Exercise 9    #
#                  #
####################
#Perform a paired t test
t.test(Postwt,Prewt,paired = TRUE)
```

```
##
##  Paired t-test
##
```

```
## data:  Postwt and Prewt
## t = 2.9376, df = 71, p-value = 0.004458
## alternative hypothesis: true difference in means is not
equal to 0
## 95 percent confidence interval:
##  0.8878354 4.6399424
## sample estimates:
## mean of the differences
##                2.763889

####################
#                  #
#    Exercise 10   #
#                  #
####################
#Interpret results
#All assumptions required were satisfied
#There were no outliers, data was normally distributed and the
t test had adequate power
#The difference in weight before and after treatment was
statistically significant at 5% LOs.
```