LAB 07

TOPICS:

- One-way ANOVA and MANOVA
- Two-ways ANOVA and MANOVA

1.

```
load("mcshapiro.test.RData")

### ------------------------------------------------------------------------
### One-way ANOVA
### (p=1, g=6)
### ------------------------------------------------------------------------
# Dataset Chicken Weights:
# Data from an experiment to measure and compare the effectiveness
# of various feed supplements on the growth rate of chickens.
# (71 observations)

help(chickwts)
```

```
head(chickwts)
```

only one factor (one treatment) with 6 levels

```
##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

We have 71 observations of 2 variables measuring the growth of chickens

```
dim(chickwts)
```

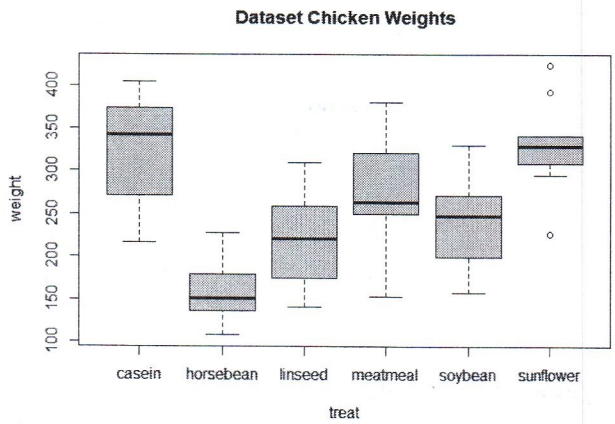
```
## [1] 71  2
```

```
summary(chickwts)
```

```
##      weight             feed
##  Min.   :108.0   casein   :12
##  1st Qu.:204.5   horsebean:10
##  Median :258.0   linseed  :12
##  Mean   :261.3   meatmeal :11
##  3rd Qu.:323.5   soybean  :14
##  Max.   :423.0   sunflower:12
```

note that we don't have the same number of observations for each level!

```
attach(chickwts)

x11()
plot(feed, weight, xlab='treat', ylab='weight', col='grey85', main='Dataset Chicken Weights')
```



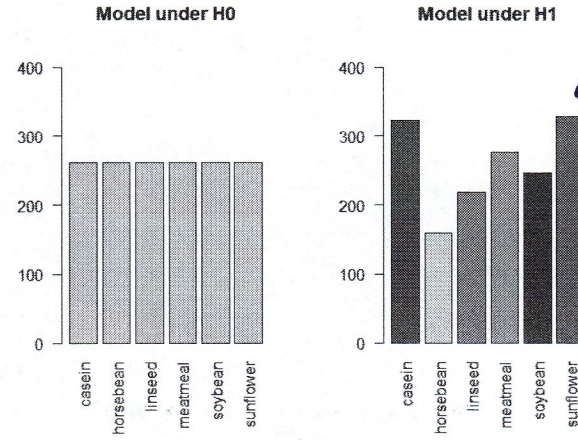

```
dev.off()
```

```
### Model: weigth.ij = mu + tau.i + eps.ij; eps.ij~N(0,sigma^2)
### Test:
### H0: tau.1 = tau.2 = tau.3 = tau.4 = tau.5 = tau.6 = 0
### H1: (H0)^c
### i.e.,
### H0: The feed supplements don't have effect
###       (= "chickens belong to a single population")
### H1: At least one feed supplement has effect
###       (= "chickens belong to 2, 3, 4, 5 or 6 populations")

x11()
par(mfrow=c(1,2))
barplot(rep(mean(weight),6), names.arg=levels(feed), ylim=c(0,max(weight)),
        las=2, col='grey85', main='Model under H0')
barplot(tapply(weight, feed, mean), names.arg=levels(feed), ylim=c(0,max(weight)),
        las=2, col=rainbow(6),main='Model under H1')
```

$$\text{Weight}_{(i,j)} = \mu + \tau_{(i)} + \varepsilon_{(i,j)}$$

random noise

$$\begin{cases} H_0: \ \tau_{(1)} = \dots = \tau_{(6)} = 0 \\ H_1: \ \exists j: \ \tau_{(j)} \neq 0 \quad j = 1, \dots, 6 \end{cases}$$



barplot with the height = $\mathbb{E}$[obs of the group]

```
dev.off()
```

```
# This is a case of one-way ANOVA: one variable (weight) observed
# over g=6 levels (feed)
n       <- length(feed)      # total number of obs.
ng      <- table(feed)       # number of obs. in each group
treat   <- levels(feed)      # Levels of the treatment
g       <- length(treat)     # number of Levels (i.e., of groups)

### verify the assumptions:
# 1) normality (univariate) in each group (6 tests)
Ps <- c(shapiro.test(weight[ feed==treat[1] ])$p,
        shapiro.test(weight[ feed==treat[2] ])$p,
        shapiro.test(weight[ feed==treat[3] ])$p,
        shapiro.test(weight[ feed==treat[4] ])$p,
        shapiro.test(weight[ feed==treat[5] ])$p,
        shapiro.test(weight[ feed==treat[6] ])$p)
Ps
```

1. gaussian distribution for the errors

we use "shapiro.test()" because it's the univariate case

```
## [1] 0.2591841 0.5264499 0.9034734 0.9611795 0.5063768 0.3602904
```

```
# 2) same covariance structure (= same sigma^2)
Var <- c(var(weight[ feed==treat[1] ]),
         var(weight[ feed==treat[2] ]),
         var(weight[ feed==treat[3] ]),
         var(weight[ feed==treat[4] ]),
         var(weight[ feed==treat[5] ]),
         var(weight[ feed==treat[6] ]))
Var
```

2. homoschedasticity

```
## [1] 4151.720 1491.956 2728.568 4212.091 2929.956 2384.992
```

```
# test of homogeneity of variances
# H0: sigma.1 = sigma.2 = sigma.3 = sigma.4 = sigma.5 = sigma.6
# H1: there exist i,j s.t. sigma.i!=sigma.j
bartlett.test(weight, feed)
```

(**notice** that this test relies on the gaussian assumptions)

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight and feed
## Bartlett's K-squared = 3.2597, df = 5, p-value = 0.66
```

```
### One-way ANOVA
### ------------------------------------------------------------------------
help(aov)

fit <- aov(weight ~ feed)

summary(fit)
```

: (output ~ treatment)

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### How to read the summary:
#               Df   Sum Sq      Mean Sq      F value   Pr(>F)
#  treat       (g-1) SStreat  SStreat/(g-1)  Fstatistic p-value [H0: tau.i=0 for every i]
#  Residuals   (n-g) SSres     SSres/(n-g)
### ------------------------------------------------------------------------
#  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
###

### We reject the test, i.e., we have evidence to state that the
### treatment (feed supplement) has an effect on the growth rate
### of chicken.
### Which supplement is responsible for this? To see this, we need to
### do g*(g-1)/2 comparisons.
### We use Bonferroni
k <- g*(g-1)/2
alpha= 0.05

Mediag  <- tapply(weight, feed, mean)
SSres <- sum(residuals(fit)^2)
S <- SSres/(n-g)

# Example: CI for the difference "casein - horsebean"
paste(treat[1],"-",treat[2])
```

since we do multiple comparisons

OVERALL level

we have to take all the possible couples of levels of treatment and see if there is significant difference

```
## [1] "casein - horsebean"
```

```
as.numeric(c(Mediag[1]-Mediag[2] - qt(1-alpha/(2*k), n-g) * sqrt( S * ( 1/ng[1] + 1/ng[2] )), Mediag[1]-Mediag[2] + qt(1-alp
ha/(2*k), n-g) * sqrt( S * ( 1/ng[1] + 1/ng[2] ))))
```

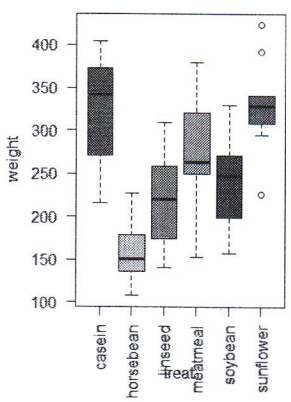
```
## [1]  91.81006 234.95661
```

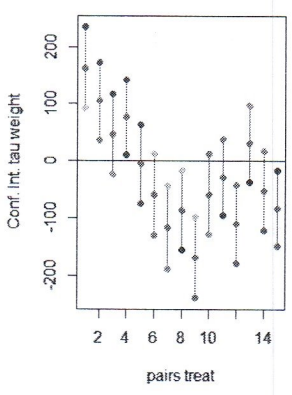
```
# CI for all the differences
ICrange=NULL
for(i in 1:(g-1)) {
  for(j in (i+1):g) {
    print(paste(treat[i],"-",treat[j]))
    print(as.numeric(c(Mediag[i]-Mediag[j] - qt(1-alpha/(2*k), n-g) * sqrt( S * ( 1/ng[i] + 1/ng[j] )),
                       Mediag[i]-Mediag[j] + qt(1-alpha/(2*k), n-g) * sqrt( S * ( 1/ng[i] + 1/ng[j] )))))
    ICrange=rbind(ICrange,as.numeric(c(Mediag[i]-Mediag[j] - qt(1-alpha/(2*k), n-g) * sqrt( S * ( 1/ng[i] + 1/ng[j] )),
                                       Mediag[i]-Mediag[j] + qt(1-alpha/(2*k), n-g) * sqrt( S * ( 1/ng[i] + 1/ng[j] )))))
  }}
```

```
## [1] "casein - horsebean"
## [1]  91.81006 234.95661
## [1] "casein - linseed"
## [1]  36.59089 173.07578
## [1] "casein - meatmeal"
## [1] -23.10193 116.45041
## [1] "casein - soybean"
## [1]  11.3947 142.9148
## [1] "casein - sunflower"
## [1] -73.57578  62.90911
## [1] "horsebean - linseed"
## [1] -130.12328   13.02328
## [1] "horsebean - meatmeal"
## [1] -189.7462  -43.6720
## [1] "horsebean - soybean"
## [1] -155.4390  -17.0181
## [1] "horsebean - sunflower"
## [1] -240.28994  -97.14339
## [1] "linseed - meatmeal"
## [1] -127.93526   11.61708
## [1] "linseed - soybean"
## [1] -93.43863  38.08149
## [1] "linseed - sunflower"
## [1] -178.40911  -41.92422
## [1] "meatmeal - soybean"
## [1] -36.86983  97.83087
## [1] "meatmeal - sunflower"
## [1] -121.7837   17.7686
## [1] "soybean - sunflower"
## [1] -148.24816  -16.72803
```

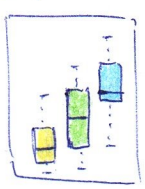
for example for this couple we found that the **CI** does not contain the 0 ⟹ there is a significant difference between these 2 groups

```
x11(width = 14, height = 7)
par(mfrow=c(1,2))
plot(feed, weight, xlab='treat', ylab='weight', col = rainbow(6), las=2)

h <- 1
plot(c(1,g*(g-1)/2),range(ICrange), pch='',xlab='pairs treat', ylab='Conf. Int. tau weight')
for(i in 1:(g-1)) {
  for(j in (i+1):g) {
    ind <- (i-1)*g-i*(i-1)/2+(j-i)
    lines (c(h,h), c(ICrange[ind,1],ICrange[ind,2]), col='grey55');
    points(h, Mediag[i]-Mediag[j], pch=16, col='grey55');
    points(h, ICrange[ind,1], col=rainbow(6)[j], pch=16);
    points(h, ICrange[ind,2], col=rainbow(6)[i], pch=16);
    h <- h+1
  }}
abline(h=0)
```

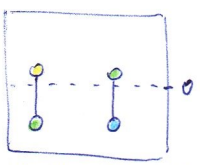




Boxplots:



CI for the differences:



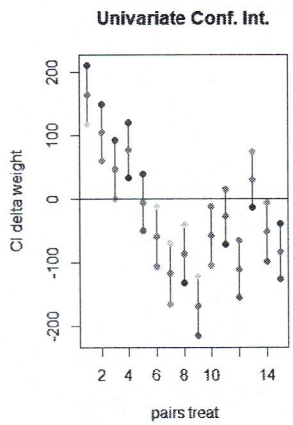
both the CI contains 0 (so there is not enough statistical evidence to say that they're ≠) but this do not imply that ▭ = ▭ !

```
### Be careful to apply transitivity in the statistical context!!
### {A not significantly (stat.) different from B} [A = B]   (red vs cyan)
###                          and
### {B not significantly (stat.) different from C} [B = C]   (cyan vs green)
###                    does NOT imply
### {A not significantly (stat.) different from C} [A = C]   (red vs green)
### Note. If we don't reject H0, we are not proving that A=B but
###       we are saying that we can't prove that A!=B

dev.off()
```

Here we're saying:
▭ and ▭ are not different enough to conclude that the means of the populations are different,
analogously ▭ and ▭.
But it can be that ▭ and ▭ are different enough to conclude that the means of the 2 populations are different.

What happen with another correction? Not Bonferroni but Benjamini?

```
# Let's change the criterion to control the univariate rejection
# (multiple testing)

# We build k one-at-a-time confidence intervals, each of level alpha
# (not including the Bonferroni correction)
Auni <- matrix(0,6,6)
for(i in 1:6) {
  for(j in i:6) {
    Auni[i,j] <- Mediag[i]-Mediag[j] + qt(1-alpha/2, n-g) * sqrt( S * ( 1/ng[i] + 1/ng[j] ) )}
  for(j in 1:i) {
    Auni[i,j] <- Mediag[j]-Mediag[i] - qt(1-alpha/2, n-g) * sqrt( S * ( 1/ng[i] + 1/ng[j] ) )}
  Auni[i,i]     <- 0
}

x11( width=14, height=7)
par(mfrow=c(1,2))
h <- 1
plot(c(1,g*(g-1)/2),range(Auni), pch='', xlab='pairs treat',
     ylab='CI delta weight', main='Univariate Conf. Int.', col='grey55')
for(i in 1:5) {
  for(j in (i+1):6) {lines (c(h,h), c(Auni[i,j],Auni[j,i]));
    points(h, Mediag[i]-Mediag[j], pch=16, col='grey55');
    points(h, Auni[i,j], col=rainbow(6)[i], pch=16);
    points(h, Auni[j,i], col=rainbow(6)[j], pch=16);
    h <- h+1
  }}
abline(h=0)
```



CI with no corrections (every interval has level α)

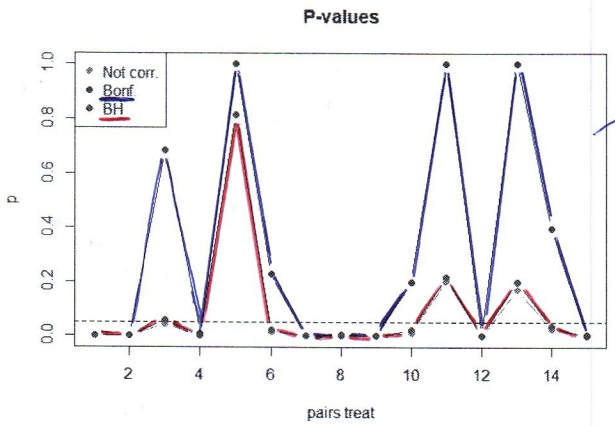
```
# We compute the p-values of the univariate tests

# Matrix of tests for the difference between all the pairs
P <- matrix(0,6,6)
for(i in 1:6) {
  for(j in i:6) {
    P[i,j] <- (1-pt(abs((Mediag[i]-Mediag[j]) / sqrt( S * ( 1/ng[i] + 1/ng[j] ) ) ), n-g))*2}
  for(j in 1:i) {
    P[i,j] <- (1-pt(abs((Mediag[i]-Mediag[j]) / sqrt( S * ( 1/ng[i] + 1/ng[j] ) ) ), n-g))*2}
  P[i,i]     <- 0
}
P
```

```
##              [,1]         [,2]         [,3]         [,4]         [,5]
## [1,] 0.000000e+00 2.067997e-09 1.493344e-05 4.556672e-02 0.0006654079
## [2,] 2.067997e-09 0.000000e+00 1.522197e-02 7.478012e-06 0.0003246269
## [3,] 1.493344e-05 1.522197e-02 0.000000e+00 1.347894e-02 0.2041446467
## [4,] 4.556672e-02 7.478012e-06 1.347894e-02 0.000000e+00 0.1725539145
## [5,] 6.654079e-04 3.246269e-04 2.041446e-01 1.725539e-01 0.0000000000
## [6,] 8.124949e-01 8.203778e-10 6.211836e-06 2.643548e-02 0.0002980438
##              [,6]
## [1,] 8.124949e-01
## [2,] 8.203778e-10
## [3,] 6.211836e-06
## [4,] 2.643548e-02
## [5,] 2.980438e-04
## [6,] 0.000000e+00
```

```
# Vector of p-values
p <- c(P[1, 2:6], P[2, 3:6], P[3, 4:6], P[4, 5:6], P[5, 6])
p
```

```
##  [1] 2.067997e-09 1.493344e-05 4.556672e-02 6.654079e-04 8.124949e-01
##  [6] 1.522197e-02 7.478012e-06 3.246269e-04 8.203778e-10 1.347894e-02
## [11] 2.041446e-01 6.211836e-06 1.725539e-01 2.643548e-02 2.980438e-04
```

```
plot(1:15, p, ylim=c(0,1), type='b', pch=16, col='grey55', xlab='pairs treat',
     main='P-values')
abline(h=alpha, lty=2)

# Bonferroni correction
p.bonf <- p.adjust(p, 'bonf')
lines(1:15, p.bonf, col='blue', pch=16, type='b')

# Correction according to the false discovery rate (Benjamini-Hockberg)
p.fdr <- p.adjust(p, 'fdr')
lines(1:15, p.fdr, col='red', pch=16, type='b')

legend('topleft', c('Not corr.', 'Bonf.', 'BH'), col=c('grey55', 'blue', 'red'), pch=16)
```



→ Bonferroni and BH have the same results but Bonferroni is more conservative

```
which(p.bonf<alpha)
```

```
## [1]  1  2  4  7  8  9 12 15
```

```
which(p.fdr<alpha)
```

```
##  [1]  1  2  4  6  7  8  9 10 12 14 15
```

```
graphics.off()
detach(chickwts)

###
### One-way MANOVA
### (p=4, g=3)
###

help(iris)
head(iris)
```

2.

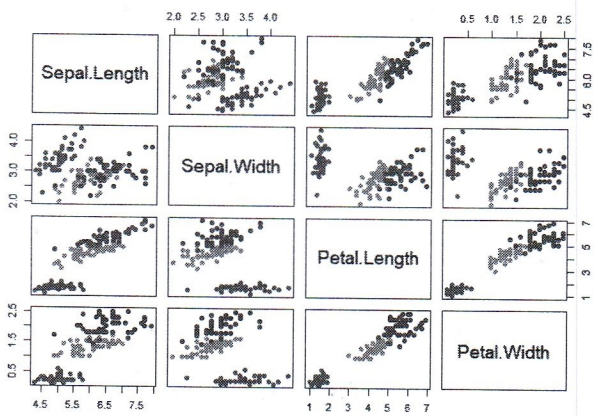
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
dim(iris)
```

```
## [1] 150   5
```

```
### Variables: Sepal and Petal Length and Sepal and Petal Width of iris
### (p = 4)
### Groups: species (setosa, versicolor, virginica; g = 3)
### n1 = n2 = n3 = 50 (balanced design)

attach(iris)

species.name <- factor(Species, labels=c('setosa','versicolor','virginica'))
iris4        <- iris[,1:4]

detach(iris)

### Data exploration
colore <- rep(rainbow(3), each = 50)

x11()
pairs(iris4, col = colore, pch=16)
```

← we consider the species as TREATMENT (g=3 ⇒ 3 levels)

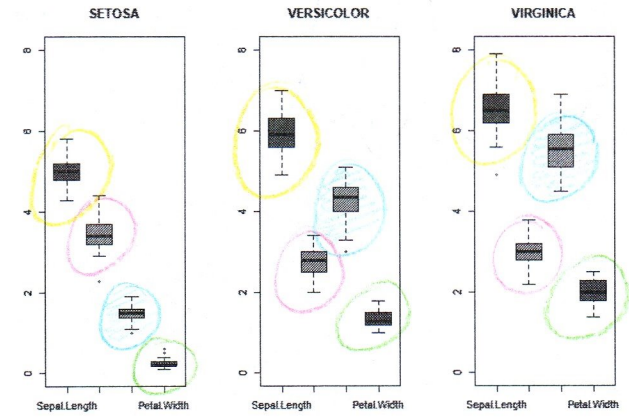



```
dev.off()
```

```
i1 <- which(species.name=='setosa')
i2 <- which(species.name=='versicolor')
i3 <- which(species.name=='virginica')

x11(width=13)
par(mfrow=c(1,3))
boxplot(iris4[i1,], main='SETOSA',     ylim=c(0,8), col = rainbow(4))
boxplot(iris4[i2,], main='VERSICOLOR', ylim=c(0,8), col = rainbow(4))
boxplot(iris4[i3,], main='VIRGINICA',  ylim=c(0,8), col = rainbow(4))
```

Different panels for different groups:

(in each panel the 4 dimensions (features))



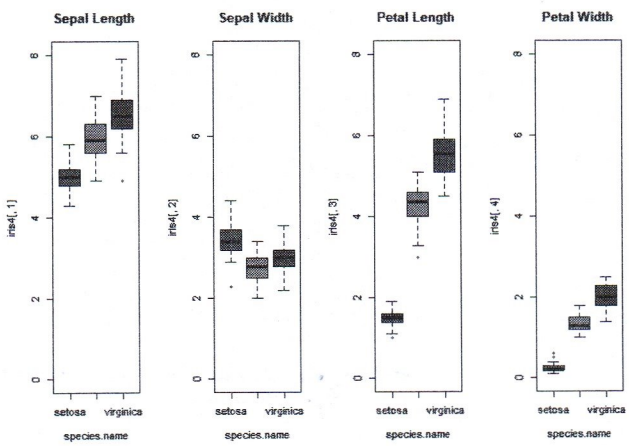
Here we see the effect of the treatment by comparing the different panels:

given the species these are the conditional distributions of sepal length, sepal width, ...

```
x11(width=13)
par(mfrow=c(1,4))
boxplot(iris4[,1]~species.name, main='Sepal Length', ylim=c(0,8), col = rainbow(3))
boxplot(iris4[,2]~species.name, main='Sepal Width', ylim=c(0,8), col = rainbow(3))
boxplot(iris4[,3]~species.name, main='Petal Length', ylim=c(0,8), col = rainbow(3))
boxplot(iris4[,4]~species.name, main='Petal Width', ylim=c(0,8), col = rainbow(3))
```

Different panels for different dimension (feature):

(in each panel the 3 groups)



given the feature these are the conditional distributions of setosa, versicolor, virginica

```
graphics.off()

### Model: X.ij = mu + tau.i + eps.ij; eps.ij~N_p(0,Sigma), X.ij, mu, tau.i in R^4
### Test:
### H0: tau.1 = tau.2 = tau.3  = (0,0,0)'
### H1: (H0)^c
### that is
### H0: The membership to an iris species hasn't any significant effect on the mean
###     of X.ij (in any direction of R^4)
### H1: There exists at least one direction in R^4 along which at least two species
###     have some feature significantly different

# This is a case of one-way MANOVA: four variables (Sepal.Length, Sepal.Width,
# Petal.Length, Petal.Width) observed over g=3 levels (setosa, versicolor, virginica)

n1 <- length(i1)
n2 <- length(i2)
n3 <- length(i3)
n  <- n1+n2+n3

g  <- length(levels(species.name))
p  <- 4

### Verify the assumptions:
# 1)  normality (multivariate) in each group (3 tests)
Ps <- NULL
for(i in 1:g)
  Ps <- c(Ps, mcshapiro.test(iris[get(paste('i',i, sep='')),1:4])$p)
Ps
```

```
## [1] 0.5660 0.1352 0.1616
```

(homoscedasticity)

```
# 2) same covariance structure (= same covariance matrix Sigma)
S  <-  cov(iris4)
S1 <-  cov(iris4[i1,])
S2 <-  cov(iris4[i2,])
S3 <-  cov(iris4[i3,])

# Qualitatively:
round(S1,digits=1)
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length          0.1         0.1            0           0
## Sepal.Width           0.1         0.1            0           0
## Petal.Length          0.0         0.0            0           0
## Petal.Width           0.0         0.0            0           0
```

```
round(S2,digits=1)
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length          0.3         0.1          0.2         0.1
## Sepal.Width           0.1         0.1          0.1         0.0
## Petal.Length          0.2         0.1          0.2         0.1
## Petal.Width           0.1         0.0          0.1         0.0
```

we check it "sentimentally"

```
round(S3,digits=1)
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length          0.4         0.1          0.3         0.0
## Sepal.Width           0.1         0.1          0.1         0.0
## Petal.Length          0.3         0.1          0.3         0.0
## Petal.Width           0.0         0.0          0.0         0.1
```

```
x11(width=21)
par(mfrow=c(1,3))
image(S1, col=heat.colors(100),main='Cov. S1', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3), (0:100)/100, na.rm=TR
UE))
image(S2, col=heat.colors(100),main='Cov. S2', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3), (0:100)/100, na.rm=TR
UE))
image(S3, col=heat.colors(100),main='Cov. S3', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3), (0:100)/100, na.rm=TR
UE))
```

Cov. S1 **Cov. S2** **Cov. S3**

→ it seems that the covariance matrices have different structure

(but we go on anyway)

```
dev.off()
```

```
# Note: We can verify the assumptions a posteriori on the residuals of
#       the estimated model

### One-way MANOVA
### ----------------------------------------------------------------------------

help(manova)
help(summary.manova)

fit <- manova(as.matrix(iris4) ~ species.name)
summary.manova(fit,test="Wilks")
```

```
##               Df    Wilks approx F num Df den Df    Pr(>F)
## species.name   2 0.023439   199.15      8    288 < 2.2e-16 ***
## Residuals    147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ almost 0

H_1: the treatment has effect (the species have effects)

```
# Exact tests for p<=2 or g<=3 already implemented in R

# Note: since g=3 the test is exact
#       (cfr. JW pag.300)

### Reject the test, i.e., we have statistical evidence to state that
### the factor "Species" has an effect on the mean features
### of the flowers.
### Who's the responsible for this?

### Via ANOVA: for each of the p=4 variables we perform an ANOVA test
###            to verify if the membership to a group has influence
###            on the mean of the variable (we explore separately the
###            4 axes directions in R^4)
summary.aov(fit)
```

(With "WILKS")

this will produce the one-way ANOVA on the 4 dimensions of the space

```
##  Response Sepal.Length :
##                Df Sum Sq Mean Sq F value    Pr(>F)
## species.name    2 63.212  31.606  119.26 < 2.2e-16 ***
## Residuals     147 38.956   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Sepal.Width :
##                Df Sum Sq Mean Sq F value    Pr(>F)
## species.name    2 11.345  5.6725   49.16 < 2.2e-16 ***
## Residuals     147 16.962  0.1154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Petal.Length :
##                Df Sum Sq Mean Sq F value    Pr(>F)
## species.name    2 437.10 218.551  1180.2 < 2.2e-16 ***
## Residuals     147  27.22   0.185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Petal.Width :
##                Df Sum Sq Mean Sq F value    Pr(>F)
## species.name    2 80.413  40.207  960.01 < 2.2e-16 ***
## Residuals     147  6.157   0.042
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Each of the 4 variables is significantly influenced by the
# factor species.
# Note: this analysis does NOT say:
#       a) which group differ
#       b) with respect to which variables the groups in (a) differ
# => As for the ANOVA, we build confidence intervals (many more!)

### Via Bonferroni
alpha <- 0.05
k <- p*g*(g-1)/2
qT <- qt(1-alpha/(2*k), n-g)

W <- summary.manova(fit)$SSS$Residuals
m  <- sapply(iris4,mean)         # estimates mu
m1 <- sapply(iris4[i1,],mean)    # estimates mu.1=mu+tau.1
m2 <- sapply(iris4[i2,],mean)    # estimates mu.2=mu+tau.2
m3 <- sapply(iris4[i3,],mean)    # estimates mu.3=mu+tau.3

inf12 <- m1-m2 - qT * sqrt( diag(W)/(n-g) * (1/n1+1/n2) )
sup12 <- m1-m2 + qT * sqrt( diag(W)/(n-g) * (1/n1+1/n2) )
inf13 <- m1-m3 - qT * sqrt( diag(W)/(n-g) * (1/n1+1/n3) )
sup13 <- m1-m3 + qT * sqrt( diag(W)/(n-g) * (1/n1+1/n3) )
inf23 <- m2-m3 - qT * sqrt( diag(W)/(n-g) * (1/n2+1/n3) )
sup23 <- m2-m3 + qT * sqrt( diag(W)/(n-g) * (1/n2+1/n3) )

CI <- list(setosa_versicolor=cbind(inf12, sup12), setosa_virginica=cbind(inf13, sup13), versicolor_virginica=cbind(inf23, su
p23))
CI
```

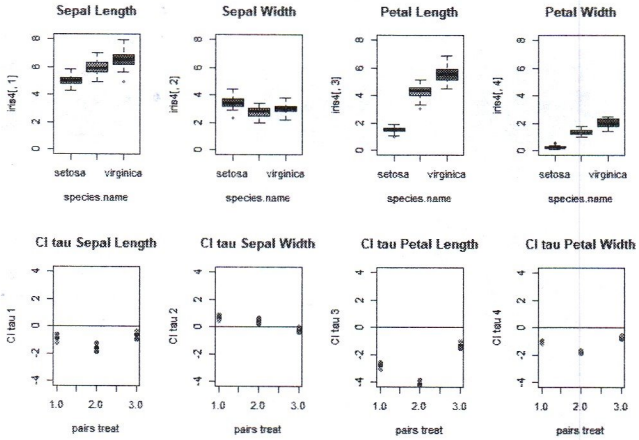
```
## $setosa_versicolor
##                    inf12      sup12
## Sepal.Length -1.2296895 -0.6303105
## Sepal.Width   0.4602476  0.8557524
## Petal.Length -3.0485232 -2.5474768
## Petal.Width  -1.1991389 -0.9608611
##
## $setosa_virginica
##                    inf13      sup13
## Sepal.Length -1.8816895 -1.2823105
## Sepal.Width   0.2562476  0.6517524
## Petal.Length -4.3405232 -3.8394768
## Petal.Width  -1.8991389 -1.6608611
##
## $versicolor_virginica
##                    inf23        sup23
## Sepal.Length -0.9516895 -0.352310544
## Sepal.Width  -0.4017524 -0.006247629
## Petal.Length -1.5425232 -1.041476789
## Petal.Width  -0.8191389 -0.580861086
```

```
# Now we have a complete frame (intervals for all the components of
# tau_1, tau_2 e tau_3): it is fault of all the groups and all the
# variables!

x11()
par(mfrow=c(2,4))
boxplot(iris4[,1]~species.name, main='Sepal Length', ylim=c(0,8), col = rainbow(3))
boxplot(iris4[,2]~species.name, main='Sepal Width', ylim=c(0,8), col = rainbow(3))
boxplot(iris4[,3]~species.name, main='Petal Length', ylim=c(0,8), col = rainbow(3))
boxplot(iris4[,4]~species.name, main='Petal Width', ylim=c(0,8), col = rainbow(3))

mg <- rbind(m1,m2,m3)
sp.name <- c('Sepal Length','Sepal Width', 'Petal Length', 'Petal Width')
for(k in 1:4){
  plot(c(1,g*(g-1)/2),ylim=c(-4,4), xlim=c(1,3), pch='',
       xlab='pairs treat', ylab=paste('CI tau',k),
       main=paste('CI tau',sp.name[k]))
  lines (c(1,1), c(CI[[1]][k,1],CI[[1]][k,2]));
  points(1, mg[1,k]-mg[2,k], pch=16);
  points(1, CI[[1]][k,1], col=rainbow(g)[2], pch=16);
  points(1, CI[[1]][k,2], col=rainbow(g)[1], pch=16);
  lines (c(2,2), c(CI[[2]][k,1],CI[[2]][k,2]));
  points(2, mg[1,k]-mg[3,k], pch=16);
  points(2, CI[[2]][k,1], col=rainbow(g)[3], pch=16);
  points(2, CI[[2]][k,2], col=rainbow(g)[1], pch=16);
  lines (c(3,3), c(CI[[3]][k,1],CI[[3]][k,2]));
  points(3, mg[2,k]-mg[3,k], pch=16);
  points(3, CI[[3]][k,1], col=rainbow(g)[3], pch=16);
  points(3, CI[[3]][k,2], col=rainbow(g)[2], pch=16);
  abline(h=0)
}
```

!! → we want to know the level (among species) that induces the difference (and on which dimension it induces it)

→ we have to compute a CI for each couple $\left(\frac{g(g-1)}{2}\right)$ and for each dimension (p)

Different panels for different feature



almost no one contains 0 (it means that every group differs on every feature)

the treatment has statistical effect for all levels and for all dimensions

```
dev.off()
```

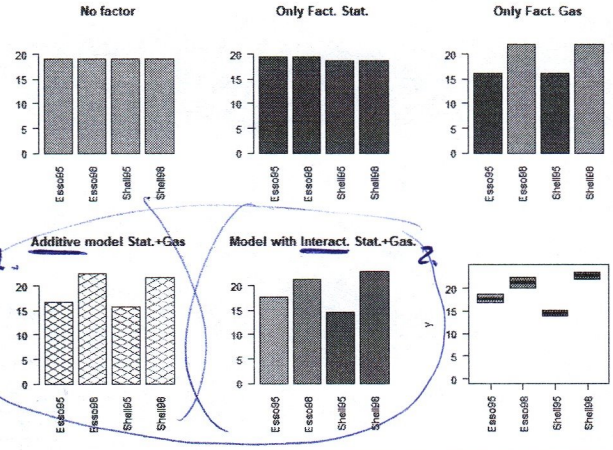
3. (2 treatments: each one with 2 factors)

```
###------------------------------------------------------------------------
### Two-ways ANOVA
### (p=1, g=2, b=2)
###------------------------------------------------------------------------

###------------------------------------------------------------------------
### Problem 4 of 14/09/06
###------------------------------------------------------------------------
# In a small village in Switzerland there are two gas stations:
# one of Essa and one of Shell. Both sell either gasoline 95 octanes
# and 98 octanes.
# A young statistician wants to find out which is the best gas station
# and the best gasoline to refuel his car, in order to maximize the
# number of kilometers covered with a single refueling.
# After 8 refuelings, the measured performances are:
# km/l  : (18.7, 16.8, 20.1, 22.4, 14.0, 15.2, 22.0, 23.3)
# distr.: ('Esso','Esso','Esso','Esso','Shell','Shell','Shell','Shell')
# benz. : ('95','95','98','98','95','95','98','98')
# (a) Via a two-ways ANOVA identify which is the best station and the
#     best gasoline for the young statistician to refuel his car.
# (b) Is there an interaction between the gas station and the gasoline?

### Variables: distance covered [km/l]
### factor1:   Gas station   (0=Esso, 1=Shell)
### factor2:   Gasoline      (0=95,   1=98)
### Balanced design

km          <- c(18.7, 16.8, 20.1, 22.4, 14.0, 15.2, 22.0, 23.3)
distr       <- factor(c('Esso','Esso','Esso','Esso','Shell','Shell','Shell','Shell'))
benz        <- factor(c('95','95','98','98','95','95','98','98'))
distr_benz  <- factor(c('Esso95','Esso95','Esso98','Esso98','Shell95','Shell95','Shell98','Shell98'))

g <- length(levels(distr))
b <- length(levels(benz))
n <- length(km)/(g*b)

M           <- mean(km)
Mdistr      <- tapply(km,      distr, mean)
Mbenz       <- tapply(km,       benz, mean)
Mdistr_benz <- tapply(km, distr_benz, mean)

x11()
par(mfrow=c(2,3),las=2)
barplot(rep(M,4), names.arg=levels(distr_benz), ylim=c(0,24), main='No factor')
barplot(rep(Mdistr,each=2), names.arg=levels(distr_benz), ylim=c(0,24),
        col=rep(c('blue','red'),each=2), main='Only Fact. Stat.')
barplot(rep(Mbenz,times=2), names.arg=levels(distr_benz), ylim=c(0,24),
        col=rep(c('darkgreen','orange'),times=2), main='Only Fact. Gas')
barplot(c(Mdistr[1]+Mbenz[1]-M, Mdistr[1]+Mbenz[2]-M, Mdistr[2]+Mbenz[1]-M,
          Mdistr[2]+Mbenz[2]-M), names.arg=levels(distr_benz), ylim=c(0,24),
        col=rep(c('darkgreen','orange'),times=2), density=rep(10,4), angle=135,
        main='Additive model Stat.+Gas')
barplot(c(Mdistr[1]+Mbenz[1]-M, Mdistr[1]+Mbenz[2]-M, Mdistr[2]+Mbenz[1]-M,
          Mdistr[2]+Mbenz[2]-M), names.arg=levels(distr_benz), ylim=c(0,24),
        col=rep(c('blue','red'),each=2), density=rep(10,4), add=T)
barplot(Mdistr_benz, names.arg=levels(distr_benz), ylim=c(0,24),
        col=rainbow(5)[2:5], main='Model with Interact. Stat.+Gas.')
plot(distr_benz, km, col=rainbow(5)[2:5], ylim=c(0,24),xlab='')
```

2 gas stations providing (each)
2 gasoline
8 observations (4 first gas station (2E, 2S)
4 second gas station (2E, 2S))

! ( ) ← just for graphical representation (it's not a one-way ANOVA, because we would lose the possibility of checking the interaction between the 2) !

Estimates of the means according to different models:



- no treatment - mean
- mean modified by treat 1
- mean modified by treat 2
- mean modified by treat 1 and 2 without interactions
- mean modified by treat 1 and 2 with interactions

1. ≠ 2.
since in 1. there is no interaction between factors

```
### Two-ways ANOVA
### ------------------------------------------------------------------------

### Model with interaction (complete model):
### X.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N(0,sigma^2),
###     i=1,2 (effect station), j=1,2 (effect gasoline)

fit.aov2.int <- aov(km ~ distr + benz + distr:benz)
summary.aov(fit.aov2.int)
```

← we start with the complete model (2 factors + interaction)

$$X_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

interaction

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## distr         1   1.53    1.53   1.018 0.37001
## benz          1  66.70   66.70  44.357 0.00264 **
## distr:benz    1  10.35   10.35   6.884 0.05857 .
## Residuals     4   6.01    1.50
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### Test:
### 1) H0: gamma.11 = gamma.12 = gamma.21 = gamma.22 = 0    vs   H1: (H0)^c
###    i.e.,
###    H0: There is no significant interaction between the factors station
###        and gasoline in terms of performances
###    H1: There exists a significant interaction between the factors station
###        and gasoline in terms of performances
###
### 2) H0: tau.1 = tau.2 = 0    vs   H1: (H0)^c
###    i.e.,
###    H0: The effect "gas station" doesn't significantly influence performances
###    H1: The effect "gas station" significantly influences performances
###
### 3) H0: beta.1 = beta.2 = 0    vs   H1: (H0)^c
###    i.e.,
###    H0: The effect "gasoline" doesn't significantly influence performances
###    H1: The effect "gasoline" significantly influences performances

# Test 1): Let's focus on the row of the summary distr:benz :
#              Df Sum Sq Mean Sq F value  Pr(>F)
# distr:benz    1  10.35   10.35   6.884 0.05857 .
# The P-value of test 1) is 0.05857. Reject at 10%, dont't reject at 1%,5% -> ?

# Test 2): Let's focus on the row of the summary distr:
#              Df Sum Sq Mean Sq F value  Pr(>F)
# distr         1   1.53    1.53   1.018 0.37001
# The P-value of test 2) is 0.37001. Don't reject at 10%, 5%, 1% -> not significant

# Test 3): Let's focus on the row of the summary benz:
#              Df Sum Sq Mean Sq F value  Pr(>F)
# benz          1  66.70   66.70  44.357 0.00264 **
# The P-value of test 3) is 0.00264. Reject at 10%, 5%, 1% -> significant

# Point b)
# From test 1): We don't have strong evidence that the interaction has effect
# => try to remove the interaction term and estimate the model without interaction

### Additive model:
### X.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N(0,sigma^2),
###     i=1,2 (effect station), j=1,2 (effect gasoline)
fit.aov2.ad <- aov(km ~ distr + benz)
summary.aov(fit.aov2.ad)
```

← $X_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## distr         1   1.53    1.53   0.468 0.52440
## benz          1  66.70   66.70  20.378 0.00632 **
## Residuals     5  16.37    3.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Remember: we have to remove the rows one-at-the-time because the p-values change every time we remove something

↓

we see that the first factor is not significant, so we remove it too (as second step)

```
# Remark: by removing the interaction, the residual degrees of freedom increase!

# Test: 2bis) H0: tau.1 = tau.2 = 0    vs   H1: (H0)^c
# From the summary:
#             Df Sum Sq Mean Sq F value  Pr(>F)
# distr        1   1.53    1.53   0.468 0.52440
# The P-value of test 2bis) is 0.52440. Don't reject at 10%, 5%, 1% -> not significant

# Test: 3bis) H0: beta.1 = beta.2 = 0    vs   H1: (H0)^c
# From the summary:
#             Df Sum Sq Mean Sq F value  Pr(>F)
# benz         1  66.70   66.70  20.378 0.00632 **
# The P-value of test 2bis) is 0.00632. Dont' reject at 10%, 5%, 1% -> significant

### Note: These aren't the only tests we can do!
### Example: global test for the significance of the two treatments
###          (model without interaction)
SSdistr <- sum(n*b*(Mdistr - M)^2)              # or from the summary: 1.53
SSbenz  <- sum(n*g*(Mbenz  - M)^2)              # or from the summary: 66.70
SSres   <- sum((km - M)^2) - (SSdistr+SSbenz)   # or from the summary: 16.37

Ftot      <- ( (SSdistr + SSbenz) / ((g-1)+(b-1)))/(SSres / (n*g*b-g-b+1))
Ptot      <- 1 - pf(Ftot, (g-1)+(b-1), n*g*b-g-b+1) # attention to the dgf!
Ptot
```

```
## [1] 0.01646126
```

```
# Test 2bis): there is no evidence that the factor "gas station" has
#             effect on the performances (don't reject at any reasonable
#             level [high p-value!])
# => we remove the variable "station" and reduce to a one-way ANOVA

### Reduced additive model (ANOVA one-way, b=2):
### X.jk = mu + beta.j + eps.jk; eps.jk~N(0,sigma^2),
###     j=1,2 (effect gasoline)
fit.aov1 <- aov(km ~ benz)
summary.aov(fit.aov1)
```

$X_{jk} = \mu + \beta_j + \varepsilon_{jk}$

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## benz         1   66.7   66.70   22.36 0.00323 **
## Residuals    6   17.9    2.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

⟹ gasoline influences the km/l

we actually chose the NO INTERACT. model

end of model selection

now we investigate how this differences are expressed (so we compute the CI's)

```
SSres <- sum(residuals(fit.aov1)^2)

### Interval at 90% for the differences (reduced additive model)
### [b=2, thus one interval only]
IC <- c(diff(Mbenz) - qt(0.95, (n*g-1)*b) * sqrt(SSres/((n*g-1)*b) *(1/(n*g) + 1/(n*g))),
        diff(Mbenz) + qt(0.95, (n*g-1)*b) * sqrt(SSres/((n*g-1)*b) *(1/(n*g) + 1/(n*g))))
names(IC) <- c('Inf', 'Sup')
IC    # IC for mu(98)-mu(95)
```

```
##      Inf      Sup
## 3.401886 8.148114
```

⟶ $0 \notin CI$ ⟹ there is significant difference between the 2 gasolines

```
### Note: we should have verified the hypotheses of normality and variance
###       homogeneity for the complete model, but with only 2 data for each
###       group we can't perform the tests.
### => we verify the assumptions on the reduced model (one-way ANOVA)
# 1) normality (univariate) in each groups (2 tests)
Ps <- c(shapiro.test(km[ benz==levels(benz)[1] ])$p,
        shapiro.test(km[ benz==levels(benz)[2] ])$p)
Ps
```

```
## [1] 0.9138954 0.6664017
```

```
# 2) homogeneity of variances
bartlett.test(km, benz)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  km and benz
## Bartlett's K-squared = 0.42658, df = 1, p-value = 0.5137
```

```
graphics.off()

### ------------------------------------------------------------------------
### Two-ways MANOVA
### (p=3, g=2, b=2)
### ------------------------------------------------------------------------

### ------------------------------------------------------------------------
### Example 6.13 JW (p.319)
### ------------------------------------------------------------------------
# The optimum conditions for extruding plastic films have been
# examined using a technique called Evolutionary Operation. In
# the course of the study, three responses were measured:
# X.1 = Tear Resistance
# X.2 = Gloss
# X.3 = Opacity
# at two levels of the factors:
# factor.1 = Rate of Extrusion     (0=Low Level, 1=High Level)
# factor.2 = Amount of an Additive (0=Low Level, 1=High Level)
# The measurements were repeated n=5 times at each combination
# of the factor levels.

plastic <- read.table('T6-4.dat',col.names=c('Ex','Ad','Tr','Gl','Op'))
plastic
```

4. (2 factors with 2 levels, 3 features) → MANOVA

2 factor
2 levels

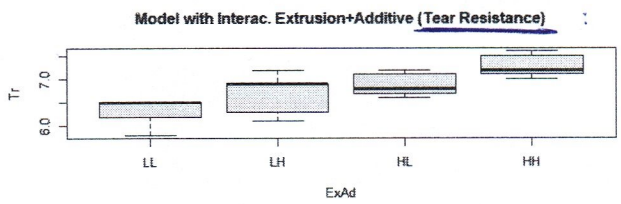
3 dimensional variable of interest
(that we want to study)

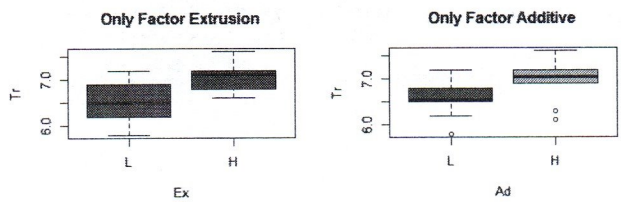
```
##    Ex Ad  Tr   Gl  Op
## 1   0  0 6.5  9.5 4.4
## 2   0  0 6.2  9.9 6.4
## 3   0  0 5.8  9.6 3.0
## 4   0  0 6.5  9.6 4.1
## 5   0  0 6.5  9.2 0.8
## 6   0  1 6.9  9.1 5.7
## 7   0  1 7.2 10.0 2.0
## 8   0  1 6.9  9.9 3.9
## 9   0  1 6.1  9.5 1.9
## 10  0  1 6.3  9.4 5.7
## 11  1  0 6.7  9.1 2.8
## 12  1  0 6.6  9.3 4.1
## 13  1  0 7.2  8.3 3.8
## 14  1  0 7.1  8.4 1.6
## 15  1  0 6.8  8.5 3.4
## 16  1  1 7.1  9.2 8.4
## 17  1  1 7.0  8.8 5.2
## 18  1  1 7.2  9.7 6.9
## 19  1  1 7.5 10.1 2.7
## 20  1  1 7.6  9.2 1.9
```

```
Ex   <- factor(plastic$Ex, labels=c('L','H')) # Treat.1
Ad   <- factor(plastic$Ad, labels=c('L','H')) # Treat.2

ExAd <- Ex
levels(ExAd) <- c('LL','LH','HL','HH')
ExAd[Ex=='L' & Ad=='L'] <- 'LL'
ExAd[Ex=='L' & Ad=='H'] <- 'LH'
ExAd[Ex=='H' & Ad=='L'] <- 'HL'
ExAd[Ex=='H' & Ad=='H'] <- 'HH'

plastic3  <- plastic[,3:5]

### Graphical exploration of the data
# effect of the treatments + their interaction on the first variable
x11()
layout(matrix(c(1,1,2,3), 2, byrow=T))
boxplot(plastic3[,1]~ExAd, main='Model with Interac. Extrusion+Additive (Tear Resistance)', ylab='Tr', col='grey95')
boxplot(plastic3[,1]~Ex,   main='Only Factor Extrusion'  , ylab='Tr', col=c('red','blue'))
boxplot(plastic3[,1]~Ad,   main='Only Factor Additive'   , ylab='Tr', col=c('forestgreen','gold'))
```

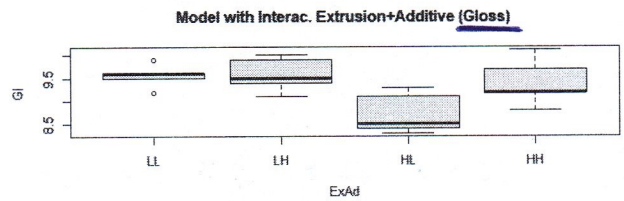
just for exploration (graphically)

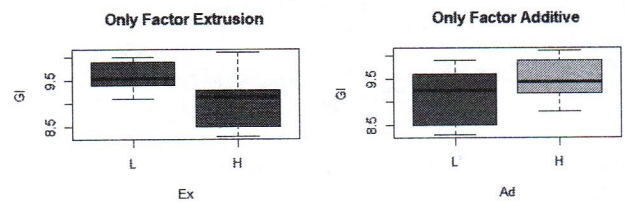
(3 sets of plots since the dimension of the variable of interest is 3)



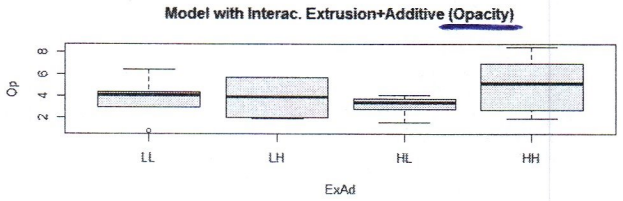


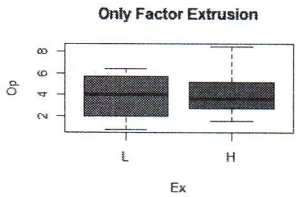

```
# effect of the treatments + their interaction on the second variable
x11()
layout(matrix(c(1,1,2,3), 2, byrow=T))
boxplot(plastic3[,2]~ExAd, main='Model with Interac. Extrusion+Additive (Gloss)', ylab='Gl', col='grey95')
boxplot(plastic3[,2]~Ex,   main='Only Factor Extrusion'  , ylab='Gl', col=c('red','blue'))
boxplot(plastic3[,2]~Ad,   main='Only Factor Additive'   , ylab='Gl', col=c('forestgreen','gold'))
```

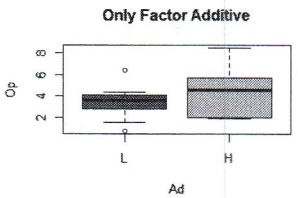





```
# effect of the treatments + their interaction on the third variable
x11()
layout(matrix(c(1,1,2,3), 2, byrow=T))
boxplot(plastic3[,3]~ExAd, main='Model with Interac. Extrusion+Additive (Opacity)', ylab='Op', col='grey95')
boxplot(plastic3[,3]~Ex,   main='Only Factor Extrusion'  , ylab='Op', col=c('red','blue'))
boxplot(plastic3[,3]~Ad,   main='Only Factor Additive'   , ylab='Op', col=c('forestgreen','gold'))
```








```
dev.off()
```

we always start with the complete (interact.) model:

$$X_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

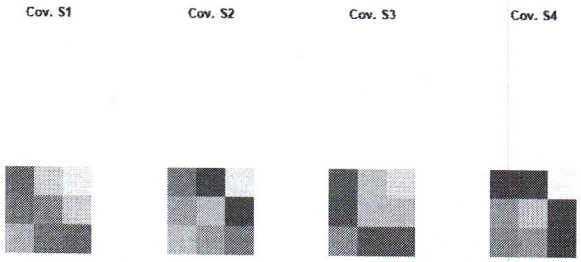
```
### Model with interaction (complete model):
### X.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N_p(0,Sigma), [p=3]
###     i=1,2 (effect Extrusion), j=1,2 (effect Additive),
###     X.ijs, mu, tau.i, beta.j, gamma.ij in R^3

### Verify the assumptions (although we only have 5 data in each group!)
# 1) normality (multivariate) in each group (4 test)
Ps <- c(mcshapiro.test(plastic3[ ExAd==levels(ExAd)[1], ])$p,
        mcshapiro.test(plastic3[ ExAd==levels(ExAd)[2], ])$p,
        mcshapiro.test(plastic3[ ExAd==levels(ExAd)[3], ])$p,
        mcshapiro.test(plastic3[ ExAd==levels(ExAd)[4], ])$p)
Ps
```

```
## [1] 0.7344 0.7652 0.0412 0.0868
```

→ we consider it verified

```
# 2) homogeneity of the covariance (qualitatively)
S1 <-  cov(plastic3[ ExAd==levels(ExAd)[1], ])
S2 <-  cov(plastic3[ ExAd==levels(ExAd)[2], ])
S3 <-  cov(plastic3[ ExAd==levels(ExAd)[3], ])
S4 <-  cov(plastic3[ ExAd==levels(ExAd)[4], ])

x11(width=21)
par(mfrow=c(1,4))
image(S1, col=heat.colors(100),main='Cov. S1', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3,S4), (0:100)/100, na.rm
=TRUE))
image(S2, col=heat.colors(100),main='Cov. S2', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3,S4), (0:100)/100, na.rm
=TRUE))
image(S3, col=heat.colors(100),main='Cov. S3', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3,S4), (0:100)/100, na.rm
=TRUE))
image(S4, col=heat.colors(100),main='Cov. S4', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3,S4), (0:100)/100, na.rm
=TRUE))
```

(only a graphical tool)




```
dev.off()
```

```
### Two-ways MANOVA
### ------------------------------------------------------------------------
### Model with interaction (complete model):
### X.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N_p(0,Sigma), [p=3]
###     i=1,2 (effect Extrusion), j=1,2 (effect Additive),
###     X.ijs, mu, tau.i, beta.j, gamma.ij in R^3
fit <- manova( as.matrix(plastic3) ~ Ex + Ad + Ex:Ad)
summary.manova(fit, test="Wilks")
```

```
##           Df   Wilks approx F num Df den Df   Pr(>F)
## Ex         1 0.38186   7.5543      3     14 0.003034 **
## Ad         1 0.52303   4.2556      3     14 0.024745 *
## Ex:Ad      1 0.77711   1.3385      3     14 0.301782
## Residuals 16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ we drop the interactions

```
### Model without interaction (additive model):
### X.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N_p(0,Sigma), [p=3]
###     i=1,2 (effect Extrusion), j=1,2 (effect additive),
###     X.ijs, mu, tau.i, beta.j, in R^3
fit2<- manova( as.matrix(plastic3) ~ Ex + Ad)
summary.manova(fit2, test="Wilks")
```

```
##           Df   Wilks approx F num Df den Df  Pr(>F)
## Ex         1 0.38684   7.9253      3     15 0.00212 **
## Ad         1 0.55384   4.0279      3     15 0.02753 *
## Residuals 17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Both the treatments have a significant effect on the mean (but not
# their interaction, that we could remove)

# Let's verify if this is true for all the variables through appropriate
# conf. int. and tests on the components:

# ANOVA on the components (we look at the 3 axes-directions in R^3
#                          separately)
summary.aov(fit2)
```

```
##  Response Tr :
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Ex           1 1.7405 1.74050  16.769 0.0007549 ***
## Ad           1 0.7605 0.76050   7.327 0.0149597 *
## Residuals   17 1.7645 0.10379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Gl :
##             Df Sum Sq Mean Sq F value  Pr(>F)
## Ex           1 1.3005 1.30050  6.9688 0.01720 *
## Ad           1 0.6125 0.61250  3.2821 0.08774 .
## Residuals   17 3.1725 0.18662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response Op :
##             Df Sum Sq Mean Sq F value Pr(>F)
## Ex           1  0.421  0.4205  0.1038 0.7513
## Ad           1  4.901  4.9005  1.2094 0.2868
## Residuals   17 68.885  4.0520
```

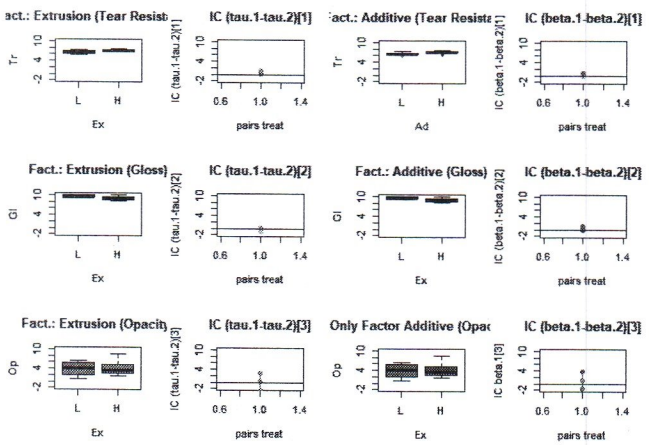
```
# Bonferroni
alpha <- 0.05
g <- 2
b <- 2
p <- 3
n <- 5
N <- n*g*b # 20

W <- summary.manova(fit2)$SS$Residuals

# how many comparisons?
k <- g*(g-1)/2*p + b*(b-1)/2*p
# because we have: g levels on the first treatment on p components
#                  b levels on the second treatment on p components
k
```

```
## [1] 6
```

```
qT <- qt(1 - alpha / (2 * k), g*b*n-g-b+1)
# the degrees of freedon of the residuals on the additive model are
# g*b*n-g-b+1

mExL  <- sapply(plastic3[Ex=='L',],mean)
mExH  <- sapply(plastic3[Ex=='H',],mean)
infEx <- mExH-mExL - qT * sqrt( diag(W)/(g*b*n-g-b+1) * (1/10+1/10) )
supEx <- mExH-mExL + qT * sqrt( diag(W)/(g*b*n-g-b+1) * (1/10+1/10) )

mAdL  <- sapply(plastic3[Ad=='L',],mean)
mAdH  <- sapply(plastic3[Ad=='H',],mean)
infAd <- mAdH-mAdL - qT * sqrt( diag(W)/(g*b*n-g-b+1) * (1/10+1/10) )
supAd <- mAdH-mAdL + qT * sqrt( diag(W)/(g*b*n-g-b+1) * (1/10+1/10) )

IC2   <- list(ExH_ExL=cbind(infEx, supEx), AdH_AdL=cbind(infAd, supAd))
IC2
```

```
## $ExH_ExL
##         infEx      supEx
## Tr  0.1600616 1.01993841
## Gl -1.0864959 0.06649592
## Op -2.3963103 2.97631028
##
## $AdH_AdL
##          infAd     supAd
## Tr -0.03993841 0.8199384
## Gl -0.22649592 0.9264959
## Op -1.69631028 3.6763103
```

```
x11(width=21, height = 14)
par(mfrow=c(3,4))
boxplot(plastic3[,1]~Ex,   main='Fact.: Extrusion (Tear Resistance)'  , ylab='Tr', col=rainbow(2*6)[c(1,2)], ylim=c(-2,10))
plot(c(1,g*(g-1)/2),range(IC2[[1]][1,]), pch='',main='IC (tau.1-tau.2)[1]',xlab='pairs treat', ylab='IC (tau.1-tau.2)[1]', y
lim=c(-2,10))
lines(c(1,1), c(IC2[[1]][1,1],IC2[[1]][1,2]), col='grey55');
points(1, (mExH-mExL)[1], pch=16, col='grey55');
points(1, IC2[[1]][1,1], col=rainbow(2*6)[1], pch=16);
points(1, IC2[[1]][1,2], col=rainbow(2*6)[2], pch=16);
abline(h=0)

boxplot(plastic3[,1]~Ad,   main='Fact.: Additive (Tear Resistance)'  , ylab='Tr', col=rainbow(2*6)[c(7,8)], ylim=c(-2,10))
plot(c(1,g*(g-1)/2),range(IC2[[2]][1,]), pch='',main='IC (beta.1-beta.2)[1]',xlab='pairs treat', ylab='IC (beta.1-beta.2)
[1]', ylim=c(-2,10))
lines(c(1,1), c(IC2[[2]][1,1],IC2[[2]][1,2]), col='grey55');
points(1, (mAdH-mAdL)[1], pch=16, col='grey55');
points(1, IC2[[2]][1,1], col=rainbow(2*6)[7], pch=16);
points(1, IC2[[2]][1,2], col=rainbow(2*6)[8], pch=16);
abline(h=0)

boxplot(plastic3[,2]~Ex,   main='Fact.: Extrusion (Gloss)'  , ylab='Gl', col=rainbow(2*6)[c(3,4)], ylim=c(-2,10))
plot(c(1,g*(g-1)/2),range(IC2[[1]][2,]), pch='',main='IC (tau.1-tau.2)[2]',xlab='pairs treat', ylab='IC (tau.1-tau.2)[2]', y
lim=c(-2,10))
lines(c(1,1), c(IC2[[1]][2,1],IC2[[1]][2,2]), col='grey55');
points(1, (mExH-mExL)[2], pch=16, col='grey55');
points(1, IC2[[1]][2,1], col=rainbow(2*6)[3], pch=16);
points(1, IC2[[1]][2,2], col=rainbow(2*6)[4], pch=16);
abline(h=0)

boxplot(plastic3[,2]~Ex,   main='Fact.: Additive (Gloss)'  , ylab='Gl', col=rainbow(2*6)[c(9,10)], ylim=c(-2,10))
plot(c(1,g*(g-1)/2),range(IC2[[2]][2,]), pch='',main='IC (beta.1-beta.2)[2]',xlab='pairs treat', ylab='IC (beta.1-beta.2)
[2]', ylim=c(-2,10))
lines(c(1,1), c(IC2[[2]][2,1],IC2[[2]][2,2]), col='grey55');
points(1, (mAdH-mAdL)[2], pch=16, col='grey55');
points(1, IC2[[2]][2,1], col=rainbow(2*6)[9], pch=16);
points(1, IC2[[2]][2,2], col=rainbow(2*6)[10], pch=16);
abline(h=0)

boxplot(plastic3[,3]~Ex,   main='Fact.: Extrusion (Opacity)'  , ylab='Op', col=rainbow(2*6)[c(5,6)], ylim=c(-2,10))
plot(c(1,g*(g-1)/2),range(IC2[[1]][3,]), pch='',main='IC (tau.1-tau.2)[3]',xlab='pairs treat', ylab='IC (tau.1-tau.2)[3]', y
lim=c(-2,10))
lines(c(1,1), c(IC2[[1]][3,1],IC2[[1]][3,2]), col='grey55');
points(1, (mExH-mExL)[3], pch=16, col='grey55');
points(1, IC2[[1]][3,1], col=rainbow(2*6)[5], pch=16);
points(1, IC2[[1]][3,2], col=rainbow(2*6)[6], pch=16);
abline(h=0)

boxplot(plastic3[,3]~Ex,   main='Only Factor Additive (Opacity)'  , ylab='Op', col=rainbow(2*6)[c(11,12)], ylim=c(-2,10))
plot(c(1,g*(g-1)/2),range(IC2[[2]][3,]), pch='',main='IC (beta.1-beta.2)[3]',xlab='pairs treat', ylab='IC beta.1[3]', ylim=c
(-2,10))
lines(c(1,1), c(IC2[[2]][3,1],IC2[[2]][3,2]), col='grey55');
points(1, (mAdH-mAdL)[3], pch=16, col='grey55');
points(1, IC2[[2]][3,1], col=rainbow(2*6)[11], pch=16);
points(1, IC2[[2]][3,2], col=rainbow(2*6)[12], pch=16);
abline(h=0)
```




```
dev.off()
```

```
### ------------------------------------------------------------------------
### Problem 3 of 18/02/09
### ------------------------------------------------------------------------
# During the austral summer three species of penguins nest in the Antarctic
# Peninsula: Chinstrap, Adelie and Gentoo.
# Some biologists of the Artowski basis measured the weight [kg] of 90 adults:
# 15 males and 15 females for each of the three species (penguins.txt file).
# a) By using an ANOVA model with two factors, claim if gender and / or species
#    belonging significantly affect the weight.
# b) Using an appropriate model (possibly reduced), provide estimates (global
#    90% confidence) of means and variances of the groups identified at point
#    (a).

penguins <- read.table('penguins.txt', header=T)
head(penguins)
```

```
##   weight   species gender
## 1   3.45 Chinstrap      M
## 2   4.82 Chinstrap      M
## 3   4.06 Chinstrap      M
## 4   5.12 Chinstrap      M
## 5   4.54 Chinstrap      M
## 6   4.74 Chinstrap      M
```

```r
attach(penguins)
### question a)
# Model with interaction (complete model):
# X.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N(0,sigma^2),
#     i=1,2 (effect gender), j=1,2,3 (effect species)
fit1 <- aov(weight ~ gender + species + species:gender, penguins)
summary(fit1)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## gender          1   0.14    0.14   0.573  0.451
## species         2  92.98   46.49 191.310 <2e-16 ***
## gender:species  2   0.41    0.21   0.849  0.432
## Residuals      84  20.41    0.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Model without interaction (additive model):
# X.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N(0,sigma^2),
#     i=1,2 (effect gender), j=1,2,3 (effect species)
fit2 <- aov(weight ~ gender + species, penguins)
summary(fit2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## gender        1   0.14    0.14   0.575   0.45
## species       2  92.98   46.49 191.986 <2e-16 ***
## Residuals    86  20.83    0.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Reduced model (one-way):
# X.jk = mu + beta.j + eps.jk; eps.jk~N(0,sigma^2),
#     j=1,2,3 (effect species)
fit3 <- aov(weight ~ species, penguins)
summary(fit3)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## species       2  92.98   46.49   192.9 <2e-16 ***
## Residuals    87  20.97    0.24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# verify assumptions (on the last model):
# 1) normality
Ps <- c(shapiro.test(weight[ species==levels(species)[1] ])$p,
        shapiro.test(weight[ species==levels(species)[2] ])$p,
        shapiro.test(weight[ species==levels(species)[3] ])$p)
Ps
```

```
## [1] 0.4721578 0.3456878 0.5124616
```

```r
# 2) homogeneity of variances
bartlett.test(weight, species)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  weight and species
## Bartlett's K-squared = 6.1048, df = 2, p-value = 0.04724
```

```r
### question b)
DF <- fit3$df # n*g*b - 1 - (g-1) = 15*3*2-1-2 = 90-3 = 87
Spooled <- sum(fit3$res^2)/DF

means <- as.vector(tapply(penguins$weight, penguins$species, mean))
names(means) <- levels(species)
means
```

```
##    Adelia Chinstrap    Gentoo
##  4.976667  4.529667  6.874333
```

```r
alpha <- 0.10
k     <- 4 # g + 1 = 4 (g Conf Int for the means and 1 for the variance)
BF    <- rbind(cbind(means - sqrt(Spooled / 30) * qt(1 - alpha / (2*k), DF),
                     means + sqrt(Spooled / 30) * qt(1 - alpha / (2*k), DF)),
               c(Spooled * DF / qchisq(1 - alpha / (2*k), DF),
                 Spooled * DF / qchisq(alpha / (2*k), DF)))
rownames(BF)[4] <- 'Var.'
BF
```

```
##               [,1]      [,2]
## Adelia    4.7722444 5.1810889
## Chinstrap 4.3252444 4.7340889
## Gentoo    6.6699111 7.0787556
## Var.      0.1758739 0.3485624
```

```r
detach(penguins)
```

LAB 07 - Additional Exercices

```
load("mcshapiro.test.RData")
### ------------------------------------------------------------------------------
### Problem 1 of 09/07/08
### (questions b and c)
### ------------------------------------------------------------------------------
# The client.txt dataset contains data on 150 customers of
# PoliBank. For each customer we are given age [years], money invested
# at low risk [thousands of euros] (safemoney variable) and money invested
# at high risk [thousands of euros] (riskymoney variable).
# [a) Using only the variable age, cluster customers in three groups and
#     describe them in terms of age. Use a hierarchical agglomerative
#     algorithm based on Euclidean distance and single linkage. Report
#     cophenetic coefficient and the size of the clusters.
#  b) Introducing the appropriate assumptions about distributions of the
#     variables safemoney and riskymoney within the three groups, perform
#     a MANOVA to see if there is statistical evidence of a difference
#     in the joint distributions of safemoney and riskymoney variables
#     in the three groups.
#  c) Comment the result of MANOVA by means of suitable Bonferroni
#     intervals with global confidence 90%.
### ------------------------------------------------------------------------------

client <- read.table('client_class.txt', header=T)  # Clustering already performed at
                                                    # point a), labels inserted as 1st
                                                    # column
head(client)
```

```
##     age safemoney riskymoney
## 1 adult    15.871     19.312
## 2 adult    15.307     21.285
## 3 young    26.618     12.123
## 4 young    33.805      7.237
## 5   old    29.086     11.775
## 6 adult    17.635     22.633
```

```
dim(client)
```

```
## [1] 150   3
```

```
# prepare data
var.risp <- client[,2:3]
group.names <- client[,1]
dim(var.risp)[2]
```

```
## [1] 2
```

```
levels(group.names)
```

```
## [1] "adult" "old"   "young"
```

```
# variables:
p <- 2
g <- 3
i1 <- which(group.names=='young')
i2 <- which(group.names=='adult')
i3 <- which(group.names=='old')
ng <- c(length(i1),length(i2),length(i3))
ng
```

```
## [1] 50 50 50
```

```
N <- sum(ng)

### question b)
### One-way MANOVA:
### Model: X.ij = mu + tau.i + eps.ij; eps.ij~N_p(0,Sigma), [p=2]
###       X.ij, mu, tau.i in R^2, i=1,2,3

# verify assumptions
# 1) normality
Ps <- c(mcshapiro.test(var.risp[ i1, ])$p,
        mcshapiro.test(var.risp[ i2, ])$p,
        mcshapiro.test(var.risp[ i3, ])$p)
Ps
```

```
## [1] 0.4068 0.5700 0.6716
```

```
# 2) homogeneity in variance
S1 <-  cov(var.risp[ i1, ])
S2 <-  cov(var.risp[ i2, ])
S3 <-  cov(var.risp[ i3, ])

x11()
par(mfrow=c(1,3))
image(S1, col=heat.colors(100),main='Cov. S1', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3), (0:100)/100, na.rm=TR
UE))
image(S2, col=heat.colors(100),main='Cov. S2', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3), (0:100)/100, na.rm=TR
UE))
image(S3, col=heat.colors(100),main='Cov. S3', asp=1, axes = FALSE, breaks = quantile(rbind(S1,S2,S3), (0:100)/100, na.rm=TR
UE))
```

Cov. S1 Cov. S2 Cov. S3

```
dev.off()
```

```
# Fit the model:
fit <- manova(as.matrix(var.risp) ~ group.names)
summary.manova(fit,test="Wilks")
```

```
##              Df   Wilks approx F num Df den Df    Pr(>F)
## group.names   2 0.20711   87.408      4    292 < 2.2e-16 ***
## Residuals   147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# who's the responsible?
summary.aov(fit,test="Wilks")
```

```
##  Response safemoney :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## group.names   2 3690.7 1845.33  219.27 < 2.2e-16 ***
## Residuals   147 1237.1    8.42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response riskymoney :
##              Df Sum Sq Mean Sq F value    Pr(>F)
## group.names   2 3184.9 1592.46  186.71 < 2.2e-16 ***
## Residuals   147 1253.7    8.53
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# all the variables

# Let's see if there is difference in the levels of the treatment -> question c)

### question c)
alpha <- 0.10
k <- p*g*(g-1)/2
k
```

```
## [1] 6
```

```
qT <- qt(1-alpha/(2*k), N-g)

# I need the diagonal of W
head(fit$res)   # residuals of the estimated model
```

```
##   safemoney riskymoney
## 1  -3.56588   -0.95882
## 2  -4.12988    1.01418
## 3  -3.62096    2.19840
## 4   3.56604   -2.68760
## 5  -0.56920    0.58426
## 6  -1.80188    2.36218
```

```
W <- diag(t(fit$res) %*% fit$res)/(N-g)
W
```

```
##  safemoney riskymoney
##   8.415747   8.528896
```

```
# mean within the groups
m1 <- colMeans(var.risp[i1,])
m2 <- colMeans(var.risp[i2,])
m3 <- colMeans(var.risp[i3,])
m1
```

```
##  safemoney riskymoney
##   30.23896    9.92460
```

```
m2
```

```
##  safemoney riskymoney
##   19.43688   20.27082
```

```
m3
```

```
##  safemoney riskymoney
##   29.65520   11.19074
```

```
Bf12 <- cbind(m1-m2 - qt(1 -alpha/(2*k), N-g) * sqrt((1/ng[1]+1/ng[2])*W), m1-m2, m1-m2 + qt(1 -alpha/(2*k), N-g) * sqrt((1/
ng[1]+1/ng[2])*W))
Bf23 <- cbind(m2-m3 - qt(1 -alpha/(2*k), N-g) * sqrt((1/ng[2]+1/ng[3])*W), m2-m3, m2-m3 + qt(1 -alpha/(2*k), N-g) * sqrt((1/
ng[2]+1/ng[3])*W))
Bf31 <- cbind(m3-m1 - qt(1 -alpha/(2*k), N-g) * sqrt((1/ng[3]+1/ng[1])*W), m3-m1, m3-m1 + qt(1 -alpha/(2*k), N-g) * sqrt((1/
ng[3]+1/ng[1])*W))

IC <- list(young_adult=Bf12, adult_old=Bf23, old_young=Bf31)
IC
```

```
## $young_adult
##                  [,1]      [,2]      [,3]
## safemoney    9.397022  10.80208 12.207138
## riskymoney -11.760692 -10.34622 -8.931748
##
## $adult_old
##                  [,1]      [,2]      [,3]
## safemoney  -11.623378 -10.21832 -8.813262
## riskymoney   7.665608   9.08008 10.494552
##
## $old_young
##                  [,1]     [,2]      [,3]
## safemoney  -1.9888178 -0.58376 0.8212978
## riskymoney -0.1483318  1.26614 2.6806118
```

```
### ---------------------------------------------------------------------------
### Problem 3 of 29/06/11
### ---------------------------------------------------------------------------
# Juan de los Euros, a known driver of Santander, has measured the duration of
# his recent trips to / from the airport (file time.txt).
# a) Having fitted a two-factor ANOVA additive model (center-aero / aero-center,
#    Weekday / weekend) provide point estimated of the means and the variances
#    of the four possible types of travel.
# b) On the basis of the model (a) perform a 90% test to test the significance
#    of the factor-center aero / aero-center.
# c) On the basis of the model (a) perform a 90% test to test the significance
#    of the factor weekday / weekend.
# d) Based on the tests (b) and (c) propose a possible reduced model and re-
#    estimate point-wise - coherently with the reduced model - the means and
#    variances of the four possible types of travel.
### ---------------------------------------------------------------------------

euros <- read.table('time.txt', header=T)
head(euros)
```

```
##   durata          AR      FF
## 1  18.02 aero_centro festivo
## 2  22.61 aero_centro festivo
## 3  20.12 aero_centro festivo
## 4  20.46 aero_centro festivo
## 5  22.53 aero_centro festivo
## 6  18.05 centro_aero festivo
```

```
attach(euros)

# question a)
### Two-ways ANOVA
### Model without interaction (additive model):
### X.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N(0,sigma^2),
###     i=1,2 (effect direction centre-aero/aero-centre),
###     j=1,2 (effect day weekday/weekend)

g <- 2
b <- 2
p <- 1
n <- 5
N <- n*g*b

# Verify the assumptions
# 1) normality (univariate) in each group
Ps <- c(shapiro.test(durata[ AR=='aero_centro' & FF=='festivo' ])$p,
        shapiro.test(durata[ AR=='aero_centro' & FF=='feriale' ])$p,
        shapiro.test(durata[ AR=='centro_aero' & FF=='festivo' ])$p,
        shapiro.test(durata[ AR=='centro_aero' & FF=='feriale' ])$p)
Ps
```

```
## [1] 0.4443545 0.2483730 0.7045298 0.5530964
```

```
# 2) homogeneity of variances
bartlett.test(list(durata[ AR=='aero_centro' & FF=='festivo' ],
                   durata[ AR=='aero_centro' & FF=='feriale' ],
                   durata[ AR=='centro_aero' & FF=='festivo' ],
                   durata[ AR=='centro_aero' & FF=='feriale' ]))
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  list(durata[AR == "aero_centro" & FF == "festivo"], durata[AR ==     "aero_centro" & FF == "feriale"], durata[AR =
= "centro_aero" &     FF == "festivo"], durata[AR == "centro_aero" & FF == "feriale"])
## Bartlett's K-squared = 0.73219, df = 3, p-value = 0.8656
```

```
# Fit the model:
fit <- aov(durata ~ AR + FF)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## AR           1   1.13    1.13   0.216 0.64831
## FF           1  62.13   62.13  11.873 0.00309 **
## Residuals   17  88.95    5.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
names(fit)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "contrasts"     "xlevels"       "call"          "terms"
## [13] "model"
```

```
# Estimate variances
W <- sum(fit$residuals^2)  # SS_res
var <- W/(g*b*n-g-b+1)     # SS_res/gdl(res)
var
```

```
## [1] 5.232579
```

```
# Estimate the great mean mu:
m <- mean(euros[,1])

# Estimate tau.i, beta.j:
tauAC  <- mean(euros[euros$AR=='aero_centro',1]) - m  # tau.1
tauCA  <- mean(euros[euros$AR=='centro_aero',1]) - m  # tau.2

betaFest <- mean(euros[euros$FF=='festivo',1]) - m  # beta.1
betaFer  <- mean(euros[euros$FF=='feriale',1]) - m  # beta.2

# Point-wise estimates of mean duration of travels
# (model without interaction!)
mAC_Fest <- m + tauAC + betaFest
mAC_Fer  <- m + tauAC + betaFer
mCA_Fest <- m + tauCA + betaFest
mCA_Fer  <- m + tauCA + betaFer

# questions b)/c)
summary(fit)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## AR           1   1.13    1.13   0.216 0.64831
## FF           1  62.13   62.13  11.873 0.00309 **
## Residuals   17  88.95    5.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# question d)
### Reduced model (one-way ANOVA):
### X.jk = mu + beta.j + eps.jk; eps.jk~N(0,sigma^2),
###     j=1,2 (effect day weekday/weekend)

fit.red <- aov(durata ~ FF)
summary(fit.red)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## FF           1  62.13   62.13   12.41 0.00243 **
## Residuals   18  90.08    5.00
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
W <- sum(fit.red$residuals^2)
var <- W/(g*b*n-b)
var
```

```
## [1] 5.004554
```

```
# point-wise estimate of the mean duration of the trips
# (one-way model!)
mAC_Fest <- m + betaFest
mAC_Fer  <- m + betaFer
mCA_Fest <- m + betaFest
mCA_Fer  <- m + betaFer
```

```
### ------------------------------------------------------------------------
### Problem 1 of 12/02/08
### ------------------------------------------------------------------------
# In a medical research center, concentrations of interferon gamma-6 and
# Interferon gamma-7 were measured in the blood of some patients who had
# been infected by Human Papilloma Virus (PV.txt file). The experiment
# included patients that have different medical profiles in terms
# Relapse and HPV-Clearance:
# Relapse: extinct infection (A) or ongoing infection (B);
# HPV-Clearance: extinct we aim to find out if Relapse and Clearance factors
# have an effect on interferon gamma-6 and interferon gamma-7 distributions.
# a) Introduce an appropriate statistical model that justifies the use of
#    MANOVA as a tool for the analysis of these data.
# b) Perform a test to verify the interaction between the factors.
# c) Identify the factor or factors that generate effects statistically
#    significant.
# c) Construct Bonferroni confidence intervals with global coverage 90%
#    that clarify the conclusions drawn at point (c).
### ------------------------------------------------------------------------

PV <- read.table('PV.txt', header = T)
PV
```

```
##    gamma6 gamma7 REL HPV
## 1   32.08  15.32   A   +
## 2   35.36   9.84   A   +
## 3   28.64  11.91   A   +
## 4   21.34   1.31   A   -
## 5   28.57   0.63   A   -
## 6   32.52  -2.03   A   -
## 7  -13.80  12.16   B   +
## 8   -4.31  11.31   B   +
## 9    6.38   5.85   B   +
## 10 -12.90  -3.22   B   -
## 11  14.59   2.33   B   -
## 12  -5.11  -8.11   B   -
```

```r
N <- dim(PV)[1]
p <- 2
g <- b <- 2
n <- N/(g*b)
n
```

```
## [1] 3
```

```r
var.risp <- PV[,1:2]

### question a)
### Two-ways MANOVA
### Model with interaction (complete model):
### X.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N_p(0,Sigma), [p=2]
###     i=1,2 (effect REL), j=1,2 (effect HPV),
###     X.ijk, mu, tau.i, beta.j, gamma.ij in R^2

### We don't verify the assumptions (very few data)

### question b)
man.int <- manova(as.matrix(var.risp) ~  HPV + REL + HPV * REL, data = PV)
summary(man.int, test = 'Wilks')
```

```
##           Df   Wilks approx F num Df den Df   Pr(>F)
## HPV        1 0.17312   16.718      2      7 0.002159 **
## REL        1 0.17763   16.203      2      7 0.002362 **
## HPV:REL    1 0.94290    0.212      2      7 0.814010
## Residuals  8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
### question c)
### Model without interaction (additive model):
### X.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N_p(0,Sigma), [p=2]
###     i=1,2 (effect REL), j=1,2 (effect HPV),
###     X.ijk, mu, tau.i, beta.j in R^2

man     <- manova(as.matrix(var.risp) ~ HPV + REL, data = PV)
summary(man, test = 'Wilks')
```

```
##           Df   Wilks approx F num Df den Df    Pr(>F)
## HPV        1 0.17337   19.072      2      8 0.0009035 ***
## REL        1 0.18524   17.594      2      8 0.0011773 **
## Residuals  9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
summary.aov(man) # interesting!
```

```
##  Response gamma6 :
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## HPV          1    2.38    2.38  0.0292 0.8680968
## REL          1 3125.35 3125.35 38.4051 0.0001594 ***
## Residuals    9  732.41   81.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response gamma7 :
##             Df Sum Sq Mean Sq F value    Pr(>F)
## HPV          1 474.77  474.77 42.8676 0.0001055 ***
## REL          1  23.13   23.13  2.0884 0.1823227
## Residuals    9  99.68   11.08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# factor relapse seems to have effect on interferon gamma-6,
# factor HPV (virus) seems to have effect on interferon gamma-7
# but we don't know in which sense!

# let's clarify with Bonferroni:
SSres <- t(man$residuals) %*% man$residuals / (N-g-b+1) # model without interaction

k <- g*(g-1)/2*p + b*(b-1)/2*p
k
```

```
## [1] 4
```

```r
qT <- qt(1 - 0.1/(2*k), N-g-b+1)

attach(PV)

m6_rel <- tapply(gamma6, REL, mean)
m7_rel <- tapply(gamma7, REL, mean)
m6_HPV <- tapply(gamma6, HPV, mean)
m7_HPV <- tapply(gamma7, HPV, mean)

m6_rel
```

```
##        A        B
## 29.75167 -2.52500
```

```r
m7_rel
```

```
##        A        B
## 6.163333 3.386667
```

```r
m6_HPV
```

```
##        -        +
## 13.16833 14.05833
```

```r
m7_HPV
```

```
##      -      +
## -1.515 11.065
```

```
REL6   <- c(diff(m6_rel) - qT * sqrt( SSres[1,1] * (1/6+1/6) ),
            diff(m6_rel) + qT * sqrt( SSres[1,1] * (1/6+1/6) ))
REL7   <- c(diff(m7_rel) - qT * sqrt( SSres[2,2] * (1/6+1/6) ),
            diff(m7_rel) + qT * sqrt( SSres[2,2] * (1/6+1/6) ))

HPV6   <- c(diff(m6_HPV) - qT * sqrt( SSres[1,1] * (1/6+1/6) ),
            diff(m6_HPV) + qT * sqrt( SSres[1,1] * (1/6+1/6) ))
HPV7   <- c(diff(m7_HPV) - qT * sqrt( SSres[2,2] * (1/6+1/6) ),
            diff(m7_HPV) + qT * sqrt( SSres[2,2] * (1/6+1/6) ))

detach(PV)

Bf <- list(B_A_6 = REL6, B_A_7 = REL7, HPVest_HPVpres_6 = HPV6, HPVest_HPVpres_7 = HPV7)
Bf
```

```
## $B_A_6
##         B         B
## -46.26096 -18.29238
##
## $B_A_7
##         B         B
## -7.935627  2.382294
##
## $HPVest_HPVpres_6
##         +         +
## -13.09429  14.87429
##
## $HPVest_HPVpres_7
##         +         +
##  7.421039 17.738961
```

```
# interf gamma 6 higher when the infection is extint (REL=A)
# interf gamma 7 higher when the virus is extint (HPVest = +)
```

```
### ------------------------------------------------------------------------------
### Problem 3 of 28/02/13
### ------------------------------------------------------------------------------
# For security reasons, the direction of the Paris Louvre museum has undertaken
# a campaign of control of the tourists flow in the museum. During the first phase
# of monitoring ("louvre.txt" file), the durations [minutes] of the visits were
# measured in the Museum for 360 tourists from Europe, USA and Japan, also recording
# the type of visit (guided, with audio guide or without guide).
# a) Formulate a suitable (complete) model for the duration of a visit of the museum
#    with respect to the two factors nationality and type of visit; in particular,
#    introduce and verify the assumptions of the model.
# b) Through a suitable test, discuss the possibility of removing the interaction
#    term and possibly reduce the model.
# c) On the basis of the model at step b), test the effect of the factors nationality
#    and type of visit on the average time of the visit and, if appropriate, propose
#    a reduced model.
# d) Provide the security managers of the four intervals museum with Bonferroni confidence
#    (globally 90%) for the mean and variance of the visit to the museum time for
#    homogeneous groups of identified visitors at step c).
### ------------------------------------------------------------------------------
```

```
museo <- read.table('louvre.txt',header=TRUE)
attach(museo)

# question a)
### Two-ways ANOVA
### Model with interaction (complete model):
### X.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N(0,Sigma),
###     i=1,2,3 (effect TYPE OF VISIT), j=1,2,3 (effect NAZIONALITY),
###     X.ijk, mu, tau.i, beta.j, gamma.ij in R

fit <- aov(tempo ~ tipo + nazione + nazione:tipo, data=museo)
summary(fit)
```

```
##               Df  Sum Sq Mean Sq  F value Pr(>F)
## tipo           2 1929030  964515 1016.568 <2e-16 ***
## nazione        2     157      79    0.083  0.921
## tipo:nazione   4     899     225    0.237  0.917
## Residuals    351  333027     949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Verify assumptions
t <- tipo:nazione
st <- NULL
for(i in 1:9)
  st<-c(st,
        shapiro.test(tempo[which(t==levels(t)[i])])$p )
st
```

```
## [1] 0.8475933 0.9835701 0.5680572 0.4039841 0.8109824 0.4587566 0.4983149
## [8] 0.9164096 0.3102867
```

```
bartlett.test(tempo,tipo:nazione)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  tempo and tipo:nazione
## Bartlett's K-squared = 0.80012, df = 8, p-value = 0.9992
```

```r
# question b)
### Two-ways ANOVA
### Model without interaction (additive model):
### X.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N(0,Sigma),
###     i=1,2,3 (effect TYPE OF VISIT), j=1,2,3 (effect NAZIONALITY),
###     X.ijs, mu, tau.i, beta.j in R

fit2 <- aov(tempo ~ tipo + nazione, data=museo)
summary(fit2)
```

```
##              Df  Sum Sq Mean Sq  F value Pr(>F)
## tipo          2 1929030  964515 1025.384 <2e-16 ***
## nazione       2     157      79    0.083   0.92
## Residuals   355  333926     941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# question c)
### Reduced model: One-way ANOVA
### X.ik = mu + tau.i + eps.ik; eps.ik~N(0,Sigma),
###     i=1,2,3 (effect TYPE OF VISIT)
###     X.ijs, mu, tau.i, beta.j in R

fit3 <- aov(tempo ~ tipo, data=museo)
summary(fit3)
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## tipo          2 1929030  964515    1031 <2e-16 ***
## Residuals   357  334084     936
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# question e)
# g=3 Groups identified by the levels of the factor TYPE
# k=4: g*(g-1)/2 comparisons between the means + 1 conf int on the variance
n <- dim(museo)[1]
g <- 3
k <- g*(g-1)/2+1
S <- sum(residuals(fit3)^2)/(n-g)

alpha<- .1

Mg  <- tapply(museo[,1], tipo, mean)

label <- levels(factor(tipo))
n1 <- length(museo[tipo==label[1],1])
n2 <- length(museo[tipo==label[2],1])
n3 <- length(museo[tipo==label[3],1])
t <- qt(1-alpha/(2*k),n-g)

# Conf int for the means
ICB1<-data.frame(L=Mg[1]-sqrt(S*(1/n1))*t,C=Mg[1],U=Mg[1]+sqrt(S/n1)*t)
ICB2<-data.frame(L=Mg[2]-sqrt(S*(1/n2))*t,C=Mg[2],U=Mg[2]+sqrt(S/n2)*t)
ICB3<-data.frame(L=Mg[3]-sqrt(S*(1/n3))*t,C=Mg[3],U=Mg[3]+sqrt(S/n3)*t)
ICB<-data.frame(rbind(ICB1,ICB2,ICB3))
ICB
```

```
##                    L        C        U
## audioguida 170.8309 177.1167 183.4024
## guida      292.1559 298.4417 304.7274
## no guida   117.1559 123.4417 129.7274
```

```r
# Conf int for variances
chi_u <- qchisq(alpha/(2*k),n-g)
chi_l <- qchisq(1-alpha/(2*k),n-g)
ICBV <- data.frame(L=(n-g)*S/chi_l,C=S,U=(n-g)*S/chi_u)
ICBV
```

```
##          L        C        U
## 1 796.2808 935.8083 1114.326
```

```r
detach(museo)
```

```r
### ---------------------------------------------------------------------------
### Pb 3 of 05/09/08
### ---------------------------------------------------------------------------
# The West Sussex Bread Association has randomly selected 60 business trade
# in which doughnuts are commonly sold. 30 activities are based
# in the city of Brighton and 30 in the town of Worthing. For each of the
# two cities, in 10 activity the price of a plain bagel was recorded, in 10
# the price of a doughnut filled with cream and in 10 the price of a doughnut
# filled with jam. The data are reported in doughnut.txt dataset.
# a) Describe the ANOVA model you deem appropriate for the analysis of these data.
# b) Identifying the factors that significantly influence the distribution
#    of the price of doughnuts, identify a possible reduced model.
# c) using Bonferroni's inequality estimate through bilateral confidence
#    intervals (with global confidence 95%) the means and variances of the
#    subpopulations associated with the reduced model identified at step (b).
### ---------------------------------------------------------------------------
```

```r
ciambelle <- read.table('doughnut.txt', header=TRUE)
head(ciambelle)
```

```
##   prezzo    citta   tipo
## 1   0.37 Brighton liscia
## 2   0.35 Brighton liscia
## 3   0.39 Brighton liscia
## 4   0.37 Brighton liscia
## 5   0.40 Brighton liscia
## 6   0.38 Brighton liscia
```

```r
attach(ciambelle)

# question a)
# ANOVA two-ways
# Model with interaction (complete model):
# X.ijk = mu + tau.i + beta.j + gamma.ij + eps.ijk; eps.ijk~N(0,sigma^2),
#     i=1,2 (effect city), j=1,2,3 (effect type)

fit.c <- aov(prezzo ~ citta + tipo + citta:tipo)
summary(fit.c)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## citta        1 0.00002 0.00002   0.022    0.883
## tipo         2 0.09026 0.04513  65.957 3.19e-15 ***
## citta:tipo   2 0.00019 0.00010   0.139    0.871
## Residuals   54 0.03695 0.00068
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
p.val <- c(shapiro.test(prezzo[which(citta==levels(citta)[1] & tipo==levels(tipo)[1])])$p,
           shapiro.test(prezzo[which(citta==levels(citta)[1] & tipo==levels(tipo)[2])])$p,
           shapiro.test(prezzo[which(citta==levels(citta)[1] & tipo==levels(tipo)[3])])$p,
           shapiro.test(prezzo[which(citta==levels(citta)[2] & tipo==levels(tipo)[1])])$p,
           shapiro.test(prezzo[which(citta==levels(citta)[2] & tipo==levels(tipo)[2])])$p,
           shapiro.test(prezzo[which(citta==levels(citta)[2] & tipo==levels(tipo)[3])])$p)
p.val
```

```
## [1] 0.07154309 0.52338976 0.39802082 0.94696817 0.87853858 0.09241245
```

```r
bartlett.test(prezzo, citta:tipo)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  prezzo and citta:tipo
## Bartlett's K-squared = 4.0808, df = 5, p-value = 0.5378
```

```r
# question b)
# Model without interaction (additive model):
# X.ijk = mu + tau.i + beta.j + eps.ijk; eps.ijk~N(0,sigma^2),
#     i=1,2 (effect city), j=1,2,3 (effect type)
fit.c2 <- aov(prezzo ~ citta + tipo)
summary(fit.c2)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## citta        1 0.00002 0.00002   0.023    0.881
## tipo         2 0.09026 0.04513  68.050 1.02e-15 ***
## Residuals   56 0.03714 0.00066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# one-way ANOVA
# X.jk = mu + beta.j + eps.ijk; eps.ijk~N(0,sigma^2),
#     j=1,2,3 (effect type)
fit.c3 <- aov(prezzo ~ tipo)
summary(fit.c3)
```

```
##             Df  Sum Sq Mean Sq F value   Pr(>F)
## tipo         2 0.09026 0.04513   69.24 5.58e-16 ***
## Residuals   57 0.03715 0.00065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# question c)
N <- dim(ciambelle)[1]
g <- length(levels(tipo))
DF <- N-g

alpha <- .05
k <- g+1

qT <- qt(1-alpha/(2*k), DF)
qCinf <- qchisq(1 - alpha / (2*k), DF)
qCsup <- qchisq(alpha / (2*k), DF)

Spooled <- (t(fit.c3$res) %*% fit.c3$res)/DF
Spooled
```

```
##              [,1]
## [1,] 0.0006518421
```

```r
m1 <- mean(ciambelle[which(tipo==levels(tipo)[1]),1])
m2 <- mean(ciambelle[which(tipo==levels(tipo)[2]),1])
m3 <- mean(ciambelle[which(tipo==levels(tipo)[3]),1])
medie <- c(m1,m2,m3)

ng <- c(length(which(tipo==levels(tipo)[1])),length(which(tipo==levels(tipo)[2])),length(which(tipo==levels(tipo)[3])))

BF    <- rbind(cbind(inf=medie - sqrt(as.vector(Spooled) / ng) * qT,
                     sup=medie + sqrt(as.vector(Spooled) / ng) * qT),
               c(inf=Spooled * DF / qCinf,
                 sup=Spooled * DF / qCsup))
BF
```

```
##               inf         sup
## [1,] 0.4292741080 0.458725892
## [2,] 0.3482741080 0.377725892
## [3,] 0.4317741080 0.461225892
## [4,] 0.0004264347 0.001098285
```

```r
detach(ciambelle)
```