```
###---------------------###
### LAB 13 (05/06/2020) ###
###---------------------###

### TOPICS:
### Exercises from past exams

library(MASS)
library(car)
library(rgl)

setwd("D:/RTDA/Didattica/Applied Statistics MATE 19-20/Lab 13 - 05062020")
load("D:/RTDA/Didattica/Applied Statistics MATE 19-20/Lab 5 - 16042020/mcshapiro.test.RData")

#_____________________________________________________________________
###### Exam of 5/09/2008
######---------------------------
#_____________________________________________________________________
### Problem 1
# Famous Dantists believe that the number of characters (NC) and the
# number of words (NP) contained in a generic sonnet of Dante follow
# approximately a jointly normal distribution with mean mu=c(400,90)
# and covariance matrix Sigma=cbind(c(100,20), c(20,10)). Recently,
# two new sonnets attributed to Dante have been discovered, and have
# been identified for the moment with the codes 2008A and 2008B.
# Assuming the number of characters and of words of the sonnet 2008A
# independent of the number of characters and words of the sonnet 2008B:
# a) identify in the plane NC x NP an ellipsoidal region in which the
#    sonnet 2008A is contained with probability 0.9.
# b) How likely only one of the two sonnets falls in the region
#    identified at point (a)?

### question a)
# a) identify in the plane NC x NP an ellipsoidal region in which the
#    sonnet 2008A is contained with probability 0.9.

mu <- c(400, 90)
Sigma <- cbind(c(100,20),c(20,10))

# Prediction ellipse of level 1-alpha=0.9
# Characterize the ellipse:
eigen(Sigma)

# Direction of the axes:
eigen(Sigma)$vectors

# Centre:
M <- mu
M

# Radius of the ellipse:
r <- sqrt(qchisq(0.9,2))
r

# Length of the semi-axes:
r*sqrt(eigen(Sigma)$values)

# Plot
plot(M[1],M[2],xlim=c(350,450), col='blue',pch=19,xlab='X.1',ylab='X.2',asp=1)
ellipse(center=M, shape=cbind(Sigma), radius=r, col = 'blue')

# b) How likely only one of the two sonnets falls in the region
#    identified at point (a)?

# It is asking to compute:
# P({2008A in RC} and {2008B not in RC}) +
#                         + P({2008A not in RC} and {2008B in RC}) =
# [independence]
# P({2008A in RC})*P({2008B not in RC}) +
#                         + P({2008A not in RC})*P({2008B in RC})=
0.9*0.1+0.1*0.9


#_____________________________________________________________________
### Problem 2
# The dataset eighteen.txt contains, for 100 Italian municipalities,
# the percentages of underage boys and underage girls with respect to
# the entire population resident in the municipality. Assuming that
# this is a sample from a bivariate normal distribution:
# a) perform a test to verify that in Italy the number of underage
#    resident boys is the same as the number of underage resident girls;
# b) knowing that 60 million people reside in Italy, provide three
#    T2-simultaneous intervals (global confidence 90%) for:
#    - The absolute number of underage boys who reside in Italy,
#    - The absolute number of underage girls who reside in Italy,
```

```
#    - The absolute number of minors who reside in Italy.

eighteen <- read.table('eighteen.txt', header=T)
head(eighteen)
dim(eighteen)

### question a)
# a) perform a test to verify that in Italy the number of underage
#    resident boys is the same as the number of underage resident girls;

n <- dim(eighteen)[1]
p <- 1

D <- eighteen[,1]-eighteen[,2]
M.d <- mean(D)
M.d

S.d <- var(D)
Sinv <- solve(S.d)

# Test: H0: delta=0 vs H1:delta!=0
delta0 <- 0

T2 <- n*t(M.d-delta0)%*%Sinv%*%(M.d-delta0)

# verify Gaussian assumption:
shapiro.test(D)

# pvalue
pvalue <- 1-pf(T2*(n-p)/((n-1)*p), p, n-p)
pvalue # Reject H0

# alternatively: D is one-dimensional
t.test(D, alternative = 'two.sided')

### question b)
# b) knowing that 60 million people reside in Italy, provide three
#    T2-simultaneous intervals (global confidence 90%) for:
#    - The absolute number of underage boys who reside in Italy,
#    - The absolute number of underage girls who reside in Italy,
#    - The absolute number of minors who reside in Italy.

mcshapiro.test(eighteen)

# Note. The data are percentages with respect to the entire population
C <- 60*10^6/100*rbind( c(1,0),     # only boys
                        c(0,1),     # only girls
                        c(1,1) )    # total
C

mu <- colMeans(eighteen)
S <- cov(eighteen)
p <- 2
n <- 100

alpha <- .1

T2 <- cbind(
  C%*%mu - sqrt(diag(C%*%S%*%t(C))/n*(p*(n-1)/(n-p))*qf(1-alpha,p,n-p)),
  C%*%mu ,
  C%*%mu + sqrt(diag(C%*%S%*%t(C))/n*(p*(n-1)/(n-p))*qf(1-alpha,p,n-p)))
colnames(T2) <- c('Inf','Mean','Sup')
T2

#_____________________________________________________________________
### Problem 3
# The West Sussex Bread Association has randomly selected 60 business
# trades 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, the price of a plain doughnut was recorded in 10
# activities, the price of a doughnut filled with cream in other 10
# activities and the price of a doughnut filled with jam in the
# remaining 10 activieties.
# The data are reported in doughnut.txt dataset.
# a) Describe the ANOVA model you deem appropriate for the analysis of
#    these data.
# b) Identifying factors that significantly influence the distribution
#    of the price of doughnuts, propose a possible reduced model.
# c) Using the Bonferroni's inequality estimate through bilateral
#    confidence intervals (with global confidence 95%) the means and the
#    variances of the subpopulations associated with the reduced model
#    identified at step (b).

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

ANOVA
$(p=1, b=2, g=3)$

```r
attach(doughnuts)

# question a)
# a) Describe the ANOVA model you deem appropriate for the analysis of
#    these data.

# 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)

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

bartlett.test(prezzo, citta:tipo)

# question b)
# b) Identifying factors that significantly influence the distribution
#    of the price of doughnuts, propose a possible reduced model.

# 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)

# 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)

# question c)
# c) Using the 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).

N <- dim(doughnuts)[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

m1 <- mean(doughnuts[which(tipo==levels(tipo)[1]),1])
m2 <- mean(doughnuts[which(tipo==levels(tipo)[2]),1])
m3 <- mean(doughnuts[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(c(Spooled) / ng) * qT,
                     sup=medie + sqrt(c(Spooled) / ng) * qT),
               c(inf=Spooled * DF / qCinf,
                 sup=Spooled * DF / qCsup))
BF

detach(doughnuts)

#_____________________________________________________________________
### Problem 4
# Due to the increased cost of oil, the airline FlyDown is interested in
# identifying a model to estimate the weight of a passenger (typically
# not available) from its age and the sex (available data). For this
# reason, the FlyDown collected, via an anonymous questionnaire, data
```

```r
# about the weight of 126 clients (63 men and 63 women) aged between 18
# and 80 years (flydown.txt dataset).
# a) Introduce a regression model quadratic in age to describe the
#    dependence of the expected weight of a passenger on their age and
#    sex.
# b) Is there statistical evidence of a dependence of the expected
#    weight on sex?
# c) Is there statistical evidence of a dependence (linear or quadratic)
#    of the expected weight on age?
# d) Identify a reduced model of the model (a) suitable to describe the
#    collected data and estimate its parameters.
# e) On the basis of the results at point (d), is there statistical
#    evidence to reject the hypothesis that the maximum of the expected
#    weight (for men and / or women) is reached at the age of 50?
# f) Identify a reduced model of the model (d) that takes into account
#    the statement in paragraph (e) and estimate its parameters.

fly <- read.table('flydown.txt', header=TRUE)
head(fly)

### question a)
# a) Introduce a regression model quadratic in age to describe the
#    dependence of the expected weight of a passenger on their age and
#    sex.

D <- ifelse(fly[,2]=='M', 1, 0)  # dummy
head(D)

Fly <- data.frame(fly[,c(1,3)], D=D)
head(Fly)

fit <- lm(peso ~ eta + I(eta^2) + D + D:eta + I(D*eta^2), data=Fly)
summary(fit)

shapiro.test(residuals(fit))

par(mfrow=c(2,2))
plot(fit)

dev.off()

### question b)
# b) Is there statistical evidence of a dependence of the expected
#    weight on sex?
A <- rbind(c(0,0,0,1,0,0),
           c(0,0,0,0,1,0),
           c(0,0,0,0,0,1))
b <- c(0,0,0)

linearHypothesis(fit, A, b)

### question c)
# c) Is there statistical evidence of a dependence (linear or quadratic)
#    of the expected weight on age?

A <- rbind(c(0,1,0,0,0,0),
           c(0,0,1,0,0,0),
           c(0,0,0,0,1,0),
           c(0,0,0,0,0,1))
b <- c(0,0,0,0)

linearHypothesis(fit, A, b)

### question d)
# d) Identify a reduced model of the model (a) suitable to describe the
#    collected data and estimate its parameters.
summary(fit)

A <- rbind(c(0,0,0,0,1,0),
           c(0,0,0,0,0,1))
b <- c(0,0)

linearHypothesis(fit,A,b)

fit2 <- lm(peso ~ eta + I(eta^2) + D , data=Fly)
summary(fit2)

shapiro.test(residuals(fit2))

par(mfrow=c(2,2))
plot(fit2)

dev.off()

# question e)
# e) On the basis of the results at point (d), is there statistical
#    evidence to reject the hypothesis that the maximum of the expected
```

linear model + dummy

H_0: max $E[W] = 50$

```r
#    weight (for men and / or women) is reached at the age of 50?

# Deriving and imposing the derivative to be 0 we obtain that the
# maximum is reached for (beta2<0):
# beta1+2*beta2*eta.max=0, i.e., eta.max=50 <=> beta1+2*beta2*50=0

A <- c(0,1,2*50,0)
b <- 0

linearHypothesis(fit2,A,b)

# question f)
# f) Identify a reduced model of the model (d) that takes into account
#    the statement in paragraph (e) and estimate its parameters.

# Constrained model:
# beta1+2*beta2*50=0  =>  beta1=-100*beta2
fit3 <- lm(peso ~ I(-100*eta+eta^2) + D, data=Fly)
summary(fit3)

coef(fit3)


#_____________________________________________________________
###### Exam of 21/09/2009
######----------------------------

#_____________________________________________________________
### Problem 1

# The sco2009.txt file collects the duration [minutes] of different
# talks at S.Co.2009 Conference (18 sessions of 4 talks each). Assuming
# a four-dimensional Gaussian distribution for the first, second, third
# and fourth talk of the same session:
# a) Is there statistical evidence to state that the mean durations of
#    the four talks are different?
# b) Using Bonferroni's inequality, provide eight intervals of 90% global
#    confidence for the mean and variance of the first, the second, the
#    third and fourth talk.
# On the basis of the introduced model, using the sample estimates of
# the mean and of the covariance matrix, and knowing that the maximum
# time available for each session is 2 hours:
# c) Estimate the probability that the time available for questions
#    is less than 15 minutes.
# d) Estimate the probability that the sum of the durations of the four
#    talks exceeds the maximum time of 2 hours.

sco <- read.table('sco2009.txt', header=TRUE)
sco

### question a)
# a) Is there statistical evidence to state that the mean durations of
#    the four talks are different?

n <- dim(sco)[1]
q <- dim(sco)[2]

M <- sapply(sco,mean)
S <- cov(sco)

# contrast matrix that looks at consecutive increments
C   <-  matrix(c(-1, 1, 0, 0,
                 0, -1, 1, 0,
                 0, 0, -1, 1), 3, 4, byrow=T)
C

mcshapiro.test(sco)

# Test: H0: C%*%mu=0 vs H1: C%*%mu!=0
delta.0 <- c(0, 0, 0)

Md <- C %*% M
Sd <- C %*% S %*% t(C)
Sdinv <- solve(Sd)

T2 <- n * t( Md - delta.0 ) %*% Sdinv %*% ( Md - delta.0 )

P <- 1-pf(T2*(n-(q-1))/((q-1)*(n-1)),(q-1),n-(q-1))
P  # Reject H0

### question b)
# b) Using Bonferroni's inequality, provide eight intervals of 90% global
#    confidence for the mean and variance of the first, the second, the
#    third and fourth talk.

k <- 8
```

```
ICmedie <- cbind('Inf'=M - sqrt(diag(S)/n) * qt(1 - alpha/(2*k), n-1),
                 'M'  =M,
                 'Sup'=M + sqrt(diag(S)/n) * qt(1 - alpha/(2*k), n-1))
ICmedie

ICBVar<-data.frame('Inf'=(n-1)*diag(S)/qchisq(1-alpha/(2*k),n-1),
                   'M'  =diag(S),
                   'Sup'=(n-1)*diag(S)/qchisq(alpha/(2*k),n-1))
ICBVar

### question c)
# c) Estimate the probability that the time available for questions
#    is less than 15 minutes.

# Compute the distribution of the time remained for questions:
# T ~ N(Mt,st) by noting that T=2*60-(X1+X2+X3+X4)
a <- c(1,1,1,1)

Mt <- 2*60-t(a)%*%M
Mt
St  <- t(a)%*%S%*%a
St

pnorm(15, Mt, sqrt(St))

### question d)
# d) Estimate the probability that the sum of the durations of the four
#    talks exceeds the maximum time of 2 hours.

a <- c(1,1,1,1)

Mi <- t(a)%*%M
Mi
Si  <- t(a)%*%S%*%a
Si

1-pnorm(2*60, Mi, sqrt(Si))

# or
pnorm(0, Mt, sqrt(St))


#_____________________________________________________________________
### Problem 2
# The file 100m.txt contains the time [seconds] used to run the 100m
# by 20 students enrolled in the athletics team of Politecnico.
# For each student 4 times are given: just back from summer holidays,
# one, two and three weeks after return. Framing the problem in the
# context of repeated measures:
# a) Is there statistical evidence to say that the mean time changes
#    over time?
# b) Using appropriate confidence intervals, describe the temporal
#    dynamics the mean times.
# c) After how many weeks one may think that the mean time has
#    stabilized?

tempi <- read.table('100m.txt', header=TRUE)
head(tempi)
dim(tempi)

### question a)
# a) Is there statistical evidence to say that the mean time changes
#    over time?

n <- dim(tempi)[1]
q <- dim(tempi)[2]

M <- sapply(tempi,mean)
S <- cov(tempi)

# contrast matrix that looks at consecutive increments
C   <-  matrix(c(-1, 1, 0, 0,
                 0, -1, 1, 0,
                 0, 0, -1, 1), 3, 4, byrow=T)
C

mcshapiro.test(tempi)

# Test: H0: C%*%mu=0 vs H1: C%*%mu!=0
delta.0 <- c(0, 0, 0)

Md <- C %*% M
Sd <- C %*% S %*% t(C)
Sdinv <- solve(Sd)

T2 <- n * t( Md - delta.0 ) %*% Sdinv %*% ( Md - delta.0 )
```

```r
P <- 1-pf(T2*(n-(q-1))/((q-1)*(n-1)),(q-1),n-(q-1))
P  # Reject H0

### question b)
# b) Using appropriate confidence intervals, describe the temporal
#    dynamics the mean times.
alpha <- .05
k <- (q-1) + 1  # we provide a conf int for the mean at time t=0

ICmedie <- cbind(M[1] - sqrt(diag(S)[1]/n) * qt(1 - alpha/(2*k), n-1),
                 M[1],
                 M[1] + sqrt(diag(S)[1]/n) * qt(1 - alpha/(2*k), n-1))

ICmedie <- rbind(ICmedie,
                 cbind(Md - sqrt(diag(Sd)/n) * qt(1 - alpha/(2*k), n-1),
                       Md,
                       Md + sqrt(diag(Sd)/n) * qt(1 - alpha/(2*k), n-1)))

rownames(ICmedie) <- c('initial mean', 'increment1', 'increment2', 'increment3')

ICmedie

### question c)
# c) After how many weeks one may think that the mean time has
#    stabilized?

ICmedie[4,]


#_____________________________________________________________________
### Problem 3
# In the summer period, hundreds of whales move to the Gulf of Maine to
# feed in views of Winter. In this period the whales are in separated
# and geographically localized colonies. The whales.txt file gives the
# geographical coordinates of sightings in the sea area in front of
# Boston that it is known to host two colonies.
# a) By using a hierarchical agglomerative clustering algorithm
#    (Euclidean metrics and linkage and Ward), identify the sightings
#    relative to each of the two colonies.
# b) Discuss about the algorithm's goodness.
# c) By assuming correct the subdivision obtained at point (a) and
#    introducing suitable assumptions, provide a point estimate and a
#    90% confidence region for the relative position of the two colonies.

whales <- read.table('whales.txt', header=TRUE)
head(whales)

plot(whales)

dev.off()

### question a)
# a) By using a hierarchical agglomerative clustering algorithm
#    (Euclidean metrics and linkage and Ward), identify the sightings
#    relative to each of the two colonies.

bdist <- dist(whales)
b.ew  <- hclust(bdist, method='ward.D')

plot(b.ew, hang=-0.1, labels=FALSE, sub='')

bal.clust <- cutree(b.ew,k=2) # k=2 dato dal testo

plot(whales, col=bal.clust+1)

dev.off()

### question b)
# b) Discuss about the algorithm's goodness.
coph <- cophenetic(b.ew)
cor(coph, bdist) # cophenetic correlation coefficient

### question c)
# c) By assuming correct the subdivision obtained at point (a) and
#    introducing suitable assumptions, provide a point estimate and a
#    90% confidence region for the relative position of the two colonies.

# Assumptions
# - 2 independent groups
# - Homogeneity of covariances
# - Gaussianity

b1 <- whales[which(bal.clust==1),]
b2 <- whales[which(bal.clust==2),]

n1 <- dim(b1)[1]
n2 <- dim(b2)[1]
```

```r
p  <- 2

M1 <- sapply(b1, mean)
M2 <- sapply(b2, mean)
M1
M2

S1 <- cov(b1)
S2 <- cov(b2)
Spooled <- ((n1-1)*S1+(n2-1)*S2)/(n1+n2-2)
Spinv <- solve(Spooled)

# Verify assumptions
S1
S2

mcshapiro.test(b1)$pvalue
mcshapiro.test(b2)$pvalue

# Point estimate
M1-M2

# Confidence region of level 1-alpha=90%
alpha <- .1
cfr.fisher <- (p*(n1+n2-2)/(n1+n2-1-p))*qf(1-alpha,p,n1+n2-1-p)
M <- M1-M2

# Characterize the ellipse:
# Direction of the axes:
eigen(Spooled)$vectors

# Center:
M

# Radius of the ellipse:
r <- sqrt(cfr.fisher)
r

# Length of the semi-axes:
r*sqrt(eigen(Spooled*(1/n1+1/n2))$values)

# Grafico
plot(M[1],M[2],pch=19)
ellipse(center=M, shape=Spooled*(1/n1+1/n2), radius=sqrt(cfr.fisher), lwd = 2, col = 'red', lty = 2, cente
r.pch = 4)
grid()

dev.off()

#_____________________________________________________________________
### Problem 4
# The BGMI.txt file contains travel times [hours] of the Bergamo-Milan
# highway, the departure time from Bergamo [hours] and the day of the
# week for 1127 vehicles leaving between 6:00 and 9:00. Assume a
# dependence at most quadratic of the mean travel time on the time of
# departure, possibly different depending on the day of the week:
# T = alpha.g + beta.g(t.partenza-7.5) + gamma.g(tpartenza-7.5)^2 + eps
# with eps ~ N (0, sigma ^ 2) and g={working, holiday}.
# a) Provide the least squares estimates of the model parameters.
# b) On the basis of a suitable test, is there statistical evidence of a
#    difference in the mean trends between weekdays and weekends?
# c) Build a suitable reduced model with 4 degrees of freedom and
#    estimate its parameters.
# d) On basis of model (c), provide interval estimates (90% global
#    confidence) for the value of the regression curve and its
#    derivative for a weekday departure and / or weekend at 7:30.

tper <- read.table('BGMI.txt', header=TRUE)
head(tper)
dim(tper)

### question a)
# a) Provide the least squares estimates of the model parameters.
D <- ifelse(tper[,3]=='festivo',1,0)

tpd <- data.frame(tper[,-3],D=D)

fit <- lm(durata ~ I(partenza-7.5) + I((partenza-7.5)^2) +
            D + D:I(partenza-7.5) + D:I((partenza-7.5)^2), data=tpd)
summary(fit)

shapiro.test(residuals(fit))

par(mfrow=c(2,2))
plot(fit)

dev.off()
```

```
### question b)
# b) On the basis of a suitable test, is there statistical evidence of a
#    difference in the mean trends between weekdays and weekends?
shapiro.test(residuals(fit))

A <- rbind(c(0,0,0,1,0,0),
           c(0,0,0,0,1,0),
           c(0,0,0,0,0,1))
b <- c(0,0,0)

linearHypothesis(fit,A,b)

### question c)
# c) Build a suitable reduced model with 4 degrees of freedom and
#    estimate its parameters.
summary(fit)

# Reduce the model:
fit2 <- lm(durata ~ I(partenza-7.5) + I((partenza-7.5)^2) +
           D + D:I((partenza-7.5)^2), data=tpd)
summary(fit2)

### question d)
# d) On basis of model (c), provide interval estimates (90% global
#    confidence) for the value of the regression curve and its
#    derivative for a weekday departure and / or weekend at 7:30.

# Confidence interval for D=0,1 and partenza=7.5 (=7:30)
k <- 3
n <- dim(tper)[1]
Z0.new <- data.frame(partenza=c(7.50,7.50), D=c(0,1))
Conf <- predict(fit2, Z0.new, interval='confidence', level=1-.1/k)
Conf

# Confidence interval for the derivative
# Curve:
# durata = b0 + b1*(partenza-7.5) + b2*(partenza-7.5)^2 +
#           + b3*D + b4*D*(partenza-7.5)^2
# Derivative:
# durata'= b1 + b2*2*(partenza-7.5) + b4*2*D(partenza-7.5)

C <- t(c(0,1,0,0,0))

c((C%*%coefficients(fit2)) - sqrt((C%*%vcov(fit2)%*%t(C))) * qt(1-0.05/(2*k), n-5),
  (C%*%coefficients(fit2)) + sqrt((C%*%vcov(fit2)%*%t(C))) * qt(1-0.05/(2*k), n-5))

#_____________________________________________________________________
###### Other excercises
######------------------


#_____________________________________________________________________
##### Problem 3 of 10/09/2010
#####--------------------------
# The file extra.txt reports the representation expenses [$] of the
# English first minister and of his vice during the first 12 months of
# 2009. Assume those data to be independent realizations of a bivariate
# Gaussian.
# a) Build an ellipsoidal region of confidence 90% for the mean of the
#    representation expenses
# b) Is there evidence of the fact that the prime minister spends in mean
#    more than twice the expences of its vice?
# c) Build a confidence interval of level 90% for the mean of the sum of
#    the expenses.

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

plot(extra, asp=1)
extra

### question a)
mcshapiro.test(extra)

n <- dim(extra)[1]
p <- dim(extra)[2]

x.mean <- sapply(extra,mean)
x.cov <- cov(extra)
x.inv <- solve(x.cov)

cfr.fisher <- (n-1)*p/(n-p)*qf(1-alpha,p,n-p)

ellipse(center=x.mean, shape=x.cov/n, radius=sqrt(cfr.fisher), lwd=2)
# centre:
x.mean
# direction of the axes
```

Gaussian

repr. of the ellipse

```
eigen(x.cov)$vector[,1]
eigen(x.cov)$vector[,2]
# radius
sqrt(cfr.fisher)
# lenght of the semi-axes
sqrt(eigen(x.cov/n)$values)*sqrt(cfr.fisher)

### question b)
# Test:   H0: mu1<=2*mu2   vs H1: mu1>2*mu2
# i.e. H0: mu1-2*mu2<=0 vs H1: mu1-2*mu2>0
# i.e H0: a'mu<=0      vs H1: a'mu>0 con a=c(1,-2)

a <- c(1,-2)
delta.0 <- 0

extra <- as.matrix(extra)
t.stat <- (mean(extra %*% a) - delta.0 ) / sqrt( var(extra %*% a) / n ) # t-statistics (statistica 1!)

# Reject for large values of t
# => compute the p-value as the probability of the right tail
# (i.e., of values >tstat)
P <- 1-pt(t.stat, n-1)
P

### question c)
a2 <- c(1,1)
alpha <- .1
cfr.t <- qt(1-alpha/2, n-1)

c(inf = mean(extra %*% a2) - cfr.t * sqrt( var(extra %*% a2) / n ),
  center = mean(extra %*% a2),
  sup = mean(extra %*% a2) + cfr.t * sqrt( var(extra %*% a2) / n ))

# Otherwise one can use the function t.test()
lc <- extra[,1] + extra[,2]
t.test(lc, alternative = 'two.sided', mu = 0, conf.level = 0.90)

#_____________________________________________________________
# Pb 2 of 26/02/2008
# The Warmer Bros produces gas boilers for domestic heating that
# exploit the chemical reaction of combustion CH4 + 2O2 -> CO2 + 2H2O.
# For 40 Warmer Bros boilers were measured (gas.txt file) the quantities
# [kmol] of H2O, CO2 and CO contained in 1 m^3 of exhaust gas.
# a) Build three T2-simultaneous confidence intervals of global level
#    90% for the mean of the three exhaust gas.
# The correct operation of the boilers assumes that the number of H2O
# kmol is twice those of CO2 and that the number of kmol of CO is equal
# to 0.
# b) Is there statistical evidence that the Warmer Bros boilers may not
#    work properly?
# c) Maintaining a global confidence of 90%, together with the intervals
#    built in (a), provide additional T2-simultaneous intervals in
#    support of the conclusions at point (b).

gas <- read.table('gas.txt')
n <- dim(gas)[1]
p <- dim(gas)[2]

M <- sapply(gas,mean)
M
S <- cov(gas)
Sinv <- solve(S)

# question a)
# a) Build three T2-simultaneous confidence intervals of global level
#    90% for the mean of the three exhaust gas.
mcshapiro.test(gas)

cfr.fisher <- ((n-1)*p/(n-p))*qf(1-alpha,p,n-p)

ICT2 <- data.frame(L=M-sqrt(diag(S)/n)*sqrt(cfr.fisher),C=M,U=M+sqrt(diag(S)/n)*sqrt(cfr.fisher))
ICT2

# question b)
# b) Is there statistical evidence that the Warmer Bros boilers may not
#    work properly?

# H0:          mu3  =0         vs      H1: !H0
#     m1-2mu2       =0

# equivalent to
# H0: C%*%mu=0 vs H1: C%*%mu!=0
C <- rbind(c(0,0,1),c(1,-2,0))
q <- dim(C)[1]
delta0 <- c(0,0)

Md <- C %*% M
```

```r
Sd <- C %*% S %*% t(C)
Sdinv <- solve(Sd)

T2 <- n * t(Md-delta0) %*% Sdinv %*% (Md-delta0)
pvalue <- 1- pf(T2*(n-q)/((n-1)*q),q,n-q)
pvalue


# question c)
# c) Maintaining a global confidence of 90%, together with the intervals
#    built in (a), provide additional T2-simultaneous intervals in
#    support of the conclusions at point (b).

alpha <- 0.1
cfr.fisher <- q*(n-1)/(n-q)*qf(1-alpha,q,n-q)

ICT2.bis <- data.frame(L=Md-sqrt(cfr.fisher*diag(Sd)/n),C=Md,U=Md+sqrt(cfr.fisher*diag(Sd)/n))
ICT2.bis

#_____________________________________________________________________
##### Problem 2 of 12/02/2008
#####--------------------------
# PoliTermos produces thermostats for the Italian market. During the last month
# some faulty thermostats have been introduced accidentally on the market;
# it is estimated that these are about 10% of sales. Lab test demonstrate that,
# in a 1°C temperature environment, the not defective thermostats detect a
# temperature normally distributed, with mean 1 and variance 1/2, while
# the defective thermostats according to an exponential law with mean 1 and
# variance 1. All the thermostats sold last month were recalled for possible
# replacement and subjected to the previous test. Taking into account that the
# replacement of a thermostat in reality not defective produces damage to the company
# of 1???, and that the failure to replace a faulty thermostat creates a loss in
# reputation estimated in 9???:
# a) Formulate a criterion for the replacement, based on the temperature measured
#    during the test, which minimizes the damage expected by the company.
# Two thermostats of your property were recalled for inspection.
# The first during the experiment revealed a temperature of -1°C, the
# second equal to 3°C:
# b) Will the first thermostat be replaced? How likely is it defective?
# c) Will the second thermostat be replaced? How likely is it defective?



# a) Formulate a criterion for the replacement, based on the temperature measured
#    during the test, which minimizes the damage expected by the company.
# Analytical solution:
# Impose c(D|N)*pN*fN(x)-c(N|D)*pD*fD(x)<0
# with: c(D|N)=1, c(N|D)=9, pN=.9, pD=.1
R <- c(1.5-.5*sqrt(5-2*log(pi)),1.5+.5*sqrt(5-2*log(pi)))
R

x <- seq(-2,10,by=.01)
plot(x, 1*0.9*dnorm(x, 1, sqrt(.5)), type='l', xlim=c(-2,10), ylim=c(-.3,1.5),
     ylab='y')
lines(x, 9*.1*dexp(x, 1), type='l', col=1, lty=2)
abline(v=R, lty=1)
abline(v=R, col='grey', lty=3)
abline(h=-.03, col='grey')
segments(R[1],-.03,R[2],-.03, col='blue')
segments(0,-.03,R[1],-.03, col='red')
segments(R[2],-.03,10,-.03, col='red')
segments(-2,-.03,0,-.03, col='blue')

# The first during the experiment revealed a temperature of -1°C, the
# second equal to 3°C:
# b) Will the first thermostat be replaced? How likely is it defective?
# c) Will the second thermostat be replaced? How likely is it defective?

abline(v=-1, col='blue', lty=4)
abline(v=3, col='red', lty=4)

x <- c(-1,3)
pD <- .1
pN <- .9
pdif <- pD*dexp(x, 1)/(pN*dnorm(x, 1, sqrt(.5))+pD*dexp(x, 1))

pdif


#_____________________________________________________________________
##### Problem 2 of 10/02/10
# The new fastfood chain Megaburger has made an experiment last month
# to choose the characteristics of his first advertising campaign. During the
# experiment 450 individuals were involved, selected in three different macro-regions
# (Europe, USA, Canada). They were asked to evaluate one of the following
# three types of burgers: Burger, Cheese-burger, Bacon-cheese-burger. The mmm.txt
# file contains the assessments (index of goodness from 0 to 10) for 450
```

ANOVA 2 ways

```
# individuals.
# a) Using an additive ANOVA model with two factors, perform three tests
#    (each of level 1%) to verify that the distribution of index of goodness
#    (i) is not dependent on the macro-region, (ii) does not depend on the type
#    of burger, (iii) is not dependent on the macro-region or on the type of hamburger.
# b) Using six overall 95% confidence intervals for appropriate differences of means,
#    identify homogeneous groups of customers (in terms of the mean index of the goodness).
# c) Based on the above confidence intervals, identify which group (or groups)
#    of customers is on average more satisfied and which less satisfied.

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

### question a)
# Modello additivo:
# X.ijk = mu + tau.i + beta.j + eps.ijk;
#     eps.ijk~N(0,sigma^2),
#     i=1,2 (effetto regione), j=1,2,3 (effetto tipo)

fit <- aov(index ~ region + sandwich, mmm)
summary(fit)

### question b)
alpha <- 0.05
g <- 3
b <- 3
p <- 1
n <- 50
N <- n*g*b

W <- sum(fit$residuals^2)

qT <- qt(1 - alpha / (2 * ( g*(g-1)/2*p + b*(b-1)/2*p )), g*b*n-g-b+1)

mE  <- mean(mmm[mmm$region=='Europe',1])
mC  <- mean(mmm[mmm$region=='Canada',1])
mU  <- mean(mmm[mmm$region=='USA',1])

mBurger  <- mean(mmm[mmm$s=='Burger',1])
mBacon  <- mean(mmm[mmm$s=='Bacon',1])
mCheese  <- mean(mmm[mmm$s=='Cheese',1])

infEC <- mE-mC - qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
supEC <- mE-mC + qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
infEU <- mE-mU - qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
supEU <- mE-mU + qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
infUC <- mU-mC - qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
supUC <- mU-mC + qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )

infBaconBurger <- mBacon-mBurger - qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
supBaconBurger <- mBacon-mBurger + qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
infCheeseBurger <- mCheese-mBurger - qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
supCheeseBurger <- mCheese-mBurger + qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
infCheeseBacon <- mCheese-mBacon - qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )
supCheeseBacon <- mCheese-mBacon + qT * sqrt( W/(g*b*n-g-b+1) * (1/150+1/150) )


IC2   <- data.frame(EC=c(infEC, supEC), EU=c(infEU, supEU), UC=c(infUC, supUC),
                    BaconBurger=c(infBaconBurger, supBaconBurger), CheeseBurger=c(infCheeseBurger, supChee
seBurger), CheeseBacon=c(infCheeseBacon, supCheeseBacon))
rownames(IC2) <- c('Inf','Sup')
IC2

matplot(1:(dim(IC2)[2]), t(IC2),pch='',axes=F,xlab='',ylab='')
for(i in 1:dim(IC2)[2])
  segments(i,IC2[1,i],i,IC2[2,i],lwd=2)
axis(1,at=1:dim(IC2)[2],labels=names(IC2), las=2, cex.axis=.8)
axis(2)
box()
abline(h=0)
points(1:(dim(IC2)[2]), sapply(IC2,mean),col='red',lwd=2)

dev.off()

-------------------------------------------------------------------------------------------

setwd("D:/RTDA/Didattica/Applied Statistics MATE 19-20/Lab 13 - 05062020")
load("D:/RTDA/Didattica/Applied Statistics MATE 19-20/Lab 5 - 16042020/mcshapiro.test.RData")

# Problem 1 of 13.09.12
# Nel file library.txt sono riportate per le 23 biblioteche comunali milanesi
# le spese (riferite all'anno 2011) relative all'acquisto libri, alla retribuzione
# del personale, al consumo di energia elettrica e infine ad altri costi. Assumendo
# iid normali i dati relativi alle 23 biblioteche:
# a) si forniscano 5 intervalli di confidenza globale 90% per la media delle quattro
# voci di spesa e per la loro somma;
# b) si forniscano 5 intervalli di confidenza globale 90% per la deviazione standard
# delle quattro voci di spesa e per la loro somma;
```

```
# c) si confermi/smentisca l'ipotesi secondo la quale la spesa media in libri copre
# la metÃ  della spesa totale.


library=read.table('library.txt', header=T)
head(library)

# a) si forniscano 5 intervalli di confidenza globale 90% per la media delle quattro
# voci di spesa e per la loro somma;
mcshapiro.test(library)

n <- dim(library)[1]
p <- dim(library)[2]

x.mean   <- sapply(library,mean)
x.cov    <-  cov(library)

alpha   <- .1
k <- 5
cfr.t <- qt(1-alpha/(2*k),n-1)
A=rbind(diag(rep(1,4)),
        c(1,1,1,1))


IC.BF <- cbind( A%*%x.mean-cfr.t*sqrt(diag(A%*%x.cov%*%t(A))/n) ,
                A%*%x.mean,
                A%*%x.mean+cfr.t*sqrt(diag(A%*%x.cov%*%t(A))/n) )

colnames(IC.BF)=c('Inf','Center', 'Sup')
IC.BF

# b) si forniscano 5 intervalli di confidenza globale 90% per la deviazione standard
# delle quattro voci di spesa e per la loro somma;
k <- 5
qCinf <- qchisq(1 - alpha / (2*k), n-1)
qCsup <- qchisq(alpha / (2*k), n-1)

BF    <- cbind(inf=diag(A%*%x.cov%*%t(A)) * (n-1) / qCinf,
               center=diag(A%*%x.cov%*%t(A)),
               sup=diag(A%*%x.cov%*%t(A)) * (n-1) / qCsup)
sqrt(BF)

# c) si confermi/smentisca l'ipotesi secondo la quale la spesa media in libri copre
# la metÃ  della spesa totale.
A=cbind( .5,-.5,-.5,-.5)
lcomb=as.matrix(library)%*%t(A)
t.test(x = lcomb, conf.level = 1-.1)

cfr.t <- qt(1-alpha/2,n-1)
IC.BF <- cbind( A%*%x.mean-cfr.t*sqrt(diag(A%*%x.cov%*%t(A))/n) ,
                A%*%x.mean,
                A%*%x.mean+cfr.t*sqrt(diag(A%*%x.cov%*%t(A))/n) )
IC.BF

#_____________________________________________________________________
# Problem 4 of 29.06.10
# Nel file energy.txt sono riportati i consumi elettrici istantanei [MW] allo
# scoccare di ogni ora nella citta' di Belfast, per il mese di Giugno. Assumendo
# le misurazioni indipendenti e normalmente distribuite con pari varianza e media
# mu = A.g +B.g*[1-cos(2pi/24*t)] dipendente dall'ora di misurazione t
# (t = 0, 1, 2, ... , 23) e dalla natura infrasettimanale o meno del giorno di
# misurazione (g in {"Lun-Ven", "Sab-Dom"}).
# a) Si stimino i parametri del modello.
# b) Si fornisca, se possibile, un opportuno modello ridotto, se ne giustifichi
#    la scelta e se ne interpretino i parametri.
# c) Sulla base del modello ridotto (b), si forniscano degli intervalli di
#    confidenza globale 90% per la media e la varianza dei consumi raggiunti
#    in data odierna (martedi' 29 giugno 2010) alle ore 12:00 ed alle ore 18:00.
# d) Sulla base del modello ridotto (b), vi e' evidenza che di venerdi' il consumo
#    medio massimo sia piu' del doppio del consumo medio minimo?
# e) Sulla base del modello ridotto (b), vi e' evidenza che di sabato il consumo
#    medio massimo sia piu' del doppio del consumo medio minimo?

energy=read.table('energy.txt', header = T, sep='\ ')
head(energy)

# Build the data matrix
t = rep(0:23,28)
y=NULL
for(i in 1:dim(energy)[1])
  y=c(y,as.numeric(energy[i,-1]))
tmp = rep(energy$giorno.della.settimana, each=(dim(energy)[2]-1))
D=ifelse(tmp %in% c('sabato', 'domenica'), 0, 1)

data = data.frame(y,t,D)

# Estimate the model parameters
```

7

```
fit = lm(y ~ D + I(1-cos(2*pi/24*t)) + I(D*(1-cos(2*pi/24*t))), data = data)
summary(fit)

# Reduce the model
shapiro.test(residuals(fit))
par(mfrow=c(2,2))
plot(fit)

A=rbind(c(0,1,0,0),
        c(0,0,0,1))
b=c(0,0)

linearHypothesis(fit, A, b)

# Reduced model
fit2 = lm(y ~ I(1-cos(2*pi/24*t)) + I(D*(1-cos(2*pi/24*t))), data = data)
summary(fit2)
shapiro.test(residuals(fit2))
par(mfrow=c(2,2))
plot(fit2)

# Confidence intervals
k=3
alpha=.1
new.data=data.frame(t=c(12,18), D=1)
predict(fit2, new.data, level = 1-alpha/k, interval = 'confidence')

sigma2=sum(residuals(fit2)^2)/669

qCinf <- qchisq(1 - alpha / (2*k), 669)
qCsup <- qchisq(alpha / (2*k), 669)

BF    <- cbind(inf=sigma2 * 669 / qCinf,
               center=sigma2 ,
               sup=sigma2 * 669 / qCsup)
BF

# question d)
# Having derived A+(B+C)*[1-cos(2pi/24*t)] we obtain t=0 and t=12, with
A=rbind(c(1,0,0),
        c(1,2,2))
A%*%as.vector(coefficients(fit2))
#t=0 is min, t=12 is max

# Test H0: 2(B+C)-A<0 vs H1: 2(B+C)-A>0
A=t(as.vector(c(-1,2,2)))
A%*%as.vector(coefficients(fit2))
A%*%vcov(fit2)%*%t(A)
n=dim(data)[1]

1-pt(A%*%as.vector(coefficients(fit2))/sqrt(A%*%vcov(fit2)%*%t(A)), n-3)
# Reject H0

# question e)
# Having derived A+B*[1-cos(2pi/24*t)] we obtain t=0 and t=12, with
A=rbind(c(1,0,0),
        c(1,2,0))
A%*%as.vector(coefficients(fit2))
#t=0 is min, t=12 is max

# Test H0: 2B-A<0 vs H1: 2B-A>0
A=t(as.vector(c(-1,2,0)))
A%*%as.vector(coefficients(fit2))
A%*%vcov(fit2)%*%t(A)
n=dim(data)[1]

1-pt(A%*%as.vector(coefficients(fit2))/sqrt(A%*%vcov(fit2)%*%t(A)), n-3)
# Reject H0
```