Prova1

Gustavo Alovisi

06/12/2020

Exercício 1

a) Obtenha os vetores coluna de diferenças. A partir deles, obtenha a matriz S de covariânciase a matriz R de correlações amostrais.

Temos que

$$X = \begin{bmatrix} y_1 & y_2 \\ 1 & 1 \\ 2 & 2 \\ 3 & 4 \\ 4 & 9 \end{bmatrix}$$

Seja o vetor de médias $\overline{x^T} = [\frac{1+2+3+4}{4}; \frac{1+2+4+9}{4}] = [\frac{5}{2}; 4]$.

O vetor de diferenças para $\mathbf{y}_1$ e $\mathbf{y}_2$ é calculado através de

$$\mathbf{d}_i = \mathbf{y}_i - \overline{x_i}\mathbf{1}$$

de forma que

$$\mathbf{d}_1 = \begin{bmatrix} 1 - 2.5 \\ 2 - 2.5 \\ 3 - 2.5 \\ 4 - 2.5 \end{bmatrix} = \begin{bmatrix} -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \end{bmatrix}$$

$$\mathbf{d}_2 = \begin{bmatrix} 1 - 4 \\ 2 - 4 \\ 4 - 4 \\ 9 - 4 \end{bmatrix} = \begin{bmatrix} -3 \\ -2 \\ 0 \\ 5 \end{bmatrix}$$

.

O comprimento ao quadrado é dado por $L_{di}^2 = \mathbf{d}_i^T\mathbf{d}_i$. Assim,

$$L_{d1}^2 = (-1.5)^2 + (-0.5)^2 + 0.5^2 + 1.5^2 = 5$$
$$L_{d2}^2 = (-3)^2 + (-2)^2 + 0^2 + 5^2 = 38$$

Ainda, com $n = 4$,

$$(n-1)S_{11} = L_{d_1}^2 \implies S_{11} = 5/3$$
$$(n-1)S_{22} = L_{d_2}^2 \implies S_{22} = 38/3$$

Temos que

$$\mathbf{d}_1^T\mathbf{d}_2 = (-1.5)(-3) + (-0.5)(-2) + 1.5 * 5 = 13$$

$$(n-1)S_{12} = \mathbf{d}_1^T \mathbf{d}_2 \implies S_{12} = 13/3$$

Assim, S se dá por

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{bmatrix} = \begin{bmatrix} 5/3 & 13/3 \\ 13/3 & 38/3 \end{bmatrix}$$

Obtemos R_{12} através de

$$R_{12} = \cos(\theta) = \frac{S_{12}}{\sqrt{S_{11}}\sqrt{S_{22}}} = \frac{13/3}{\sqrt{5/3}\sqrt{38/3}} = 0.943$$

E portanto a matriz R é

$$R = \begin{bmatrix} 1 & 0.943 \\ 0.943 & 1 \end{bmatrix}$$

b) Calcule a medida de variância generalizada a partir de S e R e mostre a relação entre elas.

A variância generalizada (VG) é dada por $VG_S = |S|$ e $VG_R = |R|$.

Assim,

$$VG_S = \begin{vmatrix} 5/3 & 13/3 \\ 13/3 & 38/3 \end{vmatrix} = 2.33$$

$$VG_R = \begin{vmatrix} 1 & 0.943 \\ 0.943 & 1 \end{vmatrix} = 0.11$$

A relação entre VG_S e VG_R é dada através de $VG_S = (S_{11}...S_{pp})VG_R$. Assim, para este caso temos $VG_S = 2.33 = (\frac{5}{3} * \frac{38}{3})VG_R$.

c) Calcule a variância total utilizando a matriz S. Comente as diferenças entre as duas medidas.

A variância total (VT) de S é simplesmente a soma da diagonal principal de S. Assim, $VT_S = S_{11} + S_{22} = \frac{5}{3} + \frac{38}{3} = 14.33$.

A Variância Total é a soma das distâncias ao quadrado dos p vetores de desvios. Porém, ao contrário da VG, ela não considera a orientação (a estrutura de correlação) dos vetores de resíduos.

Exercício 2

Seja $\mathbf{X} \sim N_3(\mu, \Sigma)$ onde

$$\mu^T = [-3, 1, 4]$$
$$\Sigma = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

a) Escreva genericamente a forma quadrática d^2:

Podemos escrever d^2 como

$$d^2 = \sum_{i=1}^{p} \frac{1}{\lambda_i}[(x-\mu)^T \mathbf{e}_i]^2$$

onde $(\lambda_i, \mathbf{e}^i)$ são os autopares de Σ.

b) Encontre o comprimento dos eixos do elipsoide formado por d^2, com 0.95 de probabilidade (Encontre os autovalores no R usando a função eigen).

Sabemos que $(x-\mu)^T\Sigma^{-1}(x-\mu) \sim \chi^2_p$. Queremos encontrar cš que contemple 95% de nossos dados tal que $(x-\mu)^T\Sigma^{-1}(x-\mu) = c^2 \sim \chi^2_3$. Assim, c^2 é o quantil da Chi² tal que $P(\chi^2_3 \leq c^2) = 0.95$.

O comprimento dos p-eixos a partir de μ são calculados como $comp_i = \pm c\sqrt{\lambda_i}\mathbf{e}_i$.

Assim, vamos primeiro encontrar os autopares de Σ:

```
sigma <- cbind(c(1,-2,0),c(-2,5,0),c(0,0,2))
mu <- c(-3, 1, 4)

autopares <- eigen(sigma)
autopares
```

```
## eigen() decomposition
## $values
## [1] 5.8284271 2.0000000 0.1715729
##
## $vectors
##            [,1] [,2]      [,3]
## [1,] -0.3826834    0 0.9238795
## [2,]  0.9238795    0 0.3826834
## [3,]  0.0000000    1 0.0000000
```

Precisamos ainda encontrar c^2 tal que $P(\chi^2_3 \leq c^2) = 0.95$.

Este valor é simplesmente o quantil da Chi-quadrado com 95% de confiança e 3 graus de liberdade:

```
qchisq(0.95, 3)
```

```
## [1] 7.814728
```

Assim, podemos encontrar os eixos da elipse:

$$comp_1 = \sqrt{7.814}\sqrt{5.828}\begin{bmatrix} -0.382 \\ 0.923 \\ 0 \end{bmatrix} = \pm 6.74$$

$$comp_2 = \sqrt{7.814}\sqrt{2}\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \pm 3.95$$

$$comp_3 = \sqrt{7.814}\sqrt{0.171}\begin{bmatrix} 0.923 \\ 0.382 \\ 0 \end{bmatrix} = \pm 1.15$$

c) Encontre a distribuição do vetor aleatório $\mathbf{Y}^T = (Y1, Y2)$, onde $Y1 = \frac{(X1+X2)}{2}$ e $Y2 = 2X1 - X2 + X3$:

Através do Resultado (4.3) do Wichern, sabemos que se $\mathbf{X} \sim N_p(\mu, \Sigma)$, as q combinações lineares seguem $AX \sim N_q(A\mu, A\Sigma A^T)$.

Assim, temos que nossa matriz A é dada por

$$A = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 2 & -1 & 1 \end{bmatrix}$$

e podemos reescrever Y como AX de forma que

$$AX = \begin{bmatrix} 1/2X_1 + 1/2X_2 + 0X_3 \\ 2X_1 - 1X_2 + 1X_3 \end{bmatrix}$$

. Assim, podemos encontrar $A\mu$:

$$\mu^T = [-3, 1, 4]$$

$$A\mu = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} -3 \\ 1 \\ 4 \end{bmatrix} = \begin{bmatrix} -1 \\ -3 \end{bmatrix}$$

E também $A\Sigma A^T$:

$$A\Sigma A^T = \begin{bmatrix} 1/2 & 1/2 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1/2 & 2 \\ 1/2 & -1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1/2 & -5/2 \\ -5/2 & 19 \end{bmatrix}$$

Exercício 3

```r
#carregando dados
require(mvShapiroTest)
```

```
## Loading required package: mvShapiroTest
```

```
## Warning: package 'mvShapiroTest' was built under R version 4.0.3
```

```r
require(carData)
```

```
## Loading required package: carData
```

```r
x=carData::Anscombe
#x1: education (Gasto per capita com educação)
#x2: income (Renda per capita)
#x3: young (Proporção da pop abaixo de 18 anos)
#x4: urban (Proporção d pop na área urbana)
```

a) Aplique o teste de Shapiro de normalidade univariada e bivariada. Informe o valor p decada teste e conclua a respeito.

```r
#############Teste de Normalidade Univariada
W <- rep(0,ncol(x))

for(k in 1:4){
        W[k]=shapiro.test(x[,k])$p.value }

print(W)
```

```
## [1] 0.016982037 0.693875450 0.001717814 0.849336144
```

```r
#[1] 0.016982037 0.693875450 0.001717814 0.849336144
#primeiro e terceiro rejeita a hipotese de normalidade univariada
```

Aplicando o Teste de Shapiro Univariado para x1, x2, x3 e x4, observamos pelo p-valor que não há indícios para rejeitar H0: os dados seguem uma distribuição normal univariada, para x2 (income) e x4 (urban). Já para x1 (education) e x3 (young), rejeitamos H0.

```r
#############Teste de Normalidade Multivariada
x <- as.matrix(x)
mvShapiro.Test(x)  ## teste Shapiro
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  x
## MVW = 0.97308, p-value = 0.2411
```

```r
#p-value = 0.2411
#Parece não haver indicios a 5% para rejeitar a hipotese de que os dados seguem uma normal 4 multivaria
```

Aplicando o teste de Shapiro Multivariado, não temos indícios para rejeitar a hipótese de que os dados seguem uma distribuição normal multivariada a 5%.

b) Teste a hipótese a 5% de que o vetor de média populacional seja $\mu^T = [210,3225,360,665]$.Informe o valor da estatística de teste, o valor crítico e conclua a respeito.

Para o Teste de Uma Média, como não temos a matriz de covariância populacional, devemos utilizar a matriz de covariância S dos dados. Assim, devemos calcular T^2 tal que $T^2 \sim$ chi² ajustada.

```r
p <- 4
alpha <- 0.05
n <- nrow(x)
x_bar <- colMeans(x)
S <- cov(x)

T2_cal<-n*mahalanobis(x_bar, c(210,3225, 360, 665), S, inverted = FALSE)
T2_cal
```

```
## [1] 12.69119
```

```r
q <- qf(1-alpha,p,n-p)*((n-1)*p)/(n-p)
q
```

```
## [1] 10.93421
```

Como $T^2 = 12.69119 > q = 10.93421$, concluímos que com significância de 5%, rejeitamos H0: $\mu^T =$ [210,3225,360,665].

c) Para cada variável construa os intervalos T e bonferroni e conclua a respeito.

```
alpha = 0.05
##Intervalos T
Ls = c()
Li = c()
for (i in 1:p) {
        Ls[i]=(x_bar[i])+sqrt(q*S[i,i]/n)
        Li[i]=(x_bar[i])-sqrt(q*S[i,i]/n)

}

Lim = rbind(Ls,Li)
Lim
```

```
##        [,1]     [,2]     [,3]     [,4]
## Ls 217.8235 3484.603 369.9805 734.5869
## Li 174.8039 2965.986 347.7921 594.4327
```

```
##Intervalos de Bonferroni
Lsb = c()
Lib = c()
for (i in 1:p) {
        Lsb[i]=(x_bar[i])+qt(1-(alpha/(2*p)),n-1)*sqrt(S[i,i]/n)
        Lib[i]=(x_bar[i])-qt(1-(alpha/(2*p)),n-1)*sqrt(S[i,i]/n)

}

Limb = rbind(Lsb,Lib)
Limb
```

```
##         [,1]     [,2]     [,3]     [,4]
## Lsb 213.1698 3428.501 367.5802 719.4256
## Lib 179.4576 3022.088 350.1923 609.5940
```

Como resultado, os ICs de Bonferroni mantém a confiança global de $1 - \alpha$ e possuem menor amplitude que os ICs T^2. Em ambas medidas, $\mu^T = [210, 3225, 360, 665]$ está dentro dos intervalos.

Exercício 4 - Um programa de reforço de estudos foi avaliado a partir de uma amstra de 15 estudantes. Cada um deles realizou provas com conteúdos de matemática, física e química antes e depois de serem submetidos ao programa de reforço.

a) Aplique o teste de Shapiro de normalidade multivariada. Informe o valor p de cada teste e conclua a respeito.

Carregando os dados:

```
mat_a=c(6.62, 5.53, 6.27, 5.49, 5.26, 6.46, 2.64, 3.89, 6.23, 4.36,
        6.08, 5.70, 6.91, 4.70, 5.14)
fis_a=c(6.08, 6.82, 5.68, 4.80, 6.62, 5.86, 7.26, 4.49, 6.44, 6.32,
```

```r
        6.28, 7.07, 5.58, 5.76, 6.22)
qui_a=c(5.43, 6.39, 5.41, 6.10, 4.71, 5.22, 5.63, 7.27, 5.92, 3.34,
        6.96, 7.77, 4.37, 4.80, 6.33)

mat_d=c(8.00, 6.87, 6.12, 5.49, 5.74, 5.94, 4.87, 7.92, 6.79, 9.59,
        5.49, 8.34, 8.29, 6.77, 7.66)
fis_d=c(6.18, 7.82, 4.68, 4.00, 3.62, 5.86, 8.26, 4.30, 6.44, 8.32,
        6.28, 7.07, 5.58, 6.76, 6.00)
qui_d=c(8.06, 8.43, 6.39, 7.14, 6.78, 6.83, 7.93, 7.31, 9.11, 7.62,
        7.72, 8.32, 9.12, 7.18, 8.98)

x=cbind(mat_a,fis_a,qui_a,mat_d,fis_d,qui_d)
```

Calculando a Diferença (d) das amostras pareadas:

```r
d_mat=mat_d-mat_a
d_fis=fis_d-fis_a
d_qui=qui_d-qui_a
d<-cbind(d_mat,d_fis,d_qui)
```

Realizando o Teste de Shapiro de Normalidade Multivariada:

```r
require(mvShapiroTest)
mvShapiro.Test(d)
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  d
## MVW = 0.91525, p-value = 0.08633
```

Podemos notar que a significância de 5%, não rejeitamos H0: d ~ Normal Multivariada. Assim, podemos continuar a inferência e realizar o teste para a diferença de média de duas amostras pareadas.

b) O programa de reforço teve influência nas notas? Teste com 1% de significância. Informe o valor da estatística de teste, o valor crítico e conclua a respeito:

Para testar a diferença de médias de duas amostras pareadas, testamos $H0 : \mu_D = \phi_0 \ vs \ H1 : \mu_D \neq \phi_0$. Verificamos se um vetor de médias amostrais p-variado d^a é oriundo de $N_p(\phi_0, \Sigma_D)$ utilizando $T\check{s}_{cal} = n(d^a - \phi_0)^T S_D^{-1}(d^a - \phi_0)$, onde S_D é o estimador não viciado de Σ_D.

```r
Sd <- cov(d)
n <- nrow(d)
p <- ncol(d)
d_bar <- colMeans(d)
alpha = 0.01
##############################testando hiptese de diferenca igual a zero#############
mu=c(0,0,0)
T2_cal<-n*mahalanobis(d_bar, mu, cov(d), inverted = FALSE)
T2_cal
```

```
## [1] 54.44693
```

```r
q <- ((n-1)*p)/(n-p)*qf(1-alpha,p,n-p)
q
```

```
## [1] 20.83391
```

Como $T^2_{cal} > q$, rejeitamos a 1% a hipótese H0 explicitada acima: há indícios que existe uma diferença multivariada nas médias com/sem reforço. Ou seja, o programa de reforço melhorou a média conjunta dos alunos.

c) Construa os intervalos T e bonferroni e conclua a respeito.

```
###########Intervalos T
Ls=c()
Li=c()
for (i in 1:p) {
        Ls[i]=(d_bar[i])+sqrt(q*Sd[i,i]/n)
        Li[i]=(d_bar[i])-sqrt(q*Sd[i,i]/n)

}


Lim=rbind(Ls,Li)
colnames(Lim) =c("Mat", "Fis", "Qui")
Lim
```

```
##           Mat       Fis       Qui
## Ls  3.4714207  1.319069 3.6508727
## Li -0.4580873 -1.333736 0.5184606
```

```
############Intervalos Univariados Bonferroni
Lsb=c()
Lib=c()
for (i in 1:p) {
        Lsb[i]=(d_bar[i])+qt(1-(alpha/(2*p)),n-1)*sqrt(Sd[i,i]/n)
        Lib[i]=(d_bar[i])-qt(1-(alpha/(2*p)),n-1)*sqrt(Sd[i,i]/n)

}

Limb=rbind(Lsb,Lib)
colnames(Limb) =c("Mat", "Fis", "Qui")
Limb
```

```
##             Mat       Fis       Qui
## Lsb  3.02598045  1.018353 3.2957895
## Lib -0.01264711 -1.033020 0.8735438
```

Observando ambos os intervalos T e Bonferroni, apenas o intervalo [,3] (Química) não contém o $d^a = 0$. Ou seja, a 1% apenas o reforço de Química possuiu um efeito significativo nas notas dos alunos, apesar de que ao analisar o programa como um todo (multivariado), o reforço ainda possui um efeito significativo nas notas.

Exercício 5

```
x <- mtcars
#Fator = forma do motor
#vs=0 V-shaped
#vs=1 Straight
# quereos buscar as variáveis mpg, hp, wt, qsec
X <- x[,c(1,4,6,7,8)]
```

```r
head(X)
```

```
##                    mpg  hp    wt  qsec vs
## Mazda RX4         21.0 110 2.620 16.46  0
## Mazda RX4 Wag     21.0 110 2.875 17.02  0
## Datsun 710        22.8  93 2.320 18.61  1
## Hornet 4 Drive    21.4 110 3.215 19.44  1
## Hornet Sportabout 18.7 175 3.440 17.02  0
## Valiant           18.1 105 3.460 20.22  1
```

```r
nrow(X)
```

```
## [1] 32
```

a) Aplique o teste de Shapiro de normalidade multivariada dentro de cada grupo (motores em V e convencionais). Informe o valor p de cada teste e conclua a respeito. Em caso de não normalidade, verifique se os dados transformados (via logaritmo natural) aderem a hipótese de normalidade. Neste caso realizar as análises subsequentes utilizando os dados transformados.

```r
require(dplyr)
```

```
## Loading required package: dplyr
```

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

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

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

```r
X1 <- X %>% filter(vs == 0) %>% select(!vs)
X2 <- X %>% filter(vs == 1) %>% select(!vs)

mvShapiro.Test(as.matrix(X1))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(X1)
## MVW = 0.95159, p-value = 0.5432
```

```r
mvShapiro.Test(as.matrix(X2))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(X2)
## MVW = 0.87951, p-value = 0.003922
```

Realizando o Tesde de Shapiro Multivariado para os grupos, notamos que no caso em que $vs = 0$, não rejeitamos a hipótese de normalidade multivariada. Já para $vs = 1$, rejeitamos. Vamos então aplicar a transformação de ln() em nossos dados e testar a normalidade multivariada:

```
X1 <- X1 %>% apply(2,log)
X2 <- X2 %>% apply(2,log)
mvShapiro.Test(as.matrix(X1))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(X1)
## MVW = 0.94314, p-value = 0.3063
```

```
mvShapiro.Test(as.matrix(X2))
```

```
##
##  Generalized Shapiro-Wilk test for Multivariate Normality by
##  Villasenor-Alva and Gonzalez-Estrada
##
## data:  as.matrix(X2)
## MVW = 0.91987, p-value = 0.1296
```

Após a realização da transformação, os dados de ambos os grupos aderem a 5% a H0: provém de uma distribuição normal multivariada.

b) Existe a 5% diferença significativa nas variáveis analisadas entre os 2 tipos de motores? Informe o valor da estatística de teste, o valor crítico e conclua a respeito.

Vamos realizar o teste para a diferença de duas médias multivariadas com amostras independentes para $\mu_1 - \mu_2 = 0$. Utilizaremos a função hotelling.test do pacote Hotelling:

```
require(Hotelling)
```

```
## Loading required package: Hotelling

## Warning: package 'Hotelling' was built under R version 4.0.3

## Loading required package: corpcor

## Warning: package 'corpcor' was built under R version 4.0.3
```

```
#X1 <- X %>% filter(vs == 1) %>% select(!vs) %>% apply(2,log)
#X2 <- X %>% filter(vs == 0) %>% select(!vs) %>% apply(2, log)
X_test <- X %>% select(!vs) %>% apply(2, log)
X_test <- cbind(X_test, X[,5])
X_test <- as.data.frame(X_test)
names(X_test)[names(X_test) == 'V5'] <- 'vs'
#X2 <- cbind(X2,X %>% filter(vs == 0) %>% select(vs))

#X_test <- rbind(X1,X2)

fit = hotelling.test(.~vs, data = X_test)
fit
```

```
## Test stat:  20.067
## Numerator df:  4
```

```
## Denominator df:  27
## P-value:  9.069e-08
```

A 5%, aplicando o teste de Hotelling temos evidência para rejeitar H0: $\mu_1 - \mu_2 = 0$.

c) Construa os intervalos T e bonferroni e conclua a respeito.

Vamos construir os intervalos T e Bonferroni utilizando o estimador da variância combinada S_{pool} como nos slides:

```
X <- X_test
x1_bar=colMeans(X[which(X$vs==0),-5])
S1=cov(X[which(X$vs==0),-5])
n1=nrow(X[which(X$vs==0),-5])
x2_bar=colMeans(X[which(X$vs==1),-5])
S2=cov(X[which(X$vs==1),-5])
n2=nrow(X[which(X$vs==1),-5])
n=n1+n2

S_pool=((n1-1)*S1+(n2-1)*S2)/(n-2)

Sco= ((1/n1)+(1/n2))* S_pool

alpha=0.05

p = 4

#Intervalos T
Ls=c()
Li=c()
for (i in 1:p) {
  Ls[i]=(x1_bar[i]-x2_bar[i])+sqrt(qf(1-alpha,p,n-p-1)*(((n-2)*p)/((n-p-1)))*Sco[i,i])
  Li[i]=(x1_bar[i]-x2_bar[i])-sqrt(qf(1-alpha,p,n-p-1)*(((n-2)*p)/((n-p-1)))*Sco[i,i])

}

Lim=rbind(Ls,Li)
#colnames(Lim)<-colnames(x)
Lim
```

```
##          [,1]      [,2]      [,3]        [,4]
## Ls -0.1146098 1.1065765 0.6862370 -0.06262254
## Li -0.6739879 0.3298875 0.0252894 -0.23076507
```

```
#Intervalos Univariados Bonferroni
Lsb=c()
Lib=c()
for (i in 1:p) {
  Lsb[i]=(x1_bar[i]-x2_bar[i])+qt(1-(alpha/(2*p)),n-2)*sqrt(Sco[i,i])
  Lib[i]=(x1_bar[i]-x2_bar[i])-qt(1-(alpha/(2*p)),n-2)*sqrt(Sco[i,i])

}

Limb=rbind(Lsb,Lib)

Limb
```

```
##            [,1]      [,2]      [,3]        [,4]
## Lsb -0.1808407 1.0146158 0.6079802 -0.08253076
## Lib -0.6077571 0.4218482 0.1035462 -0.21085685
```

Analisando ambos os intervalos T e Bonferroni, para ambos os casos o zero não está incluso em nenhuma variável. Ou seja, há diferença entre os vetores de médias para ambos os grupos vs=0 e vs=1.

Exercício 6

```r
xx=data.frame(carData::Salaries)
x=xx[which(xx[,1]=="Prof"),-1] ##Restringe a analise apenas ao #Professors# (Rank=Prof)

#Fator1: discipline A (theoretical'' departments) or B (applied departments))
#Fator2: gender(Female or Male)

#variaveis de analise:
#x1: yrs.since.phd (anos desde que completou o phd)
#x2: yrs.service (anos de servico)
#x3: Salário

head(x)
```

```
##   discipline yrs.since.phd yrs.service  sex salary
## 1          B            19          18 Male 139750
## 2          B            20          16 Male 173200
## 4          B            45          39 Male 115000
## 5          B            40          41 Male 141500
## 7          B            30          23 Male 175000
## 8          B            45          45 Male 147765
```

a) Realize a MANOVA a 5% e conclua a respeito. Informe o valor da estatística de Wilks e seu valor p.

```r
y_vars <- as.matrix(x[,c(2,3,5)])
dis <-  x[,1]
gender = x[,4]
mnv <- manova(y_vars ~ dis * gender)
summary(mnv, test = "Wilks")
```

```
##              Df   Wilks approx F num Df den Df    Pr(>F)
## dis           1 0.88454  11.3128      3    260 5.337e-07 ***
## gender        1 0.97952   1.8121      3    260    0.1453
## dis:gender    1 0.98147   1.6361      3    260    0.1814
## Residuals   262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analisando os p-valores da MANOVA para dis (discipline), gender e a interação entre ambos, concluímos que a 5% apenas a variável discipline possui um efeito significativo multivariado em relação as variáveis yrs.since.phd, yrs.service e salary.

A estatística de Wilks para discipline, gender e a intereção é respectivamente $(0.884, 0.979, 0.981)$, e os p-valores respectivamente $(5.337e-07, 0.145, 0.181)$.

b) Realize a ANOVA em cada variável a 5% e conclua a respeito. Informe o valor da estatística F e seu valor p.

Agora que temos evidências que existem variáveis que influenciam conjuntamente nossas variáveis dependentes, vamos analisar os efeitos individuais em cada variável.

```
summary.aov(mnv)
```

```
##  Response yrs.since.phd :
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## dis           1  1235.5 1235.47 12.7075 0.0004327 ***
## gender        1   369.2  369.19  3.7973 0.0524031 .
## dis:gender    1     2.7    2.68  0.0276 0.8682387
## Residuals   262 25472.6   97.22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response yrs.service :
##              Df Sum Sq Mean Sq F value  Pr(>F)
## dis           1    683  683.24  5.2331 0.02296 *
## gender        1    595  595.23  4.5590 0.03367 *
## dis:gender    1    115  114.51  0.8771 0.34987
## Residuals   262  34207  130.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response salary :
##              Df     Sum Sq    Mean Sq F value    Pr(>F)
## dis           1 1.2019e+10 1.2019e+10 16.5170 6.378e-05 ***
## gender        1 5.7391e+08 5.7391e+08  0.7887    0.3753
## dis:gender    1 3.5909e+08 3.5909e+08  0.4935    0.4830
## Residuals   262 1.9065e+11 7.2769e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para a variável resposta yrs.since.phd, a variável discipline tem um efeito significativo nas médias, com a variável gender sendo significativa apenas a 10% mas próximo de 5% (5.2%). A interação entre elas não é significativa nas médias.

Para a variável resposta yrs.service, ambos discipline e gender possuem efeitos significativos nas médias a 5%.

Para a variável resposta salary, discipline é a única variável que exibe um efeito significativo nas médias.

c) Faça uma conclusão geral sobre os resultados:

Sabemos que a variável discipline tem um efeito significativo a 5% considerando o caso multivariado para variáveis respostas yrs.since.phd, yrs.service e salary. Rodando uma ANOVA para cada variável resposta separademente, chegamos a conclusão que discipline tem um efeito significativo para yrs.since.phd, para yrs.service ambos discpline e gender são significativos e para salary apenas discipline. Para visualizar melhor esta relação, vamos plotar alguns boxplots afim de analisar a magnitude destas diferenças:

```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.3
```

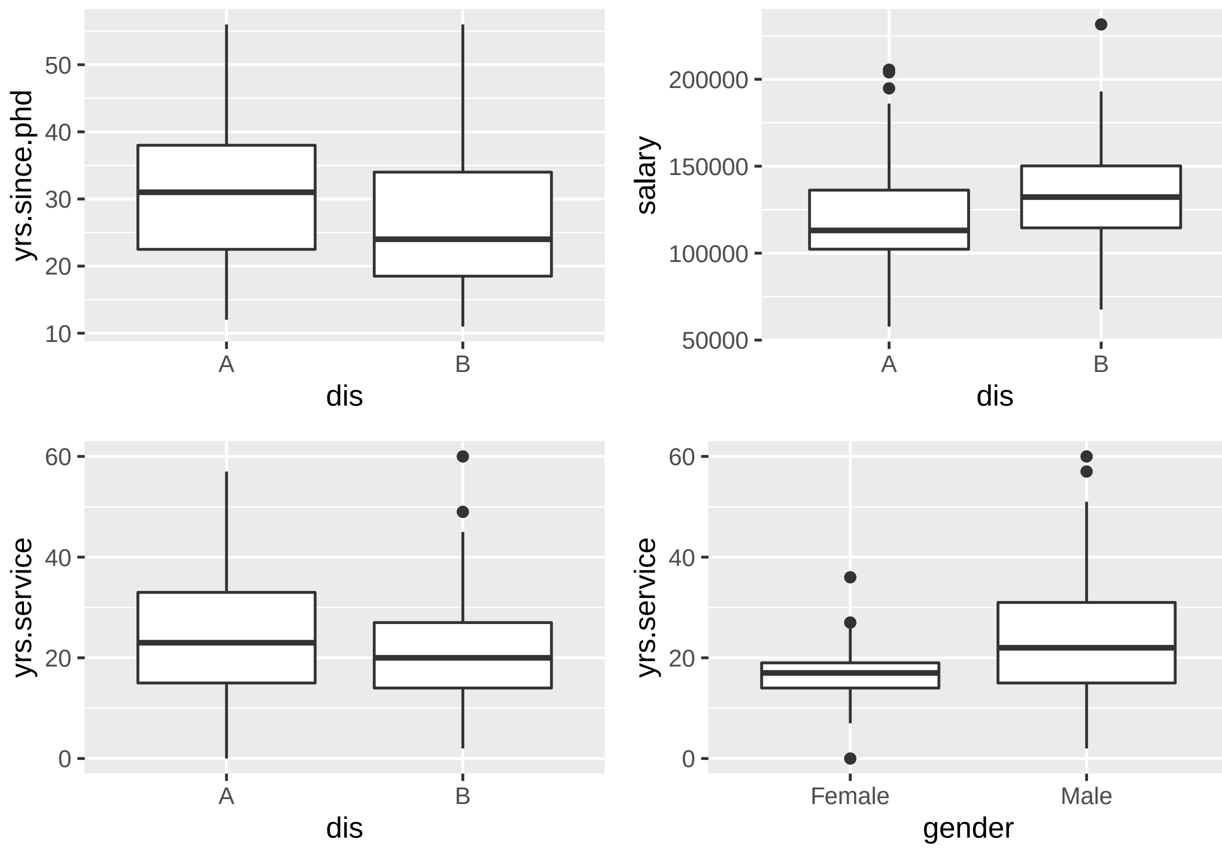
```
##
```

```
## Attaching package: 'gridExtra'

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

```r
data <- x

p1 <- ggplot(data, aes(x = dis, y = yrs.since.phd, group = dis)) + geom_boxplot()
p2 <- ggplot(data, aes(x = dis, y = yrs.service, group = dis)) + geom_boxplot()
p3 <- ggplot(data, aes(x = gender, y = yrs.service, group = gender)) + geom_boxplot()
p4 <- ggplot(data, aes(x = dis, y = salary, group = dis)) + geom_boxplot()
grid.arrange(p1,p4,p2,p3, ncol = 2)
```



Podemos notar que para yrs.since.phd, discipline = A (departamentos teóricos) possuem professores com médias maiores de carreira em relação a professores com discipline = B (departamenos aplicados). No caso de salary, departamentos aplicados possuem uma média salarial maior. Em relação a yrs.service, para professores com o efeito discipline = A, os anos de serviço são em média maiores que professores de discipline = B. Analisando gender, professores Homens tem médias maiores de anos de serviço que professoras.