MATH 537 HOMEWORK 3

Gustavo Esparza

7/19/2019

1

To create the 95% confidence interval region for the mean vector of A with $n = 25$, $p = 3$, we will use the following formula:

$p[n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \leq \frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha)] = 1-\alpha$

In particular we have:
$\frac{(n-1)p}{(n-p)}F_{p,n-p}(\alpha) = \frac{(24)3}{(22)}F_{3,22}(.95) = 9.978955$

Thus, we have a maximum distance for our 95% confidence reigion for (μ_1, μ_2, μ_3).
In particular, the 95% confidence interval region for $\vec{\mu}$ is all values $(\vec{\mu_1}, \vec{\mu_2}, \vec{\mu_3})$ such that

$n(\bar{X}-\mu)'S^{-1}(\bar{X}-\mu) \leq 9.978955$

In a more expanded format, we have the following inequality:

$$25\begin{bmatrix}6.37134-\mu_1, 3.33062-\mu_2, 3.55357-\mu_3\end{bmatrix}\begin{bmatrix}0.08558951 & -0.02480793 & 0.02710110\\ -0.02480793 & 0.11827828 & -0.03276486\\ 0.02710110 & -0.03276486 & 0.11459714\end{bmatrix}\begin{bmatrix}6.37134-\mu_1\\ 3.33062-\mu_2\\ 3.55357-\mu_3\end{bmatrix} \leq 9.978955$$

Extra

We can also manipulate our confidence region in order to create simultaneous confidence intervals for each μ_i.
The interval is provided as follows:
$\bar{x}_i - \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p(.05)}}\sqrt{\frac{S_{ii}}{n}} \leq \mu_i \leq \bar{x}_i + \sqrt{\frac{(n-1)p}{n-p}F_{p,n-p(.05)}}\sqrt{\frac{S_{ii}}{n}}$

Here, F follows an F distribution and S represents the covariance matrix of our variables X_1 - X_3

Thus, we have the following result:

```
## 95% Confidence Interval for Mu 1  = [ 4.087223 , 8.655458 ]
## 95% Confidence Interval for Mu 2  = [ 1.383053 , 5.278188 ]
## 95% Confidence Interval for Mu 3  = [ 1.559964 , 5.547185 ]
```

It is important to note that these individual intervals are wider than the combined region due to each interval having to secure space for the joint distribution, and thus don't exactly equal our confidence region.

2

In order to find the 95% Confidence Interval for a linear combination $a^T\mu$, we have the following result:

$$a^T\bar{X} \pm t_{n-1}(\alpha/2)\sqrt{\frac{a^TSa}{n}}$$

We will be using this formula for parts i -iii.

2i

μ_1

Here, a = $[1, 0, 0]$ and the corresponding Confidence Interval is as follows:

```
## 95% Confidence Interval for Mu1 = [ 4.879012 7.863668 ]
```

2ii

$\mu_2 - \mu_1$

Here, a = $[-1, 1, 0]$ and the corresponding Confidence Interval is as follows:

```
## 95% Confidence Interval for Mu2 - Mu1 = [ -4.815116 -1.266324 ]
```

2iii

$\mu_3 + \mu_2 - \mu_1$

Here, a = $[-1, 1, 1]$ and the corresponding Confidence Interval is as follows:

```
## 95% Confidence Interval for Mu3 + Mu2 - Mu1 = [ -2.028756 3.054465 ]
```

3

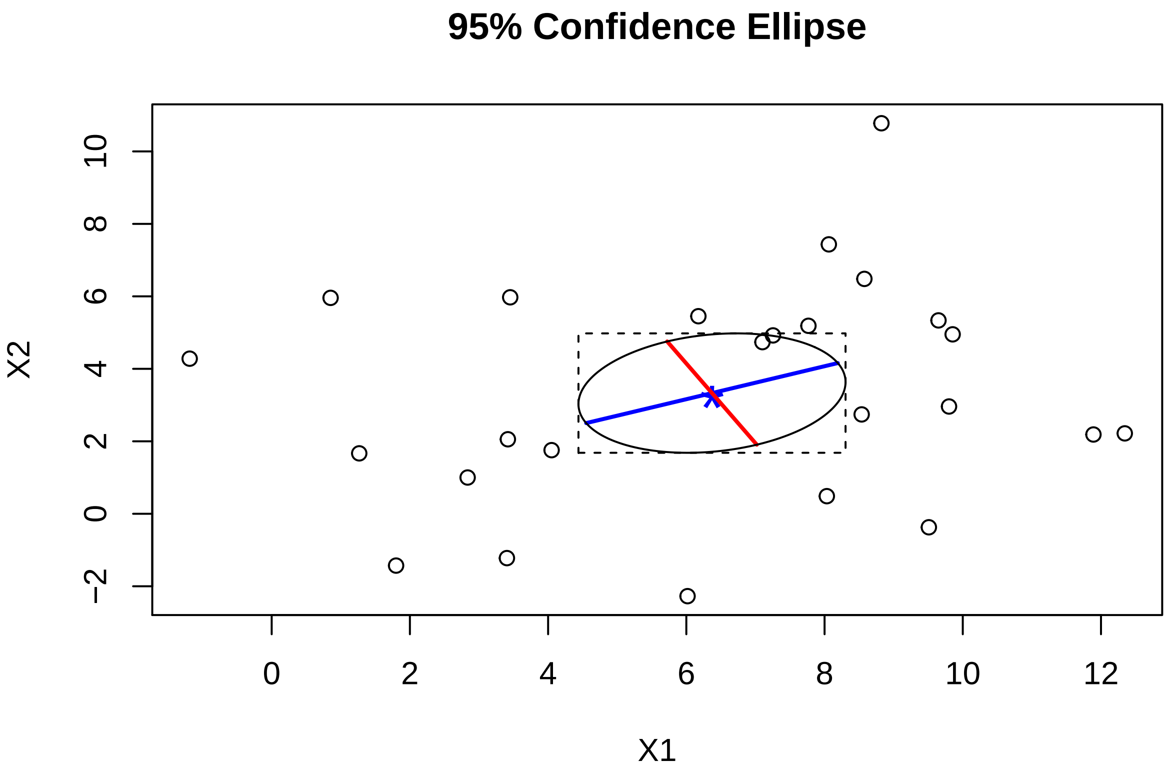
In this problem, we wish to compute and plot a 95% Confidence Interval Region for (μ_1, μ_2).

Using X_1 and X_2 from A, we will use the eigenvalues and eigenvectors of $\frac{S}{n}$, the covariance matrix for the mean of A.

Using the distance formula : $\sqrt{\frac{(n-1)p}{n-p} F_{p,n-p(.05)}}$, we can use the eigenvalues (λ) and eigenvectors (V) to create the axes for our confidence region ellipse. The respective axes are defined as follows:

$a_i = V_i \times \sqrt{\lambda_i} \times \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p(.05)}}$

Now, using the ellipse function with $\sqrt{\frac{(n-1)p}{n-p} F_{p,n-p(.05)}}$ as the boundary, we can plot (X_1, X_2) and the corresponding 95% Confidence Ellipse, along with the axes a_i.



We can see the confidence ellipse centered around our mean, with accompanying axes computed using the eigenvectors of our covariance matrix.

In addition, we have added a rectangle with corners consisting of the lower and upper simultaneous confidence bounds for μ_1 and μ_2, respectively. Again, we can see that the individual confidence intervals are wider than the joint confidence ellipse. These intervals were computed in the following manner:

$\bar{x}_i - \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p(.05)}} \sqrt{\frac{S_{ii}}{n}} \leq \mu_i \leq \bar{x}_i + \sqrt{\frac{(n-1)p}{n-p} F_{p,n-p(.05)}} \sqrt{\frac{S_{ii}}{n}}$

```
## 95% Confidence Interval for Mu 1 = [ 4.439012 8.303669 ]
## 95% Confidence Interval for Mu 2 = [ 1.683008 4.978233 ]
```

4

Before conducting any formal testing, we should review the computed means for A, and B to get a glance of any distinct differences in the data.

```
## Mean vector for A = ( 6.37134 , 3.33062 , 3.553574 )
```

```
## Mean vector for B = ( 5.743127 , 3.611131 , 4.286794 )
```

We can see that the two mean vectors are not too different from each other.

We will be using the 2 sample Hoetelling Test in order to test our hypothesis. In particular,we have

$H_o : \vec{\mu_A} = \vec{\mu_B}$ VS $H_a : \vec{\mu_A} \neq \vec{\mu_B}$
Test Statistic: F dsitribution value.

With Amat and Bmat representing matrices consisting of A and B, respectively, we will use the Hotelling Test function to compute our test statistic and P-value

```
hyp.test = hotelling.test(Amat,Bmat)
hyp.test
```

```
## Test stat:  0.34586
## Numerator df:  3
## Denominator df:  61
## P-value:  0.7922
```

We can observe that our test statistic is 0.34586 with(3,61) degrees of freedom. The corresponding p-value is .7922, thus we proceed to fail to reject the null hypothesis.

At the $\alpha = .05$ level, we do not have sufficient evidence to suggest there is a difference between the mean vectors of datasets A and B.

5

In this problem, we will be testing the following hypothesis:

$H_o : \vec{\mu_A} = \vec{\mu_B} = \vec{\mu_C}$ VS At least one $\vec{\mu_X}$ is different.
Test Statistic: F distribution value

Before conducting any formal testing, we should review the computed means for A, B and C to get a glance of any distinct differences in the data.

```
## Mean vector for A = ( 6.37134 , 3.33062 , 3.553574 )
```

```
## Mean vector for B = ( 5.743127 , 3.611131 , 4.286794 )
```

```
## Mean vector for C = ( 1.919498 , 1.558253 , 8.350298 )
```

It is apparent that the mean vector for dataset C is quite different from the other two datasets.

With *table*, we have created a dataframe that consists of X_1,X_2,X_3 for datasets A,B,C and added an additional column that designates which group the data belongs to. With this formatting, we are able to conduct a Multiple Analysis of Variance (MANOVA) test for the difference in the mean vectors between our three groups.

The MANOVA test provides the following results:

```r
hypoth = manova(cbind(table$X1,table$X2,table$X3)~Group ,data = table)
summary(hypoth)
```

```
##            Df  Pillai approx F num Df den Df   Pr(>F)
## Group       2 0.45172   9.8225      6    202 1.69e-09 ***
## Residuals 102
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can observe that our test statistic is 9.8225 with(2,102) degrees of freedom. The corresponding p-value is 1.69e-09, thus we proceed to reject the null hypothesis.

At the $\alpha = .05$ level, we have strong evidence to suggest there is a difference between the mean vectors of datasets A and B and C. The vast difference in mean values for C is likely the cause for this evidence.

APPENDIX

1

```
A = read.csv("dataA.csv")
attach(A)
A = cbind(X1,X2,X3)
p = 3
n = 25
alpha = .05

xbar = c(mean(X1),mean(X2),mean(X3))
S = cov(A)
```

To create the 95% confidence interval region for the mean vector of A with $n = 25$, $p = 3$, we will use the following formula:

$$p[n(\bar{X} - \mu)'S^{-1}(\bar{X} - \mu) \leq \frac{(n-1)p}{(n-p)} F_{p,n-p}(\alpha)] = 1 - \alpha$$

```
p=3
n=25

S=var(A)
eig = eigen(S/n)

lam1 = eig$values[1]
lam2 = eig$values[2]
lam3 = eig$values[3]

v1 = eig$vectors[,1]
v2 = eig$vectors[,2]
v3 = eig$vectors[,3]

dist = (p*(n-1)/(n-p))*qf(.95,p,n-p)
a1 = sqrt(lam1)*sqrt(dist)
a2 = sqrt(lam2)*sqrt(dist)
a3 = sqrt(lam3)*sqrt(dist)
a = c(a1,a2,a3)
```

Extra

```
f = qf(1-alpha,df1=p,df2=n-p)

dist = (p*(n-1)*f)/(n-p)

LCI = rep(0,3)
UCI = rep(0,3)

for ( i in 1:p ){
LCI[i] <- xbar[i]-sqrt((n-1)*p/(n-p)*f)*sqrt(S[i,i]/n)
UCI[i] <- xbar[i]+sqrt((n-1)*p/(n-p)*f)*sqrt(S[i,i]/n)
```

```
cat("95% Confidence Interval for Mu",i," = [",LCI[i],",",UCI[i],"]","\n")}
```

It is important to note that these individual intervals are slightly wider than the combined region due to each interval having to secure space for the joint distribution.

2

2i

μ_1

Here, a $= [1, 0, 0]$ and the corresponding Confidence Interval is as follows:

```
a = matrix(c(1,0,0),ncol =1)

L = t(a)%*%xbar - qt(.975,n-1)*sqrt(t(a)%*%S%*%a/n)
U = t(a)%*%xbar + qt(.975,n-1)*sqrt(t(a)%*%S%*%a/n)

cat("95% Confidence Interval for Mu1 = [",L,U,"]")
```

2ii

$\mu_2 - \mu_1$

Here, a $= [-1, 1, 0]$ and the corresponding Confidence Interval is as follows:

```
a = matrix(c(-1,1,0),ncol =1)

L = t(a)%*%xbar - qt(.975,n-1)*sqrt(t(a)%*%S%*%a/n)
U = t(a)%*%xbar + qt(.975,n-1)*sqrt(t(a)%*%S%*%a/n)

cat("95% Confidence Interval for Mu2 - Mu1 = [",L,U,"]")
```

2iii

$\mu_3 + \mu_2 - \mu_1$

Here, a $= [-1, 1, 1]$ and the corresponding Confidence Interval is as follows:

```
a = matrix(c(-1,1,1),ncol =1)

L = t(a)%*%xbar - qt(.975,n-1)*sqrt(t(a)%*%S%*%a/n)
U = t(a)%*%xbar + qt(.975,n-1)*sqrt(t(a)%*%S%*%a/n)

cat("95% Confidence Interval for Mu3 + Mu2 - Mu1 = [",L,U,"]")
```

3

In this problem, we wish to compute and plot a 95% Confidence Interval Region for (μ_1, μ_2).

```r
p=2
n=25
A2 = cbind(X1,X2)
X = X1
Y = X2

x.bar = mean(X1)
y.bar = mean(X2)
mean = cbind(x.bar,y.bar)
mean = c(mean)

S=var(A2)
eig = eigen(S/n)


lam1 = eig$values[1]
lam2 = eig$values[2]

v1 = eig$vectors[,1]
v2 = eig$vectors[,2]


dist = (p*(n-1)/(n-p))*qf(.95,p,n-p)
a1 = v1*sqrt(lam1)*sqrt(dist)
a2 = v2*sqrt(lam2)*sqrt(dist)
```

```r
f = qf(.95,df1=p,df2=n-p)

L = rep(0,2)
U = rep(0,2)

for ( i in 1:2 ){
L[i] = xbar[i]-sqrt((n-1)*p/(n-p)*f)*sqrt(S[i,i]/n)
U[i] = xbar[i]+sqrt((n-1)*p/(n-p)*f)*sqrt(S[i,i]/n)
}
```

```r
elps2 = t(t(ellipse(S/n, level=0.95,centre =mean,t=sqrt(dist) ,  npoints=1000)))
plot(X,Y,xlab="X1",ylab="X2",main="95% Confidence Ellipse")
points(x.bar,y.bar,pch='*',col="blue",cex=2)

lines(c(x.bar,x.bar+ a1[1]),c(y.bar,y.bar+a1[2]),lwd=2,col="blue")
lines(c(x.bar,x.bar- a1[1]),c(y.bar,y.bar-a1[2]),lwd=2,col="blue")
lines(c(x.bar,x.bar+ a2[1]),c(y.bar,y.bar+a2[2]),lwd=2,col="red")
lines(c(x.bar,x.bar- a2[1]),c(y.bar,y.bar-a2[2]),lwd=2,col="red")
lines(elps2,col="black")
rect(L[1],L[2],U[1],U[2],lty=2)
```

```r
for(i in 1:2)
{cat("95% Confidence Interval for Mu",i, "= [",L[i],U[i],"]","\n")}
```

4

```
B= read.csv("dataB.csv")
attach(B)
B = cbind(X1,X2,X3)
xbarB = c(mean(X1),mean(X2),mean(X3))
```

```
cat("Mean vector for A = (",xbar[1],",",xbar[2],",",xbar[3],")")
cat("\n\n")
cat("Mean vector for B = (",xbarB[1],",",xbarB[2],",",xbarB[3],")")
```

```
Amat = as.data.frame(A)
Bmat = as.data.frame(B)
```

```
hyp.test = hotelling.test(Amat,Bmat)
hyp.test
```

5

```
C = read.csv("dataC.csv")
attach(C)
C= cbind(X1,X2,X3)
Cmat = as.data.frame(C)

xbarC= c(mean(X1),mean(X2),mean(X3))
```

Before conducting any formal testing, we should review the computed means for A, B and C to get a glance of any distinct differences in the data.

```
cat("Mean vector for A = (",xbar[1],",",xbar[2],",",xbar[3],")")
cat("\n\n")
cat("Mean vector for B = (",xbarB[1],",",xbarB[2],",",xbarB[3],")")
cat("\n\n")
cat("Mean vector for C = (",xbarC[1],",",xbarC[2],",",xbarC[3],")")
```

```
table = rbind(A,B,C)

table = as.data.frame(table)
Group = c(rep("A",25),rep("B",40),rep("C",40))
table = cbind(table,Group)
```

The MANOVA test provides the following results:

```
hypoth = manova(cbind(table$X1,table$X2,table$X3)~Group ,data = table)
summary(hypoth)
```