**Application**

a.

> x=matrix(c(1,1,0,2,0,1,1,1,0,1,-3,0,0,0,1),ncol=3,byrow=TRUE)

> x

[,1] [,2] [,3]

[1,] 1 1 0

[2,] 2 0 1

[3,] 1 1 0

[4,] 1 -3 0

[5,] 0 0 1

> x1=x[,1]

> x2=x[,2]

> x3=x[,3]

> norme<-function(x) sqrt(x%\*%x)

> u1=x1/norme(x1)

> v2=x2-t(u1)%\*%x2\*u1

> u2=v2/norme(v2)

> v3=x3-(t(u1)%\*%x3\*u1+t(u2)%\*%x3\*u2)

> u3=v3/norme(v3)

> u=cbind(u1,u2,u3)

> u

u1 u2 u3

[1,] 0.3779645 0.346844 -0.2649065

[2,] 0.7559289 0.086711 0.3532086

[3,] 0.3779645 0.346844 -0.2649065

[4,] 0.3779645 -0.867110 -0.1766043

[5,] 0.0000000 0.000000 0.8388705

> c=t(u)%\*%x

> c

[,1] [,2] [,3]

u1 2.645751e+00 -0.3779645 0.7559289

u2 -1.110223e-16 3.2950179 0.0867110

u3 4.718448e-16 0.0000000 1.1920791

b.

**a.**

> sigma<-matrix(c(3,2,2,3),nrow=2)

> vp1<-eigen(sigma)$vectors[,1]

> vp2<-eigen(sigma)$vectors[,2]

> mu<-c(2,3)

> alpha1<-t(vp1)%\*%mu

> alpha2<-t(vp2)%\*%mu

> teta<-seq(0,2\*pi,length.out=200)

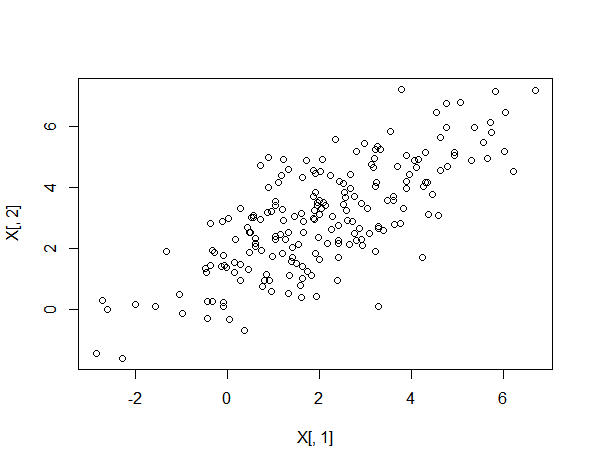
> y1<-sqrt(4\*5)\*cos(teta)+alpha1

> y2<-sqrt(4\*1)\*sin(teta)+alpha2

> z<-outer(y1,y2,function(y1,y2) 0.6\*(y1-2)^2+0.6\*(y2-3)^2-2\*0.4\*(y2-3)\*(y1-2))

**b**.

> plot(X[,1],X[,2])



> ellBase <- cbind(y1, y2)

> ellRot <- eigen(sigma)$vectors %\*% t(ellBase)

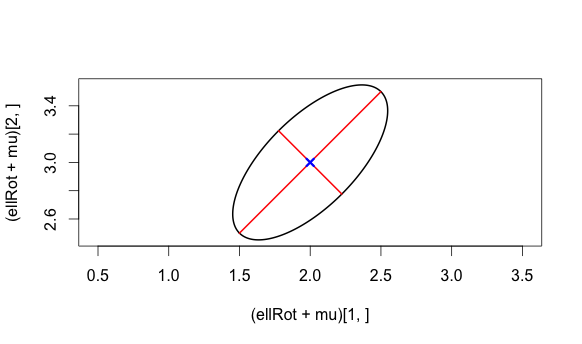
> plot((ellRot+mu)[1, ], (ellRot+mu)[2, ], asp=1, type="l", lwd=2)

> eigScl <- eigen(sigma)$vectors %\*% diag(sqrt(eigen(sigma)$values))

> xMat2 <- rbind(mu[1] + eigScl[1, ], mu[1] - eigScl[1, ])

> yMat2 <- rbind(mu[2] + eigScl[2, ], mu[2] - eigScl[2, ])

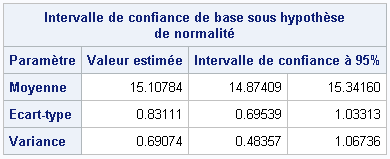
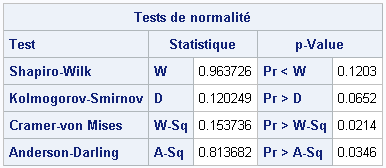
> matlines(xMat2, yMat2, lty=1, lwd=2, col="red")



**Exercice 5**

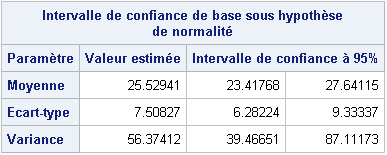
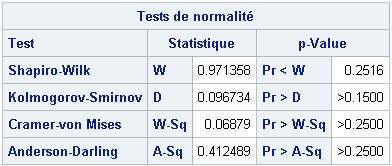
**a**. En utilisant le PROC UNIVARIATE de SAS, on peut faire le test de normalité univariée pour chaque variable. On a des sorties suivantes pour chaque variables.

Variable 1

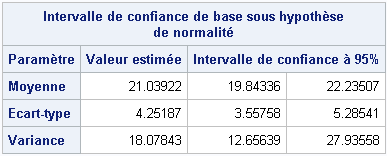
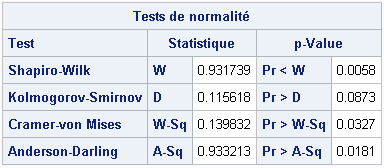
On voit que les p-values pour les tests Shapiro-Wilk et Kolmogorov-Smirnow sont respectivement 0.1203 et 0.0652 qui sont tous supérieurs à 0.05, il n’y donc pas assez d’évidence à rejetter l’hypothèse de la normalité de la variable 1.

Variable 2

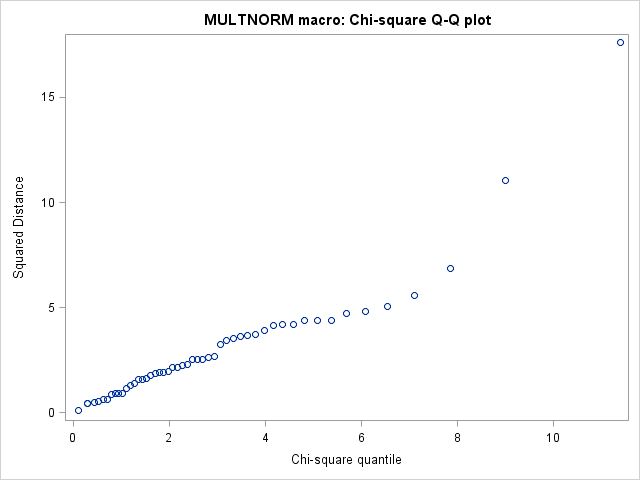
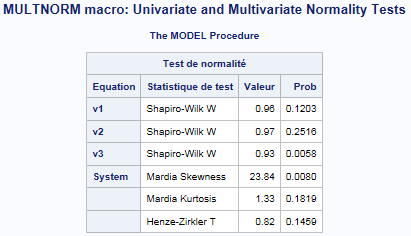
On voit que le p-values pour le test Shapiro-Wilk est 0.2516 qui est supérieur à 0.05, en plus les p-values pour tous les autres tests (Kolmogorov-Smirnow, Cramer-von Mises et Anderson-Darling) sont aussi supérieurs à 0.05, il n’y donc pas assez d’évidence à rejetter l’hypothèse de la normalité de la variable 2.

Variable 3

On voit que le p-value pour le test Kolmogorov-Smirnow est 0.0873 qui est supérieure à 0.05. Mais pour le test Shapiro-Wilk, on a un p-value égale à 0.0058 qui est inférieure à 0.05. Comme ici la taille de l’échantillon 51 est petite, on considère donc le test de Shapiro-Wilk qui est le plus puissant. On rejette donc l’hypothèse de la normalité pour la variable 3.

**b**. En appelant le macro multnorm dans SAS, on peut faire le test multivarié de la normalité. On obtient les tableaux suivants.



Par le test de Mardis Skewness on rejette l’hypothèse de normalité multivarieé car le p-value est 0.008 qui est plus petit que 0.05. Par les autres tests, il n’y pas assez d’évidence à rejeter l’hypothèse nulle de normalité multivariée. Mais selon la propriété d’un vecteur normal, tous ses variables sont aussi univariée normales. Ici ce n’est pas le cas car on a rejeté l’hypothèse de normalité pour la variable 3. On devrait donc rejeter l’hypothèse de normalité univariée.

**c**. **Les codes pour trouver la statistique de Hotelling par le SAS sont ci-dessous.**

**data** exo5;

input v1 v2 v3;

datalines;

13.4 25 17

……

14.7 34 16

;

%***multinorm***(data=exo5,var=v1 v2 v3)

**data** test;

set exo5;

v1c=v1-**14.6**;

v2c=v2-**26**;

v3c=v3-**21**;

dep=**1**;

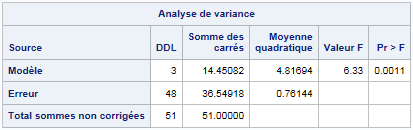
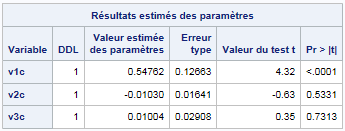
**proc** **reg** data=test;

model dep=v1c v2c v3c/noint;

**run**;

**quit;**

On a les tableaux suivants.

La statistique de Hotelling est donc 6.33, le p-value est 0.0011. Le test est donc significatif, et on rejette l’hypothèse nulle.

**Les codes pour trouver la statistique de Hotelling par le R sont ci-dessous.**

> exo5<-scan()

13.4 25 17

……

14.7 34 16

> exo5<-matrix(exo5,51,3,byrow=TRUE)

> tCarre<-51\*t(apply(exo5,2,mean)-c(14.6,26,21))%\*%solve(var(exo5))%\*%(apply(exo5,2,mean)-c(14.6,26,21))

> tCarre

[,1]

[1,] 19.769

> F=((51-3)/(3\*50))\*tCarre

> F

[,1]

[1,] 6.326081

> 1-pf(F,3,48)

[,1]

[1,] 0.001055176

On a le même résultat en R que en SAS, la statistique de Hotelling est 6.33, le p-value est 0.0011. Ainsi, on rejette l’hypothèse nulle.

**Exercice 6**

**a.** **Les codes pour trouver la statistique de Hotelling par le R sont ci-dessous.**

> exo6<-scan()

1: 1 121 22 74 223 54 254

……

274: 2 164 32 76 187 30 264

Read 280 items

> exo6<-matrix(exo6,nrow=40,byrow=TRUE)

> inge<-exo6[c(1:20),c(2:7)]

> pilot<-exo6[c(21:40),c(2:7)]

> moyenne1<-apply(inge,2,mean)

> moyenne2<-apply(pilot,2,mean)

> Sp<-(19\*var(inge)+19\*var(pilot))/(20+20-2)

> Tcarre<-(20\*20/(20+20))\*t(moyenne1-moyenne2)%\*%solve(Sp)%\*%(moyenne1-moyenne2)

> Tcarre

[,1]

[1,] 66.66044

> F<-((20+20-6-1)/(6\*38))\*Tcarre

> F

[,1]

[1,] 9.648221

> 1-pf(F,6,33)

[,1]

[1,] 3.851202e-06

Ainsi, on a la statistique de Hotelling T carré = 66.66, la statistique F associée est 9.65. Le p-value = 3.851202e-06 est très significatif. On rejette donc l’hypothèse nulle mu1 = mu2.

**b.** Utilisons R pour calculer les coefficients de la fonction discriminante.

> a<-solve(Sp)%\*%(moyenne1-moyenne2)

> a

[,1]

[1,] 0.007515209

[2,] 0.193260187

[3,] -0.129182304

[4,] -0.042785864

[5,] 0.071839177

[6,] -0.049062257

Donc la combinaison linéaire qui sépare les deux groupes sont : a'y = 0.0075y1 + 0.19y2 – 0.13y3 – 0.043y4 + 0.072y5 – 0.049y6. On voit que y2 et y3 contribuent le plus à la séparation de deux groupes.

**c**. **Les codes pour calculer la statistique F par le modèle de régression en SAS sont suivants.**

FILENAME exo6 'C:\Users\hg791104\Desktop\devoir1\_donnees\Chapitre5LivreTable5\_6\_PILOT.dat';

**DATA** exo6;

INFILE exo6;

INPUT type var1 var2 var3 var4 var5 var6;

TITLE 'exo6-Test multivarié à deux échantillons';

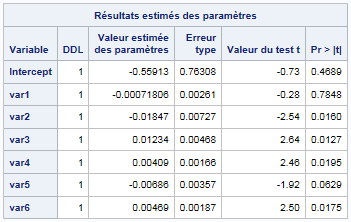
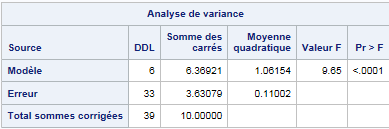
**proc** **reg** data=exo6;

model type = var1 var2 var3 var4 var5 var6;

**run**;

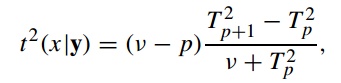
**quit**;

On a les sorties des tableaux suivantes.



On obtient la même valeur de la statistique que en utilisant SAS, F = 9.65.

**d.** Pour tester l’indépendance d’une variable en présence des autres variables, on utilise la statistique au-dessous :



On rejette l’hypothèse quand



En utilisant R, on a la matrice de la statistique et le p-value pour chaque variable :

[,1] [,2]

[1,] 0.07579313 0.9955540181

[2,] 6.45133687 0.0001974117

[3,] 6.95182122 0.0001070975

[4,] 6.03086507 0.0003346042

[5,] 3.70519667 0.0078929501

[6,] 6.26185324 0.0002500053

On rejette donc l’hypothèse nulle pour la variable 2, la variable 3, la variable 4 la variable 5 et la variable6. Donc toutes les variables sauf variable 1 contribuent à distinguer les deux groupes.

**e.**

> Sjj<-c(Sp[1,1],Sp[2,2],Sp[3,3],Sp[4,4],Sp[5,5],Sp[6,6])

> moyenneDiff<-moyenne1-moyenne2

> t<-(1/sqrt((20+20)/400))\*(1/sqrt(Sjj))\*moyenneDiff

> t

[1] -0.6556107 2.6139358 -3.2884332 -4.6315123 1.8873351 -3.2204869

La valeur critique de la correction de Bonferroni est

> qt(1-.05/(2\*6),20+20-2)

[1] 2.783474

On rejette l’hypothèse nulle quand |ti| > t critique. Ainsi, la variable 3, la variable 4 et la variable 6 sont des variables les plus importantes qui distinguent les deux groupes.

**f.** Il semble que le test univarié avec la correction est moins puissant que le test multivariée quand il n’y a pas beaucoup de corrélation entre les variables.