#################################################

####### IBI5086 – IME - 2o.Sem-2023 #######

####### Profa. Júlia Maria Pavan Soler #######

####### Análise Multivariada #######

####### Estatísticas Descritivas #######

#################################################

#Dados Iris no R

attach(iris)

iris

head(iris)

str(iris)

yg <- iris

table(yg$Species)

####ANÁLISE DESCRITIVA

pairs(yg[,-5]) #matrix de gráficos de dispersão

# plot iris com diferentes cores

class.vector <- numeric(150)

class.vector[yg$Species == 'setosa'] <- 1

class.vector[yg$Species == 'versicolor'] <- 2

class.vector[yg$Species == 'virginica'] <- 3

pairs(yg[,-5], pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

##Dados multivariados (p=4) agrupados em espécies

y <- yg[,1:4] #extrair as variáveis quantitativas

##Análise independente de espécie (Ynxp, n=150 p=4)

head(y)

mi<-colMeans(y) #vetor centróide (Total:n=150)

mi

round(mi,2)

St<-cov(y) # matriz de covariâncias para os dados totais

St

round(St,2)

Rt <- cor(y) #matriz de correlação

Rt

round(Rt,2)

#Medidas de variabilidade multivariada

vartotal <- sum(diag(St))

vartotal

vargen <- det(St)

vargen

##Correlação entre as variáveis (independente de espécie)

names(y)

par(mfrow=c(2,2))

with(y, plot(Petal.Length,Petal.Width))

with(y, plot(Sepal.Length,Sepal.Width))

with(y, plot(Petal.Length,Sepal.Length))

with(y, plot(Petal.Width,Sepal.Width))

#Comente o padrão de correlação entre as variáveis

yn <- scale(y) #variáveis normalizadas

?scale

colMeans(yn) #médias zero

round(colMeans(yn),2)

cov(yn) # variâncias unitárias

Rt # cov(yn) = cor(y)

#Medidas de Distância

#Distância Euclidiana entre observações (originais)

diste<-dist(y)

summary(diste)

diste #não está em formato matricial

diste <- as.matrix(diste)

diste

dim(diste)

diste[1:6,1:6] #distância entre as 6 primeiras observações

#Distância Euclidiana Padronizada

#Distância de Pearson entre observações

distp <- dist(yn) #yn: variável padronizada

summary(distp)

distp #não está em formato matricial

distp <- as.matrix(distp)

distp

dim(distp)

distp[1:6,1:6] #compare com diste

##Diagnóstico de obs atípicas

##Critério Univariado

boxplot(y)

out <- boxplot.stats(y[,2])$out

out

out.ind <- which(y[,2] %in% c(out))

out.ind

mtext(paste("Outliers: ", paste(out.ind, collapse = ", ")))

#Boxplot univariado

par(mfrow=c(2,2))

boxplot(y[,1],main="Sepal.Length")

boxplot(y[,2],main="Sepal.Width")

mtext(paste("Outliers: ", paste(out.ind, collapse = ", ")))

boxplot(y[,3],main="Petal.Length")

boxplot(y[,4],main="Petal.Width")

#Boxplot bivariado

#Copiar a função bivbox proposta por Everitt:

bivbox<-function(a, d = 7, mtitle = "Bivariate Boxplot",

method = "robust",xlab="X",ylab="Y")

{

#a is data matrix

#d is constant(usually 7)

p <- length(a[1, ])

if(method == "robust") {

param <- biweight(a[, 1:2]); m1 <- param[1]; m2 <- param[2]

s1 <- param[3]; s2 <- param[4]; r <- param[5]

}

else {

m1 <- mean(a[, 1]); m2 <- mean(a[, 2]);

s1 <- sqrt(var(a[, 1])); s2 <- sqrt(var(a[, 2])); r <- cor(a[, 1:2])[1, 2]

}

x <- (a[, 1] - m1)/s1; y <- (a[, 2] - m2)/s2

e <- sqrt((x \* x + y \* y - 2 \* r \* x \* y)/(1 - r \* r))

e2 <- e \* e; em <- median(e); emax <- max(e[e2 < d \* em \* em])

r1 <- em \* sqrt((1 + r)/2); r2 <- em \* sqrt((1 - r)/2); theta <- ((2 \* pi)/360) \* seq(0, 360, 3)

xp <- m1 + (r1 \* cos(theta) + r2 \* sin(theta)) \* s1; yp <- m2 + (r1 \* cos(theta) - r2 \* sin(theta)) \* s2

r1 <- emax \* sqrt((1 + r)/2); r2 <- emax \* sqrt((1 - r)/2); theta <- ((2 \* pi)/360) \* seq(0, 360, 3)

xpp <- m1 + (r1 \* cos(theta) + r2 \* sin(theta)) \* s1; ypp <- m2 + (r1 \* cos(theta) - r2 \* sin(theta)) \* s2

maxxl <- max(xpp); minxl <- min(xpp); maxyl <- max(ypp); minyl <- min(ypp)

b1 <- (r \* s2)/s1; a1 <- m2 - b1 \* m1; y1 <- a1 + b1 \* minxl; y2 <- a1 + b1 \* maxxl

b2 <- (r \* s1)/s2; a2 <- m1 - b2 \* m2; x1 <- a2 + b2 \* minyl; x2 <- a2 + b2 \* maxyl

maxx <- max(c(a[, 1], xp, xpp, x1, x2)); minx <- min(c(a[, 1], xp, xpp, x1, x2))

maxy <- max(c(a[, 2], yp, ypp, y1, y2)); miny <- min(c(a[, 2], yp, ypp, y1, y2))

plot(a[, 1], a[, 2], xlim = c(minx, maxx), ylim = c(miny, maxy), xlab =xlab, ylab =ylab,

lwd = 2, pch = 1)

lines(xp, yp, lwd = 2); lines(xpp, ypp, lty = 2, lwd = 2)

segments(minxl, y1, maxxl, y2, lty = 3, lwd = 2); segments(x1, minyl, x2, maxyl, lty = 4, lwd = 2)

}

par(mfrow=c(2,3))

bivbox(y[,c(1,2)], method ="O")

bivbox(y[,c(1,3)], method ="O")

bivbox(y[,c(1,4)], method ="O")

bivbox(y[,c(2,3)], method ="O")

bivbox(y[,c(2,4)], method ="O")

bivbox(y[,c(3,4)], method ="O")

#Distância de Mahalanobis das observações ao centróide

#Outliers multivariado (p=4)

dm2 <- mahalanobis(y, colMeans(y), cov(y))

dm2

plot(density(dm2, bw = 0.5),

main="Dist Mahalanobis ^2, n=150, p=4") ; rug(dm2))

qqplot(qchisq(ppoints(150), df = 4), dm2,

main = expression("Q-Q plot de Mahalanobis:" \* ~dm^2 \*

" vs. quantis de" \* ~ chi[4]^2))

abline(0, 1, col = 'gray')

ymh.o<-cbind(seq(1:150),dm2)

plot(ymh.o, main="Observações Atípicas - p=4")

abline(h=qchisq(c(0.9), df = 4, lower.tail = TRUE))

out.dm2 <- dm2[dm2>=qchisq(c(0.9), df = 4, lower.tail = TRUE)]

out.dm2

out.inddm2 <- which(dm2 %in% c(out.dm2))

out.inddm2

mtext(paste("Outliers: ", paste(out.inddm2, collapse = ", ")))

#Explore as facilidades da library(distances) para cálculos de distâncias

##Decomposição espectral da matriz de covariância S\_4x4

vdv.dat<-eigen(cov(y))

round(vdv.dat$values,2) #autovalores

round(vdv.dat$vectors,2) #matriz de autovetores

sum(vdv.dat$values[1:2])/sum(vdv.dat$values) #%VarTotalExplicada

#CP1=0.36Y1-0.08Y2+0.86Y3+0.36Y4

#CP2=-0.66Y1-0.73Y2+0.17Y3+0.08Y4

##Decomposição espectral da matriz de covariância R\_4x4

vdv.datp<-eigen(cor(y))

round(vdv.datp$values,2) #autovalores

round(vdv.datp$vectors,2) #matriz de autovetores

sum(vdv.datp$values[1:2])/sum(vdv.datp$values) #%VarTotalExplicada

#CP1=0.52Y1p-0.27Y2p+0.58Y3p+0.56Y4p

#CP2=-0.38Y1p-0.92Y2p-0.02Y3p-0.07Y4p

#Análise de CP dos dados originais

cps<-prcomp(y,scale=F)

summary(cps)

cps$sdev #raiz quadrada dos autovalores

cps$rotation #autovetores

biplot(cps)

names(cps)

plot(cps$x[,1:2], pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

#Análise de CP dos dados padronizados

cpr<-prcomp(y,scale=TRUE)

#Comando equivalente: princomp(y,cor=TRUE)

summary(cpr)

cpr$sdev #raiz quadrada dos autovalores

cpr$rotation #autovetores

biplot(cpr)

plot(cpr$x[,1:2], pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

par(mfrow=c(1,2))

plot(cps$x[,1:2], main="CP.Y")

text(cps$x, labels=seq(1:150), lwd=3)

plot(cpr$x[,1:2], main="CP.Y\*")

text(cpr$x, labels=seq(1:150), lwd=3)

##identificando as espécies

par(mfrow=c(1,2))

plot(cps$x[,1:2], main="CP.Y", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cps$x, labels=seq(1:150), lwd=3)

plot(cpr$x[,1:2], main="CP.Y\*", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cpr$x, labels=seq(1:150), lwd=3)

##Compare as duas análises

#Escalonamento Multidimensional - Solução Métrica

#cmdscale(d, k = 2, eig = FALSE, add = FALSE, x.ret = FALSE)

#Adotando a dist Euclidiana (equivale à análise dos dados originais)

cops<- cmdscale(diste)

round(cops,2)

#x <- cops[,1]

#y <- cops[,2]

#plot(x,y)

plot(cops, main="CoP.Y")

text(cops, labels=seq(1:150), lwd=3)

##identificando as espécies

plot(cops, main="CoP.Y", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cops, labels=seq(1:150), lwd=3)

#Adotando a dist de Pearson (equivale à análise dos dados padronizados)

copr<- cmdscale(distp)

round(copr,2)

#x <- copr[,1]

#y <- copr[,2]

#plot(x,y)

plot(copr, main="CoP.Y\*", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(copr, labels=seq(1:150), lwd=3)

#Comparando CP e CoP

par(mfrow=c(2,2))

plot(cps$x[,1:2], main="CP.Y", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cps$x, labels=seq(1:150), lwd=3)

plot(cops, main="CoP.Y", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cops, labels=seq(1:150), lwd=3)

plot(cpr$x[,1:2], main="CP.Y\*", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cpr$x, labels=seq(1:150), lwd=3)

plot(copr, main="CoP.Y\*", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(copr, labels=seq(1:150), lwd=3)

###NOTE que às vezes é preciso trocar o sinal das coordenadas no gráfico

par(mfrow=c(2,2))

plot(cps$x[,1],(-1)\*cps$x[,2], main="CP.Y", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cps$x[,1],(-1)\*cps$x[,2], labels=seq(1:150), lwd=3)

plot(cops, main="CoP.Y", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cops, labels=seq(1:150), lwd=3)

plot(cpr$x[,1],(-1)\*cpr$x[,2], main="CP.Y\*", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(cpr$x[,1],(-1)\*cpr$x[,2], labels=seq(1:150), lwd=3)

plot(copr, main="CoP.Y\*", pch=23, bg=c('red', 'blue', 'yellow')[class.vector])

text(copr, labels=seq(1:150), lwd=3)

##Análise de Correspondência

boxplot(y)

soma<-rowSums(y)

boxplot(soma)

summary(soma)

taman<-c(rep(NA,150))

i=1

for(i in 1:150){

if(soma[i] <= 10.7){taman[i]<-1}

if(soma[i]> 10.7 & soma[i]<=14.3){taman[i]<-2}

if(soma[i]>14.3 & soma[i]<=16.25){taman[i]<-3}

if(soma[i]>16.25){taman[i]<-4}

i=i+1

}

table(taman)

tabi<-table(yg[,5],taman)

tabi

chisq.test(tabi)

library(ca)

fit.ca<-ca(tabi)

fit.ca

plot(fit.ca,main="Biplot Simétrico") #biplot simétrico

plot(fit.ca, map="rowprincipal", main="Rowprincipal")

#Colunas em Coord Padrão, Linhas em Coord Principais

plr<-plot(fit.ca, map="rowprincipal", main="Rowprincipal",arrows=c(FALSE,TRUE))

round(plr$row,2) #autovetor\_linha\*sqrt(autovalor)

round(plr$col,2) #autovetor\_col

##Ver Análise Discriminante realizada em Aula

####################################

## MANOVA - Iris

##Iris: p=4 e G=3

fit.m <- manova(as.matrix(iris[,-5]) ~ as.matrix(as.factor(iris[,5])))

summary(fit.m, test="Wilks")

summary(fit.m, test="Pillai")

summary(fit.m, test="Hotelling-Lawley")

summary(fit.m, test="Roy")

summary(fit.m)

aov(fit.m)

## Hipóteses sob teste: H0: mi\_G1 = mi\_G2 = mi\_G3 para mi o vetor de médias das 4 variáveis

library(DiscriMiner)

#Matriz de Soma de Quadrados e Produtos Cruzados Dentro de Grupos

#É a matriz E: é uma fonte de variabilidade devido ao Erro

SSw<-withinSS(as.matrix(iris[,-5]),as.factor(as.matrix(iris[,5])))

round(SSw,2)

#Matriz de Soma de Quadrados e Produtos Cruzados Entre Grupos

#É a matriz H: é uma fonte de variabilidade devido ao Grupo (o fator sob estudo)

SSb<-betweenSS(as.matrix(iris[,-5]),as.factor(as.matrix(iris[,5])))

round(SSb,2)

round(SSw+SSb,2)

##Qual a importância de cada variável para o ajuste?

##Decomposição espectral de: (H - lambda E)L=0

E<-SSw

H<-SSb

deE<-eigen(E)

diagE<-diag(deE$values)

solve(diagE)

Eim<-(deE$vectors)%\*%solve(diagE)^0.5%\*%t(deE$vectors)

Eim #E^(-0.5)

M<-Eim%\*%H%\*%t(Eim)

L<-Eim%\*%eigen(M)$vectors

L

t(L)%\*%E%\*%L

L[,1:2]

eigen(M)$values

eigen(M)$vectors

#o primeiro autovetor indica que a variável 3 seguida da 4

# são as que mais contribuem para a diferença entre os grupos

#Isso é confirmado pela ANOVA

summary.aov(fit.m)