SY09 TP1

1) Babies.txt

>babies<-read.table("babies23.data",header=F)

lit le fichier de donnée comme une table et le met dans babies. Header defini si R doit nommer les colonnes avec la premiere ligne de la table.

>babies<-babies[c(7,5,8,10,12,13,21,11)]

selectionne les colonnes defnies de la table babies.

>names(babies)<-c("bwt","gestation","parity","age","height","weight","smoke","education")

nomme les colonnes avec les champs definis

>babies$smoke<-factor(c("NonSmoking","Smoking","NonSmoking","NonSmoking")[babies$smoke+1])

associe chaque chaine de caractère avec la valeur du champs contenu dans la colonne babies$smoke ( le +1 met les valeurs entre 1 et 4 au lieu de 0 et 3 ). Ceci transforme une variable quantitative en qualitative.

>babies$education<-factor(babies$education,ordered=T)

defnit un ordre sur la variable qualitative education

> babies[babies$bwt == 999, 1] <-NA

> babies[babies$gestation == 999, 2] <-NA

> babies[babies$age == 99, 4] <-NA

> babies[babies$height == 99, 5] <-NA

> babies[babies$weight == 999, 6] <-NA

Ceci normalise les données suivant les conventions du fichier.

Questions

Vous vous attacherez à répondre aux questions suivantes.

1. Quelle est la différence de poids entre les bébés nés de mères qui fumaient durant leur grossesse et celles qui ne fumaient pas ?

2. Est-ce qu’une mère qui fume durant sa grossesse est encline à avoir un temps de gestation plus court qu’une mère qui ne fume pas ?

3. Le niveau d’étude a-t-il une influence sur le fait que la mère soit fumeuse ?

1. On calcule la moyenne des poids à la naissance des bébés issues de mere fumeuse et non fumeuse. Pour e faire, il a fallu creer des vecteurs de données correspondant aux poids des bébés à la naissance pour les deux cas ( fumeurs et non fumeurs) et enlever les valeurs NA. Ensuite il suffit d’executer la fonction mean pour calculer les moyennes des deux tableaux.

> mean( babies[babies$smoke==”Smoking”,1], na.rm=T)

> mean( babies[babies$smoke==”NonSmoking”,1], na.rm=T)

Moyenne :

* non fumeurs : 123,0472 oz.
* Fumeurs : 114,1095 oz.

Difference de poids moyen : 8,937666 oz.

Ecart type :

* Non fumeurs :17, 39869.
* Fumeurs : 18.09895.

Grace à une etude plus poussée sur les boxplot du poids à la naissance pour les mere fumeuses et non fumeuses. On constate que les intervalles de confiances sur la mediane pour chaque distribution sont disjoint. Le fait qu’une mere fume intervient donc significativement sur le poids du bébé à la naissance.

1. temps de gestation en jour : ( meme commande sur colonne differente ).

Moyenne :

* non fumeur : 280,1869 jours.
* Fumeur : 277,9792 jours.

Difference en moyenne : 2 jours.

Ecart type :

* Non fumeurs : 16, 62761.
* Fumeurs : 15,07364.

Apres comparaison des medianes de chaque distribution des gestation de mere fumeuse et non fumeuse, on constate, que les intervalles de confiances sur la madiane(95%) sont quasi similaires. De plus, les moyennes empiriques et ecart types sont relativement proches. Le fait qu’une mere fume n’influe donc pas signifivativement sur son temps de gestation.

[un histogramme n’utilise que des variables quantitatives, pour que R puisse faire un histo sur une qualitative, il faut la caster en quantitative avec : as.numeric(vecteur).

1. Niveau d’etude et fumeur, non fumeur :

II données crabs

1)

Cependant, en effectuant les codes R suivant :

> pairs(crabsquant, col=c("blue","red")[crabs$sex])

> pairs(crabsquant, col=c("blue","orange")[crabs$sp])

On remarque des tendances significativement différentes en ce qui concerne le sexe ou l’espce. Il sera donc possible d’identifier l’espece ou le sexe d’un crabe par rapport à une serie de mesure.

2) On remarque que toutes les données sont fonctions approximativement lineaires les unes des autres. Ceci degage une correlation entre les differente jeu de données. De plus le calcul du coefficient de correlation entre les différentes variables confirme cette importante correlation :

> cor(crabsquant$FL, crabsquant$BD)

[1] 0.9876272

> cor(crabsquant$FL, crabsquant$CW)

[1] 0.9649558

> cor(crabsquant$FL, crabsquant$CL)

[1] 0.9788418

> cor(crabsquant$RW, crabsquant$CL)

[1] 0.892743

> cor(crabsquant$RW, crabsquant$FL)

[1] 0.9069876

> cor(crabsquant$RW, crabsquant$BD)

[1] 0.8892054

> cor(crabsquant$CW, crabsquant$CL)

[1] 0.9950225

IV ACP :

1. créer la matrcie X
2. Centrer X
3. Calculer la matrice de variance S = X’DpX. Ici Dp = matrice 3,3 avec pi = 1/n = ¼.

Soit S = ¼ \* X’X.

1. Diagonaliser S : valeur propre, vecteur propre.
2. Ordonner les vecteurs propres selon les valeurs propres decroissantes.

1° Resultat :

> U

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

[1,] -0.2078616 0.9561407 -0.2063700

[2,] 0.3087780 0.2643297 0.9136662

[3,] 0.9281431 0.1261936 -0.3501793

On a donc deltaU1 = (-0.2078616, 0.3087780, 0.9281431).

Pourcentage d’inertie expliquée par chacun des axes :

> inerties <- eig$values / sum(eig$values)

> inerties

[1] 0.779056425 0.216559509 0.004384067

2° Composantes principales :

> C <- X%\*%M%\*%U

> C

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

[1,] 2.686414 4.831346 1.8550684

[2,] 3.370178 3.018622 2.3515562

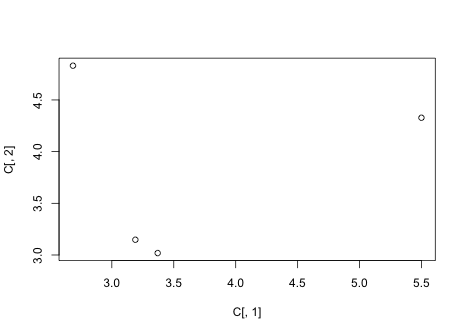
[3,] 5.499137 4.327833 0.1713040

[4,] 3.189926 3.147581 -0.9576550

/\* Cette matrice contient tous les vecteurs composante principale de la matrice de depart. Chaque vecteur colonne de cette matrice correspond à un axe factoriel. On peut donc en dégager des plans factoriels en utilisant par contre l’axe 1 et l’axe 2 ou l’axe 1 et l’axe 3. La representation des données à travers le premier plan factoriel ( axe 1 et 2 ) donne le resultat suivant : \*/

Chaque colonne de la matrice C correspond aux coordonnées des variables de la matrice de depart X projetés sur la nouvelle base constituée des vecteurs propres ordonnés de la matrice de depart. Représenter les 4 individus en fonction des deux premiers axes correspond donc à plotter la premiere colonne de C en fonction de la deuxieme. Resutlat :

plot(C[,1], C[,2])



ACP reliées aux individus par rapport aux axes 1 et 2.

