TD10

Présentation des données

Les données correspondent aux réponses à l'enquête *Histoire de Vie* de l'INSEE (Institut National de la Statistiques et des Études Économiques) de 2003. Nous nous intéressons ici à la segmentation sociale des pratiques de loisirs et de son ancrage par rapport à l'équilibre loisir-travail, dans une perspective de sociologie des pratiques culturelles.

Outre l'intérêt sociologique de ces données, le but du TD est aussi de montrer les ajustements nécessaires en termes de recodage pour réaliser une ACM lorsqu'on mobilise un jeu de données.

Nettoyage des données

1. Ouvrir la base. Elle se trouve dans le package "questionr" et peut être ouverte comme suit :

```
#On liste tous les packages nécessaires au TD et on les installe si nécessaire
list.of.packages <- c("questionr", "FactoMineR","factoextra","dplyr","gtsummary","Hmi
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,"Packa
if(length(new.packages)) install.packages(new.packages)


library(questionr)
data("hdv2003")
```

Examiner les variables (en reprenant les fonctions R que nous avons étudiées), leur type (variable quantitative ou qualitative) à l'aide des variables . Quelles sont les variables quantitatives ?

Note : la variable poids ne désigne pas la charge pondérale des répondants mais correspond à la variable de pondération. En effet, issue d'un échantillon représentatif de la société française,

la pondération permet de corriger des biais de sur- ou sous-représentation de certains groupes de l'échantillon. Il faut de ce fait inclure cette variable dans les manipulations statistiques comme nous allons le voir.

2. Filtrer la base sur les individus qui exercent un emploi grâce à la variable occup.

```
bdd<-subset(hdv2003,subset = occup=="Exerce une profession")
```

3. Sur cette nouvelle base, faire un tri à plat de l'ensemble des variables grâce à la tbl_summary du package et noter les modalités des variables qui caractérisent moins de 5% de l'échantillon. Noter également si les variables ont des modalités manquantes (affiché "Unknown" pour NA), des modalités "Ne sait pas" ou "Non répondu" par exemple.

```
var<-names(bdd)[-1]
library(gtsummary)
library(dplyr)
bdd %>% tbl_summary(include=var,missing_text="NA")
```

4. Créer une nouvelle variable âge catégorielle en quartiles (4 classes représentant 25 % de la population à chaque fois).

```
summary(bdd$age)
#Ici on a les quartiles non pondérés
library(Hmisc)
wtd.quantile(bdd$age,weights=bdd$poids)
#Les quartiles pondérés sont légèrement différents.
#La fonction cut permet de catégoriser une variable numérique
bdd$agec<-cut(bdd$age,breaks=c(18,30,41,50,67),include.lowest=T)
#on regarde si ça c'est bien passé
table(bdd$agec)
summary(bdd$agec)
```

5. Créer une nouvelle variable heures de visionnage de télévision qui, de la variable numérique, la transforme en catégorielle (classe "factor"), en utilisant les quartiles. Même chose pour la variable nombre de frères et soeurs

```
summary(bdd$heures.tv)
table(bdd$heures.tv)
wtd.quantile(bdd$heures.tv,weights=bdd$poids)
#Les quartiles pondérés sont légèrement différents.
#La fonction cut permet de catégoriser une variable numérique
```

```r
bdd$heures.tvc<-cut(bdd$heures.tv,breaks=c(0,1,2,3,10),include.lowest=T)
table(bdd$heures.tvc)
summary(bdd$heures.tvc)

wtd.quantile(bdd$freres.soeurs,weights=bdd$poids)
#La fonction cut permet de catégoriser une variable numérique
bdd$freres.soeursc<-cut(bdd$freres.soeurs,breaks=c(0,1,2,4,22),include.lowest=T)
table(bdd$freres.soeursc)
summary(bdd$freres.soeursc)
```

6. Transformer toutes les modalités “Ne sait pas”, “Rejet”, “Autre”… en NA.

```r
#La commande ci dessous est ineffective car les variables concernées sont de classe factor
bdd<-replace(bdd, c("Ne sait pas","Rejet","NSP ou NVPR"), NA)
#Solution :
bdd$relig<-factor(bdd$relig, exclude = c("Rejet","NSP ou NVPR"))
summary(bdd$relig)
#ci dessous sentiment d'appartenance à une classe sociale :
bdd$clso<-factor(bdd$clso, exclude = c("Ne sait pas"))
summary(bdd$clso)

#On peut refaire tourner le tri à plat non pondéré des variables pour voir si les changeme
bdd %>% tbl_summary(include=var,missing_text="NA")
```

7. Créer une nouvelle variable à partir de nivetud qui réunit ensemble les modalités “N’a jamais fait detudes” et “A arrete ses etudes…”. Créer une nouvelle variable à partir de trav.imp qui réunit les modalités “Le plus important” et “Aussi important que le reste”.

```r
levels(bdd$nivetud)
levels(bdd$nivetud)[levels(bdd$nivetud) %in% c("N'a jamais fait d'etudes","A arrete ses et
#Tant qu'à faire recodons les autres levels qui sont un peu longs
levels(bdd$nivetud)[levels(bdd$nivetud) %in% "Derniere annee d'etudes primaires"]<-"Primai
levels(bdd$nivetud)[levels(bdd$nivetud) %in% "Enseignement technique ou professionnel cour
levels(bdd$nivetud)[levels(bdd$nivetud) %in% "Enseignement technique ou professionnel long
levels(bdd$nivetud)[levels(bdd$nivetud) %in% "Enseignement superieur y compris technique s
levels(bdd$nivetud)

levels(bdd$trav.imp)
levels(bdd$trav.imp)[levels(bdd$trav.imp) %in% c("Le plus important","Aussi important que
levels(bdd$trav.imp)
```

8. Créer une nouvelle variable poids qui normalise cette variable de telle sorte que la somme des poids soit égale au nombre total d'individus dans la base (démarche nécessaire pour ensuite l'utiliser et faire des tests statistiques).

```
sum(bdd$poids)
bdd$poidsnorm<-bdd$poids*nrow(bdd)/sum(bdd$poids)
sum(bdd$poidsnorm)
#C'est bon.
```

Statistiques descriptives

9. Étudier l'association entre le sexe et la pratique du bricolage grâce aux pourcentages en ligne, à un test du chi-2 et au V de Cramer.

```
#Il faut utiliser une autre fonction que table car elle n'accepte pas la pondération
#Il sagit de xtabs
tab<-xtabs(poidsnorm~sexe+bricol,data=bdd)
tab
lprop(tab)
l<-lprop(tab)
library(kableExtra)
kable(l,digits=1) %>% kable_styling()

chisq.test(tab)

cramer.v(tab)
```

10. Étudier l'association entre le sexe et la pratique de la pêche et de la chasse grâce aux pourcentages en ligne, à un test du chi-2 et au V de Cramer.

```
#Utiliser les mêmes fonctions
```

11. Étudier l'association entre le niveau d'études et la pratique de la pêche et de la chasse grâce aux pourcentages en ligne, à un test du chi-2 et au V de Cramer.

```
#Utiliser les mêmes fonctions
```

12. Conclure sur l'importance relative des déterminants sociaux de la pêche et de la chasse.

13. Étudier l'association entre le niveau d'études et le cinéma grâce aux pourcentages en ligne, à un test du chi-2 et au V de Cramer.

```
#Utiliser les mêmes fonctions
```

14. Étudier l'association entre le niveau d'études et le nombre d'heures de visionnage de télévision grâce aux pourcentages en ligne, à un test du chi-2 et au V de Cramer.

```
#Utiliser les mêmes fonctions
```

Statistique multidimensionnelle

15. Justifier l'intérêt de conduire une Analyse des Correspondances Multiples (ACM) sur ce jeu de données. Dans la perspective d'étude d'un "espace social des pratiques de loisirs et de l'équilibre loisir-travail", quelles seraient les variables actives et les variables supplémentaires ?

16. Créer une nouvelle base de données en sélectionnant les variables actives et supplémentaires qui seront mobilisées dans l'analyse (écarter la variable hard-rock dont l'écoute est trop rare).

```
names(bdd)

bdacm<-subset(bdd,select=c("agec","sexe","nivetud","qualif","freres.soeursc","clso","relig
```

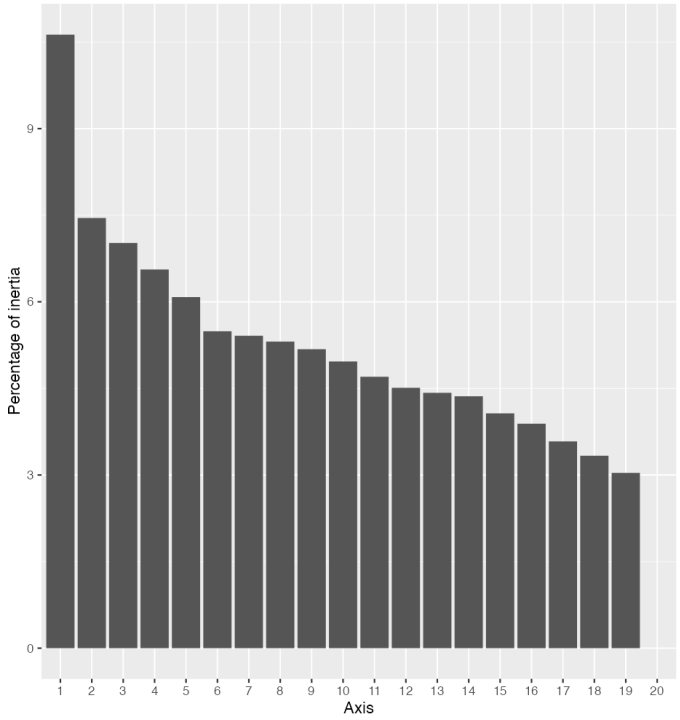
17. Lancer l'ACM, en n'oubliant pas de pondérer les individus avec la variable de poids normalisée. Avec la fonction explor, observer le nuage des individus et des modalités. Quelle observation pouvez-vous en faire ? Comment résoudre ce problème ?

```
library(FactoMineR)
names(bdacm)
acm1<-MCA(bdacm,quali.sup = c(1:6),row.w=bdd$poidsnorm,graph = F)
library(explor)
explor(acm1)
```

Dimensions to plot

20

Eigenvalues histogram



Eigenvalues table

Axis	%	Cum. %
1	10.6	10.6
2	7.5	18.1
3	7.0	25.1
4	6.6	31.7
5	6.1	37.7
6	5.5	43.2
7	5.4	48.6
8	5.3	53.9
9	5.2	59.1
10	5.0	64.1
11	4.7	68.8

18. Repérer les modalités qui peuvent être mises en supplémentaire.

```
mod<-data.frame(acm1$var$coord)
row.names(mod)
#5,9,13,30
```

19. Relancer une ACM en excluant ces modalités.

```
acm2<-MCA(bdacm,quali.sup = c(1:6),excl=c(5,9,13,30),row.w=bdd$poidsnorm,graph = F)
#explor(acm2)
```

20. Réaliser le graphique du carré des corrélations sur les deux premiers axes. Par quelles variables actives les deux premiers axes sont-ils structurés ? Y-a-t-il des variables supplémentaires qui semblent également expliquer ces axes ?

```
plot(acm2,choix="var")
```

21. Suivre pas à pas les différentes étapes de l'analyse de l'ACM pour l'interpréter. Pour les variables actives, retenir les modalités qui ont une contribution supérieure à la moyenne sur chacun des axes. Pour les variables supplémentaires, retenir les modalités qui sont significativement associées à chacun des axes.

```
#On peut utiliser explor
#On peut aussi utiliser factoextra

library(factoextra)
#Sur l'axe 1 on sélection les modalités actives qui ont une contribution sup à la contribu
contrib1<-rownames(acm2$var$contrib)[acm2$var$contrib[,1]>100/nrow(acm2$var$contrib)]
#Sur l'axe 1 on sélection les modalités supplémentaires qui sont significatives
sigsup1<-rownames(acm2$quali.sup$v.test)[abs(acm2$quali.sup$v.test[,1])>1.96]
#On projette:
fviz_mca_var(acm2,axes=c(1,2),
             select.var=list(name = c(sigsup1,contrib1)))
```

```
#On peut refaire la même chose sur l'axe 2...
```

22. Dessiner les ellipses de concentration des modalités supplémentaires des variables qualif et nivetud.

```
fviz_mca_ind(acm2,
             label = "none", # hide individual labels
```

```
             axes = c(1, 2), #les axes de projection
             habillage = bdacm$qualif, # color by groups
             addEllipses = TRUE # Concentration ellipses
)
```

```
fviz_mca_ind(acm2,
             label = "none", # hide individual labels
             axes = c(1, 2), #les axes de projection
             habillage = bdacm$nivetud, # color by groups
             addEllipses = TRUE # Concentration ellipses
)
```

23. Créer une nouvelle variable qui croise le sentiment d'appartenance à une classe (clso) et le niveau de qualification professionnel (clso). Faire un tri à plat pondéré de cette variable. Projeter cette nouvelle variable croisée en supplémentaire dans une nouvelle ACM. Cette variable explique-t-elle mieux un des axes que les variables considérées indépendamment l'une de l'autre ? Quel est l'indicateur statistique pertinent pour l'affirmer ?

```
summary(bdacm$qualif)
summary(bdacm$clso)

bdacm$qualifclso<-ifelse(!is.na(bdacm$qualif) & !is.na(bdacm$clso),paste(bdacm$qualif,bdac

tab<-xtabs(bdd$poidsnorm~qualifclso,data=bdacm)
freq(tab)

names(bdacm)

acm3<-MCA(bdacm,quali.sup = c(1:6,17),excl=c(5,9,13,30),row.w=bdd$poidsnorm,graph = F)
#explor(acm3)
```

24. Conclure sur l'espace social des pratiques de loisir à partir de cette analyse.