Introduction à R

Introduction

- Logiciels click-souris : Modalisa, SPAD
- Logiciel de programmation : SAS, SPSS, STATA, R
- Avantage de R : libre, en développement permanent grâce au Packages, graphiques

Validation

- Dossier entre 10 et 20 pages sur un thème de votre choix à partir de l'enquête TeO
- Doit contenir une problématique, des statistiques descriptives (tableaux de contingence, graphiques)
- Une analyse statistique multivariée (régression linéaire, logistique ou analyse factorielle)

Créer et manipuler des objets

- Assignation :

a<-2 ou a=2

a=2

b=2

c=a+b

- Créer un vecteur :

a= c(1,2,3,4) ou a= c(1:4)

b=c(5:8)

- Combiner des vecteurs

d=a+b

d2=a*b

d3=c(a,b)

Vecteurs caractères et logiques

- Vecteurs numériques

v=(1:4)

- Vecteurs caractères

v= c(“a”, “b”, “c”)

- Vecteurs logiques

v=c(T, T, F, T)

Les fonctions

- v=c(1:10)
- length(v)
- Les fonctions statistiques classiques
  mean(v)
  min(v)
  median(v)
  summary
  var(v)
  sd(v)
- Trouver de l'aide sur une fonction
  help(sd)

Fonctions et arguments

- Un argument peut prendre une valeur logique (T ou F), une valeur numérique, ou une valeur caractère
- v=c(1:10, NA)
- mean (v) =NA
- mean (v, na.rm=T)
- c=seq(from=1, to =101, by=2)

Packages

- Ils constituent des extensions : ils ajoutent des données, des fonctions, des interfaces graphiques, etc.

- Télécharger / Installer
  - install.packages ("nom_package", dep=T)
- Charger le package
  - library (nom_package)

Inspecter un tableau

- Le tableau hdv2003 du package questionr
  - data(hdv2003)
  - d=hdv2003
- Inspecter le tableau
  - colnames (d)
  - rownames(d)
  - nrow(d)
  - ncol(d)
  - str(d)
  - names(d)
  - head(d)

Manipuler les variables d'un tableau (1)

```
d$age
head(d$age)
tail(d$age,20)
hist ( d$heures.tv,
        probability=T,
        breaks=c(0,1,4,6,12),
        col=c("red","blue","yellow"","green"),
        main="Nombre d'heures passées devant la
        télévision",
        xlab="Proportion en %",
        ylab="Nombre d'heures"
        )
```

Manipuler les variables d'un tableau (2)

- Indexer
  - Un vecteur : v [5]
  - Une variable : d$heures.tv [5]
  - Un tableau
    - Une observation : d[5,20]
    - Une ligne : d[5,]
    - Une colonne : d[,20]
  - mean (d[,20], na.rm=T)
  - mean (d[1:5,20], na.rm=T)
  - mean (d[c(1:5, 20:40),20], na.rm=T)
  - mean (d[-c(1:5, 20:40),20], na.rm=T)

Les opérateurs logiques

- == égal à
- != différent de
- & et
- | ou

```
mean(d[d$qualif=="Ouvrier specialise",20], na.rm=T)
mean(d[d$qualif=="Ouvrier specialise", colnames(d)=="heures.tv"],
    na.rm=T)

mean(d[d$qualif=="Ouvrier specialise"|d$qualif=="Ouvrier
    qualifie",20], na.rm=T)

mean(d[d$qualif=="Ouvrier specialise" & d$sexe=="Homme",20],
    na.rm=T)
```

2 Variables qualitatives

- Créer un tableau (1 à n dimensions)

```
table(d$sexe)
table(d$sexe,d$occup)
addmargins(table(d$sexe,d$occup))
```

- Faire du tableau un objet

```
tab1=table(d$sexe,d$occup)
margin.table(tab1) ## total
margin.table(tab1,1) ## Marges en ligne
margin.table(tab1,2) ## Marges en colonne
addmargins(tab1)  ## Ajouter les marges au tableau
summary(tab1) ## test du chi²
```

% lignes et colonnes

- prop(tab1) ## %
- lprop(tab1) ## % lignes
- cprop(tab1,percent=T, digits=3) ## % colonnes
- Ajouter une colonne et la renommer
  - tab2=lprop(tab1,percent=T)
  - tab3=cbind(tab2,c(margin.table(tab1,1),sum(tab1)))
  - colnames(tab3)[9]="Effectif"

Export

- La fonction copie se trouve dans le package questionr
- copie(tab3) ## coller sur excel
- copie(tab3,

  file=T,

  filename= "mon_dossier/tableau.html")
  - Utiliser des / et non des \

2 Variables quantitatives

- Calculer une corrélation

```
data(rp99) ## données du recensement
d2=rp99
cor(d2$tx.chom,d2$proprio) ## corrélation linéaire
```

- Faire un graphique de la corrélation

```
plot(d2$tx.chom,d2$proprio,)
plot(d2$tx.chom,d2$proprio,
        pch=16,
        cex=0.7,
        main=« Chômage et accès à la propriété »,
        xlab=« Taux de chômage »,
        ylab= «Taux de propriétaires » )
```

Faire un ajustement linéaire

- Utiliser la fonction lm()

```
M1=lm(d2$proprio~d2$tx.chom)
summary(M1)
abline(M1, col="red")
```

Ajouter les labels

- ville=tolower(as.character(d2$nom)) ## la variable nom est un facteur
- text(d2$tx.chom,d2$proprio+1,ville, cex=0.4)
- Utilisation du package maptools
  - install.packages("maptools",dep=TRUE)
  - library(maptools)
  - pointLabel(d2$tx.chom,d2$proprio,labels=c(ville), cex=0.4) ## permet un meilleur placement des labels

Sélectionner certains labels

- Création d'une nouvelle table

```
d3=d2[d2$proprio<52 | d2$proprio>84,]
ville2=tolower(as.character(d3$nom))
plot(d2$tx.chom,d2$proprio,pch=16,cex=0.5)
abline(M1, col='red')
pointLabel(d3$tx.chom,d3$proprio,
   labels=c(ville2), cex=0.4)
```

Matrice de graphiques

```
par(mfrow=c(2,2))

    ### Graph 1
plot(d2$tx.chom,d2$proprio,pch=16,cex=0.5)
M1=lm(d2$proprio~d2$tx.chom)
abline(M1, col="red")

    ### Graph 2
plot(d2$tx.chom,d2$hlm,pch=16,cex=0.5)
M2=lm(d2$hlm~d2$tx.chom)
abline(M2, col="red")

    ### Graph 3
plot(d2$ouvr,d2$hlm,pch=16,cex=0.5)
M3=lm(d2$hlm~d2$ouvr)
abline(M3, col="red")

    ### Graph 4
plot(d2$cadres,d2$proprio,pch=16,cex=0.5)
M4=lm(d2$proprio~d2$cadres)
abline(M4, col="red")
```

1 variable quali et une variable quanti

- Moyenne de la v. quanti selon les modalités de la v. quali
  - tapply(dage,dhard.rock,mean)

- Test de différence des moyennes
  - t.test(d$age ~d$hard.rock)

- Graphiques pour illustrer les différences
  - boxplot(d$age ~d$hard.rock)

- Construire deux histogrammes
  - d.hard=subset(d,hard.rock=="Oui")
  - d.non.hard=subset(d,hard.rock=="Non")

  - par(mfrow=c(1,2))
  - hist(d.hard$age,col='red')
  - hist(d.non.hard$age,col='red')

Importer des données

- a=read.table
  (« mondossier/EB_country_trend.csv",
  sep="," , header=T)
- Représenter par une courbe les attitudes des français à l'égard de l'Europe entre 1973 et 2012
  - Utiliser les variables Gd_thing1973 à Gd_thing2012

- colnames (a) pour connaître les variables
  - Variables Gd_Thing1973 à 2012 : colonnes 1 à 40
- rownames(a) pour connaître les lignes
  - Lignes 12 pour la France
- Utiliser la fonction plot

- plot(1973:2012, c(a[12,1:40]),
  type='l',
  ylim=c(0,100),
  main="Pour mon pays, faire partie de l'Union européenne est une bonne chose 73-11",
  xlab="Année",
  ylab="%",
  col='blue',
  cex.main=1)

- Ajouter des lignes

```
lines(1973:2012,c(a[15,1:40]),col='deepskyblue1') ### Grande Bretagne
lines(1973:2012,c(a[20,1:40]),col='violet')     ### Italie
lines(1973:2012,c(a[27,1:40]),col='orange') ### Pays-Bas
```

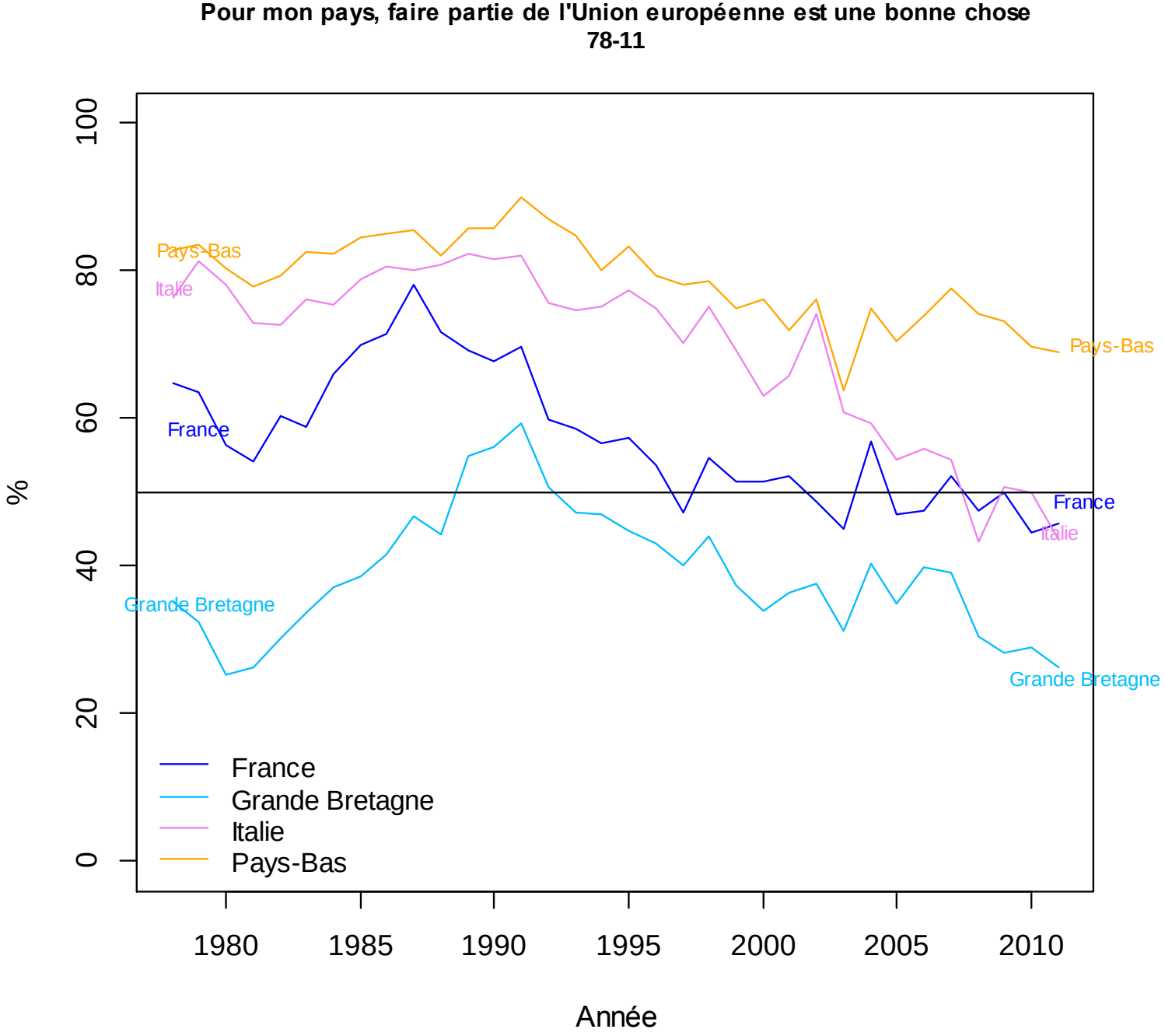
- Ajouter une légende

```
legend("bottomleft", legend = c("France" ,"Grande Bretagne","Italie","Pays-Bas"),
       col = c('blue', 'deepskyblue1','violet', 'orange'),
       lty=1,
       horiz = FALSE,
       bty='n',
       cex=0.9 )
```

- par(xpd=T)

- text(2013,70, label='Pays-Bas',cex=0.7,col='orange')
- text(1979,83, label='Pays-Bas',cex=0.7,col='orange')
- text(2012,49, label='France',cex=0.7,col='blue')
- text(1979,59, label='France',cex=0.7,col='blue')
- text(2011,45, label='Italie',cex=0.7,col='violet')
- text(1978,78, label='Italie',cex=0.7, col='violet')
- text(2012,25, label='Grande Bretagne',cex=0.7,col='deepskyblue1')
- text(1979,35, label='Grande Bretagne',cex=0.7,col='deepskyblue1')

- par(xpd=F)


Pour mon pays, faire partie de l'Union européenne est une bonne chose
78-11
%
100
80
60
40
20
0
1980
1985
1990
1995
2000
2005
2010
Année
Pays-Bas
Italie
France
Grande Bretagne
France
Grande Bretagne
Italie
Pays-Bas
Pays-Bas
France
Italie
Grande Bretagne

Rcommander

- Interface graphique pour utiliser R en click-souris
- Install.packages(« Rcmdr », dep=T)

Recodage

- Créer une nouvelle variable d$qualifreg à partir de d$qualif et :
  - Recoder technicien en Profession intermédiaire
  - Regrouper Ouvrier qualifié et Ouvrier spécialisé dans la catégorie Ouvrier

- Créer var d$act_manuelles à partir de d$cuisine et d$bricol, créer les modalités
  - "Cuisine et Bricolage "
  - "Cuisine seulement "
  - "Bricolage seulement "
  - "Ni Cuisine ni bricolage"

Recodage, regrouper les modalité d'une variable qualitative

- d$qualif.reg=as.character(d$qualif)
- d$qualif.reg[d$qualif=="Ouvrier specialise"]="Ouvrier"
- Utilisation de l'opérateur logique |
  - d$qualif.reg[d$qualif=="Ouvrier specialise" | d$qualif=="Ouvrier qualifie"]="Ouvrier«
  - d$qualif.reg[d$qualif=="Profession intermediaire"| d$qualif=="Technicien"]="PI"

- ## avec l'utilisation de %in%

```
d$qualif.reg[d$qualif %in% c("Ouvrier
   specialise", "Ouvrier qualifie")]="Ouvrier"

d$qualif.reg[d$qualif %in% c("Profession
   intermediaire", "Technicien")]="PI"
```

- ## on reconvertit la variable en facteur

```
d$qualif.reg=factor(d$qualif.reg)
```

- ### Combiner plusieurs variables

```
d$act_manuelles = "NA"
d$act_manuelles[d$cuisine=="Oui" &
   d$bricol=="Oui"]="Cuisine et Bricolage"
d$act_manuelles[d$cuisine=="Oui" &
   d$bricol=="Non"]="Cuisine seulement"
d$act_manuelles[d$cuisine=="Non" &
   d$bricol=="Oui"]="Bricolage seulement"
d$act_manuelles[d$cuisine=="Non" &
   d$bricol=="Non"]="Ni Cuisine ni bricolage"

table(d$act_manuelles)
```

Découper une variable numérique en classe

```
d$age5cl = cut(d$age,5)
table(d$age5cl)

d$age5cl = cut(d$age, c(0,20,30,40,60,80,100))
range(d$age)

d$age5cl = cut(d$age,c(0,20,40,60,100),
   labels=c("<20", "21-40","41-60", "60-100"))
```

L’enquête Trajectoires et Origines

- Enquête réalisé par l’INED et l’INSEE
  - A entrainé beaucoup de controverses sur la question des statistiques ethniques
  - Permet d’évaluer le nombre d’enfant d’immigrés sur la première génération

- Enquête à obtenir auprès de l’ADISP, équipe du réseau Quételet

Population

- Ensemble des personnes d'âge actif vivant dans un ménage ordinaire en France métropolitaine en 2008.
- Cinq sous-populations sont distinguées
  - Les immigrés nés entre 1948 et 1990 (18-60 ans en 2008)
  - Personnes nés dans les DOM, 1948 et 1990 (18-60 ans en 2008)
  - les descendants d'un parent immigré, c'est-à-dire nés en France métropolitaine d'une personne née étrangère à l'étranger, nés entre 1958 et 1990 (âgés de 18 à 50 ans)
  - les descendants d'un parent né dans un Dom, nés entre 1958 et 1990 (âgés de 18 à 50 ans) ;
  - Les autres personnes, nées entre 1948 et 1990 (âgées de 18 à 60 ans en 2008), n'appartenant à aucun de ces groupes. Ces "autres" constituent un groupe que l'on appelle "témoin" dans la documentation de l'enquête, mais que l'on peut aussi désigner comme "population de référence" ou "groupe ou population majoritaire". Il faut noter que ce groupe inclut des personnes qui ont aussi un lien à la migration vers la France métropolitaine : les Français nés à l'étranger et leurs enfants, les rapatriés de l'empire colonial et leurs enfants, les personnes originaires des COM et leurs enfants, les personnes qui ont une ascendance immigrée lointaine (dont au moins un aïeul était immigré)…
  - P. 5 du dictionnaire des codes

Thèmes de l’enquête

- Tableau des habitants du logement (THL) ;
- Revenus (du ménage) ;
- Nationalité et origine des parents ;
- Langues ;
- Trajectoires migratoires et transnationalisme ;
- Enfants (de l'enquêté) ;
- Image de soi et regard des autres ;
- Éducation ;
- Vie professionnelle ;
- Religion ;
- Vie matrimoniale ;
- Logement et cadre de vie ;
- Vie citoyenne ;
- Santé ;
- Discriminations ;
- Relations sociales.

Exercice

- Construire une variable de nationalité
  - Par ex : Algérien, Chinois, Français, Laotien, Marocains
  - Vous pouvez aussi faire les regroupements que vous considérez pertinents (Amérique du Nord, Asie de Sud Est, Maghreb, etc.)
- Construire une variable d'origine

- Utiliser la variable inat, p. 253

- NATE1NIV2 (p. 323) avec la nomenclature nationalité (p. 504)