Introduction à l'analyse des statistiques des données : travaux pratiques avec le logiciel "R".

Christophe Pallier[1] Christophe Lalanne[2]

25 février 2005

[1]www.pallier.org
[2]christophe.lalanne.free.fr

Résumé

Ce document est destiné à accompagner des travaux pratiques d'introduction à l'analyse des données expérimentales avec le logiciel R. Le traitement des mêmes exemples sous Statistica est également présenté.

Ce document est disponible en version pdf et en version html, à partir de l'adresse http ://www.pallier.org/ressource

Table des matières

1 Introduction à R

R est un logiciel pour l'analyse statistique des données. Il fournit les procédures usuelles (t-tests, anova, tests non paramétriques...) et possède des possibilités graphiques performantes pour explorer les données. Pouvant être utilisé aussi bien en mode interactif qu'en mode batch, R est un logiciel libre, dont le code source est disponible et qui peut être recopié et diffusé gratuitement. Des versions compilées de R sont disponibles pour Linux, Windows et Mac OS X.

Au moment de la rédaction de ce document (Octobre 2004), la version courante de R est la 2.0.

1.1 Installation du système de base

Le site principal du logiciel R est www.r-project.org.

Le téléchargement de R se fait à partir d'un des sites du "*Comprehensive R archive Network*" (CRAN), par exemple cran.cict.fr (cf. Fig. 1.1).

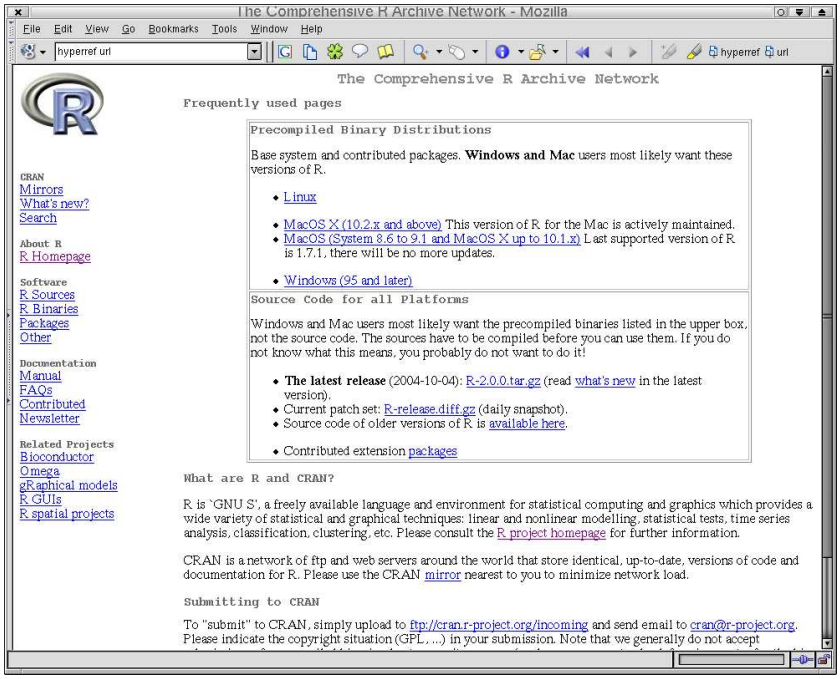

FIG. 1.1 – Site de téléchargement de R

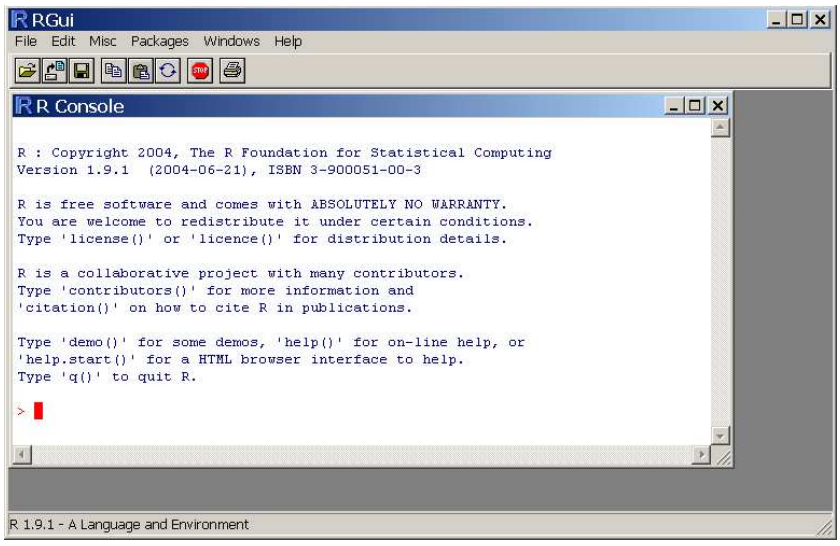

FIG. 1.2 – RGui : l'interface graphique de R sous Windows

Installation sous Windows : Le programme d'installation pour Windows est accessible en suivant les liens "Windows", puis "Base". Le nom de ce programme dépend de la version, il s'agit, par exemple, de "`Rw2000.exe`" pour la version 2.0. Téléchargez ce fichier sur votre disque, puis cliquez-le pour installer le logiciel. Si vous acceptez l'option par défaut "Create a desktop icon", une icône représentant une lettre "R" en bleu est ajoutée sur le bureau. Cliquez dessus, pour voir apparaître la fenêtre 'RGui' ("R Graphical User Interface", voir figure 1.2)

Installation sous Mandrake Linux : Pour une installation sous Linux, vérifiez s'il existe un paquetage "rpm" adapté à votre distribution et prêt à être installé. Si tel est le cas, télécharger-le et installez le, en tant qu'administrateur, avec la commande "`rpm -i R*.rpm`".

Si vous utilisez Mandrake Linux 10.x, R fait partie de la distribution de base (il est sur les CD), et il suffit de taper "`urpmi R-base`" pour l'installer.

En l'absence de binaire précompilé, il vous faudra récupérer le code source (`R-2.0.0.tar.gz`) et le compiler avec une commande '`configure && make && make install`' (en tant qu'utilisateur 'root'). Cela ne doit pas poser de problème mais nécessite que les outils de compilation soient bien installés sur votre système (notamment le compilateur fortran g77).

Pour lancer R sous Linux, il suffit de taper "`R`" dans un terminal (cf. Fig. 1.3).

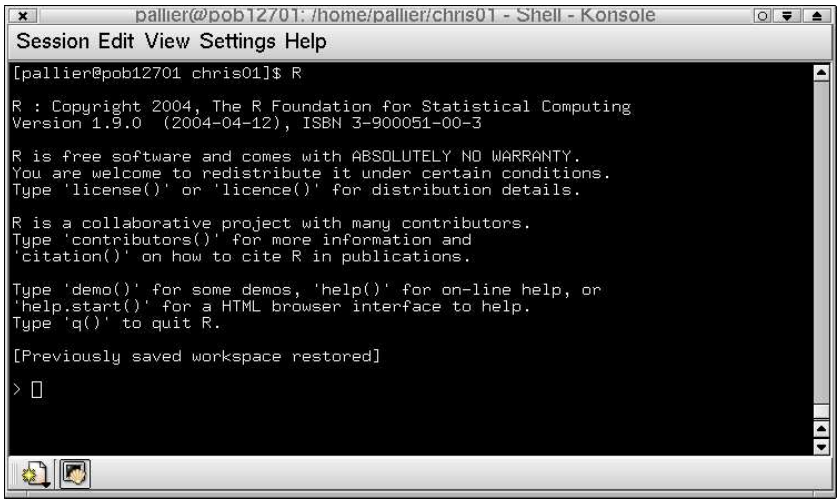

FIG. 1.3 – R dans un terminal sous Linux

1.2 Modules additionnels

Après avoir installé le système de base, vous pouvez installer des modules supplémentaires, parfois appelés "paquetages" (packages), qui ajoutent des fonctions à R.

Pour l'analyse des données d'expériences, les paquetages `car`, `gregmisc`, `vcd`, `psy`, `multcomp` fournissent des fonctions supplémentaires intéressantes. Par exemple, "`multcomp`" fournit diverses procédures pour effectuer des comparaisons multiples (Dunnett, Tukey, Sequen, AVE, Changepoint, Williams, Marcus, McDermott, Tetrade).

Ces modules sont disponibles sur les sites CRAN dans la section "*Contributed extension packages*".

Pour installer un module sous Windows, dans RGui, utiliser le menu 'Package/Install package from CRAN' (il faut être connecté a Internet).

Pour installer un module sous Linux, il faut d'abord télécharger le fichier `package.tar.gz` du CRAN, puis, en tant que root, exécuter :

```
R CMD INSTALL package.tar.gz
```

1.3 Interfaces graphiques

Rest un programme avec lequel on communique en tapant des commandes plutôt qu'en cliquant dans des menus ou sur des icônes.

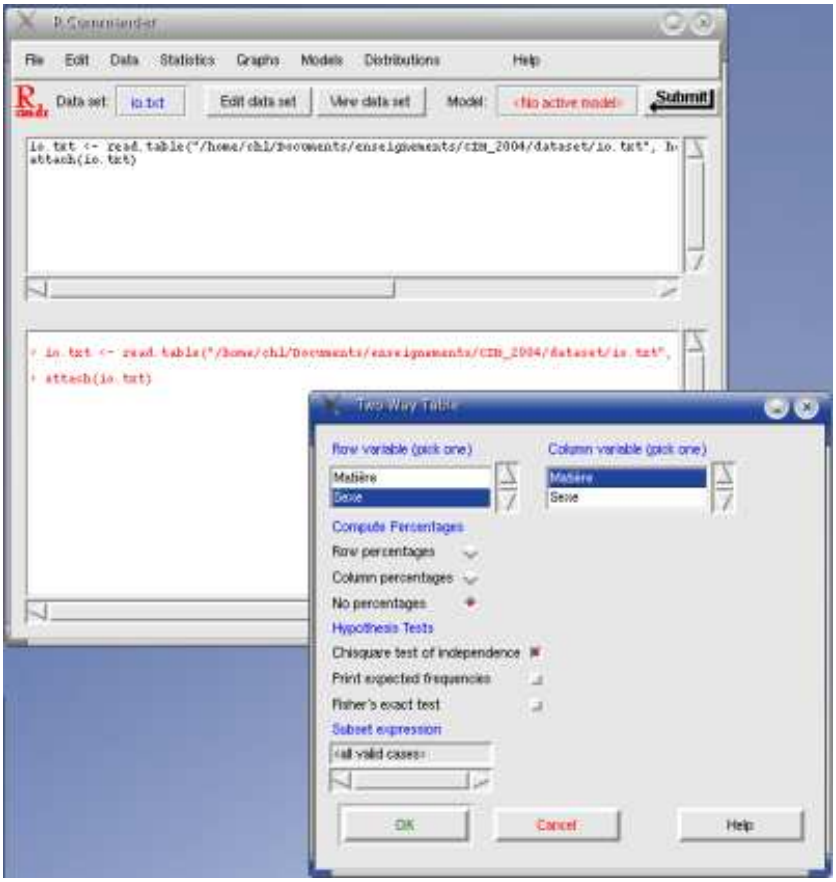

FIG. 1.4 – L'interface Rcommander sous Linux

Il existe cependant des systèmes à bases de menu et d'icônes (des "cliquodrômes") qui gèrent l'interaction avec R, et permettent, plus ou moins, d'éviter de taper des commandes. Citons, entre autres, les interfaces graphiques "Rcommander" (Linux + Windows, cf. figure 1.4), et "SciViews" (Windows).

Néanmoins, il nous paraît qu'apprendre les commandes de R permet de mieux comprendre ce qu'on fait et autorise finalement plus de flexibilité. Pour ces TPs, nous avons fait le choix de vous enseigner les rudiments du langage R.

1.4 Documentation

De nos jours, beaucoup de gens trouvent naturel de pouvoir utiliser les logiciels sans lire de documentation. Si cela est raisonnable pour les logiciels qui réalisent des opérations assez simples, c'est dangereux avec les logiciels qui effectuent des opérations conceptuellement compliquées.

Dans le cas de R, qui comprend de nombreuses commandes, il est illusoire d"envisager utiliser ce logiciel sans lire un minimum de documentation. Notre expérience est que les premières heures d'analyse de données avec R nécessitent de fréquents recours aux documentations, mais lorsqu'on est devenu à l'aise, alors il n'y a pratiquement plus besoin de s'y référer.

Il est donc utile de savoir où chercher l'information a propos de R.

Pour les débutants, on trouve sur Internet un bon nombre de documents sur R, notamment dans la section "Documentation/Contributed" du site www.r-project.org. Mentionnons en particulier :

- R pour les débutants par Emmanuel Paradis.
- Introduction au système R par Yves Brostaux.
- le site *Statistiques avec R*, réalisé par Vincent Zoonekynd, à l'adresse suivante : http ://zoonek2.free.fr/UNIX/48_R/all.html.
- Introduction to analysis of variance with "R", (qui bien qu'inachevé, vous sera sans doute utile).
- Notes on the use of R for psychology experiments and questionnaires par Jonathan Baron and Yuelin Li.

R possède aussi une documentation officielle, sous forme de fichiers pdf et html, qui est copiée sur votre disque dur lors de l'installation du logiciel. Dans l'interface graphique sous Windows, les manuels au format pdf sont accessibles dans les menus `Help/Manuals`. Il est fortement conseillé de parcourir, au minimum, les deux documents "*An Introduction to R*" et "*R Data Import et Export*".

Les manuels sont également accessibles sous forme html, dans le menu `Help/Html help` sous Windows, et en tapant `help.start()` sous Linux. Cela ouvre votre navigateur Internet sur une page web locale qui contient divers liens, entre autres vers ces manuels. Par exemple, le lien `Packages/base` liste les commandes de bases de R.

Il existe plusieurs livres publiés qui traitent de R. Pour les débutants, les deux livres suivants peuvent offrir une aide utile :

- *Introductory statistics with R* par Peter Dalgaard édité par Springer-Verlag.
- *An R and S-plus companion to applied regression* par John Fox, édité par Sage publications.

Pour un niveau plus avancé :

- *Modern Applied Statistics with S-PLUS* par Venables et Ripley.
- *Mixed-Effects Models in S and S-PLUS* par Pinheiro et Bates

2 Premiers pas

L'interaction avec R se fait en tapant des commandes dans la fenêtre `R Console`.

2.1 Entrer des commandes dans la console R

Pour commencer, vous pouvez utiliser R comme une calculatrice. Cliquez dans la fenêtre 'R Console', puis tapez :

```
2+3
```

Le résultat, '5', doit s'afficher.

Poursuivez avec :

```
a=5
a+8
```

RGui doit se présenter comme sur la Figure 2.1 page suivante.

Le principe de R est le suivant : vous entrez une ligne de commande, et quand vous tapez sur 'Entrée', R lit cette ligne et effectue l'opération demandée.

Essayez maintenant les commandes suivantes :

```
a=1:10
a
b=rnorm(10)
plot(a,b)
plot(a,b,pch=16,col=2)
```

La commande `plot` provoque l'affichage d'une fenêtre graphique (Fig. 2.2).

Cliquez à nouveau dans la fenêtre `R Console`, puis tapez :

```
a=c(3,4,6,7,8,9)
a
length(a)
b=c('alpha','beta')
b
length(b)
```

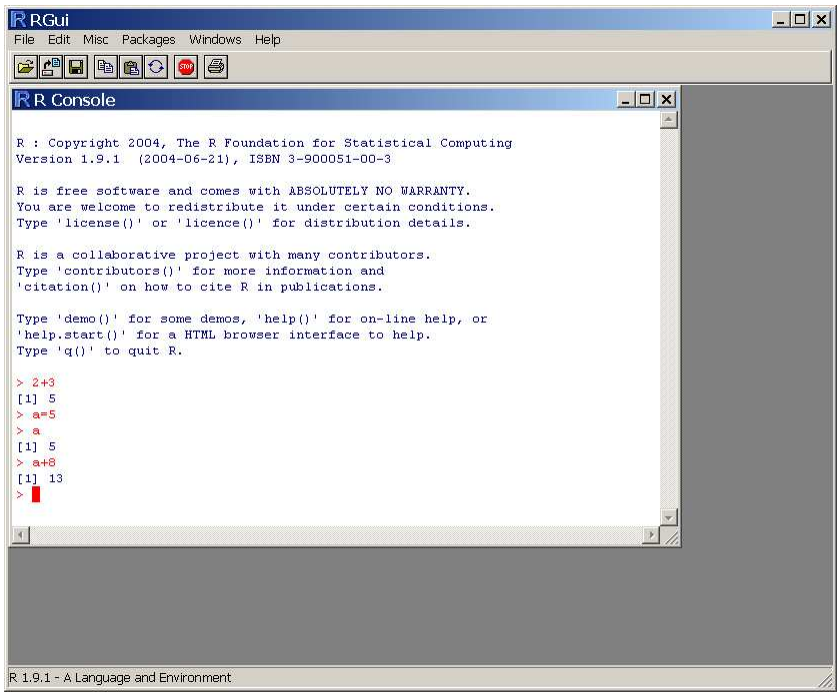

FIG. 2.1 – De simples additions

La variable 'a' contient un vecteur numérique à six éléments.

La variable 'b' contient un vecteur contenant deux chaînes de caractères.

Les concepts de *vecteur* et de *variable* sont essentiels dans R. On y reviendra plus tard ; pour le moment, retenez que :

– un vecteur n'est rien d'autre qu'un suite d'items qui ont tous le même type (numérique, chaîne de caractères, ...). C'est l'objet de base dans R.
– une variable contient un objet, et permet de le retrouver sans le ré-écrire en entier.

Comme on l'a déjà vu, la liste des variables peut être affichée par "`ls()`.", et un variable peut être détruite par la commande "`rm(nom)`".

Entrez les commandes suivantes, pas à pas, et observez le résultats :

```
a=rnorm(20,mean=55,sd=10)
mean(a)
sd(a)
max(a)
summary(a)
hist(a)
boxplot(a)
```

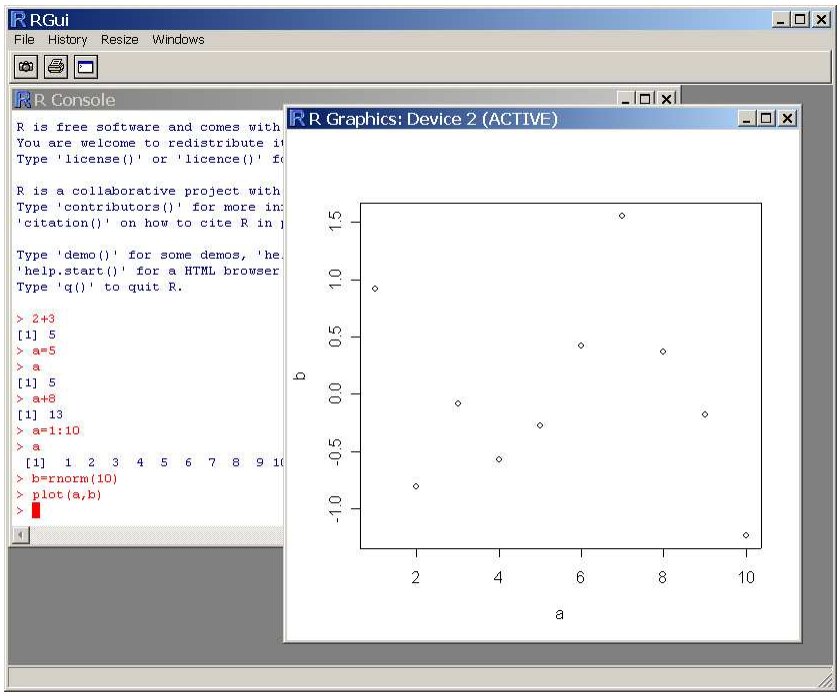

FIG. 2.2 – Fenêtre graphique

```
stripchart(a)
stripchart(a,pch=16,cex=2,col=2,method='jitter',vertical=T)

x1=rnorm(10,mean=100,sd=10)
x2=rnorm(10,mean=110,sd=10)
boxplot(x1,x2)
t.test(x1,x2)
plot(x1,x2)
summary(lm(x2~x1)
```

2.2 Aide en ligne

A tout moment, une aide en ligne est disponible à l'aide de la commande *help.search('mot clé')*. La description détaillée d'une commande s'obtient en tapant ' *?nom_de_la_commande*'.

Essayez :

```
?t.test
help.search("test")
help.start()
```

2.3 Quitter R

La fenêtre "R Console" étant active, sélectionnez "File/Exit" et répondez "Oui" à la question "Save workspace image ?"

Et voilà...

Tout votre travail est-il perdu ?

Non. Redémarrez R, et remarquez la ligne :

```
[Previously saved workspace restored]
```

Tapez "`ls()`" et constatez que vos variables sont toujours là.

Le "*workspace*" ("espace de travail"), c'est à dire l'ensemble des variables, a été sauvegardé sur le disque. Cela permet de reprendre une analyse de données au point où on l'a laissée quand on a quitté R.

Si vous voulez "nettoyer" le workspace, c'est à dire supprimer toutes les variables qu'il contient, tapez la commande "`rm(list=ls())`".

Il est possible de choisir le nom de fichier où est sauvegardé le workspace (par défaut "`.RData`"). Cela permet de faire plusieurs analyses indépendantes sans les mélanger. (Voir les menus File/Load workspace/ Save Workspace). Une alternative plus recommandée et de créer un dossier pour chaque analyse de données indépendantes.

2.4 Sauvegarder les commandes dans un script

Tapez la commande `history()`. Une fenêtre s'affiche listant les dernières commandes que vous avez tapées (voir figure 2.3 page suivante).

La manière la plus efficace de travailler avec R consiste à sauvegarder les commandes au fur et à mesure dans un fichier texte. Pour cela, en parallèle avec R, ouvrez un éditeur de fichier texte (le plus simple d'entre eux, bien qu'il soit très limité, est le bloc-notes de Windows disponible dans les accessoires).[1]

En utilisant le copier/coller, copier dans le fichier texte les commandes qui font l'essentiel de l'analyse. A la fin de votre session de travail, sauvez ce fichier avec un nom explicite (par exemple le nom de l'expérience) et une extension ".R".

[1] Pour ceux qui emploient l'éditeur Emacs, il existe un package appelé ESS qui fournit la colorisation syntaxique des commandes R, et plein d'autres fonctions utiles (voir stats.ethz.ch/ESS.

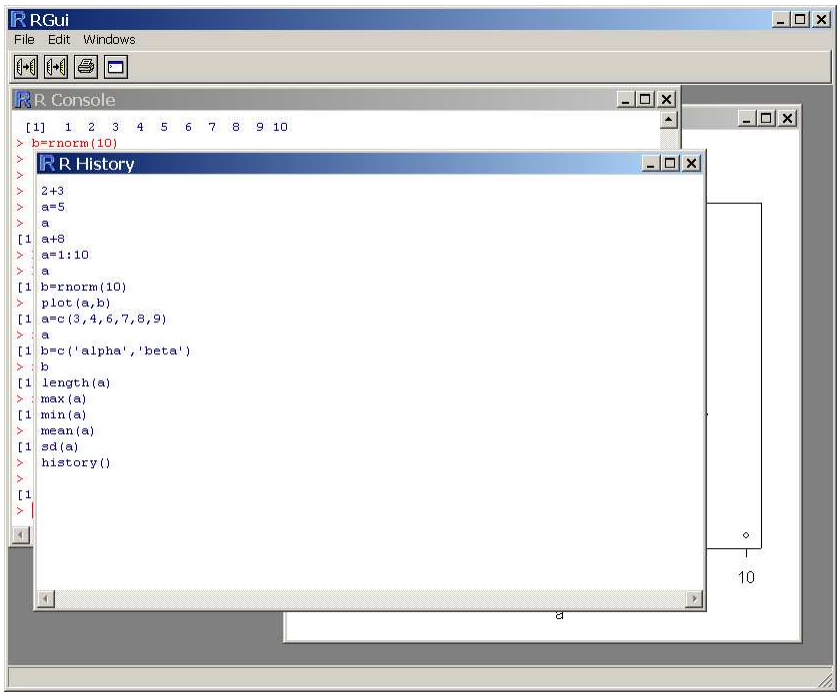

FIG. 2.3 – Historique des commandes affiché par `history()`

Quand vous reprendrez cette analyse quelques jours ou mois plus tard, vous pourrez réutiliser ce fichier, qu'on appelle habituellement un script. R vous permettra de ré-executer les commandes de ce script en utilisant la commande `source`.

Faites un essai : créez un fichier qui contient les lignes suivantes :

```
a=rnorm(100)
b=rnorm(100)
summary(a)
summary(b)
cor.test(a,b)
```

Sauvez-le dans "Mes documents", sous le nom "`test.R`".

Dans R, utilisez le menu "File/Change Dir" pour aller dans "Mes Documents". Puis tapez :

```
source('test.R',echo=T)
```

Vérifiez que cela marche.

Sous Linux, il n'est pas nécessaire de démarrer R : on peut entrer "R BATCH script.R" sur une ligne de commande dans un terminal et les résultats sont écrits automatiquement dans le fichier 'script.Rout'.

2.5 Sauver les résultats d'une analyse

Les commandes et les résultats des analyses statistiques et les graphiques peuvent être copiés/collés dans un document.

Les résultats (sans les commandes) peuvent être copiés automatiquement dans un fichier texte grâce à "`sink`". Tapez :

```
sink('monanalyse.txt',split=T)
a=1:10
mean(a)
summary(a)
sink()
```

Puis ouvrez le fichier "`monanalyse.txt`".

Les graphiques peuvent être sauvés directement dans des fichiers graphiques en utilisant les commandes `postscript`, `jpeg` ou `png` (voir l'aide en ligne de ces fonctions).

Mentionnons le paquetage `R2HTML` qui permet de créer des rapports au format html de façon semi-automotique.

2.6 Organisation du travail

L'expérience prouve que la meilleure stratégie est de créer un répertoire (dossier) par analyse de données, et d'y disposer : (a) les fichiers de données brutes ; (2) le fichier script contenant les commandes R ; (3) le workspace et le(s) fichiers(s) résultats (textes et graphiques).

3 Manipulations de base

3.1 Objets

L'objet de base en R est le vecteur. Un vecteur peut contenir des valeurs numériques, des valeurs de vérité (True or False), des chaînes de caractères... Les fonctions les plus utilisées pour créer des vecteurs sont `c`, `rep` et `seq` :

```
c(1,2,3,4,5,6)
c(T,T,F,F)
c('a','b')
rep(55,10)
rep(c(1,2),10)
rep(c('a','b'),c(2,7))
seq(1,10,by=.1)
```

Un type de vecteur particulièrement utile est le type *factor*. Les facteurs sont des vecteurs utilisés pour *classifier* les valeurs d'autres vecteurs (les facteurs sont des "variables indicatrices"). Par exemple, étant donné 100 scores provenant de plusieurs groupes de sujets, une variable facteur peut désigner ces sous-groupes.

```
(a=factor(c(rep('alpha',10),rep('beta',10))))

(b=gl(3,4,48,labels=c('a','b','c')))

(x=rnorm(48))
tapply(x,b,mean)
boxplot(x~b)
stripchart(x~b,method='jitter')
stripchart(x~b,method='jitter',vertical=T)
```

On peut créer un facteur à partir d'un vecteur grâce à la fonction `factor`, ou directement avec la fonction "`gl`".

3.2 Accéder aux éléments d'un vecteur

```
(a=rnorm(50))
a[1]
a[2]
a[c(1,3,5)]
```

```
a>0
a[a>0]

(b=gl(2,25,labels=c('g1','g2')))
a[b=='g1']
```

Une particularité de R est que les éléments d'un vecteur peuvent avoir des noms :

```
v=c(1,2,3,4)
names(v)=c('alpha','beta','gamma','delta')
v['beta']
```

Cela s'avère très utile pour créer des dictionnaires. Par exemple, un vecteur '`freq`' donnant la fréquence d'usage des mots peut avoir les mots comme 'names' ; il suffit alors de taper "`freq['aller']`" pour obtenir la fréquence du mot 'aller'.

```
mots=c('aller','vaquer')
freq=c(45,3)
freq
freq[mots=='aller']
names(freq)=mots
freq
freq['aller']
```

3.3 Arrays, listes et data.frames

D'autres objets de R sont les les listes, les arrays (vecteurs multidimensionnels) et les data.frames.

Les data.frames sont des listes de vecteurs qui ont tous la même longueur. Les data.frames sont très bien adaptés pour stocker des données présentées sous forme de tableau bi-dimensionnel.

```
(a=array(1:20,dim=c(4,5)))
a[2,4]

(b=list(alpha=1:3,
        beta=c('a','b','c','d')))
names(b)
b$alpha
b$beta

(c=data.frame(a=gl(2,5,10),b=1:10,x=rnorm(10)))
c$a
c$b
c$x
c[1:2,]
```

3.4 Variables

Les objets peuvent être enregistrés dans des variables avec l'opérateur = (ou <-). Pour voir le contenu de l'objet représenté par une variable, il suffit de taper le nom de celle-ci.

```
a<-c(1,2,3)
a
ls()
rm(a)
ls()
```

Les vecteurs contenus dans une liste ou dans un data.frame sont accessibles avec le symbole $. Un data.frame peut être "attaché" pour que ses vecteurs soient directement accessibles.

```
mydata<-data.frame(a=gl(2,5,10),b=1:10,x=rnorm(10))
names(mydata)
mydata$a
mydata$b
mydata$x
attach(mydata)
a
b
x
detach(mydata)
```

3.5 Lire des données

Quand les données sont très peu nombreuses, on peut les entrer directement dans un vecteur (comme on l'a fait jusqu'ici) avec la fonction 'c'.

Les fonctions `scan` et `read.table` permettent de lire des données enregistrées dans des fichiers textes.

`scan` lit une suite de données dans un vecteur.

Avec un éditeur de texte, créez un fichier `datafile1.txt` contenant :

```
3.4 5.6 2.1 6.7 8.9
```

Puis, dans R, entrez :

```
scores<-scan('datafile1.txt')
```

On peut également entrer des données directement en ligne de commande :

```
scores<-scan('')
```

La fonction *read.table* lit des données présentées sous forme tabulaire (par ex. les fichiers .csv enregistrés par Excel) et renvoie un data.frame.

Créez un fichier `datafile2.txt` contenant :

```
sujet groupe score
s1    exp    3
s2    exp    4
s3    exp    6
s4    cont   7
s5    cont   8
```

Puis importez le dans R :

```
a<-read.table('datafile2.txt',header=T)
a
```

R dispose d'un éditeur de data.frame très limité :

```
scores<-edit(data.frame(a))
```

scan et read.table ne lisent que des fichiers textes, Le package 'foreign' permet de lire directement certains fichiers de données binaires provenant de SPSS, SAS, ...

```
library(help='foreign')
```

Mentionnons également l'existence de packages permettant d'accéder à des informations stockées dans des bases de données (MySQL, Oracle...).

4 Statistiques élémentaires

Cette section a pour but d'illustrer quelques concepts fondamentaux de la statistique inférentielle, et de présenter les principales fonctions de R pour le traitement statistique des données recueillies lors d'un protocole expérimental.

4.1 Manipulation des distributions de probabilités

4.1.1 Distributions univariées

Différentes fonctions permettent de générer des nombres aléatoires suivant une certaine distribution de probabilité :

```
runif(10) # distribution uniforme
rnorm(10) # distribution normale
rnorm(10,mean=100)
rbinom(10,size=1,prob=.5) # distribution binomiale
```

La fonction `rnorm` génère des nombres aléatoires distribués selon une loi normale. En augmentant le nombre d'échantillons générés (de 10 à 10000), on constate que la distribution des valeurs obtenues se rapproche de plus en plus d'une distribution normale continue :

```
s1=rnorm(10,mean=2)
summary(s1)
s2=rnorm(100,mean=2)
summary(s2)
s3=rnorm(10000,mean=2)
summary(s3)

par(mfrow=c(3,3))       # organisation des graphiques selon une matrice 3 x 3

hist(s1)                # histogrammes
hist(s2)
hist(s3)

# graphes en évitant le chevauchement des points de même coordonnées
stripchart(s1,method='jitter',vert=T,pch=16)
stripchart(s2,method='jitter',vert=T,pch=16)
stripchart(s3,method='jitter',vert=T,pch='.')


plot(density(s1))       # fonction de densité
```

```
x=seq(-5,5,by=.01)      # vecteur de coordonnées normées pour les abscisses
lines(x,dnorm(x,mean=2),col=2)
plot(density(s2))
lines(x,dnorm(x,mean=2),col=2)
plot(density(s3))
lines(x,dnorm(x,mean=2),col=2)
```

En première approximation, la distribution théorique de la taille des individus de sexe masculin, français, et dans la tranche d'âge 20-35 ans,suit une loi normale de moyenne 170 et d'écart-type 10.

On peut donc non seulement situer un individu, ou un groupe d'individus, dans cette distribution, mais également évaluer la probabilité qu'un individu choisi au hasard parmi la population entière mesure moins de 185 cm, ou plus de 198 cm, ou ait une taille comprise entre 174 et 186 cm.

Lorsque l'on ne dispose pas des tables de lois normales $N(\mu; \sigma^2)$ (il y en a une infinité puisqu'il y a 2 paramètres libres), on utilise la loi normale centrée-réduite $N(0; 1^2)$ (encore appelée loi Z), dont la table est disponible la plupart des manuels ou bien sur le web. Cependant R fournit directement les tables des lois normales, par l'intermédiaire de la commande `pnorm`, qui prend en arguments la valeur repère, la moyenne et l'écart-type théoriques.

```
taille=seq(130,210,by=1)
plot(taille,dnorm(taille,mean=170,sd=10),type='b',col="red")
pnorm(185,mean=170,sd=10)
abline(v=185,col=4)
text(185,.012,paste("P(X<185)=",signif(p,3)),col=4,pos=2,cex=.6)
p=pnorm(198,mean=170,sd=10)
abline(v=198,col=4)
text(198,.002,paste("P(X>198)=",round(1-p,3)),col=4,pos=4,cex=.6)
```

La probabilité qu'un individu choisi au hasard parmi la population entière mesure moins de 185 cm ($P(X < 185)$) est de 0.933 (obtenu par `pnorm(185,mean=170,sd=10)`). La probabilité qu'un individu mesure plus de 198 cm est de 0.003 ($1 - P(X < 198)$, et la probabilité que sa taille soit comprise entre 174 et 186 est 0.290 ($P(X < 186) - P(X < 174)$).

On constate que la probabilité qu'un individu choisi aléatoirement dans une population de moyenne 170 ± 10 mesure plus de 198 cm est très faible. C'est sur la base de ce calcul de probabilités que repose le test de typicalité, ou "test Z" : un groupe d'individus (i.e. un échantillon) sera déclaré atypique ou non représentatif de la population parente dont il est issu, lorsqu'il a une position au moins aussi extrême qu'une certaine position de référence, correspondant en général à la probabilité 0.05.

R permet également de générer d'autres distributions de probabilités, notamment la loi binomiale, les lois statistiques telles que le t de Student, le F de Fisher-Snedecor, le chi-deux (χ^2), etc. On peut ainsi voir dans l'exemple qui suit que la distribution du t de Student tend vers la loi normale lorsque la taille de l'échantillon est suffisamment grande (dans cet exemple, on a manipulé le degré de liberté `df`, donné en argument de la fonction `dt`).

```
?pnorm
?pt
?pbinom
help.search('distribution')
pnorm(2)
pt(3,df=10)             # fonction de répartition de la loi du t de Student
qnorm(.99)              # donne la valeur associée au 99ème centile d'une distribution normale
t<--50:50/10
plot(dnorm(t),type='l',col='red')
par(new=T)              # le prochain graphe sera superposé au précédent
plot(dt(t,df=5),type='l')
```

Imaginons que vous disposiez d'une pièce dont vous vous demandez si elle est baisée. Vous prévoyez de la lancer 10 fois à pile ou face. A partir de quelle proportion relative d'essais face/pile (ou l'inverse) considérerez- vous que la pièces est truquée ?

Si la pièce n'est pas truquée, le nombre de "pile" suit une loi binomiale.

```
plot(dbinom(0:10,rep(10,11),prob=1/2),type='h')
hist(rbinom(100,10,.5))
hist(rbinom(1000,10,.5))
hist(rbinom(10000,10,.5))
```

Supposez que vous tiriez à pile ou face 10 fois de suite, et que la pièce retombe 8 fois sur 'pile'. Quelle la probabilité d'observer cela si la pièce n'est pas biaisée ?

```
binom.test(8,10)
prop.test(8,10,1/2) # test approché
```

4.1.2 Distributions conjointes

Si l'on reprend l'exemple précédent des tailles de la population française masculine (20-25 ans), on a une distribution similaire (i.e. suivant une loi normale de moyenne 70 et d'écart-type 7) pour les poids. On peut bien évidemment se poser les mêmes questions que précédemment, mais on peut également s'intéresser à la relation entre ces deux variables quantitatives. En représentant le poids en fonction de la taille, on peut évaluer la liaison linéaire entre ces deux variables à l'aide du coefficient de corrélation de Bravais-Pearson.

Pour illustrer cela, nous allons utiliser les données issues d'une population d'enfants de sexe masculin âgés de 11 à 16 ans.

```
taille<-scan('')        # saisie manuelle des données
1: 172 155 160 142 157 142 148 180 167 165
11:
```

```
Read 10 items           # indicateur de fin d'entrée-sortie généré par R
poids<-scan('')
1: 50.5 38.1 57.3 39.3 46.1 37.1 45.9 66.3 60 50.5
11:
Read 10 items
plot(poids~taille)
r<-lm(poids~taille)     # modèle linéaire (x,y)
summary(r)                      # diagnostic de la régression
abline(r)                       # tracé de la droite de régression
-55.1963626 + 175 * 0.6568411   # "prédiction" pour taille=175 cm
predict(r,list(taille=c(175))
```

Ensuite, à partir de la connaissance de cette liaison linéaire, on peut se demander quelle serait le poids théorique (non observé) d'un individu dont on ne connaît que la taille : c'est le domaine de la régression linéaire. L'affichage des paramètres de la droite de régression donne la relation poids = 0.657 x taille - 55.196. Ainsi, on peut *prédire* que le poids d'un enfant mesurant 175 cm sera de 59.8 kg.

4.2 Résumés numériques et représentations graphiques

4.2.1 Résumés numériques

Le résumé statistique des principaux indicateurs descriptifs de position et de dispersion peut être obtenu à l'aide des fonctions `mean`, `sd`, `median` ; la fonction `summary` donne un résumé plus complet – par exemple, lorsqu'il s'agit d'un vecteur, elle indique la moyenne et la médiane, ainsi que l'étendue et les valeurs des premier et troisième quartiles.

```
a<-rnorm(100)
mean(a)
sd(a) # écart-type corrigé
summary(a)
boxplot(a)
mean(a,trim=.1) # moyenne sans les 10 % d'observations en fin de vecteur
```

4.2.2 Représentations graphiques

Les fonctions graphiques standard en 2D – `boxplot`, `plot`, `hist` – ont été vues dans les sections précédentes. La création de graphiques personnalisés sous R est facilitée par son extrême souplesse quant au paramétrage des graphiques (positionnement, symboles et type de tracés, etc.). L'utilisation de l'aide en ligne est vivement recommandée.

Pour les graphiques en trois dimensions (z étant une matrice de dim 3), on pourra utiliser les fonctions `image` et `contour` :

```
x=1:10
y=1:10
z=outer(x,y,"*")
persp(x,y,z)
image(z)
contour(z)
```

4.3 Définition de fonctions

Il est possible de définir ses propres fonctions sous Ret d'enrichir ainsi le langage.

Par exemple, Rne possède pas de fonction pour calculer l'erreur-type ($\sigma/\sqrt{(N)}$). On peut en définir une de la manière suivante :

```
se <- function (x) { sd(x)/sqrt(length(x)) }
```

L'exemple suivant permet de calculer la moyenne arithmétique après suppression des valeurs atypiques, i.e. supérieures à 2 écart-types de la moyenne :

```
clmean <- function (x) {
  m<-mean(x)
  d<-sqrt(var(x))
  threshold<-2
  mean(x[(x-m)/d<threshold])
}

a<-c(rnorm(100),5)
mean(a)
clmean(a)
```

On peut lire le code des fonctions existantes :

```
clmean
ls
t.test
methods(t.test)
getAnywhere(t.test.default)
```

5 Tests statistiques

Ce chapitre a pour but de présenter de manière non exhaustive certains tests statistiques employés fréquemment en statistique inférentielle.

Comme on l'a vu précédemment (voir section 4.1), la détermination des seuils de significativité (p) se fait grâce aux fonctions associées à chaque distribution (voir section 4.1).

```
1-pnorm(167,mean=150,sd=10)
1-pbinom(8,10,0.5)
```

5.1 Test du khi-deux

Soit le tableau de contingence A x B suivant à analyser :

	A1	A2	A3
B1	13	24	20
B2	10	7	18

Le calcul du test du χ^2 associé à ce tableau s'effectue de la manière suivante :

```
a<-scan('')
1: 13 24 20
4: 10 7 18
7:
Read 6 items
chisq.test(matrix(a,2,3,byrow=T))
```

5.2 Estimation de la moyenne d'un groupe

L'intervalle de confiance de la moyenne peut être obtenu à l'aide de la fonction `t.test` :

```
a<-10+rnorm(10,sd=10)
t.test(a,conf.level=.01)
```

Si l'hypothèse de normalité n'est pas soutenable, le test de Wilcoxon (non-paramétrique) peut être utilisé à l'aide de la fonction `wilcox.test` : ce test des signes permet de déterminer si la médiane du groupe peut être considérée comme significativement différente de 0.

5.3 Comparaison de deux groupes

Ce sont les mêmes fonctions – `t.test` (test paramétrique) et `wilcox.test` (test non paramétrique) – qui permettent la comparaison entre deux groupes ; dans ce cas, on passe en arguments les deux groupes :

```
a<-rnorm(10)
b<-rnorm(10,mean=1)
t.test(a,b)
wilcox.test(a,b)

c<-c(a,b)
x<-gl(2,10,20)
t.test(c~x)
wilcox.test(c~x)
```

5.4 Analyse de variance sur un facteur

Lorsque l'on est en présence d'un ensemble de k observations indépendantes (un seul facteur inter-sujets), on peut comparer leurs moyennes respectives à l'aide de la fonction `aov` (ou selon un modèle linéaire général, avec la fonction `lm`).

```
x<-rnorm(100)
a<-gl(4,25,100)
plot(x~a)
r<-aov(x~a)
anova(r)
pairwise.t.test(x,a)
t.test(x[a==1],x[a==2])
```

5.5 Anova sur deux facteurs

Avec deux facteurs inter-sujets, le principe d'analyse est le même, mais on étudie également l'interaction entre les deux facteurs.

```
x<-rnorm(100)
a<-gl(2,50,100)
b<-gl(2,25,100)
plot(x~factor(a:b))
interaction.plot(a,b,x)
l<-aov(x~a*b)
anova(l)
```

5.6 Anova sur des protocoles de mesures répétées

Avec un seul facteur intra-sujet, on procèdera ainsi :

```
subject<-gl(10,3,30)
cond<-gl(3,1,30)
x<-rnorm(30)
interaction.plot(cond,subject,x)
summary(aov(x~cond+Error(subject/cond))
```

Avec deux facteurs intra, la démarche est à peu près identique :

```
subject<-gl(10,4,40)
cond1<-gl(2,1,40)
cond2<-gl(2,2,40)
table(cond1,cond2)
x<-rnorm(40)
plot(x~factor(cond1:cond2))
interaction.plot(cond1,cond2,x)
interaction.plot(cond1,subject,x)
interaction.plot(cond2,subject,x)
summary(aov(x~cond1*cond2+Error(subject/(cond1*cond2))))
```

5.7 Régression linéaire

Comme nous l'avons vu dans le cas des distributions conjointes (cf. section 4.1.2), la démarche pour effectuer de la régression linéaire est la suivante :

```
a<-rnorm(100)
b<-2*a+rnorm(100)
plot(b~a)
r<-lm(b~a)
anova(r)
abline(r)
```

```
a<-rnorm(100)
b<-2*a+rnorm(100)
c<-5*a+rnorm(100)
pairs(cbind(a,b,c))
summary(lm(c~a*b))
```

6 Exemples d'analyses de données

Ces exemples proviennent principalement du site web "Analyse Statistique des Données en Psychologie (ASDP)" de l'UFR de Psychologie de l'université Paris 5 (piaget.psycho.univ-paris5.fr/, lien "Analyse des Données" puis "Données").

6.1 Dossier sommeil

Lors d'une expérimentation médicale, on a relevé le temps de sommeil T de 10 patients (facteur Sujet) sous l'effet de deux médicaments (d'où le facteur Médicament M). Chaque sujet a pris successivement l'un et l'autre des deux médicaments.

Source Student (1908) The probable error of a mean, *Biometrika*, VI, 1-25.

Données Fichier sommeil.txt

Question Ces données ont été recueillies pour tester l'hypothèse que le médicament m2 est plus efficace que le médicament m1. Est-ce le cas ?

Une solution Voir l'exemple de script listé en A.1 et B.1 (Statistica).

6.2 Dossier pédagogie

Lors d'une expérimentation pédagogique, on désire comparer l'efficacité de quatre méthodes d'enseignements.

On dispose des notes obtenues à un examen par quatre groupes d'élèves ayant chacun reçu un des 4 types d'enseignements.

Source : Données fictives.

Données Fichier pedago.txt

Questions Comparer les résultats obtenus en fonction des méthodes.

Une solution Voir l'exemple de script listé en A.2 et B.2 (Statistica).

6.3 Dossier négligence

Une recherche a porté sur la "pseudo-négligence" qu'on observe chez des sujets normaux. Ce nom provient des similarités qu'elle présente avec l'hémi-négligence (atteinte de la moitié du champ visuel) de sujets atteints d'une lésion cérébrale. La tâche des sujets consiste à déterminer le milieu subjectif d'une baguette de 24cm avec la seule aide d'informations kinesthésiques. La pseudo-négligence se traduit par une déviation systématique vers la droite (pour les droitiers) de ce milieu subjectif par rapport au milieu objectif de la baguette.

Les données portent sur 24 femmes droitières (facteur S) réparties selon 2 conditions (12 sujets pour chacune) : active (c1) où le sujet peut librement déplacer son doigt posé sur un curseur mobile le long de la baguette ; ou passive (c2) où le sujet commande un moteur déclenchant le mouvement de la baguette dans un sens ou dans l'autre, alors que son doigt ne bouge pas (facteur C). Chaque sujet exécute cette tâche dans 6 situations expérimentales obtenues par le croisement de : la main utilisée, gauche (m1) ou droite (m2) ; et l'orientation du regard, 30° à gauche (o1), 0° (o2) ou 30° à droite (o3) (facteurs M et O). Pour chaque sujet et chaque situation on mesure la déviation en cm entre le milieu subjectif et le milieu objectif de la baguette. Une déviation à droite est notée par une valeur positive, à gauche par une valeur négative.

On s'intéresse ici à l'effet de la condition (C) lorsque le sujet utilise sa main habituelle (m2) (Rappel : tous les sujets sont droitiers) et lorsqu'il se trouve en face du milieu de la baguette (avec l'orientation à 0 degrés)

Source Chokron, Imbert (1993) - Egocentric reference and asymmetric perception of space, Neuropsychologia, 31, 3, 267-275. D'après J.M. Bernard (1994) - Structure des données, d

Données fichier neglige2.txt

Questions Importer ces données, les visualiser, comparer les groupes. Conclusion ?

Une solution Voir l'exemple de script listé en A.3 et B.3 (Statistica).

6.4 Dossier family

Étude réalisée au USA sur les origines des stéréotypes liés au sexe. 35 familles choisies au hasard et ayant une fille ainée (ou fille unique) en "ninth grade" (Troisième).

Le père a répondu à un questionnaire sur ses intérêts pour le sport, noté sur une échelle numérique de 0 à 50 (FATH)

La mère a répondu au même questionnaire (MOTH)

Le professeur d'éducation physique de chacune des filles a noté les performances physiques générales de la fille de 0 à 20 (PROF).

La fille a répondu également au questionnaire d'inérêt pour le sport (GIRL).

Source Hays, W.L. (1994) - *Statistics*, Fort Worth : Harcourt Brace College Publishers (5ème édition), p.671-672

Données family.txt

Questions Que faire avec ces données ?

Une solution Voir l'exemple de script listé en A.4 et B.4 (Statistica).

6.5 Dossier IO

En 1980, on a interrogé des lycéens (garçons et filles) sur leurs intentions d'orientation après le bac (études scientifiques, littéraires ou techniques).

Source Il s'agit de données en partie fictives, inspirées d'un exemple de M. Reuchlin.

Données Fichier io.txt

Questions Peut-on dire que l'orientation envisagée est liée au sexe chez l'ensemble des lycéens de cette année 1980 ?

Une solution Voir l'exemple de script listé en A.5 et B.5 (Statistica).

A Solutions sous R

Nous proposons ici des scripts pour analyser les exemples du chapitre 6. Il y a plusieurs manières de résoudre le même problème avec R. Par conséquent, vos scripts peuvent différer.

A.1 Dossier sommeil

```
sommeil<-read.table('sommeil1.txt',header=T)
sommeil
attach(sommeil)
summary(M1)
summary(M2)

plot(M1,M2,xlim=c(0,10),ylim=c(0,10),col=2)
identify(M1,M2,SOMMEIL)
abline(0,1)

stripchart(M2-M1,method='stack')
t.test(M2-M1)

t.test(M1,M2,paired=T)
detach()
```

A.2 Dossier pedago

```
a<-read.table('pedago.txt')
attach(a)
boxplot(notes~pedago)
stripchart(notes~pedago,method='stack',vertical=T)
tapply(notes,pedago,mean)
tapply(notes,pedago,sd)
tapply(notes,pedago,summary)
barplot(t(tapply(notes,pedago,mean)))

m<-aov(notes~pedago)
summary(m)
TukeyHSD(m)
plot(TukeyHSD(m))
```

A.3 Dossier négligence

```
d<-read.table('neglige4.txt')
x<-d$V1
a<-gl(2,12,24)
b<-gl(2,6,24)
table(a,b)
tapply(x,list(a=a,b=b),mean)
interaction.plot(a,b,x)
l<-aov(x~a*b)
summary(l)
model.tables(l,se=T)
t.test(x[a==1 & b==1],x[a==1 & b==2])
t.test(x[a==2 & b==1],x[a==2 & b==2])
```

A.4 Dossier family

```
fam<-read.table('family.txt',header=T)
fam
attach(fam)
data<-as.matrix(fam[,-1])
pairs(data,panel=panel.smooth)
cor(data)
cor.test(FATH,GIRL)
cor.test(MOTH,GIRL)
cor.test(INST,GIRL)

l<-lm(GIRL ~ FATH + MOTH + INST)
summary(l)
detach(fam)
```

A.5 Dossier IO

```
a<-read.table('io.txt',header=T)
attach(a)
table(Sexe,Matière)
chisq.test(table(Sexe,Matière))
```

B Solutions sous Statistica

Cette section présente la façon de traiter les problèmes présentés dans le chapitre 6.

De manière générale, lorsque l'on dispose de simples fichiers texte pour les données, l'importation des données se fait à l'aide de la commande **Fichier ▷ Importer des données ▷ Rapide**. Le cas échéant, lorsque le fichier de données est déjà sous le format Statistica (extension .sta), il suffit simplement d'utiliser la commande **Fichier ▷ Ouvrir des données**.

B.1 Dossier Sommeil

Une fois les données chargées, une boîte de dialogue `Tables et Statistiques Elémentaires` s'affiche. On choisit `Test T` pour des échantillons appariés—. Une nouvelle boîte de dialogue s'affiche dans laquelle on va définir la première et la seconde variable à l'aide du bouton `VARIABLES`.

Il est intéressant de visualiser les données sous forme de "boîtes à moustaches" (cf Fig. B.2) ; pour cela, il suffit de cliquer sur le bouton `BOITE A MOUSTACHES`, et de sélectionner ensuite l'option `Médiane/Quartile/Etendue` dans la boite de dialogue suivante.

Ensuite, on peut lancer le test t sur le panneau initial (s'il n'est plus visible, cliquer sur la petite boîte de dialogue `Reprendre analyse` ou `SUITE` si vous êtes sur la dernière fenêtre graphique), en appuyant sur le bouton `TESTS`.

Le résultat de l'analyse s'affiche dans une nouvelle fenêtre de sortie (cf Fig. B.1). A la lecture des résultats, on voit que le test t est significatif : t = -4,06213, p/2 = ,0014165 (p = ,002833, dl = 9).

On pourra remarquer que l'analyse aurait abouti au même résultat en dérivant le protocole par différence, et en effectuant un test t contre une moyenne théorique $\mu = 0$ (p = ,0016). Sous Statistica, aller dans `Autres tests de significativité`, `Différence entre deux moyennes`, cocher `Moyenne du cas 1 vs. Moyenne de la population 2`, M1=-1,56, s1=1,24, n=10, ap-

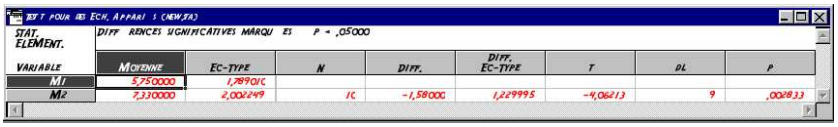

FIG. B.1 – Résultat du test t pour échantillons appariés sous Statistica

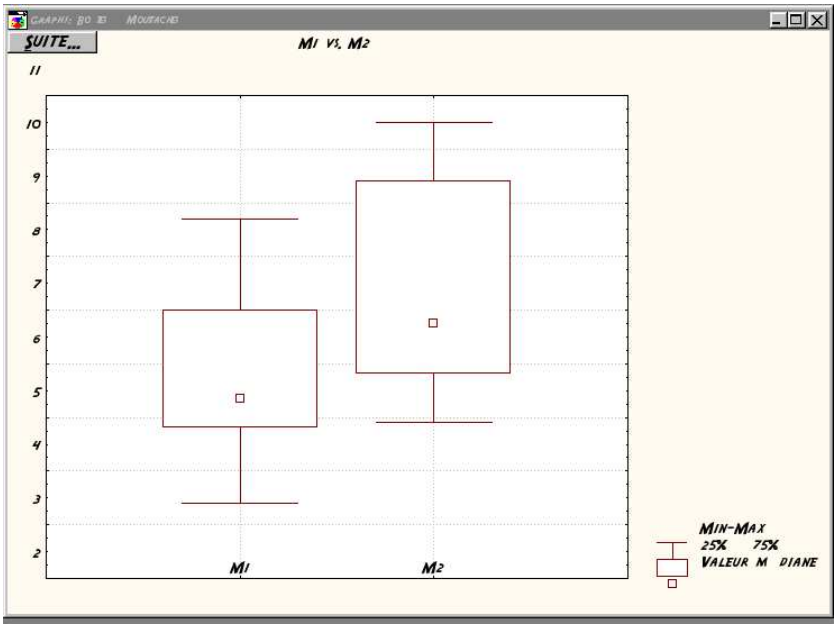

FIG. B.2 – Affichage des différences de groupe sous forme de boîte à moustaches sous Statistica

puyer sur `Calculer` et lire le seuil p correspondant[1].

B.2 Dossier Pédago

Il s'agit ici de mettre en oeuvre une ANOVA d'ordre 1 (un seul critère de classification, ou un seul facteur de groupe à 4 modalités). Après avoir saisi les données, ou importé le fichier, l'analyse de variance s'effectue soit via le module `ANOVA/Décomposition de la variance`, soit via le module plus général pour les analyses de variance (groupes indépendants et mesures répétées) `ANOVA/MANOVA`.

En supposant que le tableau de données aît été correctement saisi (2 colonnes comprenant la VD et la VI sous forme indicée – 1, 2, 3, 4 – par exemple ; les observations en ligne), il suffit de sélectionner les variables de l'analyse en cliquant sur le bouton `Variables` et d'indiquer la colonne contenant la variable indépendante et celle contenant la variable dépendante. Après avoir validé, on revient sur l'écran précédent, et on indique la liste des facteurs inter (la VI est un facteur de groupe) que l'on veut prendre en compte dans l'analyse en cliquant sur le bouton `Liste fact. inter` et en indiquant `Tous`[2].

[1] Une autre solution consiste à effectuer le calcul à la main : $t_{(n-1)} = (\bar{x} - \mu)/(s/\sqrt{(n)}) = -3.97$, et à comparer la valeur aux valeurs repères de la table du t : p ¡ .003 (p/2 ¡ .0015)

[2] On pourrait vouloir restreindre l'analyse à deux conditions seulement, auquel cas on indiquerait les conditions individuelles

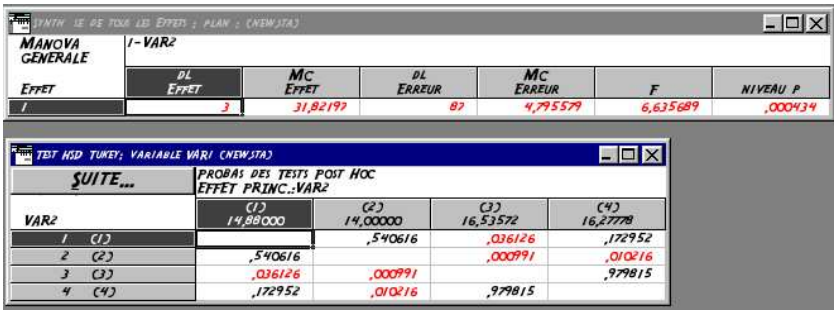

FIG. B.3 – Résultat de l'analyse de variance d'ordre 1 sous Statistica

Lorsque l'on valide en appuyant sur la touche `OK`, le panneau d'analyse de variance s'affiche, le plan d'analyse considéré (facteurs systématique inter, intra etc.) étant indiqué dans la partie supérieure. Lorsqu'on clique sur le bouton `Tous les effets`, Statistica lance l'analyse de variance d'ordre 1, et une fenêtre de résultats s'affiche. Cette dernière comprend un tableau d'ANOVA d'ordre 1 classique, avec le carré moyen de l'effet et celui de l'erreur (MC effet et MC error), les degrés de libertés associés aux sommes des carrés (3 et 86), la valeur du F (6,635689), ainsi que le seuil p associé (0,000434) (cf. Fig. B.3). Par défaut[3], Statistica affiche en rouge les valeurs significatives par rapport aux seuils repères (que l'on peut redéfinir dans les options Statistica).

Il est également possible d'avoir une représentation graphique des moyennes des groupes de sujet en cliquant sur le bouton `Comparaison moy.`, puis en sélectionnant la sortie `graphique`, mais par défaut ce n'est pas une boîte à moustache qui est affichée (cf. Fig. B.4).

B.3 Dossier Négligence

Le protocole doit être analysé à l'aide d'une ANOVA d'ordre 2 (deux groupes indépendants de sujets), avec comme facteurs principaux (systématiques) de l'analyse les conditions "Active" et "Main", à deux modalités chacune.

Nous utiliserons comme dans le dossier précédent le module `ANOVA/MANOVA`, en supposant les données déjà disponibles au bon format et présentes dans le tableau de données (3 colonnes comprenant la VD et les 2 VI sous forme indicée – 1, 2 – par exemple ; les observations en ligne). On définira comme précédemment les variables dépendantes et indépendantes, ainsi que les facteurs inter à prendre en compte dans l'analyse (i.e. tous).

En répétant les mêmes étapes que celles effectuées dans le dossier Pédago, on obtient le tableau d'ANOVA d'ordre 2 avec l'effet des deux facteurs systématiques et l'interaction entre ces deux facteurs (cf. Fig. B.5). L'interaction entre les deux facteurs, et son seuil de significativité peuvent

[3] ce n'est le cas lors des comparaisons multiples

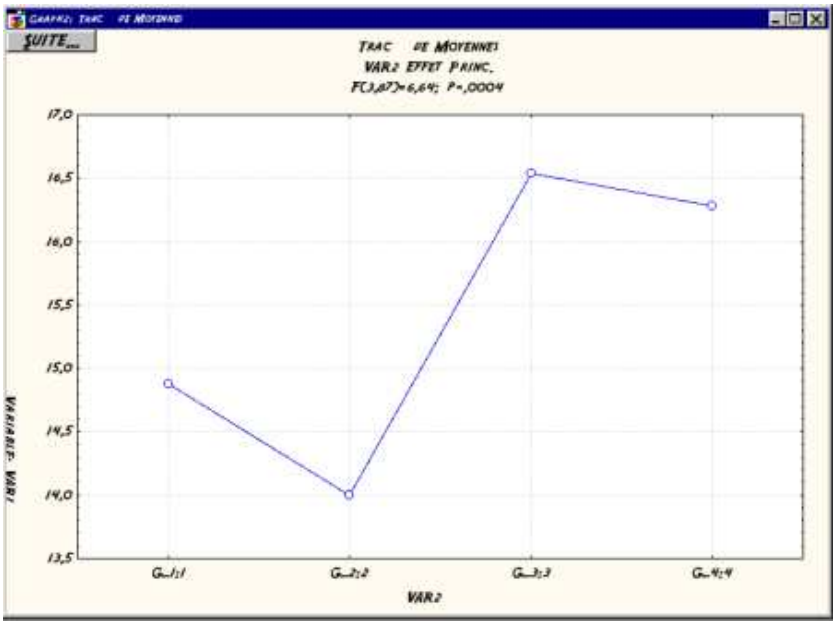

FIG. B.4 – Comparaison de moyennes sous Statistica

être visualisé en cliquant sur le bouton `Comparaison moy.`, puis en sélectionnant la sortie `graphique`. Etant donné qu'il y a deux variables, il faudra indiquer quelle variable sera reprise sur l'axe des abscisses (cf. Fig. B.6).

En revanche, puisqu'on est dans un cas d'ANOVA à plusieurs facteurs, il faudra analyser les moyennes qui sont significativement différentes prises deux à deux : on utilisera pour cela les comparaisons multiples (non planifiées) qui sont accessibles en cliquant sur le bouton `Tests post-hoc`. Le test de Tukey-HSD peut-être utilisé, et on sélectionnera l'option `Différences significatives`. Le résultat du test s'affiche dans une nouvelle fenêtre, sous forme d'un tableau indiquant en ligne les comparaison par paire de modalités des deux variables[4].

On notera que que sous Statistica, le résultat du test de Tukey-HSD indique les seuils p pour les différences significatives et non les intervalles de confiance à 95 % comme sous R.

B.4 Dossier family

Il s'agit ici d'un problème classique de régression multiple. La matrice des corrélations peut être obtenue dans le menu **Analyse ▷ Statistiques élémentaires rapides ▷ Matrice de corrélations**.

[4] le tableau étant une matrice symétrique de seuils de significativité p, on peut se contenter de lire la moitié des valeurs...

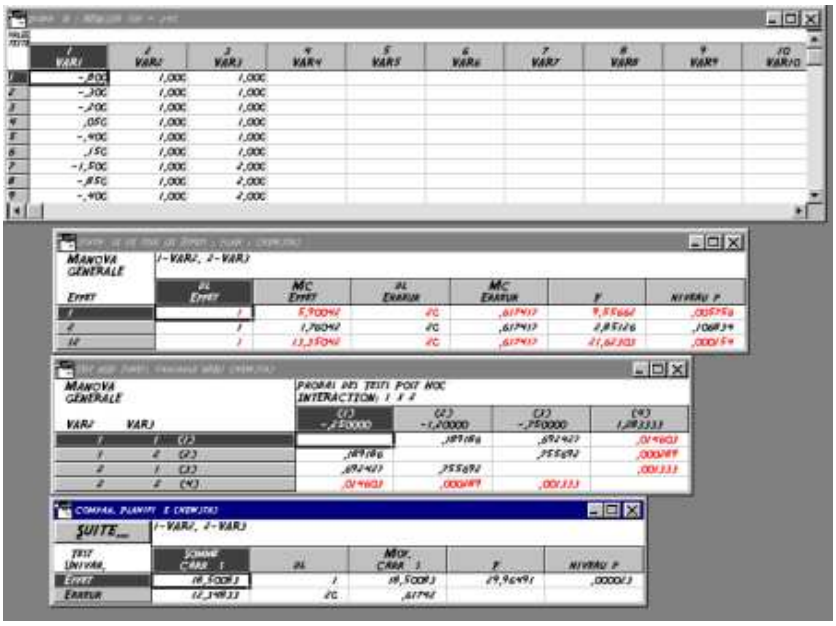

FIG. B.5 – Résultat de l'analyse de variance d'ordre 2 sous Statistica

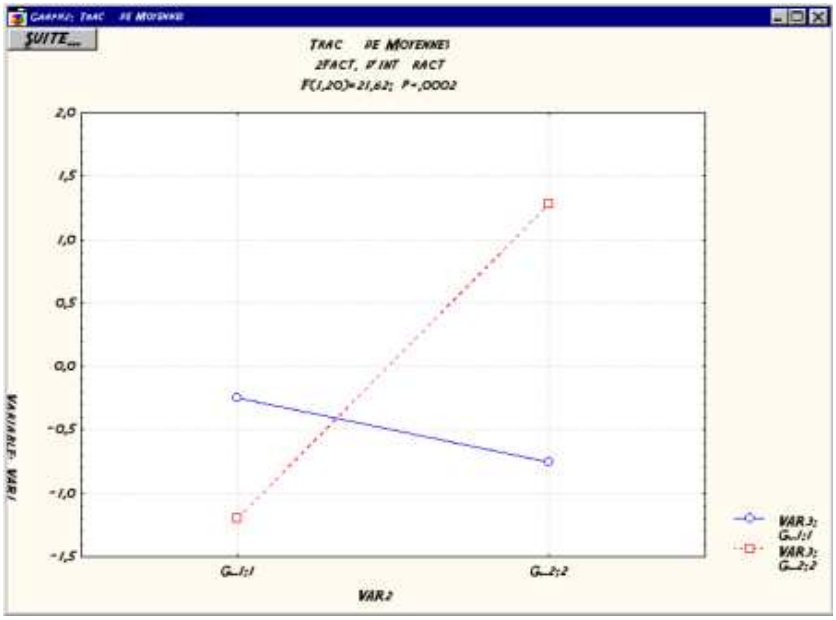

FIG. B.6 – Graphique de l'interaction des deux facteurs sous Statistica

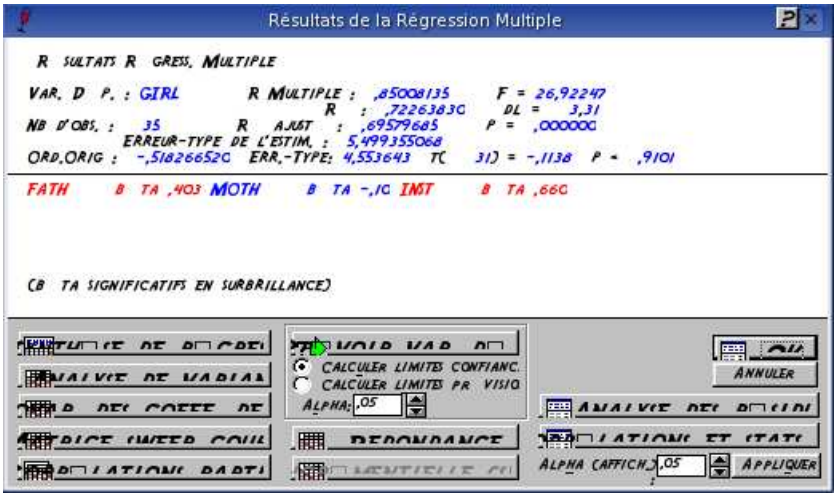

FIG. B.7 – Résultat de la régression multiple sous Statistica

Après avoir cliquer dans le commutateur de modules de Statistica `Régression multiple` (ou menu **Analyse** ▷ **Autres statistiques** ▷ **Régression multiple**), il faut spécifier la variable dépendante (ici les données de la fille), et les variables indépendantes, i.e. les variables prédictrices (les 3 autres séries de données – père, mère, prof). Une fois le codage des variables effectué, valider en appuyant sur `OK`. Un tableau indiquant les résultats de la régression multiple – coefficients β, R^2 – s'affiche (cf. Fig. B.7).

L'analyse des valeurs prédites et des résidus est obtenue grâce à la commande `Analyse des résidus` dans la fenêtre `Analyse de la régression multiple`, puis en sélectionnant `Afficher Résidus & Prév.`.

B.5 Dossier IO

Ce dossier comprend un tableau de contingence 3 x 2 pour lequel il faut mettre en oeuvre une procédure d'analyse des tableaux de contingence (module `Analyse des Correspondances`). Lorsque la boîte de dialogue s'affiche, on sélectionne l'option `Analyse des Correspondances (AC)`, ainsi que `Fréquences sans var. de classement`, puis on sélectionne toutes les modalités de la variable en colonnes.

Lorsque l'on clique sur `OK`, un panneau intitulé `Résultats de l'Analyse des Correspondances` apparaît et indique dans la partie supérieure le résultat du test du chi-deux : ici, $\chi^2 = 6.66667$, $dl = 2$, $p = 0.357$. Les indicateurs descriptifs concernant le tableau de contingence sont accessibles via le panneau de contrôle dans la partie inférieure :
– les distributions marginales (lignes et colonnes) sont accessibles en cliquant sur le bouton

`Pourcentages Lignes` ou `Pourcentages Colonnes`

– les effectifs théoriques sont accessibles en cliquant sur le bouton `Théorique (Chi)`
– les écarts à l'indépendance (différence entre effectifs observés et effectifs théoriques) sont accessibles en cliquant sur le bouton `Obs. moins Théorique`
– le carré moyen de contingence (ϕ^2) est indiqué dans toutes les fenêtres de résultats précédentes et désigné par le terme `Inertie Totale`

Les contributions au χ^2 sont indiquées pour chaque croisement des modalités des deux variables en cliquant sur le bouton `Contrib. au Chi-deux`.

C Prise en main de Statistica

C.1 Introduction

Etant donné la richesse de l'interface, ou plutôt des interfaces (cf. *infra*) de Statistica, nous nous contenterons d'évoquer quelques-unes de ses principales fonctionnalités, afin que le lecteur soit à même : (1) d'ouvrir ou d'importer un fichier de données, (2) d'effectuer des statistisques descriptives élémentaires, (3) de créer des représentations graphiques et (4) d'analyser des protocoles simples (cf. également la section B). De plus amples informations peuvent être obtenues grâce aux manuels de Statistica, à l'aide en ligne, ou aux nombreux tutoriels disponibles sur le web.

Statistica est un logiciel très puissant permettant de faire de l'analyse descriptive et inférentielle. Statistica est organisé en différents modules – `Statistiques Elémentaires`, `ANOVA/MANOVA`, etc. –, accessibles au travers du commutateur de modules (cf. Fig. C.1), qui est automatiquement lancé au démarrage. Il demeure ensuite accessible dans le menu **Analyse ▷ Autres Statistiques**.

Chaque module correspond en fait à un environnement d'analyse particulier, et l'interface de Statistica (boutons, menus) est spécifique de chaque module, et des fenêtres actives (feuille de données, graphique). Lorsque l'on bascule d'un module à l'autre, par exemple de celui des `Statistiques Elémentaires` à celui dédié à l'analyse de variance `ANOVA/MANOVA`, il est préférable de fermer le module précédent : utiliser pour cela le bouton `Fermer & Basculer vers` ; cela évitera d'avoir plusieurs fenêtres Statistica ouverte en même temps.

FIG. C.1 – Commutateur de modules de Statistica

C.2 Organisation des données

Statistica contient un gestionnaire de données intégré, mais nous allons nous limiter à exposer brièvement l'organisation des données dans la feuille de données. Celle-ci est équivalente à un tableur (comme Excel) dans lequel les observations sont disposées en lignes, et les variables en colonnes. En fait, les observations sont le plus généralement les sujets, et les variables les modalités des variables indépendantes (V.I.). L'intersection ligne-colonne contient ainsi la valeur de la variable dépendante (V.D.).

Lorsqu'il n'y a qu'une seule V.I. à plusieurs modalités, on peut coder ses modalités dans une autre colonne-variable, qui sert alors de variable de classement. Dans le cas où on a plusieurs V.I. à plusieurs modalités (cas par exemple d'un protocole de mesures répétées), les variables correspondent en fait au croisement de chaque modalité de chaque variable. Par exemple, si l'on a 2 V.I. A et B à 2 niveaux (i.e. un plan de type $S * A_2 * B_2$), il y aura 4 colonnes disposées *précisemment*[1] comme suit : a1b1 a1b2 a2b1 a2b2. Il est utile de s'assurer de la bonne disposition des données en affichant un graphique, car si l'ordre des facteurs est inversés par exemple, il risque d'y avoir des problèmes lors de l'interprétation de l'interaction A x B.

L'aide en ligne est généralement bien rédigée et indique dans chaque situation (plan avec groupes indépendants, groupes appariés, mesures répétées, plan factoriel, "split-plot" etc.) comment organiser les données. N'hésitez pas à vous y référer, même pour vérification.

C.3 Statistiques descriptives

C.3.1 Résumé numérique

Les statistiques descriptives constitue un module à part entière – `Statistiques Elémentaires` – , accessible depuis le commutateur de modules (le fichier permettant le lancement direct de ce module est `Sta_bas.exe`). Le choix des résultats à afficher se fait dans le menu **Analyse** ▷ **Panneau de démarrage**, puis **Statistiques Descriptives**, ou dans le menu **Analyse** ▷ **Statistiques élémentaires rapides** ▷ **Autres** (cf. Fig. C.2). Par défaut, le résumé numérique indique le nombre d'observations, la moyenne, le minimum, le maximum et l'écart-type, pour la ou les variable(s) sélectionnée(s) (bouton **Variables**).

Il est également possible de paramétrer le type de résultats à afficher (moyenne, médiane, quantiles, écart-type etc.), à l'aide du bouton **Davantage de Statistiques**. En revanche, il faut garder à l'esprit que `Statistica` ne calcule que des écart-types et variances corrigés[2].

[1] Il faut faire attention à l'ordre, en effet, car Statistica va déterminer les modalités des facteurs à partir de l'ordre dans lequel ils sont rangés dans la feuille de données ; ainsi, si on rangeait les données sous la forme a1b1 a2b1 a1b2 a2b2, le plan correspondant serait $S * B_2 * A_2$, et pire encore si on intervertissait 2 colonnes, on aurait un plan incorrect puisque mélangeant les facteurs !

[2] Pour obtenir des écart-types et variances non corrigés, il faut multiplier les valeurs obtenues par $\frac{n-1}{n}$.

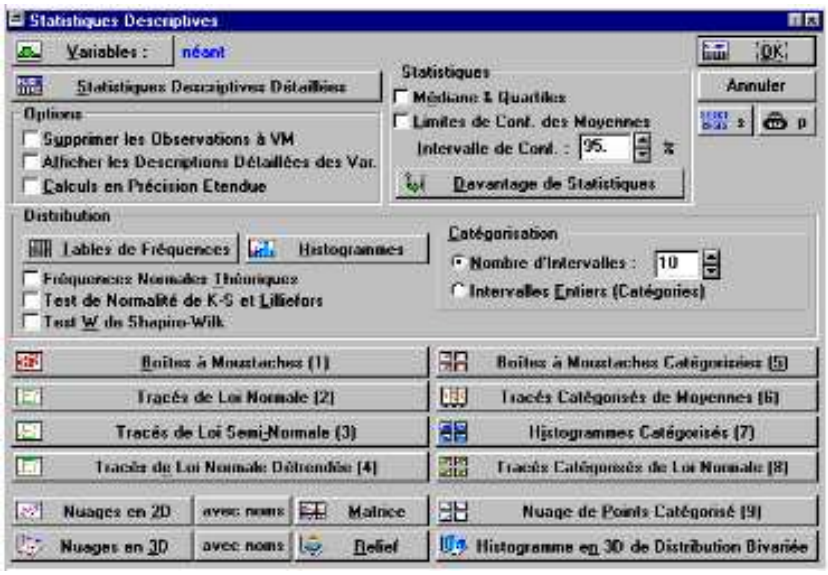

FIG. C.2 – Panneau de Statistiques Descriptives

C.3.2 Remarque

Par ailleurs, toutes les fonctions décrites dans les paragraphes précédents sont généralement accessibles à l'aide d'un clic droit effectué dans la colonne de la ou les variable(s) du tableau de données (**Statistiques Rapides** ▷ **Autres**).

C.4 Représentations graphiques

Les outils graphiques se trouvent dans le même module que celui utilisé pour les statistiques descriptives ; ils apparaissent dans la partie inférieure de la boîte de dialogue **Statistiques Descriptives**. Les graphiques les plus couramment utilisés sont les histogrammes, les boîtes à moustaches catégorisées et les nuages en 2D (nuages bivariés). Lorsque l'on sélectionne un type de graphique, une nouvelle boîte de dialogue apparaît et permet de préciser ou paramétrer le graphique sélectionné. Si aucune variable n'a été sélectionnée, il est possible de le faire à l'aide du bouton **Variables**.

C.4.1 Histogrammes

Pour les histogrammes (cas des variables catégorisées), on peut choisir entre des histogrammes simple (cas d'une seule variable) ou multiple (cas de plusieurs variables), et y associer une courbe

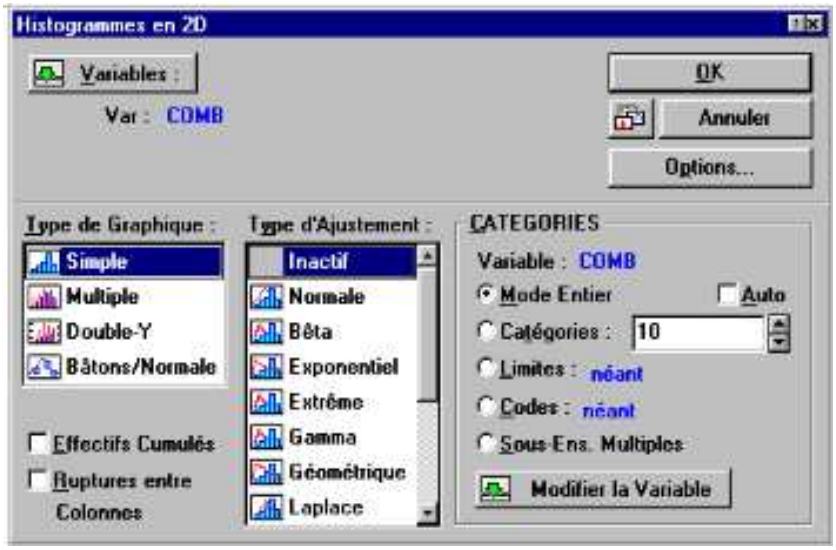

FIG. C.3 – Panneau de Statistiques Descriptives

d'ajustement normale par exemple (cf. Fig. C.3). Il est également possible d'afficher les effectifs cumulés, plutôt que les effectifs simples (cela peut permettre d'obtenir la fonction de répartition discrète de la variable).

C.4.2 Boîtes à moustaches

Les boîtes à moustaches sont très utiles pour les données catégorisées. Elles peuvent être présentées sous forme classique (médiane, 1er et 3ème quartiles, étendue), ou à l'aide d'autres indicateurs (moyennes, erreur-type ou écart-type et intervalles de confiance à 95 %) (cf. Fig. C.4).

C.4.3 Nuages de points en 2D

Pour les nuages en 2D (cas des variables numériques), on peut également choisir des nuages simple ou multiple, associés à une fonction d'ajustement pré-définie ou paramétrable. Des boutons radio permettent de sélectionner le type d'axes du repère, qui peuvent être de type cartésien ou polaire. On peut également inclure les ellipses et bandes de confiance (en général au seuil $p = .95$) de la moyenne.

D'autres options très utiles sont disponibles ne cliquant sur le bouton **Options**. La nouvelle boîte de dialogue qui s'affiche permet en effet de sélectionner le nombre d'observations à inclure dans le graphique, de spécifier l'affichage des étiquettes d'observations (ce qui permet de repérer directement une observation sur le graphique par une étiquette de type i1, i2, ..., sans avoir à le

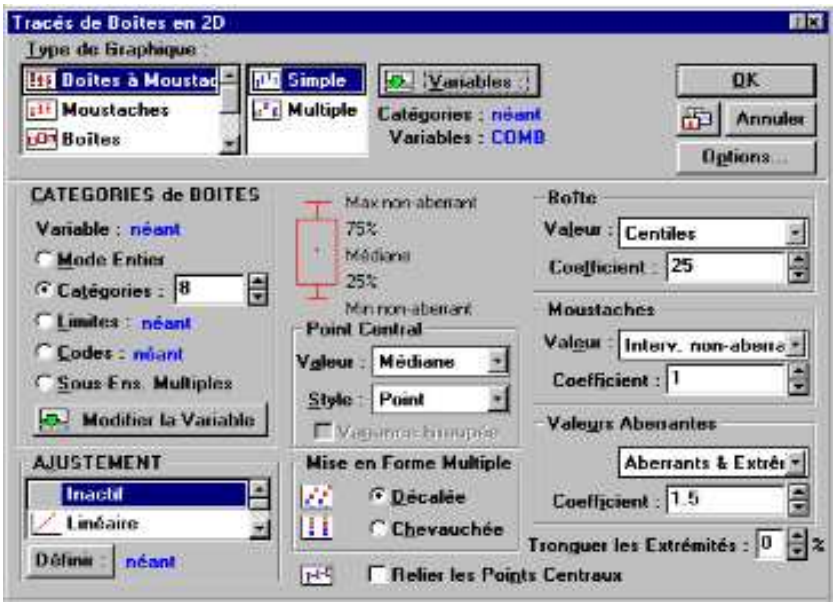

FIG. C.4 – Panneau de Statistiques Descriptives

faire à l'aide de ses coordonnées).

C.4.4 Remarque

Par ailleurs, toutes les fonctions décrites dans les paragraphes précédents sont généralement accessibles à l'aide d'un clic droit effectué dans la colonne de la ou les variable(s) du tableau de données (**Graphiques Rapides ▷ (...)**).