Méthodes statistiques appliquées aux questions internationales
Cours de Mayeul Kauffmann
IEP de Grenoble, Master OIG-ONG - 2008-2009

La notation de l'examen devrait à peu près suivre le barème suivant :

Thèmes abordés en cours magistral	8 à 11 points
Interprétation de graphiques et de résultats statistiques	6 à 9 points
Compétence en écriture et débogage du langage de script R	environ 4 points

Pour le cours magistral, la référence écrite est constituée par les chapitres 1 à 3 du manuel (Kauffmann, Mayeul, *Méthodes statistiques appliquées aux questions internationales*, L'Harmattan, 2009), en excluant les encadrés intitulés « Pour aller plus loin » et les exercices.
On y trouvera aussi des exemples d'interprétation de graphiques et de résultats statistiques.
En ce qui concerne le programme pour la compétence en écriture du langage de script R, il s'agit d'une partie des exercices faits en cours (ce qui est reproduit ci-après) et d'extraits du manuel de Julien Barnier.
Extraits du manuel de Julien Barnier, version du 6 mars 2009 en ligne ici:
http://otan.ecoledelapaix.org/spip.php?article105
Pour le groupe intermédiaire : pages 1 à 37 du manuel de Julien Barnier. Pour le groupe avancé : pages 1 à 53 (ne pas faire les sections « Valeurs manquantes dans les conditions » page 50 ni « tapply » page 53).

Sauf erreur, les symboles et fonctions vus dans ces pages et dans les exercices faits en cours sont les suivants :

Symboles:	c	lm	read.delim2
- + * / ^ ( )	class *	ls	read.table
; , . " :	coplot	max	rnorm
[ ] $	cor	mean	sd
#	data	median	segments
<- ->	dim *	min	seq *
== !=	dotchart	mosaicplot *	setwd
< > <= >=	download.file	names	sort
& (et) | (ou)	dput *	ncol	sqrt
~ | (condition)	for *	nrow	str
T TRUE F FALSE NA	fix	par(new=T)	subset *
NULL *	getwd	paste	sum
Fonctions:	help	pie	summary
abline	hist	plot	sunflowerplot
as.factor *	histogram	predict *	table
attach	ifelse *	print *	tabulate
axis	image *	q	unique
barplot	length	range *	var
boxplot	library	read.csv	which
bwplot	log	read.csv2	xyplot

* les fonctions marquées d'un astérisque sont surtout destinées aux étudiants du groupe avancé.

Une courte description de la plupart de ces fonctions est disponible en annexe du manuel (*Méthodes statistiques appliquées...*) et reproduite ici:
http://otan.ecoledelapaix.org/spip.php?article107
Si ce n'est déjà fait, vous êtes invités à lire au moins une fois l'aide de chacune de ces fonctions. Rappel: l'aide de *toutes* les fonctions est accessible par le moteur de recherche de l'aide html ; pour l'aide des fonctions contenues dans des paquets additionnels, ceux-ci doivent être chargés avant que cette aide ne soit accessible

avec la fonction help(). Par exemple, faites «library(lattice)» avant de faire «help(histogram)».
Pour la compétence en écriture du langage de script R, l'aide nécessaire sur les fonctions (mais pas sur les symboles) sera fournie lors de l'examen (elle sera généralement tirée de l'aide de R ou d'un des documents cités ci-dessus) . Le sujet ne devrait pas nécessiter d'utiliser une fonction autre qu'une fonction citée ci-dessus ; dans le cas contraire (exceptionnel), la fonction concernée sera introduite avec une explication claire en français et un exemple simple.

```
################
## EXERCICE 1 ##
################

## Question 1. Télécharger le fichier "Intra-State Wars (V 3-0).csv" du jeu de
   données de Sarkees sur le site http://www.correlatesofwar.org/, les lire dans
   R. Définir le nom des colonnes, en utilisant les noms trouvés dans la
   documentation des données.

# Créons un objet pour contenir le chemin du fichier
# A l'Ecole de la paix:
# fichier <- "/home/commun/ecole/REC/UNESCO/Grenoble-
   stats/livre_UPMF/exercices/Intra-StateWars_V 3-0.csv"

# Par exemple pour Windows:
fichier <- "G:/Intra-StateWars_V 3-0.csv"
# Par exemple pour MacOSX et GNU/Linux:
#fichier<-"~/Intra-StateWars_V 3-0.csv"

# Le code internet utilisé pour les espaces est "%20" (ceci est aussi le cas
   dans les navigateurs internet):
download.file(url="http://www.correlatesofwar.org/cow2%20data/WarData/IntraStat
   e/Intra-State%20Wars%20(V%203-0).csv",
 destfile=fichier)
GuerCiv<-read.csv( file=fichier) # lire les données
# équivaut à la ligne suivante:
# GuerCiv<-read.csv( file="G:/Intra-StateWars_V 3-0.csv") # lire les données

GuerCiv[1,] # On regarde la première ligne
names(GuerCiv) # il y a probleme
GuerCiv<-read.csv( file=fichier, header=FALSE) # lire les données
names(GuerCiv) # c'est mieux
GuerCiv[1:5,] # les lignes 1 à 5

# Pour définir le nom des colonnes, utiliser les noms trouvés sur la page:
# http://www.correlatesofwar.org/cow2%20data/WarData/IntraState/Intra-State War
   Format (V 3-0).htm

# un vecteur:
c("WarNo", "WarName", "WarType", "YrBeg1", "MonBeg1", "DayBeg1", "YrEnd1",
   "MonEnd1", "DayEnd1", "YrBeg2", "MonBeg2", "DayBeg2", "YrEnd2", "MonEnd2",
   "DayEnd2", "Insurgnt", "Winner", "Interven", "MinDur", "MaxDur", "StDeaths",
   "ToDeaths", "InCenSub", "InMajor", "WestHem", "Europe", "Africa", "MidEast",
   "Asia", "Oceania", "Version")

names(GuerCiv) <- c("WarNo", "WarName", "WarType", "YrBeg1", "MonBeg1",
   "DayBeg1", "YrEnd1", "MonEnd1", "DayEnd1", "YrBeg2", "MonBeg2", "DayBeg2",
   "YrEnd2", "MonEnd2", "DayEnd2", "Insurgnt", "Winner", "Interven", "MinDur",
   "MaxDur", "StDeaths", "ToDeaths", "InCenSub", "InMajor", "WestHem", "Europe",
   "Africa", "MidEast", "Asia", "Oceania", "Version")
names(GuerCiv)
GuerCiv[1,] # ce n'est pas mal
```

```
## Question 2. Afficher un aperçu des données.

# Trois aperçus utiles
summary(GuerCiv) # résumé
# on voit qu'il y a un problème avec les NA, il faut réimporter les données:


GuerCiv<-read.csv( file=fichier, header=FALSE,na.strings=c("-999","-888")) #
   lire les données
names(GuerCiv)<-c("WarNo", "WarName", "WarType", "YrBeg1", "MonBeg1",
   "DayBeg1", "YrEnd1", "MonEnd1", "DayEnd1", "YrBeg2", "MonBeg2", "DayBeg2",
   "YrEnd2", "MonEnd2", "DayEnd2", "Insurgnt", "Winner", "Interven", "MinDur",
   "MaxDur", "StDeaths", "ToDeaths", "InCenSub", "InMajor", "WestHem", "Europe",
   "Africa", "MidEast", "Asia", "Oceania", "Version")
summary(GuerCiv) # C'est bon: les NA sont correctement interprétés
str(GuerCiv) # structure
GuerCiv[1:5,] # les 5 premières lignes.


## Question 3. Nombre de guerres civiles dans le jeu de données complet
nrow(GuerCiv)

## Question 4. Permettre l'accès aux données sans avoir à mentionner le nom du
   tableau.
#attach(GuerCiv)# Permettre l'accès aux données sans avoir à mentionner le nom
   du tableau

## Question 5. Quel est le nombre de débuts de guerres civiles pour chaque
   année? Faites en un diagramme en bâtons.

table(GuerCiv$YrBeg1)

table(YrBeg1)
plot(table(YrBeg1)) # Diagramme en bâtons (l'objet transmis à plot() est une
   table, donc par défaut plot() suppose le paramètre graphique: type = "h"

## Question 6. Faites une courbe du nombre de débuts de guerres civiles par an

# Attention, la fonction table() ne garde pas les années pour lesquelles il n'y
   a pas de guerre. La commande suivante relie donc par des traits horizontaux
   les années non-consécutives séparées par des années en guerre, ce qui est
   faux!
plot(table(YrBeg1), type="l") # Mauvais!
# On peut superposer les deux graphes pour s'en rendre mieux compte:
plot(table(YrBeg1))
par(new=T);plot(table(YrBeg1), type="l", col="red") # Mauvais!

# Réponse pour le groupe intermédiaire:
plot(tabulate(YrBeg1), type="l", xlim=c(1820,1997), xlab="Année", ylab="Nombre
   de débuts de guerres civiles")


# Réponse pour le groupe avancé:

# La commande suivante fait le bon calcul, mais donne le résultat pour chaque
   année de l'an 1 à 1997.
tabulate(YrBeg1)
# Il faut donc extraire une partie puis refaire l'axe des abscisses (car
   l'origine de l'axe est 1 et non pas 1820)...
plot(tabulate(YrBeg1)[1820:1997], type="l",xaxt="n")
```

```
axis(side=1, at=seq(1820,1990, by=10)-1819, lab=seq(1820,1990, by=10))

# Plus simple: n'afficher qu'une partie du résultat, avec le paramètre xlim ():
plot(tabulate(YrBeg1), type="l", xlim=c(1820,1997), xlab="Année", ylab="Nombre
   de débuts de guerres civiles")

##Question 7 (groupe avancé seulement: essayez de comprendre la réponse).
   Calculez puis superposez au graphique précédant une moyenne mobile du nombre
   de débuts de guerres sur 3 ans.

NbDebGuer<-tabulate(YrBeg1)[1820:1997]
# Pour chaque année, on fera la moyenne de trois valeurs: l'année en cours,
   l'année d'avant et l'année d'après. Ceci est impossible pour les extrémités
   (on n'a pas de moyenne mobile pour 1820 ni pour 1997).
An<-NbDebGuer[-(c(1,178))] # La liste des valeurs pour les années restantes (on
   enlève la première et la dernière année)
AnMoinsUn<-NbDebGuer[-(177:178)] # la liste des valeurs pour "l'année en cours
   moins un": on enlève donc les deux dernières valeurs
AnPlusUn<-NbDebGuer[-(1:2)] # la liste des valeurs pour "l'année en cours plus
   un": on enlève donc les 2 premières valeurs
NbDebGuerMoyMob3<-(AnMoinsUn + An + AnPlusUn)/3 # on fait la somme des 3
   vecteurs: la valeur de chaque année se trouve additionnée à celle de l'année
   précédente (AnMoinsUn) et celle de l'année suivante (AnPlusUn)
lines(1821:1996, NbDebGuerMoyMob3, col="blue")

## Les questions 8 à 12 (du manuel) ne sont pas à réviser pour l'examen

## Question 13. Faites deux diagrammes en boite à moustache comparant la
   distribution des valeurs de MinDur et MaxDur, d'une part, et StDeaths et
   ToDeaths, d'autre part.
boxplot(MinDur,MaxDur,StDeaths,ToDeaths)
boxplot(StDeaths,ToDeaths)

## Question 14. Comparez avec les histogrammes de chacune de ces 4 variables.
hist(StDeaths)
hist(ToDeaths)
hist(MinDur)
hist(MaxDur)


## Question 15. Faites un diagramme en barre et un diagramme circulaire de
   chacune des deux variables Europe et Winner
barplot(table(Europe))
barplot(table(Winner))
pie(table(Europe))
pie(table(Winner))

## Question 16. Calculez le tableau croisé (table de contingence) des variables
   Winner et Europe (effectifs des combinaisons de valeurs). Faites en deux
   graphiques de types différents.
table(Europe, Winner)
mosaicplot(table(Europe, Winner))
image(table(Europe, Winner))
sunflowerplot(Europe, Winner) #nombre de pétales des "fleurs" = nombre de
   points superposés


## Question 17. Faites un diagramme de la variable Winner en abscisses et
   StDeaths en ordonnées, pour chacune des valeurs de Europe, en une seule
   commande graphique.
```

```
library(lattice) # chargement du package lattice
xyplot(StDeaths~Winner|Europe)


## Question 18.  Faites un diagramme de la variable MinDur en abscisses et
   StDeaths en ordonnées
plot(MinDur,StDeaths  )
# ou:
library(lattice) #facultatif si on l'a déjà fait. Ne sera pas répété pour la
   suite.
xyplot(StDeaths ~ MinDur)


## Question 19. Même chose, mais faire deux graphiques, selon la valeur de
   Africa, en une seule commande graphique.

xyplot(StDeaths ~ MinDur | Africa) # Africa=0: à gauche, Africa=1: à droite,
   cf. le petit trait plus foncé dans la barre coloré "Africa"

## Question 20. Même chose en logarithme pour les variables StDeaths et MinDur

xyplot(log(StDeaths)~log(MinDur)|Africa) # tendance bien plus lisible

## Question 21. Même chose en faisant un graphique pour chaque combinaison des
   variables Africa et Winner
xyplot(log(StDeaths)~log(MinDur)|Winner*Africa) # Africa=1 en bas, Africa=0 en
   haut; Winner augmente de gauche à droite


## Question 22. En groupant la variable YrBeg1 en intervalles de temps se
   recoupant, faire 6 histogrammes de MinDur en abscisses et StDeaths en
   ordonnées, en une seule commande graphique.
coplot(StDeaths ~ MinDur | YrBeg1) # de gauche à droite et de bas en haut,
   chacun des 6 graphiques du bas correspond à un intervalle de YrBeg1 (indiqué
   de gauche à droite par une bande gris dans la fenêtre du haut).

## Question 23. Faites l'équivalent de hist(MinDur) pour chaque valeur de
   Winner
histogram(~MinDur|Winner)

## Question 25. Faites des boîtes à moustache montrant la distribution de
   MinDur selon les combinaisons de valeurs de Europe et Winner

# comparez ces 3 solutions:
bwplot(Europe~MinDur|Winner)
bwplot(Winner~MinDur|Europe) # proche
bwplot(~MinDur|Europe*Winner)# presque identique

## Question 26. Comparez ces graphiques à des histogrammes
histogram(~MinDur|Europe*Winner) # par exemple


################
## EXERCICE 2 ##
################

# Question 1. Chargez les données de Anscombe, les afficher.
data(anscombe) # ce jeu de données est fourni avec le logiciel R
anscombe
```

```
# Question 2. Reproduire la Figure 1 (page 48 du manuel) représentant le nuage
   de points ajusté par un modèle linéaire
attach(anscombe)
model1 <- lm(y1 ~ x1)
model1
plot(y1 ~ x1)
abline(3.001, 0.5001) # ajoute une droite
# mieux:
abline(model1)
segments(x1, y1 , x1, model1$fitted.values) # résidus représentés par des
   segments
# NB: la description des sous-objets du modèle (tel le sous-objet
   "fitted.values") est faite dans la section "Value" de l'aide de la fonction
   "lm" (faites "help(lm)")

# Question 3. Affichez les résultats pour le modèle 1
summary(model1)

# Question 4. Refaites la même chose avec les autres jeux de données de
   Anscombe
# Réponse: il suffit de remplacer x1 par x2 et y1 par y2 par exemple. Le début
   de la réponse à la question 2 devient alors:
model2 <- lm(y2 ~ x2)

# élément de réponse pour le groupe avancé:
for (i in 1:4) {
  print(paste("i =", i))
  print(names(anscombe[,c(i,i+4)]))
  print(summary(lm(anscombe[,i+4]~anscombe[,i])))
 }

########################################
## EXERCICE 3 (groupe avancé seulement) ##
########################################

## Question 1. Téléchargez et lisez dans R le jeu de données du Correlates of
   War mesurant la puissance des Etats.

# Par exemple pour Windows:
dossier <- "C:/Mes Documents/"
# Par exemple pour MacOSX et GNU/Linux:
fichier<-"~/"

fichier<-paste(dossier, "NMC_3.02.csv", sep="")
download.file("http://www.correlatesofwar.org/COW2%20Data/Capabilities/NMC_3.02
   .csv", destfile=fichier) # un espace dans une adresse s'écrit %20 dans un
   navigateur
# NMC<-read.csv(fichier)  # Mauvais, les NA ne sont pas traités
# l'auteur des données mentionne: "Missing values are indicated by the value “-
   9”. Users must ensure that their statistical analysis software takes this
   coding into account."

NMC<-read.table(fichier, header=T, sep=",",na.strings = "-9")

## Question 2. Faire un tableau ne contenant que les données pour 2001
NMC2001<-NMC[NMC$year == 2001,]

## Question 3. Faites un modèle linéaire essayant d'expliquer les dépenses
   militaires en 2001 par le niveau de développement économique de chaque pays
   (mesuré par un indicateur de dépenses énergétiques). Interprétez lez
```

```
   résultats.
summary(lm(milex ~ energy ,data=NMC2001))

## Question 4. Idem en rajoutant la population totale. Comparez les deux
   modèles.
summary(lm(milex ~  energy + tpop, data=NMC2001)) # notez que les coefficients
   de energy changent, car il y a une corrélation entre energy et tpop, bien que
   faible en apparence, cf.:
cor(NMC2001$milex, NMC2001$tpop, use="complete")

## Question 5. Même chose pour l'ensemble des années, en ajoutant une variable
   indicatrice (0 ou 1) pour chaque année (mesurant l'effet spécifique à chaque
   année et non pris en compte par tpop ni energy). Vous ferez un graphique des
   paramètres estimés pour ces variables indicatrice
lm2 <- (lm(milex ~ tpop + energy + as.factor(year) ,data=NMC))
summary(lm2)
plot(1817:2001, lm2$coefficients[-(1:3)], type="b")

## Question 6. Les dépenses militaires sont mesurées en livre sterling jusqu'en
   1913, en dollars à partir de 1914. Vous corrigerez la réponse à la question
   précédente en multipliant les livres sterling par 4,5

lm3 <- lm( ifelse(year>=1914, milex, milex*4.5)   ~ tpop + energy  +
   as.factor(year), data=NMC)
summary(lm3)
plot(1817:2001, predict(lm3, data.frame(year=1817:2001, tpop=1000,
   energy=1000)), type="b")

## Question 7. Faites un graphique montrant pour 2001 le log de energy et celui
   de milex
plot(log(NMC2001$energy), log(NMC2001$milex+1)) # milex peut être nul, et
   log(0) n'existe pas: on rajoute 1

# Pour aller plus loin, vous pouvez modifier certains des modèles en
   introduisant le log de certaines variables, puis comparer les résultats.
```