Démonstrations/ séance du 2 octobre 2012

I- Tableau, Matrice, Base de données statistiques

Tableau : (lignes, colonnes) <- on peut mettre n'importe quoi (chiffres, des booléens, des signes)
Matrice : (lignes, colonnes) <- tableau de coefficients (peut faire des calculs)
Base de données :
(**Colonnes**, lignes)= (**variables**, observations)= {**(quantitatives, qualitatives)**, observations} = {(**(continues, discrètes)**, **discrètes**), observations}.

II- Variables continues, variables discrètes

Les variables à caractère quantitatif = continues ou discrètes (discontinues)
Les variables à caractère qualitatif = discrètes ou discontinues

Variables (quantitatives) continues :
Les observations peuvent être n'importe quelle valeur dans une fourchette donnée. Par exemple si on dispose des observations sur le poids à la naissance des étudiants qui suivent ce cours, on aura des valeurs comme 1.50 ;2 ; 2.900 ;3.500 ;3 ; 4.300… La représentation de ces observations sur un axe donne une ligne continue.
Variables (quantitatives) discrètes :
Les observations ne peuvent prendre que des valeurs isolées (ou discrètes). Par exemple si on dispose des observations sur l'âge des étudiants qui suivent ce cours, on aura des valeurs comme 19 ; 21 ; 20 ; 22 ; 27…. La représentation de ces observations sur un axe donne une série de points.
Variables qualitatives :
On peut disposer, par exemple, des observations sur le niveau d'études, le fait de travailler ou de ne pas travailler pendant les études, l'état civil, catégorie socioprofessionnelle des parents, des étudiants de l'UQAM.
Considérons les deux premières variables (niveau d'étude et travail) :
Niveau d'étude = (''Baccalauréat'', ''Maîtrise'', ''Doctorat '')
Travail = (''Oui'', ''Non'')= (''travailler pendant les études'', '' Ne pas travailler pendant les études'')

On voit que les variables qualitatives peuvent prendre k modalités distinctes.
Si k=2, on parle d'une variable dichotomique. Le cas ici de la variable **''Travail'',**
Si k>2 (k appartient à N*) on parle de variable polytomique.

La question naturelle qui se pose est de savoir : comment représenter ce genre de variables dans une base de données?
La réponse à cette question est d'associer une variable quantitative (ou codage) à la variable qualitative en question.

III- Codage d'une variable qualitative

Considérons la variable **Y= ''Niveau d'étude''** = (''B ac'', ''Maîtrise'', ''Doctorat'')'.
Il y a plusieurs choix pour coder cette variable. La première consiste à associer à Y une variable X quantitative pouvant prendre trois valeurs entiers (a, b, c) appartenant à N, suivant les modalités de Y.
Le choix de (a, b, c) est a priori non contraint. On peut, par exemple, avoir (1, 2, 3) ou (3, 5, 8) en référence au nombre d'années d'études.

$$X = \begin{cases} 1 & si \quad Y = ''Bac'' \\ 2 & si \quad Y = ''Mts'' \\ 3 & si \quad Y = ''Doc'' \end{cases} ; X = \begin{cases} 3 & si \quad Y = ''Bac'' \\ 5 & si \quad Y = ''Mts'' \\ 8 & si \quad Y = ''Doc'' \end{cases}$$

Un autre codage peut consister à associer à Y trois variables dichotomiques Z1, Z2, Z3 telle que :

$$Z_1 = \begin{cases} 1 & si \quad Y = ''Bac'' \\ 0 & sinon \end{cases}; Z_2 = \begin{cases} 1 & si \quad Y = ''Mts'' \\ 0 & sinon \end{cases}; Z_3 = \begin{cases} 1 & si \quad Y = ''Doc'' \\ 0 & sinon \end{cases}$$

Les Zi sont appelées en économétrie, des variables dummies ou muettes.
Ainsi, de façon générale, les représentations quantitatives de Y s'écrivent sous la forme d'une application injective de {''Y1'', ''Y2'', ''Y3''} -> Rp, p appartenant à N*

La différence entre les variables quantitatives et les variables qualitatives est que les variables qualitatives ne peuvent prendre que des valeurs isolées (ou discrètes).

IV- L'intérêt du codage

Autre que le fait d'avoir une représentation de ces variables dans la base de données, c'est aussi de pouvoir se ramener à des lois discrètes sur $\mathbb{R}^p$. Ainsi, si on considère l'exemple précédent, la loi de Z est une loi multinomiale de paramètre $(1, p_1, p_2, p_3)$ où les p_k représentent les probabilités d'observer respectivement, ''Bac'', ''Maîtrise'' et ''Doctorat'', et 1 = nombre de sondages réalisés sur les mêmes étudiants. Les Z_k suivent chacune une loi de Bernouilli de paramètres $(1, p_k)$, k=1, 2, 3.
Il faut noter que les lois obtenues avec ce genre de représentation dépendent forcément du choix du codage et ne sont pas interprétables. Aussi, la moyenne, la variance etc… pour des telles représentations ne sont pas importantes car elles dépendent du choix de codage. Seules les probabilités sont interprétables.

Ainsi, lorsqu'on veut réaliser une étude sur une variable à caractère qualitative, on utilise d'autres modélisations linéaires probabilistes que vous ne verrez pas en cours.
Le but de ce cours n'est donc pas de modéliser ce genre de variable mais plutôt d'en savoir utiliser dans un modèle.

Manipulation de données sous R.

R est un logiciel statistique libre qui offre un environnement dans lequel beaucoup de techniques statistiques classiques et modernes sont mises en œuvre. Quelques-unes d'entre elles sont construites dans la base de l'environnement R, mais beaucoup sont fournis comme des paquets.
Il y a environ 25 paquets standards et les autres sont disponible sur le site CRAN (via : http: // CRAN.R-project.org).

Téléchargez et installer R.
Le logiciel R, et ses packages sont disponibles librement sur le site :
http://www.r-project.org/

Ouvrir R sous window :
Je recommande de créer un dossier au préalable sur votre pc où sera enregistré tout le travail :
Vous lancez directement R en cliquant sur l'icône R de votre bureau.

I- Importation des données sur R

Nous utiliserons pendant ces démonstrations des fichiers de type ''.csv '' (ou '' .txt ''), récupérables sous Excel, ces fichiers séparent chaque variable par des '' ; '' ou '','.
La fonction **read.table** permet d'importer une base de données externe sur R.

```
>demo1=read.table(''DossierDuFichier/NomDuFichier.csv'',sep='';'', dec='','',header=TRUE)
>demo1=read.table(''DossierDuFichier/NomDuFichier.txt'',sep='' ;'',dec=''.'',header=TRUE).
```

Nous introduisons grâce à cette commande, dans notre environnement, l'objet demo1 qui contient des données issues du NomDuFichier.csv ou NomDuFichier.txt (respectivement).

Remarques : si on ne met pas header=TRUE, la première ligne sera importée comme des observations et non comme des noms des séries.

Nous récupérons un jeu de données disponible sur le blog personnel de Mr. Arthur Charpentier (http://freakonometrics.blog.free.fr/). Ce jeu de données provient de la documentation SAS, il s'agit des statistiques de criminalité dans les 50 états américains (sans Washington DC). Dans chaque état, sept crimes ou délits sont repérés par leurs nombres annuels de fait constatés rapportés sur 100 000 habitants.
Renommez les variables dans la base de données :
Meurtre (**M**) ; Viol (**Viol**) ; vols avec violence (**VaV**) ; Agression (**Agr**) ;Combriolage (**Comb**) ; Escroquerie (**Esc**) ; Vols de voiture (**Vauto**) ; Peine de Mort (**PM**)

Associez à la variable qualitative Peine de Mort une variable quantitative discrète (**PM1**) qui prend 1 si Peine de Mort égale à oui et 0 si Non.

Importation de données
>demo1=read.table("C:/Documents and Settings/Ben/Mes documents/TD_UQAM/crime.csv'',sep='' ;'',dec=''.'',header=TRUE)

Pour afficher les noms des variables présentes dans la base

>names(demo1)

>head(demo1) ****** affiche les premières observations présentes dans l'objet démo1.

Les variables sont directement liées à l'objet demo1 (base de données).

Les variables "M", " Viol", "VaV", "Agr", "Comb", "Esc", "Vauto" et" PM" sont définies par les noms:

demo1$**M**, demo1$**Viol**, demo1$**VaV**, demo1$**Agr**, demo1$**Comb**, demo1$**Esc**, demo1$**Vauto** et demo1$**PM**

>head(demo1$M) *** affiche les premières observations de la variable **M**
>demo1[1,2] **affiche la valeur présente à la 1ère ligne et 2ème colonne de la matrice demo1
>demo1[2,] **affiche la 2ème ligne de la matrice demo1
>demo1[,1] *** correspond à demo1$M

Si nous voulons créer un objet indépendant de la base, à partir d'une variable:
>variable1=demo1$M *****(ou demo1[,1]) ou bien
>assign(''variable1 '',demo1$M)

NB: si on modifie variable1 par la suite, cela ne modifiera pas demo1$M ou demo1[,1]. Il s'agit bien d'un objet indépendant!!!!!!

On peut appliquer des transformations sur les variables:
>log(demo1$M) ***** permet d'afficher les valeurs en logarithme de variable Meurtre (**M**)
>logM=log(demo1$M) ***** permet de les stocker dans un objet indépendant de la base
>demo1$logM=log(demo1$M) ***** permet de le stocker dans la matrice demo1
>demo1$sqrtM=sqrt(demo1$M) ***** permet de stocker dans les valeurs en racine carrée de la variable M dans la matrice demo1, les valeurs de Y en racine carrée.
>demo1$M2=demo1$M*demo1$M **** permet de stocker le carré des valeurs de M dans la matrice demo1

Consultez la documentation R pour connaitre les opérateurs liés aux différentes opérations (+,-,*,/,sqrt,^,log,exp…)-

Les statistiques descriptives simples :

>dimDemo1=dim(demo1) ***** la dimension de la matrice demo1 = 50 7
>nbosM=length(demo1$M) ****** le nombre d'observations de la variable M
>sumM=sum(demo1$M) ******* affiche la somme des observations de M
>mean(demo1$M) ***** affiche la moyenne de la variable M
>meanM=sumM/nbosM ********** affiche aussi la moyenne de M
>median(demo1$M) ****** affiche la médiane de la variable M
>var(demo1$M) ********** affiche la variance de la variable M
>varM=(sum((demo1$M-mean(demo1$M))^2)/(nbosM-1) *** variance de M
>cor(demo1$M,demo1$Viol) ******* affiche le coefficient de corrélation entre M et Viol

Graphiques (simples) sous R.

>boxplot(log(demo1$Comb)) ***** ce graphique (boite à moustache) permet de repérer les observations qui s'écartent de l'ensemble des observations de la variable Combriolage (**Comb**)
>resul=boxplot(demo1$Comb) ****** enregistre le résultat dans l'objet resul
>va=resul$out ******* enregistre les valeurs aberrantes dans l'objet va
>which(demo1$Comb%in%va) ******* affiche les positions où se trouve les observations aberrantes de la variable Comb dans la base de données

Les graphiques du type ¨nuage de points¨ se construisent sur R à l'aide de la fonction plot.
>plot(demo1$Viol,demo1$M) ***** trace le nuage de points avec **Viol** sur l'axe ox et **M** sur l'axe oy.
>plot(demo1$Viol,demo1$M,abline(lm(demo1$M~demo1$Viol),col="red"))) **** trace une droite lineaire
>hist(demo1$M) ******* histogramme de M

Il est possible de nommer les graphs et les axes :
>plot(demo1$Viol,demo1$M,xlab= ''Viol'',ylab=''Meurtre'',main=''Nuage de points entre M et Viol '')

construire des matrices à partir de la base de données :

>X=cbind(demo1$M,demo1$Viol,demo1$VaV,demo1$Esc,demo1$Comb,demo1$Vauto,demo1$PM1) **** matrice dimension (50 7)
>Y=demo1$Agr ******* creation d'un objet Y qui contient les observations de la variable Agression (Agr)
>tX=t(X) ***** transposée de X
>tXX=tX%*%X ***** produit matricielle de X et sa transposée tX – donne une matrice carrée
>tXy=tX%*%Y **** produit matrice- vecteur
>tXY=crossprod(X,Y) *** autre façon, plus efficace
>invtXX=solve(tXX) ***** l'inverse de cette matrice
>MCO=invtXX%*%tXY ***** estimateur des MCO