FORMATION
LOGICIEL STATA

Dakar du 12 au 23 septembre 2011

Mme CAMARA Fatou Bintou Niang, Démographe

Mr Komla Mawulom AGUDZE, Statisticien Economiste

TABLE DES MATIERES

INTRODUCTION

Stata a été créé le 1er janvier 1985 par Stata Corporation (STATCORP) : version 1.0. Il est beaucoup utilisé dans les universités américaines et occupe une place de plus en plus importante en Europe. Stata est un logiciel complet permettant l'analyse statistique et économétrique. Il n'est pas le seul logiciel d'analyse ni d'économétrie existant. Ce logiciel en est aujourd'hui à la version 11. Stata est un logiciel rapide puisqu'il utilise les données en mémoire. L'objectif est de vous faire découvrir Stata et de vous donner les bases pour une utilisation efficace.

PRESENTATION DU COURS

Jour	**Libellé de la formation**
1ier jour	PRESENTATION DE STAT TRANSFER ET STATA
2ième jour	LES OPERATIONS ET FONCTIONS DANS STATA ET COMMANDES DE GESTION DES VARIABLES
3ième et 4ième jours	Suite et fin COMMANDES DE GESTION DES VARIABLES(fin)
5ième et 6ième jours	FUSION DES BASES DE DONNEES et STATISTIQUES DESCRIPTIVES
7ième jours	STATISTIQUES DESCRIPTIVES : tests d'hypothèses
8ième et 9ième jours	PONDERATION ET GRAPHIQUES
10ième jour	LES REGRESSIONS & EVALUATION DE LA FORMATION

PARTIE I : Présentation des logiciels Stat Transfer et Stata

Dans cette partie le logiciel stat Transfer qui permet de convertir les fichiers de bases de données dans un format compatible sous Stata sera présenté dans une première section. Ensuite, dans la seconde section, ce sera au tour du logiciel Stata d'être présenté en particulier sa barre d'outils et ses différentes icônes.

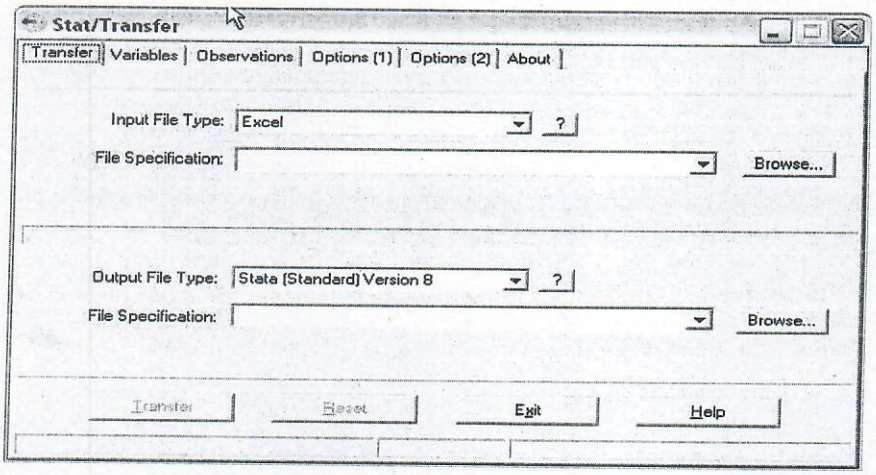
1.1 Comment transférer une base de données au format lisible par Stata

1.1.1 Utilisation de stat Transfer

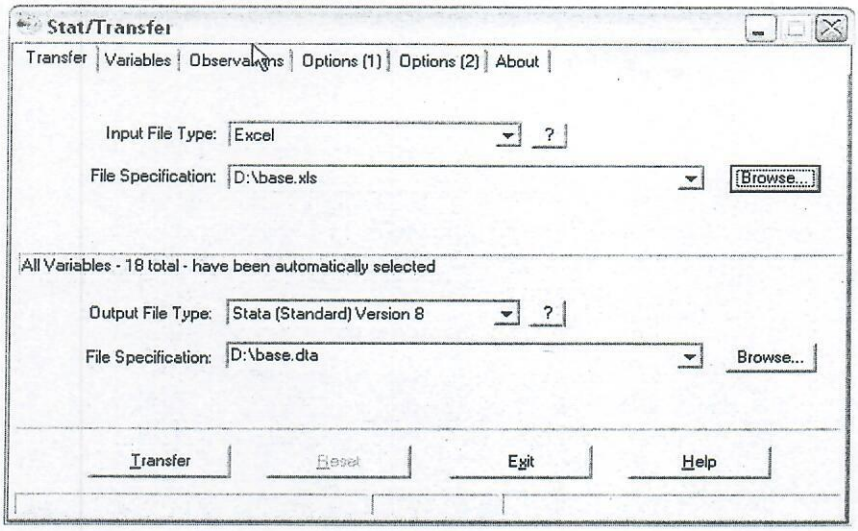
Les bases de données utilisables sous Stata doivent être dans un format spécifique (.dta). Le logiciel stat Transfer permet d'obtenir ce format. Les différentes étapes pour convertir un fichier en format .dta :

PRATIQUE

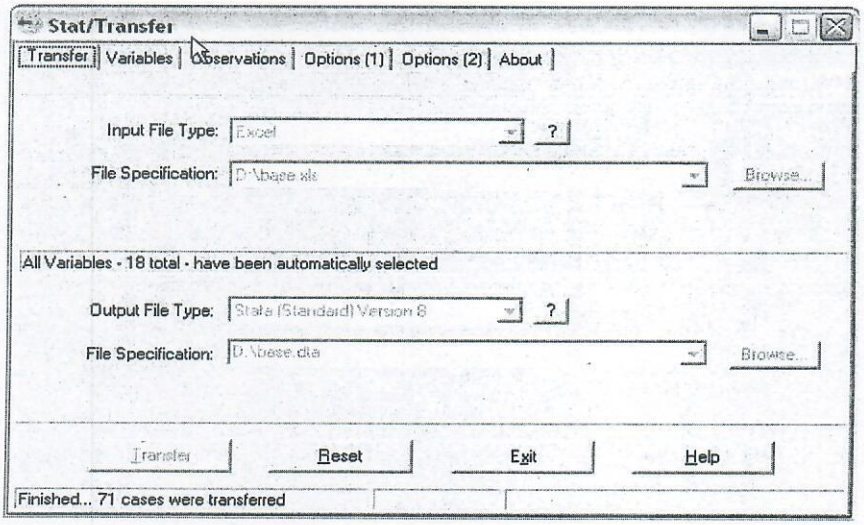
Etape1 : démarrer stat Transfer vous obtiendrez la capture d'écran ci-dessous :



Etape 2 : spécifier le chemin d'accès du fichier *.xls* à convertir en cliquant sur ***Browse*** (dans ce cas D : \Base.xls). Stat Transfer enregistrera par défaut le fichier .dta dans le même répertoire que le fichier source (D : \Base.dta dans la capture d'écran ci-dessous)



Etape 3 : cliquer sur l'icône Transfer pour convertir le fichier. Dans la barre d'état inférieure apparait « finished…71 cases were transfered » le chiffre 71 correspond au nombre d'observations dans la base de données. Cette indication permet de s'assurer que toutes les observations ont été bien prises en compte. Dans la barre d'état supérieure le nombre de variables transférées est également indiqué « *All variables - 18 total – have been automatically selected* ».



1.2. Prise en main de Stata

1.2.1. Principes de bases de Stata

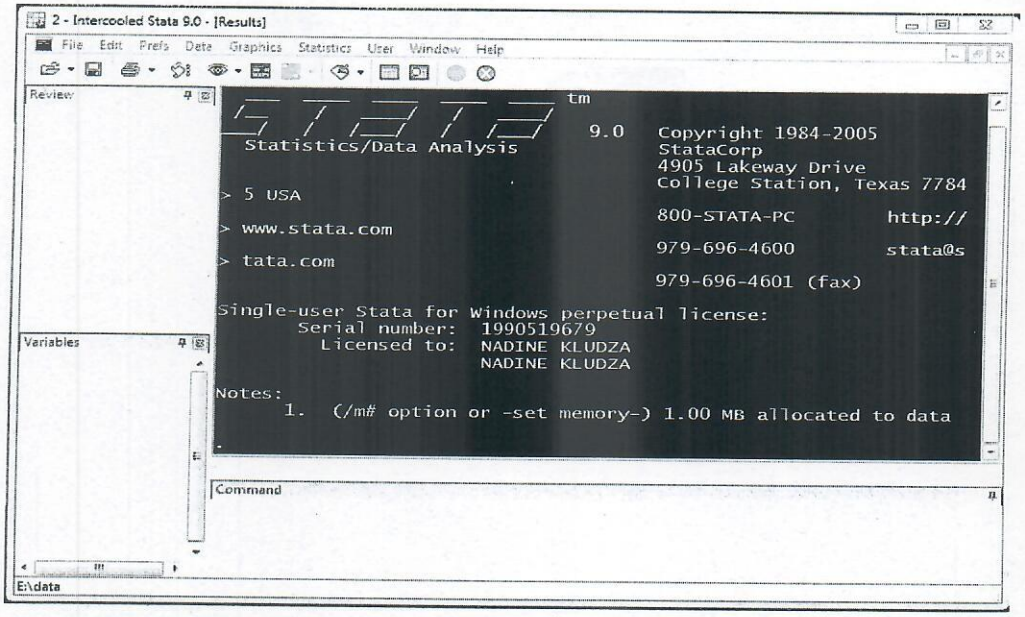
1.2.1.1. L'installation de STATA

L'installation de Stata pour Windows s'effectue comme celle de n'importe quel autre logiciel.

PRATIQUE

1.2.1.2. Les fenêtres

Au démarrage on obtient la capture d'écran suivante:



La fenêtre **Review** qui affiche l'historique des commandes tapées par l'utilisateur et permet d'en rappeler une facilement. La fenêtre **Results** est celle qu'utilise Stata pour afficher tous les résultats des commandes tapées par l'utilisateur. La fenêtre **Variables** détaille toutes les variables présentes dans la base de données actuellement ouverte dans Stata. La fenêtre **Command** permet à l'utilisateur d'enter les commandes. Chaque commande de Stata doit être validée en appuyant sur la touche entrée. Un clic sur une variable dans la fenêtre **Variables** permet d'afficher le nom de cette variable dans la fenêtre **Command**.

1.2.1.3. La barre d'outils

La barre d'outils rassemble les raccourcis des commandes de bases pour ouvrir créer ou enregistrer les fichiers créés pas Stata (fichiers de base de données fichiers de résultats et les fichiers de programmes).

Dans l'ordre l'icône ouvrir (1) ; enregistrer ; imprimer ; visualiser ou créer un fichier log (2) ; afficher l'aide et diverses options (mise à jour lien vers le site Internet de Stata etc.) (3) ; afficher les résultats (4) ; afficher un graphique ; ouvrir ou créer un fichier do (5) ; modifier la base de données (6) ; voir la base de données (7) ; faire défiler les résultats ; arrêter l'exécution d'une commande.

1- A l'ouverture d'une base de données Stata charge cette dernière dans la mémoire vive le fichier ouvert n'est plus relié à la source. Par défaut 10 MB sont alloués à la mémoire vive mais cet espace peut s'avérer insuffisant pour de grosses bases de données. Sinon la base ne s'ouvrira pas sauf si on augmente la quantité de mémoire disponible pour Stata. Pour vérifier la taille de la base à charger, on utilise la commande **describe using** *mabase.dta*. Si celle-ci nécessite qu'on augmente la taille de 50Mo la commande est la suivante : **set memory 50m.** Lorsque la base est plus grosse que la taille de la mémoire vive installée, il faut recourir à la

mémoire virtuelle **set virtual on.** Cela ralentit considérablement Stata. Pour économiser de l'espace mémoire on peut compresser la base **compress.**

2- **fichier log**

log using *nomfichier*: pour conserver les commandes et les résultats qui apparaissent à l'écran dans un fichier. Le fichier donc est enregistré sous format .smcl

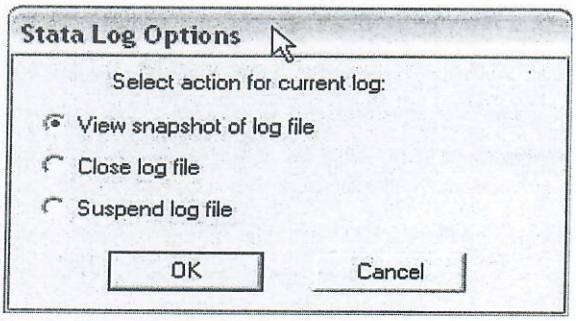
log close : pour fermer ce fichier texte

log off : pour interrompre temporairement l'écriture sur le fichier

log on pour réactiver le fichier

log using *nomfichier* **append**: pour ajouter au fichier résultats de la veille les résultats de la nouvelle session

Un fichier log est un fichier (format texte) d'impression des commandes et des résultats de ces commandes au cours d'une session de stata. Il permet de garder une trace des résultats. Un premier clic sur cette icône permet de créer un fichier log en spécifiant son chemin d'accès et le format .log (au lieu du format .smcl par défaut). Le fichier log est alors ouvert et prés à enregistrer l'historique des commandes et des résultats. Avec un second clic sur cette icône, on a les options suivantes comme l'indique la capture d'écran ci-dessous : (a) la première permet de voir le contenu du fichier log (b) la seconde permet de le fermer (c) et la troisième permet de le suspendre pour un enregistrement ultérieur. Par la suite vous verrez que la programmation rend plus simple la gestion des fichiers log.



3- L'aide de stata est précieuse car on ne peut maitriser qu'une infime partie des capacités du logiciel. De même l'aide est utile pour maitriser la syntaxe de chaque commande ainsi que ses subtilités. L'aide de stata est facile d'utilisation si on connait déjà la commande dont on a besoin mais dans le cas contraire on a la probabilité de faire des recherches par mots clés. il est également facile de parcourir la table des matières de l'aide et de repérer la commande recherchée. Pour afficher l'aide dans la fenêtre des résultats de stata il suffit de taper dans la fenêtre de commande : **help** suivi de la commande (par exemple **help graph** permet d'obtenir de l'aide sur les graphiques). Vous pouvez également utiliser le menu *help* dans la barre des menus.

4- Cette icône . permet de faire disparaitre toute fenêtre qui vient masquer celle des résultats (par exemple : fenêtre d'aide graphiques etc.).

5- Un fichier *do (do file)* est le fichier (format texte) dans lequel sont incluses les commandes de stata sous forme de programme. Grâce à ce fichier on garde une trace des commandes exécutées par stata. Ce fichier contient donc un ensemble cohérent de commandes à exécuter par stata dans l'ordre de leur apparition. Cliquer sur cette icône revient à ouvrir l'éditeur du fichier *do* comme le montre la capture ci-dessous. Dans l'éditeur du fichier *do* on a encore une barre des menus et une barre d'outils. Dans la barre d'outils il assez intuitif de savoir ce à quoi correspondent les différentes icones à l'exception des deux dernières probablement. Ces deux dernières icones correspondent aux commandes du menu *Tools*. les icones *Do* et *Run* permettent d'exécuter les commandes du fichier *do* en entier à la différence que l'icône *Run* ne permet pas de visualiser les résultats (l'intérêt est de voir si le fichier *do* s'exécute correctement et qu'il ne comporte pas d'erreur qui en bloque l'exécution). La création des fichiers do sera abordée dans la section 4.

6- Cette commande permet de visualiser la base de données avec la probabilité de modifier les données. On peut obtenir le même résultat avec la commande **edit.**

7- Cette commande permet de visualiser la base de données sans possibilité de modifier les données. La commande **browse** permet d'avoir le même résultat.

1.3. Configuration et importation

- **set** permet de configurer Stata.

Exemple:

set memory *30* **m** : Stata utilise 30 Mo de mémoire (1Mo sinon)

set *type* **double**: les variables créées après cette commande sont en double précision

set *more* **off** : le résultat d'exécution d'une commande se déroule sans arrêt dans la fenêtre d'output

set obs *n* : configurer la taille de l'échantillon égale à n observations

set obs *500* crée un fichier de 500 observations

Stata peut lire les données sous format ASCII (ou fichier texte ou extension .txt) et les fichiers sous extension .dta.

- **use** permet d'importer un fichier de données dans Stata

Syntaxe: **use** *filename[clear nolabel]*

Exemple:

use "C:\Users\Fatou\Desktop\FORMATION \CI_individu1.dta" clear

- **edit**: pour visualiser ou éditer la base
- **list**: afficher la liste des variables pour tous les individus.
- **list in** f/n: pour afficher les variables pour les n premiers individus de la base

- **list in** –n/l: pour afficher les variables pour les n derniers individus de la base
- **list** *var** : affiche toutes les variables dont les noms commencent par var
- **list** **var* : affiche toutes les variables dont les noms se terminent par var
- **list** x-y : affiche les variables x et y ainsi que toutes celles se trouvant entre les deux veriables dans l'ordre d'apparition dans la base de données

Exemples:

list in 1/5 ; list sexe in 1/10; list in -5/l

- **describe**: pour avoir une description de chacune pour chacune des variables de la base.
- **describe** *age*: pour avoir une description de la variable age
- **ds**: pour avoir une liste succincte des variables (sans détail)
- **codebook** *age*: détail de la variable âge (missing borne moyenne etc.)
- **lookfor** *at*: pour lister tous les noms ou libellé des variables contenant ''at''
- **clear**: pour effacer la mémoire

PARTIE II : LES OPERATIONS ET FONCTIONS DANS STATA ET COMMANDES DE GESTION DES VARIABLES

2.1 Les opérateurs et fonctions mathématiques

Addition	+	**Soustraction**	-
Multiplication	*	**Division**	/
Egalité	=	**Inégalité**	~= ou!=
Exposant	^	**Partie entière**	**int ()**
Racine	**sqrt ()**	**Exponentielle**	**exp ()**
Logarithme	**log ()**	**Valeur absolue**	**abs ()**
Sup.(resp.inf)	**>(resp.<)**	**Sup.(resp.inf) ou égal**	**>=(resp.<=)**
Ou	|	**Et**	**&**
Minimum	**min()**	**Maximum**	**max ()**

2.1.1. Les opérateurs arithmétiques (+ - / * ^)

Exemples :

generate *y=x*^2 crée une nouvelle variable y telle que y soit le carré de x.

dis 23*7 ou combiner avec les commandes

2.1.2. Les opérations de relations (de comparaison)

Remarque : il existe une exception pour le signe d'égalité. En effet lorsque la commande **if** précède une condition d'égalité il faut utiliser le signe « == » au lieu du signe « = » pour exprimer cette égalité.

Exemple : **list if** x==10 liste les observations dont la valeur de x est égale à 10.

list if x>. liste les observations dont les valeurs sont manquantes

2.1.3 Les fonctions mathématiques

Exemple : **egenerate** y=*log*(*sqrt*(*abs*(x))) crée une variable qui est égale au logarithme naturel de racine carrée de la valeur absolue de x.

2.1.4. Les opérateurs logiques

Ou | (combinaison de la touche *altgr* et la touche « 6 » du pavet alphanumérique)

Et &

Exemple : list if x>3 & x<20 liste toutes les observations dont la valeur est comprise entre 3 et 20 bornes non comprises.

Remarque : l'operateur & est prioritaire sur l'operateur |

List if x>50 | (x>30 & z<2.5) équivaut à écrire List if x>50 | x>30 & z<2.5

- **sort**: pour trier les données suivant une ou plusieurs variables

exemple: **sort** *age*: trier par ordre croissant

- **gsort–***age*: trier par ordre décroissant

sort *age sexe*: trier par âge et par sexe

- **drop**: supprimer des variables

drop_all = clear: effacer la mémoire

Exemples:

drop if *age*<15 : supprime les individus d'âges<15 ans

drop in 1/5 : supprime les 5 premiers individus de la base

drop if *age*==15 | *age*==60 : on supprime les individus d'âge égal à 15 ans et 60 ans seulement

drop if *age*>15 & *age*<60 compris entre 15 et 60 ans

- **keep**: garder des variables

La commande a la même syntaxe que la commande drop.

2.2.2 Commandes de base: les expressions if by et in

by permet de répéter une commande pour chaque valeur (ou modalité) d'une variable donnée. Syntaxe générale pour ***by***

by *variables : commande*

Avant d'utiliser ***by*** il faut d'abord classer les observations en fonction des valeurs de la variable à laquelle la commande ***by*** va s'appliquer la commande ***sort*** permet d'effectuer le classement par ordre croissant.

if permet de spécifier les conditions dans lesquelles une commande doit être exécutée. Syntaxe générale pour **if** :

commande **if** *conditions*

Exemple : **generate** *y=x^*(0 5) **if** x>=0 crée une variable y qui est égale à la racine carrée de x si x est positif.

recode *menb3* 16/24=1 25/44=2 45/59=3 nonmiss=4 , gen(grage) : crée une variable nommée grage en regroupant les âges des chefs de ménages en classes (16 à 24)=1 (25à 44)=2 (45 à 59)=3 (60+)=4

gen *agegpe* = *recode*(*age* ,14, 24, 45, 59, 99) regroupe l'âge en 5 classes dont de (1 à 14) (15 à 24) …..(60 à 99)

replace *age* = 98 if *age* == . Transformer les valeurs manquantes de l'âge en 98

replace *sexe* = . **in** 100/110 recode sexe à missing pour les observations 100 à 110

2.2.5 Traitement des données manquantes

mvencode _all, mv(999) recode les missing de toutes les variables à 999

mvdecode *sexe*, **mv**(0 -1) recode les 0 et les -1 de la variable sexe à missing

2.2.6 Création d'étiquettes: label define label value

Prenons l'exemple précédent du point **2.2.4**

recode *taille* 0/4=1 5/9=2 10/14=3 15/19=4 20/24=5 nonmiss=6 gen(grtaille)

lab def *agegpe* 1« moins de 5ans» 2«5_9ans » …… 6 «25 ans et plus

Pour afficher les étiquettes

Lab val *agegpe agegpe*

Puis

tabulate *agegpe*

Pour enlever les étiquettes: **tabulate** *agegpe*, **nolabel**

Exercice 1

1. Convertissez le fichier ménage en Stata
2. Vérifiez la taille de la base et augmentez si nécessaire une taille
3. Enregistrez les résultats dans un fichier approprié et programmez les commandes
4. Faîtes la tabulation de la variable région
5. Visualisez ou éditez la base
6. Affichez la liste des variables pour tous les ménages.
7. Affichez les variables pour les 10 premiers ménages de la base

8. Affichez les variables pour les 10 derniers ménages de la base
9. Affichez toutes les variables dont les noms commencent par a
10. Affichez toutes les variables dont les noms se terminent par men
11. Affichez les variables a1et deptet ainsi que toutes celles se trouvant entre les deux variables dans l'ordre d'apparition dans la base de données
12. Utilisez la commande qui permet de décrire la variable âge.
13. Créez les nouvelles variables region, milieu (urbain rural), departement, sexe à partir de leurs correspondances dans la base
14. Créez une nouvelle variable quintile à partir de deptet découpée en 5 classes
15. Donnez les valeurs correspondantes à chaque modalité de la variable quintile
16. Mettez des étiquettes à quintile (1er quintile…5ème quintile)
17. Utilisez la commande appropriée pour faire des comparaisons entre les différents quintiles
18. Créez une nouvelle variable date qui prend en compte le jour, le mois et l'année de l'enquête
19. Créez une nouvelle variable grapage à partir de la variable âge de la base et regroupez les modalités par groupe d'âges décennaux
20. Mettez des étiquettes à la nouvelle variable créée

Correction1

1. Utiliser Stata transfer
2.

set more off

describe using "C:\.....\Formation Stata\ESPS_2005**menage_esps-2005-2006.dta** "

set mem 30m

use "C:\.....\Formation Stata\ESPS_2005**menage_esps-2005-2006.dta**"

3.

log using "C:\Documents and Settings\Utilisateur\Bureau\Formation Stata\ESPS_2005**resultat.smcl** "

4. **tab** a1

5. **browse** , ou **edit**

6. **list** _all

7. **list** in f/10

8. **list** in -10/l

9. **list** a*

10. **list** *men

11. **list** a1-deptet

12. **codebook** menb3

13. **gen** region = a1

    gen milieu = a6

    gen departement = a2

    gen sexe = menb2

14. **egen** quintile=cut(deptet) , group(5)

15. **tab** quintile

16. **lab def** quintil 0 " Q1 " 1 " Q2 " 2 " Q3 " 3 " Q4 " 4 " Q5 "

    lab val quintile quintil

17. Soit on décrit les dépenses par tête de la variable quintile pour chacune de ses modalités

    codebook quintile deptet if quintile==0

    codebook quintile deptet if quintile==1

    codebook quintile deptet if quintile==2

    codebook quintile deptet if quintile==3

    codebook quintile deptet if quintile==4

    Soit on procède en une commande comme suit :

    table quintile ,content (min deptet max deptet sd deptet)

18. egen date = concat(a13a a13b a13c)

19.

recode menb3 min/24 =1 25/34 =2 35/44 =3 45/54 =4 55/64 =5 65/max=6, **gen**(grapage)

20.

lab def grapage 1 "moins de 24 ans" 2"25-34 ans" 3"35-44 ans" 4"45-54ans" 5"55-64ans" 6"plus de 65ans"

lab val grapage grapage

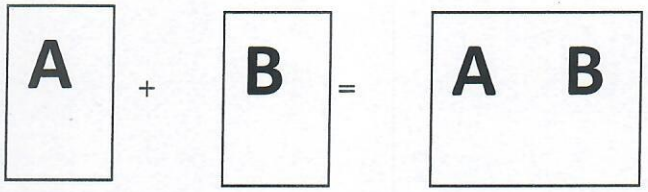
PARTIE III : FUSION DE BASES DE DONNEES

3.1. les commandes merge et append

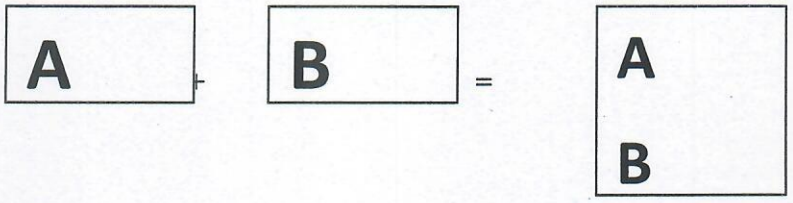
Les commandes **merge** et **append** permettent respectivement de fusionner horizontalement et verticalement deux bases de données. De façon générale la commande merge vous permet d'ajouter

de nouvelles variables à la base de données et la commande append vous permet d'ajouter de nouvelles observations.

merge :



append :



Les différentes étapes pour fusionner deux bases de données.

3.2. La commande merge

Tout d'abord il faut avoir une variable commune aux bases **A** et **B** qui permettra de faire la fusion prenons par exemple les noms de régions contenus dans la variable nommée *region*.

1. Ouvrir la base A et classer les observations par nom des region : **sort** *region*

2. Ouvrir la base B classer les observations par nom des régions : **sort** *region*

Enregistrer puis fermer la fenêtre stata de la base B

3. Revenir dans la fenêtre de la base A pour appliquer la commande suivante :

merge region **using** ''chemin d'accès de la base B'' puis enregistrer ensuite la nouvelle base obtenue.

Remarque : la variable de fusion (nom des *region* dans cet exemple) de la base A doit être rigoureusement identique à celle de la base B.

3.3 La commande append

Pour la commande append la syntaxe est plus simple. Il suffit d'ouvrir la première base (base A) et dans la fenêtre de la commande de stata taper la ligne de commande suivante :

append using ''*chemin d'accès de la base* B''

Exercice 2

1-a) A partir du *fichier ménage ESPS1* converti en fichier .dta créer deux nouveaux fichiers **fusion_merge1** et **fusion_merge2**. Le fichier fusion1 comportant les variables de a1 à menb7 et le fichier fusion2 les variables a1 a2 a6 a7 a8 et age5 à educ117.

b) Fusionner les deux fichiers en un seul fichier **fusion_merge** en veillant à avoir un identifiant unique

2-a) A partir du *fichier ménage* ESPS converti en fichier .dta créer un nouveau fichier **fusion_append1** comportant les variables a1 a2 a6 a7 a8 et age5 à educ9. Compter le nombre d'observations de la base créée **fusion_append1** .

b) A partir du *fichier population* ESPS converti en fichier .dta créer un nouveau fichier **fusion_append2** comportant les variables de a1 à menb7. Compter le nombre d'observations de la base créée **fusion_append2.**

c) Fusionner les deux fichiers en un seul fichier **fusion_append** en veillant à avoir un identifiant unique. Compter le nombre d'observations de la base créée **fusion_append.**

Correction 2 1-a)&b)

```
. use "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\menage_esps-2005-2006.dta", clear

. keep  a1 a2 a6 a7 a8 a10 a11 a12 a13a a13b a13c a14a a14b a14c a15a a15b a15c a16 a17 a18 menb0 menb1 menb2 menb3 menb4 menb5 menb6 menb7

. sort  a1 a2 a6 a7 a8

. save "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_merge1.dta"
file C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_merge1.dta saved

. use "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\menage_esps-2005-2006.dta", clear

. keep  a1 a2 a6 a7 a8 age5 educ1 educ2 educ3 educ4 educ5 educ6 educ7 educ8 educ9 educ101 educ102 educ103 educ104 educ105 educ106 educ111 educ112 educ113 educ114 educ115 educ116 educ117

. sort  a1 a2 a6 a7 a8

. save "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_merge2.dta"
file C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_merge2.dta saved

. merge a1 a2 a6 a7 a8 using "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_merge1.dta"
(label menb7 already defined)
(label menb6 already defined)
(label menb5 already defined)
(label menb4 already defined)
(label menb2 already defined)
(label menb1 already defined)
(label a15b already defined)
(label a14b already defined)
(label a13b already defined)
(label a12 already defined)
(label a6 already defined)
(label a1 already defined)
```

2-a)&b)&c)

```
. use "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\menage_esps-2005-2006.dta", clear

. keep  a1 a2 a6 a7 a8 age5 educ1 educ2 educ3 educ4 educ5 educ6 educ7 educ8 educ9

. count
13568

. save "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_append1.dta"
file C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_append1.dta saved

. use "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\population_esps-2005-2006.dta", clear

. keep  a1 a2 a6 a7 a8 age5 educ1 educ2 educ3 educ4 educ5 educ6 educ7 educ8 educ9

. count
123558

. save "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_append2.dta"
file C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_append2.dta saved

. append using "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_append1.dta"
(label educ9 already defined)
(label educ8 already defined)
(label educ7 already defined)
(label educ6 already defined)
(label educ5 already defined)
(label educ4 already defined)
(label educ3 already defined)
(label educ2 already defined)
(label educ1 already defined)
(label age5 already defined)
(label a6 already defined)
(label a1 already defined)

. count
137126

. save "C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_append.dta"
file C:\Users\Hp\Desktop\Formation Stata\ESPS_2005\fusion_append.dta saved
```

PARTIE IV : LES STATISTIQUES DESCRIPTIVES

4.1. Les fréquences et les tableaux

4.1.1 La commande *summarize*

La commande ***summarize*** (ou ***sum*** en abrégé) calcule pour une variable ou une liste de variables la moyenne l'écart-type le minimum et le maximum de l'échantillon sélectionné.

Syntaxe générale : **sum** *noms_variables* (**if in**)

Exemples :

sum *y* : la commande **sum** s'applique à la variable *y*

sum *xy* : la commande **sum** s'applique aux variables *x* et *y*

Bysort *region* : **sum** *deptet* : la commande **sum** s'applique séparément à chaque modalité de la variable *region*

sum *deptet* **if** *region* ==1 : la commande Sum s'applique à la variable *deptet* mais uniquement pour les observations dont la variable *region* est égale 1

Remarque1 : lorsqu'aucune variable n'est spécifiée à la suite de la commande sum alors les statistiques descriptives sont faites pour toutes les variables de la base de données.

Remarque2 : **sum** y, **detail** (avec l'option **detail** la commande **sum** donne en plus des statistiques standard le *skewness* et le *kurtosis* de la variable y)

Remarque3 : la commande **tabstat** permet également de faire des statistiques descriptives. Elle offre plus de flexibilité que la commande **sum** en ce sens qu'elle permet de faire un tableau unique de statistiques descriptives pour plusieurs variables et elle offre une plus large panoplie de statistiques.

mean permet de calculer la moyenne d'une variable, l'écart-type et l'intervalle de confiance de l'estimation de la moyenne: **mean** *age*

4.1.2. La commande *tabulate*

La commande tabulate (ou tab en abrégé) calcule les fréquences des observations d'une variable et permet de faire des tableaux croisés pour deux variables.

Exemples :

tab *y* : permet de faire un tableau des valeurs de *y* avec leurs fréquences

tab *x y* : fait un tableau croisé des valeurs de *x* et *y*

tab *y x* row : tableau croisé de *y* et *x* avec les fréquences en lignes

tab *y x* col : tableau croisé de y et *x* avec les fréquences en colonnes

Il existe d'autres variantes de la commande tab il s'agit de *tab1* et *tab2*

tab1 y x : crée non pas un tableau croisé de y x mais un tableau séparé pour chacune de ces variables.

tab2 y x z : crée un tableau croisé pour chaque combinaison possible de deux variables de cette liste de variables (xy yz et xz)

On peut combiner tab et ses variantes avec **by if** et **in**.

❖ La commande **tabulate** permet de dichotomiser .

Commande: **tab** *var,* **gen** (*var)*

<u>Exemples</u> :

tab *instruction,* **gen** (*instruction*)

❖ **tabulate** permet d'effectuer un test du chi-2

Test: Ho: Indépendance contre H1: dépendance

tabulate *pauvrete alphabetisation,* **col chi2**

```
. tab pov educ1 , col chi2

+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+

           | c1.sait-il lire et
    menage |       ecrire
    pauvre |       oui        non |     Total
-----------+----------------------+----------
non pauvre |     3,782      3,970 |     7,752
           |     67.74      49.72 |     57.13
-----------+----------------------+----------
    pauvre |     1,801      4,015 |     5,816
           |     32.26      50.28 |     42.87
-----------+----------------------+----------
     Total |     5,583      7,985 |    13,568
           |    100.00     100.00 |    100.00

         Pearson chi2(1) = 435.7937   Pr = 0.000
```

La P-value est inférieure à 1% donc on rejette l'hypothèse nulle d'indépendance entre le sexe et la pauvreté. On note une association fort significative entre la pauvreté de ménage et l'alphabétisation.

❖ **collapse** permet de créer une base de données ne contenant que des statistiques descriptives

Exemple

collapse mean *(age)* median(*deptet*), by(*a1*) : calcule l'âge moyen et les dépenses par tête médianes et ne conserve que ces statistiques dans la base.

4.1.3. Les corrélations

correlate *deptet taille*: calcule le coefficient de corrélation entre la taille et deptet

Plus simplement: **corr** *deptet taille*

```
. corr deptet ta
(obs=13568)

             |   deptet   taille
-------------+------------------
      deptet |   1.0000
      taille |  -0.3009   1.0000
```

Les dépenses par tête sont négativement corrélées à la taille du ménage. Plus la taille du ménage augmente, plus les dépenses par tête diminuent.

pwcorr: permet également de calculer le coefficient de corrélation

comparer **pwcorr** *deptet taille* et comparer **corr** *deptet taille*

Pour savoir si cette corrélation est significative, on peut procéder à un test de nullité du coefficient de corrélation à l'aide de la commande pwcorr et de son option sig:

Exemple

pwcorr *deptet educ1 taille*

```
. pwcorr deptet educ1 taille

             |   deptet    educ1   taille
-------------+---------------------------
      deptet |   1.0000
       educ1 |  -0.2366   1.0000
      taille |  -0.3009   0.0441   1.0000
```

Les coefficients de corrélation sont tous significatifs. On a une chance sur 1000 de se tromper en affirmant que le coefficient de corrélation est différent de 0 entre (deptet, taille).

4.1.4. La commande centile

centile permet de calculer les quantiles spécifiés

centile *deptet*, c(25 50 75): pour calculer les quartiles de la variable deptet.

centile *deptet*, c(10 25 50 60): pour calculer le premier décile, le premier quartile, la médiane et le sixième décile de la variable deptet

centile *deptet*, c(10 20 30 40 50 60 70 80 90): calcule les déciles

4.1.4. La commande ci

ci *deptet*, level(95): permet de donner un intervalle de confiance à 95% pour la variable deptet.

4.2. La pondération

4.2.1. La commande fweight

- ❖ C'est la pondération naïve non basée sur des estimateurs mais plutôt sur la structure de la population.
- ❖ **fweight** prend en compte, dans un calcul statistique, la fréquence d'apparition des modalités d'une variable donnée.

Syntaxe : **command** *variable* [**fweight** =variable],

<u>Exemples</u> :

table avof1 [**fweight** =menb2] , content(mean aln11_2 mean menb3 mean avof2)

```
. table  avof1 [fweight =menb2] , content(mean aln11_2 mean menb3 mean avof2 )
```

f1.statut occup.logement	mean(aln11_2)	mean(menb3)	mean(avof2)
propriétaire/copropriétaire	25777.55	52.30501	4.517616
locataire/colocataire	33132.34	42.9552	2.464952
loge gratuite par un tiers	21324.13	47.39874	2.874214
autre à présiser	26325.68	48.09459	3.45946

NB:

Avof1 désigne le statut d'occupation du ménage dans notre base ménage ESPS1

menb2 est la variable genre de notre base ménage ESPS1

aln11_2 représente les dépenses en gaz

menb3 désigne l'age et avof2 le nombre de pièces occupées par le ménage

[**fweight** =menb2] signifie que les statistiques pour chaque genre seront pondérées par le pourcentage de genre

table est une extension de **tabulate** tout comme **egen** est une extension de **generate**

4.2.2. La commande pweight

- ❖ La commande **pweight** fait référence au poids d'échantillonnage : C'est l'inverse de la probabilité d'appartenir à l'échantillon.

- ❖ Généralement ce poids est déjà calculé et disponible même avant la collecte pour les données recueillies sur la base des méthodes probabiliste. Dans le cas de l'ESPS1, il s'agit de la variable *poid_fin*

command *variable* [**pweight** =inverse de la probabilité d'inclusion],

Exemples :

table avof1 [**pweight** = poid_fin] , content(mean aln11_2 mean menb3 mean avof2)

```
. table  avof1 [pweight = poid_fin] , content(mean aln11_2 mean menb3 mean avof2 )
```

f1.statut occup.logement	mean(aln11_2)	mean(menb3)	mean(avof2)
propriétaire/copropriétaire	27824.53	52.39878	4.617301
locataire/colocataire	42913.38	44.45884	2.333329
loge gratuite par un tiers	26994.38	46.70374	2.859147
autre à présiser	29178.68	48.76066	3.249052

4.2.3. La commande iweight

- ❖ La commande **iweight** fait référence au poids des individus. Par exemple un ménage de taille 5 n'aura pas le même poids individuel qu'un ménage de taille 10 même s'ils ont des probabilités d'appartenir à l'échantillon identiques.

- ❖ Généralement ce poids est déjà calculé et disponible après la collecte dans une base de données dont la collecte est fondée sur les méthodes probabiliste et dont les caractérist. Dans le cas de l'ESPS1, il s'agit de la variable *pds*

command *variable* [**iweight** =pondération_individu],

Exemples :

table avof1 [**iweight** = pds] , content(mean aln11_2 mean menb3 mean avof2)

```
. table  avof1 [iweight = pds] , content(mean aln11_2 mean menb3 mean avof2 )
```

f1.statut occup.logement	mean(aln11_2)	mean(menb3)	mean(avof2)
propriétaire/copropriétaire	31675.65	54.12609	5.510982
locataire/colocataire	49644.59	47.4811	2.881893
loge gratuite par un tiers	33034.65	48.76436	3.462545
autre à présiser	24412.73	47.73293	3.812958

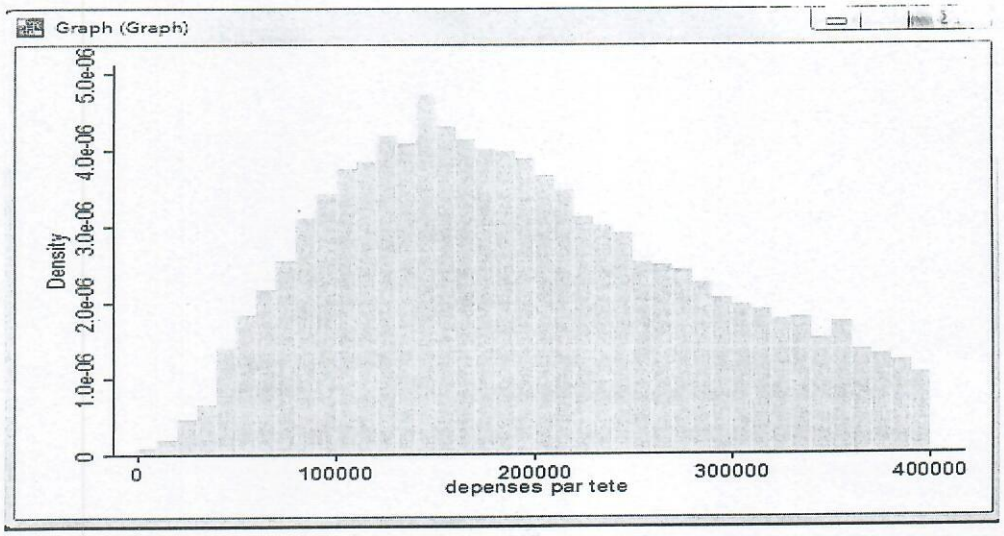
PARTIE V : LES GRAPHIQUES

5. 1. Les histogrammes

Nous construisons un histogramme de la distribution de la variable *deptet*

histogram deptet

histogram deptet **if** deptet <400000



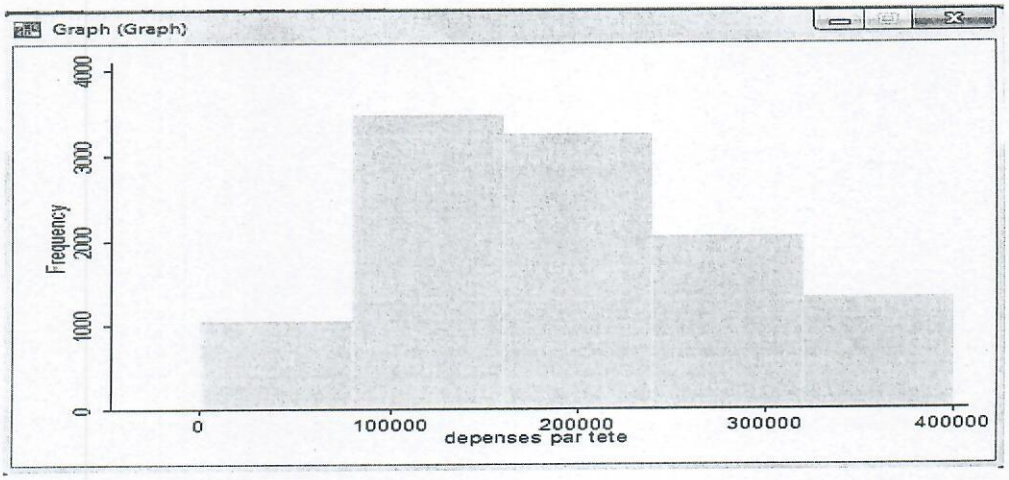
On peut également faire usage des options pour rendre plus conviviale notre histogramme.

L'option bin(5) permet d'avoir un histogramme en effectif de 5 classes

histogram deptet **if** deptet <400000, **bin**(5)

L'option frequency permet d'afficher les effectifs sur l'axe des ordonnees

histogram deptet **if** deptet <400000, **bin**(5) **freq**

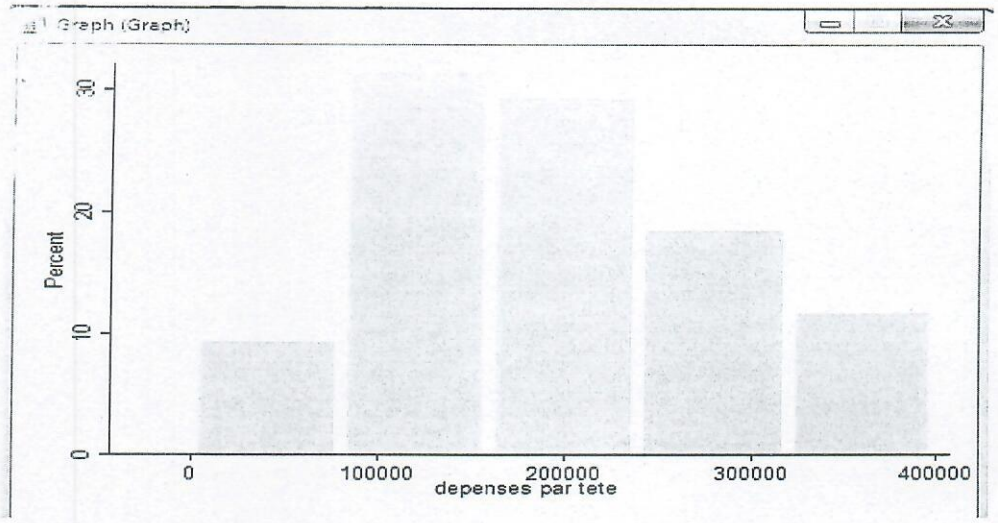


L'option percent permet d'afficher les pourcentages sur l'axe des ordonnées

histogram deptet **if** deptet <400000,**bin**(5) **percent**

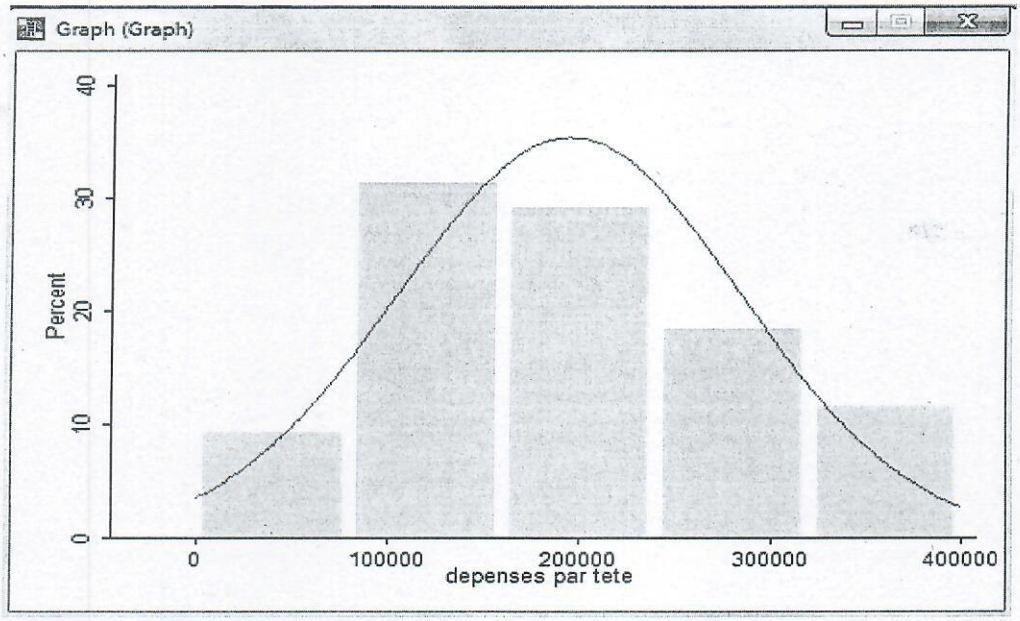
L'option gap(10) permet de creer un espacement de 10 entre les classes

histogram deptet **if** deptet <400000,**bin**(5) **percent gap**(10)



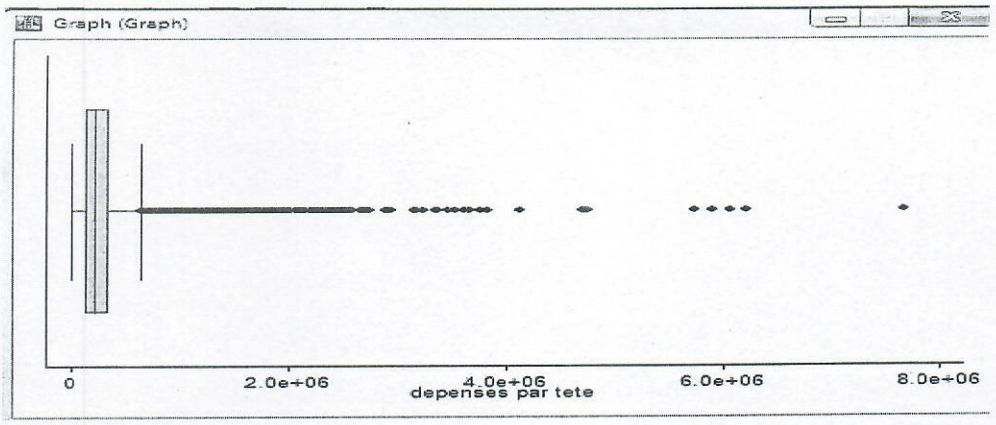
L'option normal ajoute la densite de la loi normale a l' histogramme

histogram deptet **if** deptet <400000,**bin**(5) **percent gap**(10) **normal**

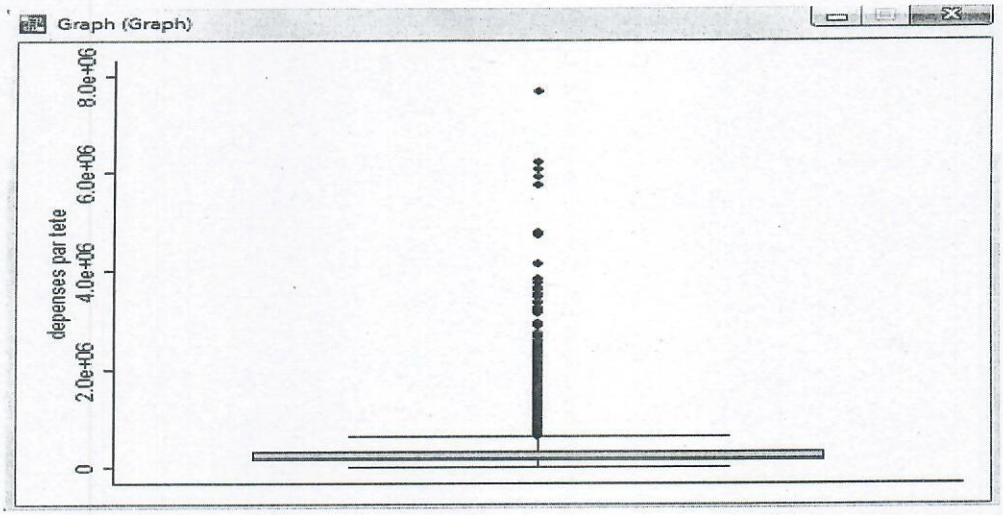


5.2. Boite à moustaches

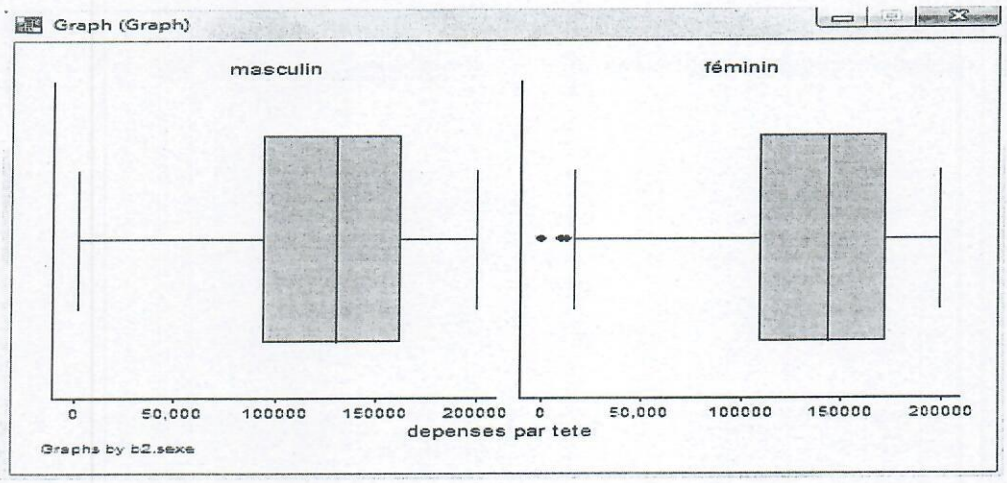
graph hbox deptet : boite à moustaches horizontale



graph box deptet : boite à moustaches verticale



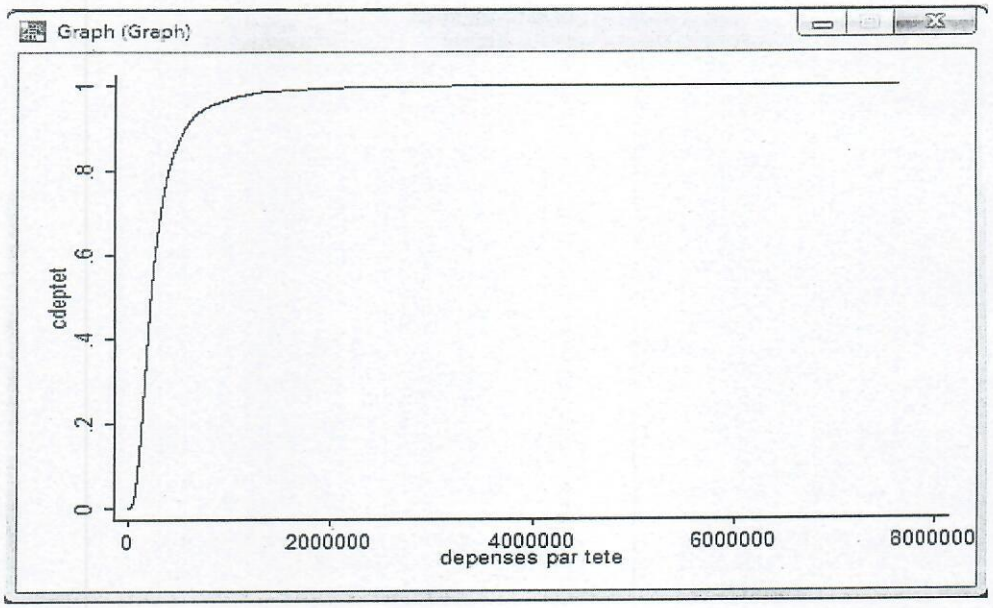
graph hbox deptet if deptet <200000, by(menb2) : boîte à moustache à sexe



5. 3. Les fonctions de répartition

cumul deptet, gen(cdeptet). crée la variable cumulée de deptet

line cdeptet deptet, sort : permet de tracer la fonction de répartition à l'aide de la variable cdeptet



Exercice3

Construire la fonction de répartition des variables dépenses en santé (depsant) et dépenses en logement (deplog) et les afficher sur un même graphique.

Correction3

cumul depsant, **gen**(cdepsant)

cumul deplog, **gen**(cdeplog)

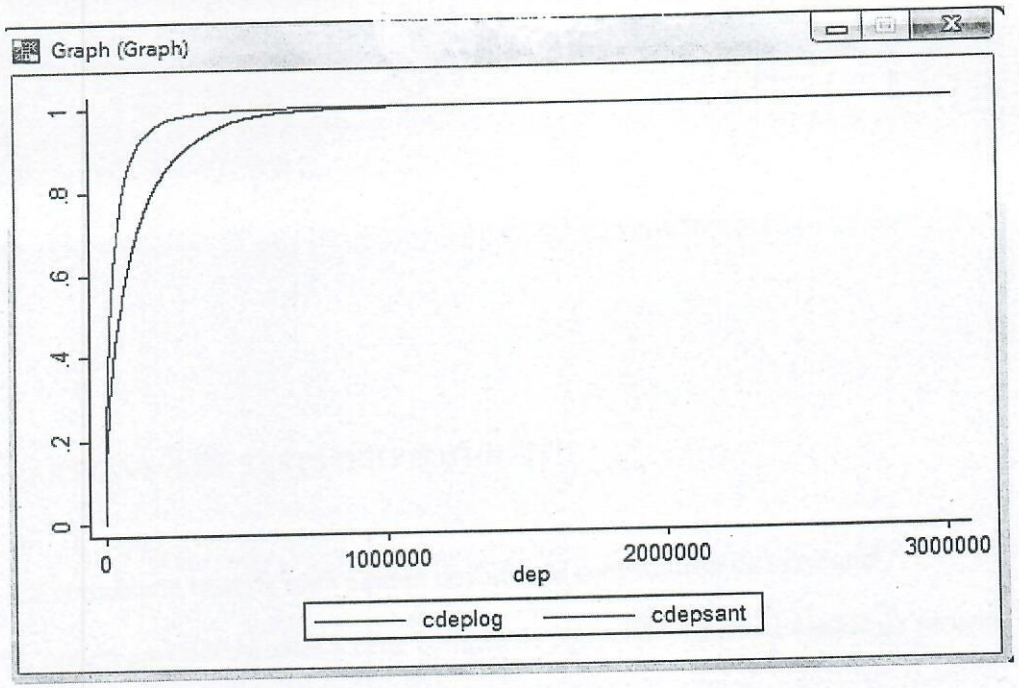
stack cdepsant depsant cdeplog deplog, **into**(c dep) **wide clear**

sum c dep cdepsant depsant cdeplog deplog

```
. sum  c dep cdepsant depsant cdeplog deplog
```

Variable	Obs	Mean	Std. Dev.	Min	Max
c	27136	.5000369	.2886805	.0000737	1
dep	27136	73494.48	120395.5	0	3027000
cdepsant	13568	.5000369	.2886858	.0000737	1
depsant	13568	42873.27	75724.62	0	1589590
cdeplog	13568	.5000369	.2886858	.0000737	1
deplog	13568	104115.7	146224.3	0	3027000

line cdeplog cdepsant dep, **sort**



L'option *unpaired* associée à cette commande signifie que les deux variables ne représentent pas la même chose.

Les résultats se présentent comme suit :

```
. ttest aln11_19==aln11_5, unp
Two-sample t test with equal variances
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
aln11_19	13568	50172.33	750.0823	87370.91	48702.07	51642.6
aln11_5	13568	42770.96	1274.126	148412.4	40273.5	45268.43
combined	27136	46471.65	739.5876	121832.3	45022.02	47921.28
diff		7401.369	1478.52		4503.395	10299.34

```
    diff = mean(aln11_19) - mean(aln11_5)                         t =   5.0059
Ho: diff = 0                                     degrees of freedom =    27134

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Cas3

Grâce à la commande ttest, l'on peut également tester si la moyenne d'une variable est égale à une valeur donnée.

Commande : **ttest variable = #**

Exemple : **ttest menb3 = 50 if menb2==1 in -1000/l** permet de tester l'hypothèse selon laquelle l'âge des chefs de ménages hommes pour les 1000 dernières observations est égale à 50 ans.

6. 2. Comparaison de variance

La commande **sdtest** de stata permet de procéder à des comparaisons de variances. On retrouve les trois cas similaires à ceux de la commande **ttest**

Cas1

Soit la variable deptet de notre base ESPS I. Nous allons tester la différence de variance de dépenses par tête entre le groupe des chefs de ménage femme et celui des chefs de ménage homme.

La commande et les résultats se présentent comme suit:

```
. sdtest deptet , by(menb2)

Variance ratio test

Group          Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
masculin     10686    290303.1    3219.161    332774.7    283993    296613.3
féminin       2882    319471.5    5600.849    300677.5    308489.4  330453.5
combined     13568    296498.8    2802.401    326428.6    291005.7  301991.9

    ratio = sd(masculin) / sd(féminin)                 f =    1.2249
Ho: ratio = 1                     degrees of freedom = 10685, 2881

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
 Pr(F < f) = 1.0000         2*Pr(F > f) = 0.0000          Pr(F > f) = 0.0000
```

La variance des dépenses par tête des hommes qui vaut 332775 est supérieure en valeur à celle des femmes qui vaut 300678.

D'après les résultats du test, on peut conclure avec une précision de 95% que la variance des dépenses par tête est plus élevée dans le groupe des hommes que dans celui des femmes.

Cas 2

Test d'égalité de variances entre deux variables différentes.

Commande : **sdtest variable1 = variable2**

Cas 3

Test d'égalité de variances entre une variable et une valeur donnée.

Commande : **sdtest variable1 = #**

6. 3. Test de la médiane

median (var1), by(var2): permet de tester la médiane

Exemple

median (deptet), by(a6)

Median test

Greater than the median	a6.milieu urbain	rural	Total
no	3,061	3,723	6,784
yes	5,515	1,269	6,784
Total	8,576	4,992	13,568

Pearson chi2(1) = 1.9e+03 Pr = 0.000

Continuity corrected:
Pearson chi2(1) = 1.9e+03 Pr = 0.000

6. 4. Test d'indépendance de deux variables continues ou ordinales

Les test de Kendall ou de Spearman permettent de comparer les distributions de deux variables.

Commande : **ktau** variable variable

spearman variable variable

Pour changer de niveau de significativité , on utilise l'option st(#)

Exercice 4

1-Peut-on postuler qu'en moyenne les hommes chefs de ménage font plus de dépenses scolaires que les femmes chefs de ménages avec un niveau de confiance de 95%

2-Peut-on postuler que les hommes chefs de ménage sont plus scolarisés que les femmes chefs de ménages avec un niveau de confiance de 99%

3-a)Tester si la moyenne d'âge des 20% des hommes chefs de ménage les plus riches est égale à 50

b)Tester si la moyenne d'âge des 20% des hommes chefs de ménage les plus pauvre est égale à 50

c)Que peut-on conclure?

4-Vérifier si la fièvre palustre (edud51), les problèmes dentaires (edud54), sont indépendants de la région (a1)

5-Vérifier si l'inefficacité des traitements (edud108) est lié au type du personnel de santé (edud105)

Correction 4

1-Peut-on postuler qu'en moyenne les hommes chefs de ménage (menb2==1) font plus de dépenses scolaires (depscol) que les femmes chefs de ménages (menb2==2) avec un niveau de confiance de 95% ?

Réponse :

```
. ttest depscol, by(menb2)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
masculin |   10686    22961.18    547.8827    56636.34    21887.23    24035.13
féminin  |    2882    25063.57    1120.073     60130.3    22867.35     27259.8
---------+--------------------------------------------------------------------
combined |   13568    23407.75     492.785    57400.47    22441.83    24373.68
---------+--------------------------------------------------------------------
    diff |           -2102.393     1204.72               -4463.811    259.0246
------------------------------------------------------------------------------
    diff = mean(masculin) - mean(féminin)                         t =  -1.7451
Ho: diff = 0                                     degrees of freedom =    13566

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0405         Pr(|T| > |t|) = 0.0810          Pr(T > t) = 0.9595
```

Au seuil de 5%, on ne rejette pas l'hypothèse selon laquelle les dépenses scolaires des hommes chefs de ménage sont inférieures à celles des femmes chefs de ménages.

2-Peut-on postuler que les hommes chefs de ménage sont plus scolarisé (edud4) que les femmes chefs de ménages avec un niveau de confiance de 99%

Réponse :

On dichotomise la variable *educ4* qui est la variable scolarisation de la base grâce à la commande suivante :

```
. tab educ4, gen(educ4)

c4.scolaris
      ation |      Freq.     Percent        Cum.
------------+-----------------------------------
        oui |      4,364       32.16       32.16
        non |      9,204       67.84      100.00
------------+-----------------------------------
      Total |     13,568      100.00
```

Nous faisons alors un test de proportion concluant que la proportion des chefs de ménage homme scolarisés est supérieure à celle des femmes.

```
. prtest educ41, by(menb2)

Two-sample test of proportion                masculin: Number of obs =    10686
                                             féminin: Number of obs =     2882
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    masculin |    .3414748   .0045873                      .3324839    .3504658
     féminin |    .2480916   .0080453                      .2323231    .2638601
-------------+----------------------------------------------------------------
        diff |    .0933832   .0092612                      .0752316    .1115349
             |  under Ho:    .0098043     9.52   0.000
------------------------------------------------------------------------------
    diff = prop(masculin) - prop(féminin)                         z =   9.5247
Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 1.0000    Pr(|Z| < |z|) = 0.0000     Pr(Z > z) = 0.0000
```

3-a)Tester si la moyenne d'âge des 20% des hommes chefs de ménage les plus riches est égale à 50

On retrouve les 20% des chefs de ménage les plus riches. C'est la modalité 4 de la variable *quintile* qu'on vient de créer :

```
. egen quintile=cut(deptet),group(5)

. tab quintile

   quintile |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,713       20.00       20.00
          1 |      2,714       20.00       40.00
          2 |      2,713       20.00       59.99
          3 |      2,714       20.00       80.00
          4 |      2,714       20.00      100.00
------------+-----------------------------------
      Total |     13,568      100.00
```

Premièrement, on trie notre base suivant la variable deptet. Ensuite, on procède au test pour les 20% des hommes chefs de ménage plus riches.

```
. sort deptet

. ttest  menb3 = 50 if menb2==1 in -2714/l

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   menb3 |    2044    46.38796    .3154074    14.25976    45.76941    47.00652
------------------------------------------------------------------------------
    mean = mean(menb3)                                            t = -11.4520
Ho: mean = 50                                     degrees of freedom =     2043

    Ha: mean < 50                Ha: mean != 50                 Ha: mean > 50
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

On conclut que l'âge moyen des 20% plus riches des hommes chef de ménage est inférieur à 50 ans

3-b) Tester si la moyenne d'âge des 20% des hommes chefs de ménage les plus pauvres est égale à 50 ans

```
. sort deptet

. ttest  menb3 = 50 if menb2==1 in f/2713

One-sample t test
```

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
menb3	2362	52.43438	.2892457	14.05746	51.86718	53.00158

```
    mean = mean(menb3)                                     t =   8.4163
Ho: mean = 50                                 degrees of freedom =     2361

   Ha: mean < 50                 Ha: mean != 50                  Ha: mean > 50
Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

On en déduit que l'âge moyen des 20% plus pauvre des hommes chef de ménage est supérieur à 50 ans.

c) Que peut-on conclure?

Les 20% des hommes chefs de ménage les plus pauvres sont plus âgés que les 20% des hommes chefs de ménage les plus riches.

4-Vérifier si la fièvre palustre (edud51), les problèmes dentaires (edud54), sont indépendants de la région (a1)

```
. ktau edud51 a1

  Number of obs =    3368
Kendall's tau-a =      -0.0328
Kendall's tau-b =      -0.0495
Kendall's score = -185704
    SE of score =   55273.714   (corrected for ties)

Test of Ho: edud51 and a1 are independent
     Prob > |z| =       0.0008  (continuity corrected)

.
```

La fièvre palustre n'est pas indépendante de la région.

```
. ktau edud54 a1

  Number of obs =    3368
Kendall's tau-a =       0.0037
Kendall's tau-b =       0.0159
Kendall's score =   21112
    SE of score =   19605.875   (corrected for ties)

Test of Ho: edud54 and a1 are independent
     Prob > |z| =       0.2816  (continuity corrected)
```

Les problèmes dentaires sont indépendants de la région.

5-Vérifier si l'inefficacité des traitements (edud108) est lié au type du personnel de santé (edud105)

```
. ktau edud108 edud105

  Number of obs =    2772
Kendall's tau-a =       0.0050
Kendall's tau-b =       0.0557
Kendall's score =   19016
    SE of score =    6490.382   (corrected for ties)

Test of Ho: edud108 and edud105 are independent
     Prob > |z| =       0.0034  (continuity corrected)
```

L'inefficacité des traitements est liée au type du personnel de santé.

CONCLUSION SUR LES STATISTIQUES DESCRIPTIVES ET LES TESTS D'HYPOTHESES

Les résultats obtenues des analyses descriptives ne suffisent pas. Elles ne sont en réalité que de simples corrélations isolées, des apparitions des faits qui ne permettent pas d'établir des liens de causalité. Il faudra donc rechercher les effets intrinsèques de chacun des facteurs à l'explication d'un phénomène (exemple pauvreté des ménages) et les mettre ensemble afin de pouvoir vérifier les hypothèses.

PARTIE VII : LES ANALYSES EXPLICATIVES

7.1. La commande logit

La régression logistique, qui est une méthode multivariée (ou multidimensionnelle), nous semble appropriée pour rechercher les facteurs explicatifs. Elle est d'autant plus indiquée que les variables à expliquer (variables dépendantes) sont dichotomiques. En ce qui concerne les variables indépendantes, elles doivent être qualitatives ou catégorielles. Avant d'introduire ces dernières dans le modèle d'analyse, on doit dichotomiser toutes les variables y afférentes. Toutefois, la modalité de référence, en général celle qui a le plus grand effectif, est écartée de l'équation de la régression pour éviter les problèmes de multicollinéarité qui ne permettent pas de calculer les coefficients de régression.

Soient p la probabilité ou risque de subir l'événement, et 1-p la probabilité de ne pas subir l'événement. Le modèle de régression logistique permet de mettre l'équation $Z=\log[(p/(1-p)]$

- soit sous la forme linéaire $Z=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn$,
- soit sous la forme multiplicative $e^z=p/(1-p)$ ←→ $p=e^z/(1+e)$. Et $e^z=p/(1-p)$=odds ratio, c'est à dire le rapport de chances.

Pour hiérarchiser les variables, c'est à dire la contribution de chaque variable à l'explication, on procède au modèle pas-à-pas. Le modèle pas-à-pas s'inscrit dans le cadre des variables intermédiaires.

Exemple $Y=\beta 1X1+\varepsilon$

$Y=\beta 1X1+...+\beta 4X4+\varepsilon$

Si l'effet de X1 disparaît lorsqu'on introduit X4, c'est que l'effet de X1 sur y passe par X4. Il peut arriver que l'effet de X1 se renforce ou que l'effet de X4 diminue, en partie, la différence qui était observée était due à X1.

7.1.1. **Les effets bruts** de la région

```
. gen region=a1
. tab region,gen(region)
```

region	Freq.	Percent	Cum.
1	1,598	11.78	11.78
2	1,200	8.84	20.62
3	1,184	8.73	29.35
4	1,199	8.84	38.19
5	1,200	8.84	47.03
6	1,199	8.84	55.87
7	1,190	8.77	64.64
8	1,200	8.84	73.48
9	1,200	8.84	82.33
10	1,198	8.83	91.16
11	1,200	8.84	100.00
Total	13,568	100.00	

```
. gen pauvrete=pov
. tab pauvrete, gen(pauvrete)
```

pauvrete	Freq.	Percent	Cum.
0	7,752	57.13	57.13
1	5,816	42.87	100.00
Total	13,568	100.00	

```
. logit pauvrete1 region2 region3 region4 region5 region6 region7 region8 region9 region10 region11,or

Iteration 0:   log likelihood = [illegible]
Iteration 1:   log likelihood = [illegible]
Iteration 2:   log likelihood = [illegible]
Iteration 3:   log likelihood = [illegible]

Logistic regression                               Number of obs   =      13568
                                                  LR chi2(10)     =     653.97
                                                  Prob > chi2     =     0.0000
Log likelihood = -8939.0406                       Pseudo R2       =     0.0353

------------------------------------------------------------------------------
   pauvrete1 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     region2 |   .3493596   .0282727   -13.00   0.000     .2981174    .4094096
     region3 |   .3997861   .0324857   -11.28   0.000     .3409268    .4688072
     region4 |   .4668966   .0379537    -9.37   0.000      .398132    .5475379
     region5 |   .3887002   .0314664   -11.67   0.000      .331671    .4555352
     region6 |    .281367   .0228621   -15.61   0.000      .239944    .3299411
     region7 |   .5593133    .045926    -7.08   0.000     .4761698    .6569744
     region8 |   1.089664   .0954902     0.98   0.327     .9176979    1.293856
     region9 |   .4297623    .034847   -10.42   0.000     .3666141    .5037876
    region10 |   .6406497    .052981    -5.38   0.000     .5447876    .7533798
    region11 |   .2493089   .0203492   -17.02   0.000     .2124519      .29256
------------------------------------------------------------------------------
```

Le pouvoir prédictif du modèle comprenant la région de résidence est de 3,53 %. On remarque que les ménages de la region2 (Diourbel) Sud courent 3 fois plus de risque de subir la pauvreté que ceux de Dakar. Il en est de même des autres régions à l'exception de region8 (ST louis). Le risque est plus élevé à Ziguinchor où il est multiplié par quatre. A St Louis, on ne note pas de différences significatives.

7.1.2. **Les effets bruts du milieu de** résidence

```
. gen milieu=a6

. tab milieu,gen(milieu)

     milieu |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      8,576       63.21       63.21
          2 |      4,992       36.79      100.00
------------+-----------------------------------
      Total |     13,568      100.00
```

```
. logit  pauvrete2 milieu2,or

Iteration 0:   log likelihood = -9266.0257
Iteration 1:   log likelihood = -8907.0427
Iteration 2:   log likelihood = -8906.8336

Logistic regression                               Number of obs   =      13568
                                                  LR chi2(1)      =     718.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -8906.8336                       Pseudo R2       =     0.0388

------------------------------------------------------------------------------
   pauvrete2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     milieu2 |   2.637117   .0965182    26.49   0.000     2.454571     2.83324
------------------------------------------------------------------------------
```

La contribution de la variable milieu de résidence à l'explication de la pauvreté des ménages est de 4%. Le milieu de résidence discrimine les ménages en matière de pauvreté. Les ménages ruraux courent un risque de 164% supérieurs aux ménages urbains.

7.1.3. Les effets bruts du sexe

```
. gen sexe=menb2

. tab sexe , gen(sexe)
```

sexe	Freq.	Percent	Cum.
1	10,686	78.76	78.76
2	2,882	21.24	100.00
Total	13,568	100.00	

```
. logit  pauvrete2  sexe2,or

Iteration 0:   log likelihood = -9266.0257
Iteration 1:   log likelihood =  -9212.812
Iteration 2:   log likelihood =  -9212.763

Logistic regression                             Number of obs   =      13568
                                                LR chi2(1)      =     106.53
                                                Prob > chi2     =     0.0000
Log likelihood =  -9212.763                     Pseudo R2       =     0.0057
```

pauvrete2	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sexe2	.6402523	.0280077	-10.19	0.000	.5876457	.6975683

Les ménages dirigés par les hommes courent moins de risque que ceux dirigés par les femmes. En effet, ce risque est de 36% inférieurs.

7.1.4 Les effets bruts de l'âge du CM

```
. recode menb3 min/19 =1 20/35 =2 36/45 =3 46/59 =4 60/max=5 , gen(gpage)
(13568 differences between menb3 and gpage)

. lab def gpage 1"moins de 19ans" 2"20 à 35ans" 3"36 à 45ans" 4"46 à 59ans" 5"plus de 60ans"

. lab val gpage gpage

. tab gpage, gen(gpage)
```

RECODE of menb3 (b3.age)	Freq.	Percent	Cum.
moins de 19ans	33	0.24	0.24
20 à 35ans	2,151	15.85	16.10
36 à 45ans	3,232	23.82	39.92
46 à 59ans	4,388	32.34	72.26
plus de 60ans	3,764	27.74	100.00
Total	13,568	100.00	

```
. logit  pauvrete2   gpage1 gpage2 gpage3 gpage5,or

Iteration 0:   log likelihood = -9266.0257
Iteration 1:   log likelihood = -9138.4817
Iteration 2:   log likelihood = -9138.3116
Iteration 3:   log likelihood = -9138.3116

Logistic regression                               Number of obs   =      13568
                                                  LR chi2(4)      =     255.43
                                                  Prob > chi2     =     0.0000
Log likelihood = -9138.3116                       Pseudo R2       =     0.0138

------------------------------------------------------------------------------
   pauvrete2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gpage1 |   .7818273   .2795428    -0.69   0.491     .3879388    1.575645
      gpage2 |   .5512545   .0305766   -10.74   0.000     .4944681    .6145625
      gpage3 |   .7360845   .0347843    -6.48   0.000     .6709704    .8075175
      gpage5 |   1.236511   .0550528     4.77   0.000     1.133183     1.34926
------------------------------------------------------------------------------
```

Au niveau de l'âge, on constate que les rapports de chance entre les ménages dont le CM est âgé entre 20-35 ans et 36-45 sont respectivement 36% et 26% inférieurs. Par contre, dans les ménages dont le CM est âgé de 60ans ou plus, le risque est de 23% supérieurs que ceux avec un CM d'âges compris entre 46-59ans. Autrement dit, le risque qu'un ménage subisse la pauvreté augmente avec l'âge du CM.

7.1.5 Les effets bruts du type de formation du CM

```
. gen typeformation=educ2

. tab typeformation, gen(typeformation)

typeformati |
         on |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     10,587       78.03       78.03
          2 |      1,790       13.19       91.22
          3 |        347        2.56       93.78
          4 |        237        1.75       95.53
          5 |        287        2.12       97.64
          6 |        320        2.36      100.00
------------+-----------------------------------
      Total |     13,568      100.00
```

```
. logit pauvrete2 typeformation2 typeformation3 typeformation4 typeformation5 typeformation6 ur
Iteration 0:  log likelihood = -9266.0257
Iteration 1:  log likelihood = -8954.6506
Iteration 2:  log likelihood = -8938.0638
Iteration 3:  log likelihood = -8937.2907
Iteration 4:  log likelihood = -8937.2859

Logistic regression                               Number of obs     =      13568
                                                  LR chi2(5)        =     657.48
                                                  Prob > chi2       =     0.0000
Log likelihood = -8937.2859                       Pseudo R2         =     0.0355

------------------------------------------------------------------------------
   pauvrete2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
typeformat~2 |   .6286858   .0332631    -8.77   0.000     .5667579    .6973805
typeformat~3 |   .2559999   .0355735    -9.81   0.000     .1949655    .3361413
typeformat~4 |   .1023637   .0240033    -9.72   0.000     .0646469    .1620856
typeformat~5 |   .2119905   .0343578    -9.57   0.000      .154298    .2912544
typeformat~6 |    .081994   .0181838   -11.28   0.000     .0530898    .1266348
------------------------------------------------------------------------------
```

La contribution du type de formation du CM à l'explication de la pauvreté du ménage est de 3,55%. Le modèle est parfaitement adéquat aux données utilisées, la probabilité de khi² associée au modèle est de 0.0000 (inférieur à 1%). Comme attendu, l'avantage est en faveur de ceux qui ont reçu une formation. Il est plus visible chez les CM avec une formation après le bac quel que soit le type (tech ou professionnel). Le rapport de chances est presque à 100% quand on a eu une formation professionnel après le bac. Donc le risque pour ces ménages de subir la pauvreté est presque nul.

7.1.6. Les effets bruts de la taille du ménage

7.1.7. Les effets nets

Variables et modalités	Effets bruts	Odds ratio associés au modèle
		M0
Région de résidence	***	
Dakar	Ref	Ref
Diourbel	2,86***	2,20***
Fatick	2,50***	1,90***
Kaolack	2,14***	1,62***
Kolda	2,57***	1,97***
Louga	2,55***	2,77***
Matam	1,79***	1,30***
Saint louis	0,92ns	0,77***
Tamba	2,32***	1,77***
Thiès	1,56***	1,16*
ziguinchor	4,01***	3,14***
Milieu de résidence		***
Urbain	Ref	Ref
rural	2,64***	2,56***
Proba chi²	***	***
R²	**3,58**	**6,90**

Dans le modèle d'analyse de la régression logistique, le pouvoir prédictif est de 7%. La région de résidence est discriminante. Les risques ont diminué dans pratiquement toutes les régions. Par contre, l'influence de region8 se renforce et devient significative au seuil de 1%. Ce qui montre son influence sur la pauvreté était due au milieu de résidence. L'effet de la region10 sur la variable dépendante diminue, donc l'influence était due à la region10.

Variables et modalités	Effets bruts	Odds ratio associés au modèle	
		M0	M1
Région de résidence	***		
Dakar	Ref	Ref	Ref
Diourbel	2,86***	2,20***	2,19***
Fatick	2,50***	1,90***	1,90***
Kaolack	2,14***	1,62***	1,61***
Kolda	2,57***	1,97***	1,95***
Louga	2,55***	2,77***	2,74***
Matam	1,79***	1,30***	1,29***
Saint louis	0,92ns	0,77***	0,66***
Tamba	2,32***	1,77***	1,73***
Thiès	1,56***	1,16*	1,16*
ziguinchor	4,01***	3,14***	3,18***
Milieu de résidence		***	
Urbain	Ref	Ref	
rural	2,64***	2,56***	
Sexe du CM			***
Masculin	Ref		Ref
Féminin	0,64***		2,47***
Proba chi²	***	***	***
R²	**3,58**	**6,90**	**7,09**

L'introduction de la variable sexe du CM dans le modèle ne fait augmenter le pouvoir prédictif que de 0,19%. En outre, par rapport au modèle précédent, les effets des variables explicatives sur la dépendante restent les mêmes.

Variables et modalités	Effets bruts	Odds ratio associés au modèle		
		M0	M1	M2
Région de résidence	***			
Dakar	Ref	Ref	Ref	Ref
Diourbel	2,86***	2,20***	2,19***	2,16***
Fatick	2,50***	1,90***	1,90***	1,87***
Kaolack	2,14***	1,62***	1,61***	1,68***
Kolda	2,57***	1,97***	1,95***	2,05***
Louga	2,55***	2,77***	2,74***	2,70***
Matam	1,79***	1,30***	1,29***	1,36***
Saint louis	0,92ns	0,77***	0,66***	0,66***
Tamba	2,32***	1,77***	1,73***	1,86***
Thiès	1,56***	1,16*	1,16*	1,11ns
Ziguinchor	4,01***	3,14***	3,18***	3,18**
Milieu de résidence		***		
Urbain	Ref	Ref		Ref
rural	2,64***	2,56***		
Sexe du CM			***	
Masculin	Ref		Ref	Ref
Féminin	0,64***		2,47***	0,74***
Age du CM				****
Au plus 20 ans	0,93ns			0,79ns
21-35 ans	0,54***			0,49***
36-45 ans	0,73***			0,70***
46-59 ans	Ref			Ref
60 ans+	1,23***			1,18***
Proba chi²	***	***	***	***
R²	**3,58**	**6,90**	**7,09**	**8,56**

Variables et modalités	Effets bruts	Odds [illegible]	M1	M2	M3	M4
Région de résidence	***					
Dakar	Ref	Ref	Ref	Ref	Ref	Ref
Diourbel	2,86***	[illegible]	[illegible]	[illegible]	1,80***	1,63***
Fatick	2,50***				1,66***	1,73***
Kaolack	2,14***	1,6_ ***	1,6_***	1,68***	1,52***	1,40***
Kolda	2,57***	1,97***	1,95***	2,05***	1,82***	1,70***
Louga	2,55***	2,77***	2,74***	2,70***	2,25***	2,46***
Matam	1,79***	1,30***	1,29***	1,36***	1,16*	1,08ns
Saint louis	0,92ns	0,77***	0,66***	0,66***	0,56***	0,52***
Tamba	2,32***	1,77***	1,73***	1,86***	1,61***	1,73***
Thiès	1,56***	1,16*	1,16*	1,11ns	1,02ns	0,93ns
Ziguinchor	4,01***	3,14***	3,18***	3,18**	3,13***	3,85***
Milieu de résidence		***				
Urbain	Ref	Ref	Ref	Ref	Ref	Ref
rural	2,64***	2,56***	2,52***	2,47***	2,15***	2,03***
Sexe du CM			***			
Masculin	Ref		Ref	Ref	Ref	Ref
Féminin	0,64***		2,47***	0,74***	0,64***	0,74***
Age du CM				****		
Au plus 20 ans	0,93ns			0,79ns	0,71ns	1,33ns
21-35 ans	0,54***			0,49***	0,48***	0,71***
36-45 ans	0,73***			0,70***	0,69***	0,80***
46-59 ans	Ref			Ref	Ref	Ref
60 ans+	1,23***			1,18***	1,09	1,01ns
Type de formation	***				***	
Aucune	Ref				Ref	Ref
Sur le tas	0,63***				0,75***	0,73***
Technique avant bac	0,26***				0,28***	0,26***
Technique après bac	0,10***				0,13***	0,14***
Professionnel avant bac	0,21***				0,25***	0,22***
Professionnel après bac	0708***				0,11***	0,11***
Taille du ménage	***					***
1-3pers	0,29***					0,32***
4-6 pers	0,52***					0,51***
7-10 pers	Ref					Ref
+ 10 pers	1,80***					1,81***
Proba chi²		***	***	***	***	***
R²		6,90	7,09	8,56	10,83	15,28

7.2. Les moindres carrés ordinaires (MCO)

La méthode la plus utilisée dans le cadre de l'économétrie linéaire est les MCO. Sa mise en œuvre consiste à utiliser la commande ***regress*** ou ***reg*** suivie de la variable dépendante, des variables indépendantes et éventuellement des options.

Syntaxe : ***regress*** *var_dep var_indep…,* ***[options]***

Une constante est insérée par défaut et il faut utiliser l'option *noconstant* pour faire une régression sans constante si c'est nécessaire. Avec les commandes ***describe*** et ***codebook*** on a les informations sur les variables. La première étape consiste à représenter graphiquement les variables de la régression à laide de la syntaxe : ***graph matrix*** *var_dep var_indep …* La commande ***predict*** s'applique à la régression la plus récente et permet d'obtenir sur la base des coefficients estimés, entre autres, la valeur prédite de la variable dépendante ainsi que les résidus de la régression.La syntaxe générale est : ***predict*** *variable* ***(if, in), option***

Notons que Stata offre la possibilité d'inclure automatiquement dans la régression les variables muettes correspondant à chaque modalité d'une variable qualitative spécifiée. Dans ce cas la commande ***xi*** suivi de deux points (:) précède la régression. On peut également estimer le modèle avec contraintes. Dans ce cas il faut définir la contrainte et exécuter ensuite la commande ***cnsreg*** en lieu et place de ***reg***.

EXEMPLE DE SIMULATION

Loi normale

- set obs 100 : on fixe la taille de l'échantillon à 100

gen x= 3 + sqrt(4)*invnorm(uniform()) : générer 100 observations d'une variable aléatoire x de loi normale de moyenne 3 et de variance 4, i.e. x ≃N(1, 4).

- set obs 1000 : on fixe la taille de l'échantillon à 1000

gen y = 0 + 1 * invnorm(uniform()) : générer1000 observation d'une variable aléatoire y de loi normale centrée réduite, i.e y ≃N(0, 1) Loi normale.

- set obs 100 : on fixe la taille de l'échantillon à 100

genx= 3 + sqrt(4)*invnorm(uniform()) : générer 100 observations d'une variable aléatoire x de loi normale de moyenne 3 et de variance 4, i.e. x ≃N(1, 4).

- set obs1000 : on fixe la taille de l'échantillon à 1000

gen y = 0 + 1 * invnorm(uniform()) : générer1000 observation d'une variable aléatoire y de loi normale centrée réduite, i.e y ≃N(0, 1)

Loi uniforme

set obs 300

gen z = uniform(): crée300 observations d'une variable z de loi uniforme sur[0,1], i.e z≃ U[0,1]

Pour les simulations des autres lois, consulter le help de Stata.

BIBLIOGRAPHIE

B. CHITOU, 2006 "Modèles logistiques appliqués"; les cahiers de l'ENSEA

J. SCOTT LONG & J. FREESE, 2006 "Regression Models for Categorical Dependent Variables Using Stata"; Second edition, Stata press

K. KPODAR, janv. 2005 « Manuel d'initiation à Stata (version 8) ; CERDI

N. COUDERC, « Econométrie appliquée avec Stata », université Paris Pantheon-sorbone1

N° d'ordre	Prénoms et Noms	Structure/provenance	E mail	Téléphone	Emargement
07	Mme DIOP NDeye Farma MBODJ	DCEF / MEF	farmas@hotmail.fr	776551385	
08	Papa [illegible] Seck	DCEF / MEF	seck[illegible]@yahoo.fr	77 5447803	
09	Mme BA Aïssatou Fall	DCEF / MEF	aichafal@yahoo.fr	776325266	
10	Ibrahima DIENG	UCSPE/MEF	ibdieng@yahoo.fr	776457794	
11	Sanor DIEYE	UCSPE / MEF	sanor71@yahoo.fr	779711440	
12	Ndeye Naye Diouf	DCEF / MEF	mayediouf@hotmail.com	776540287	
13	Mamadou Moustapha BA	DCEF/MEF	bosquieron@yahoo.fr	776562183	
14	Camara Mayacine	UCSPE/MEF	camaramayacine@[illegible]Y.fr		
15	Fanta SAKHO SECK	DCEF / MEF	fanfaseck@yahoo.fr	775780446	
16	Issekha Seck	DCEF / MEF	issekhaseck@yahoo.fr	775321628	
17	Khalifa [illegible] Fall	DCEF / MEF	[illegible]	[illegible]	
18	Abdoulaye DIENG	DCEF / MEF	layemadior@yahoo.fr	776629784	
19	Cheikh Dioum	UCSPE/MEF	[illegible]@ucspe.sn	77 6402687	
20	Fatou Ndiaye	UCSPE-MEF	fatou[illegible]y.fr	772770423	

N° d'ordre	Prénoms et Noms	Structure/provenance	E mail	Téléphone	Emargement
21	Aissatou SARR	clutou23@yahoo.fr	UCSPE Nat	766667555	Aissa
22	Mahi Amadou DEME	mademe@ucspe.com	UCSPE/MEF	776787471	
23	Gora BEYE	UCSPE/MEF	beyegora2005@gmail.com	77 635 74 64	
24	[illegible]	UCSPE/MEF	[illegible]	776479269	
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					