

TRAITEMENT DE DONNEES

APPLICATION SOUS LE LOGICIEL STATA v12

MOUSSA K. RICHARD, PhD
Maître Assistant, ENSEA

Table de matière

Table de matière	
I- Présentation du logiciel Stata.....	1
1- Interface utilisateur de Stata	1
2- Types de fichiers	2
a- Fichier do	2
b- Fichier log	5
II- Travaux préparatoires sur les données	6
1- Importation de données.....	6
a- Fichier de type dta	6
b- Autres fichiers.....	7
c- Saisie directe.....	9
2- Fusion	11
a- Ajout d'individus.....	11
b- Ajout de variables.....	13
III- Travaux préliminaires	20
1- Organisation de la base de données	20
a- Ordonner la base	20
b- Visualiser la base	23
c- Suppression	24
2- Etiquettes et noms	25
a- Renommer.....	25
b- Etiquette de variables et de modalités	27
3- Transformation de données	29
a- Création de variables.....	29
b- Changement de format de variable	31
c- Transformation des observations	32
IV- Contrôles et apurement	33
1- Traitement de doublons.....	33
a- Unicité de l'identifiant.....	33
b- Recherche de doublons suivant des critères.....	34
2- Traitement de données manquantes.....	35
a- Recherche de données manquantes.....	35
b- Imputation des données manquantes	36

3-	Contrôle de cohérence	37
a-	Sauts de variables	37
b-	Cohérence des données	39
V-	Statistique descriptive	39
1-	Tableaux à une ou plusieurs entrées.....	39
a-	Tableaux de fréquences	39
b-	Tableaux de statistiques.....	41
2-	Graphiques	43
a-	Histogrammes	43
b-	Courbes.....	44
c-	Graphiques à bandes.....	45
d-	Diagrammes circulaires	46
e-	Boîtes à moustaches.....	47
3-	Analyse de liaisons entre variables	49
a-	Test de khi 2	49
b-	Test de corrélation	49
c-	Analyse de la variance (Anova).....	51
d-	Test d'égalité de moyennes.....	52
e-	Test d'égalité de proportions	54
	Conclusion	55

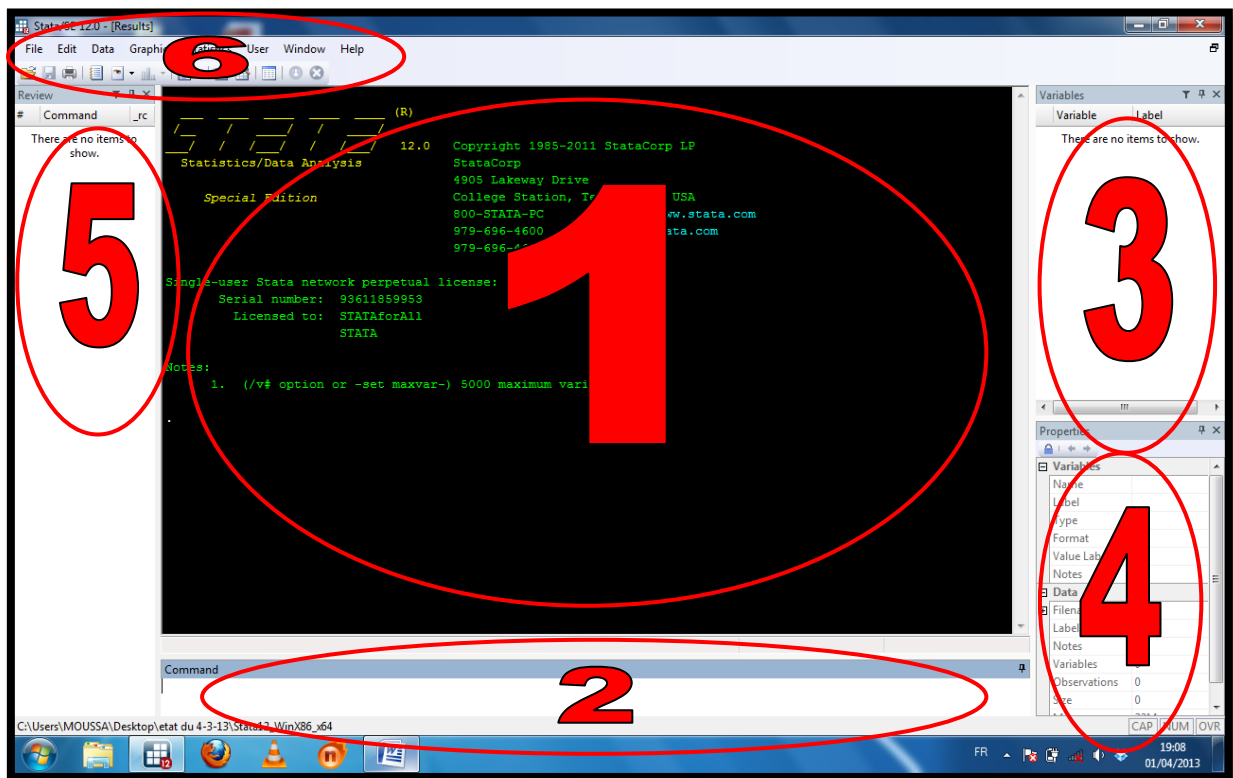
I- Présentation du logiciel Stata

Stata (Statistics data Analysis) est un puissant logiciel de traitement et d'analyse de données statistiques. Il est utilisé pour effectuer des contrôles de cohérence sur les données, d'effectuer des corrections sur les données incohérentes, de produire des statistiques sous forme de tableaux ou graphiques. Stata n'est certes pas dédié à la saisie de données, mais il offre un cadre pour réaliser la saisie au besoin.

Stata est développé par StataCorp et était à sa version 12 en 2012, laquelle version sera utilisée pour ce manuel de traitement de données.

1- Interface utilisateur de Stata

L'interface utilisateur, l'écran qui s'affiche au lancement de Stata 12, comprend 6 fenêtres, chacune dédiée à un usage spécifique.



Partie 1 : c'est la fenêtre d'affichage des résultats des travaux effectués sous Stata. Le fond noir de cette fenêtre peut être modifié selon la convenance de l'utilisateur. Il suffit pour cela de faire un clic droit dans la fenêtre puis choisir « preferences » et changer le « color scheme ». Cette fenêtre est totalement effacée lorsque l'on ferme Stata.

Partie 2 : c'est la fenêtre des commandes. Elle sert à saisir les instructions de traitement ou d'analyse de données qu'on demande à Stata d'exécuter.

Partie 3 : c'est la fenêtre qui affiche les variables de la base de données ainsi que leur label respectif (le chargement d'une base de données sera présentée dans la section II-1 de ce manuel).

Partie 4 : c'est la fenêtre qui affiche dans sa partie inférieure les informations sur la base de données (nom et le chemin d'accès, label, notes, nombre de variables, nombre d'observations, taille de la base) et la mémoire Stata disponible pour les données, et dans sa partie supérieure, les informations sur chacune des variables que l'on aura sélectionné (nom, label, type, format, étiquette des modalités et les notes).

Partie 5 : c'est une fenêtre d'historique. Elle enregistre toutes les commandes exécutées depuis l'ouverture de Stata et permet de les exécuter à nouveau. Cet historique est automatiquement supprimé quand on ferme le logiciel.

Partie 6 : c'est le menu de Stata. Il permet à l'utilisateur de faire un traitement ou une analyse interactive. Il faut à ce niveau souligner que toutes les commandes de Stata ne sont pas sous forme interactive (donc dans le menu).

2- Types de fichiers

Deux types de fichiers de Stata permettent de faciliter énormément les travaux sous Stata notamment en ce qui concerne l'enregistrement des commandes et des résultats des travaux.

a- Fichier do

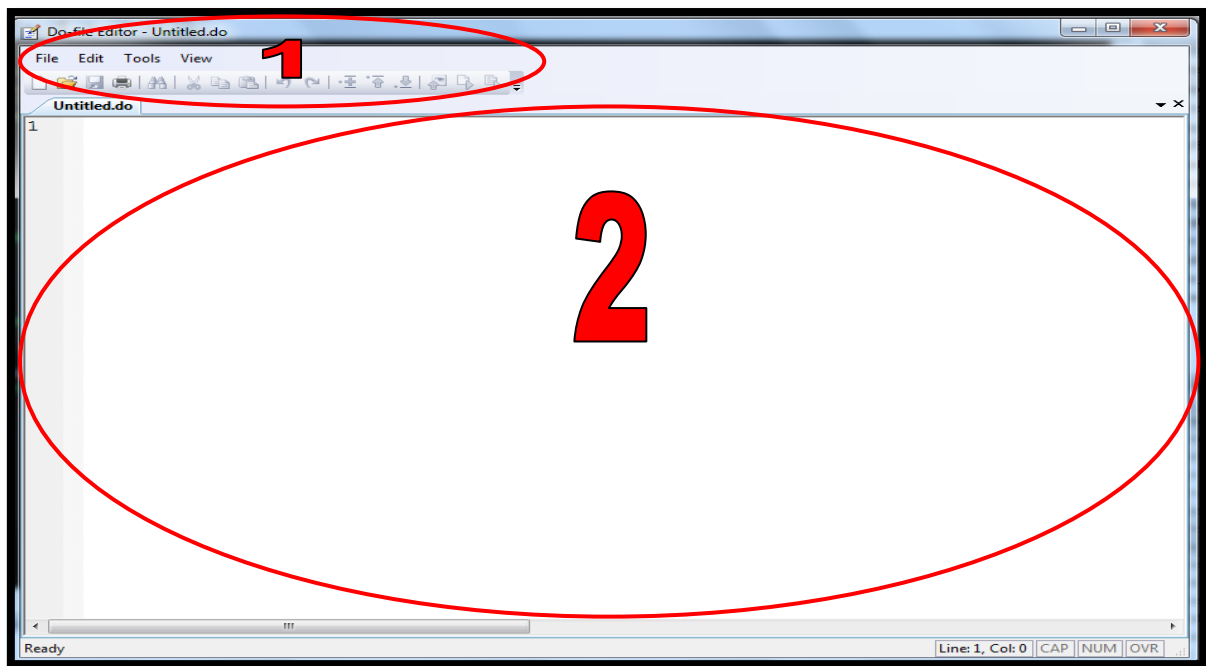
Le fichier do a plusieurs rôles. Mais le rôle principal est l'enregistrement des commandes. Un fichier do est un éditeur de commandes qui permet à l'utilisateur, non seulement d'enregistrer les commandes à exécuter mais également de les exécuter sur place sans passer par la fenêtre de commande de Stata. Ainsi en constituant un fichier do, l'on pérennise les commandes exécutées. Ce qui permet de les exécuter à nouveau à chaque fois que l'on souhaite revoir les résultats des travaux effectués. Pour avoir accès à l'éditeur do, il suffit de cliquer sur l'icône do file dans le menu de Stata.



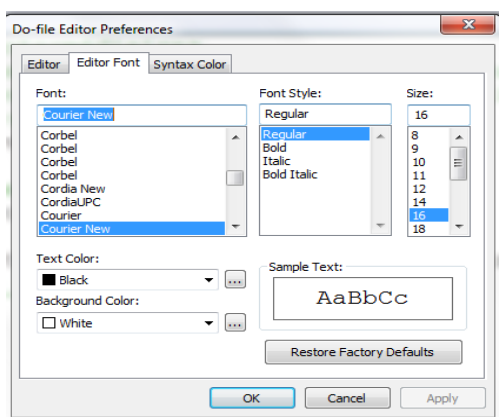
Une fois ouvert, l'on peut écrire les syntaxes ou commandes à l'intérieur et les exécuter quand on le souhaite.

Le fichier do devra par la suite être enregistré pour une réutilisation éventuelle ou une transmission à une tierce personne. Il permet ainsi de palier la difficulté liée à la fenêtre d'historique des commandes qui elle se vide dès la fermeture de Stata.


L'éditeur de fichier do se présente comme suit :



La partie 1 est le menu de l'éditeur et la partie 2, le corps de l'éditeur. Dans la partie 1, il est possible d'utiliser le bouton « edit » pour choisir la mise en forme du texte écrit dans l'éditeur. Pour cela, il faut cliquer sur « edit » puis « preferences » et on obtient le menu suivant qui permet de faire la mise en forme. Notons que dans « syntax color », on peut également choisir les couleurs des polices relatives aux commentaires, commandes, string, macro...



C'est dans la seconde partie que l'utilisateur devra saisir toutes les séquences de commandes à exécuter. L'exécution d'un fichier do peut se faire en sélectionnant une ligne ou plusieurs

lignes à exécuter puis en faisant « CTRL »+ « D » ou encore en cliquant sur le bouton  de l'éditeur do. Aussi lorsqu'aucune sélection n'est effectuée, cliquer sur le bouton « execute » ci-dessus indiqué ou faire « CTRL »+ « D » revient à exécuter toutes les lignes de commandes du fichier do.

Afin de rendre plus compréhensible le fichier do, l'éditeur offre à l'utilisateur la possibilité de commenter le fichier en cours de création. Ainsi toute ligne que l'utilisateur précèdera d'un

astérisque (*) sera considérée comme un commentaire et Stata la mettra en vert. Aussi lorsque le commentaire peut tenir sur plusieurs lignes, l'utilisateur a possibilité de commencer par un slash suivi d'un astérisque (/*) et de terminer par un astérisque et un slash (*/). Cette dernière approche pour les commentaires permet également d'insérer des commentaires à tout endroit dans une ligne de commande ou dans l'éditeur.

Aussi dans l'écriture des commandes, on peut utiliser un double slash pour aller à la ligne dans une même commande. Cette technique permet d'écrire sur plusieurs lignes une même commande.

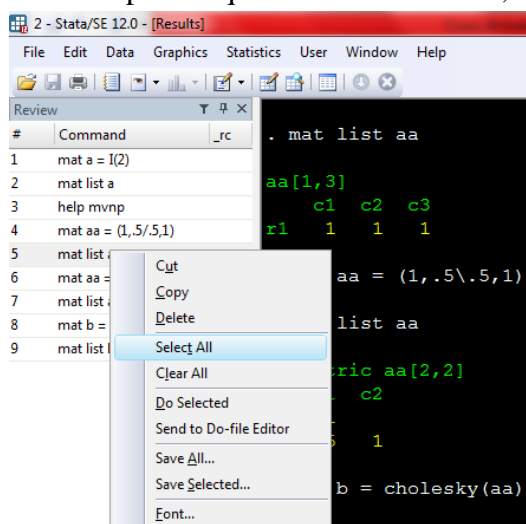
```

1  /* variable couts des intermédiaires
2  1- identifier combien d'acteurs attribuent des marchés au chauffeur
3  2- calculer le ratio coxeur sur total et utiliser ce ratio pour
4  calculer le coût des intermédiaires */
5  gen interm_new = 49500 * voyag
6
7  ***calcul du cout variable par voyage
8  gen cout_vvm_new = gc48*voyag
9
10 replace marge_new = recet_new - ( cout_vvm_new + salaire_corr //
11 + frais_gen_corr /*+ cout_expl*/ + interm_new )
12
13

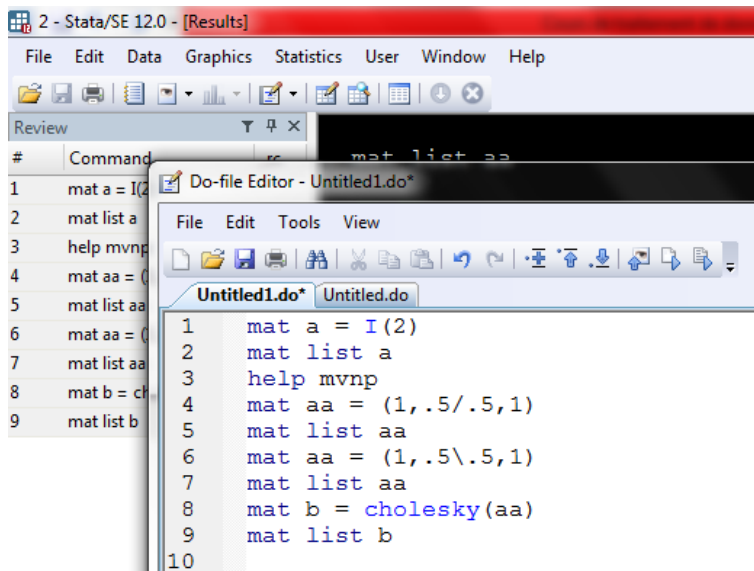
```

Aussi, faudrait-il noter que l'utilisateur peut directement exécuter les commandes dans la fenêtre des commandes sans se soucier dans un premier temps de les enregistrer dans un fichier do. Ces commandes ainsi exécutés s'enregistrent automatiquement dans la fenêtre d'historique des commandes. Une fois que le travail est terminé, on peut procéder à l'enregistrement des commandes dans un fichier do en procédant comme suit :

- Aller dans l'historique des commandes et faire un clic droit sur l'une des commandes exécuter puis cliquer sur « select all » ;

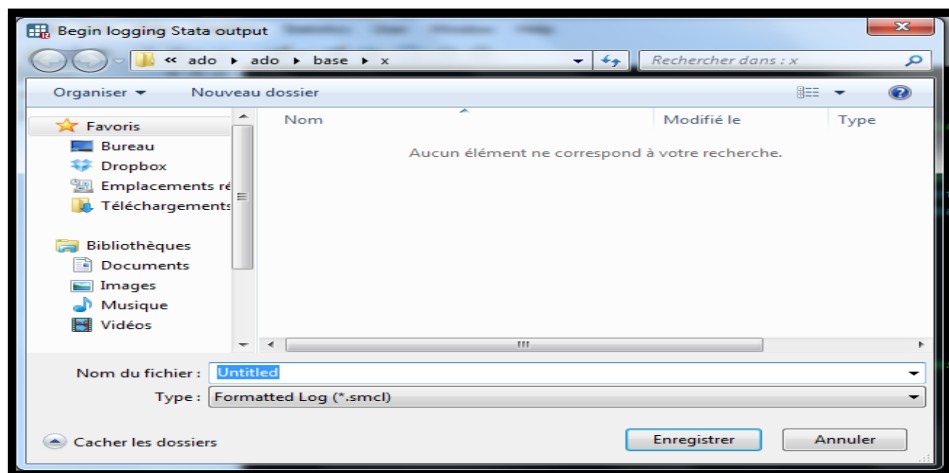


- Après avoir cliqué sur « select all », vous faites à nouveau clic droit dans l'historique des commandes sur l'une des commandes, puis vous faites « send to do-file editor » et on obtient un fichier do qu'il suffira par la suite d'enregistrer.



b- Fichier log

Le fichier log permet à l'utilisateur de Stata de garder une trace de tous les résultats issus des analyses effectuées. Il permet ainsi de palier le problème de la fenêtre résultat qui s'efface à la fermeture de Stata. Pour ouvrir un fichier log, il faut aller dans le menu de Stata et cliquer sur « file », puis cliquer sur « log » et sur « begin ». La fenêtre suivante s'affiche :



A ce niveau, il faut donner un nom au fichier et spécifier le chemin d'accès, le dossier dans lequel l'on souhaite enregistrer le fichier log puis on mettra ok. Dès cet instant, tous les résultats issus des commandes exécutées sont consignés dans le fichier do. A la fin des travaux, il faut cliquer sur « file », puis cliquer sur « log » et sur « end ».

Il est important de noter qu'on peut suspendre l'enregistrement des résultats dans le fichier log et le reprendre par la suite au cas où l'on souhaite ne pas voir certains résultats enregistrés.

Pour cela, il faut cliquer sur « file », puis cliquer sur « log » et sur « suspend » pour suspendre et « resume » pour reprendre l'enregistrement.

La syntaxe est la suivante :

log using "chemin_accès\nom_fichier.log" ou si le fichier existe déjà et que l'on souhaite le remplacer **log using "chemin_accès\nom_fichier.log",replace**

II- Travaux préparatoires sur les données

Les travaux préparatoires concernent l'importation des données sous Stata, l'organisation de ces données et les transformations qu'on souhaite faire en vue de préparer la base au traitement et analyse qu'on souhaite faire.

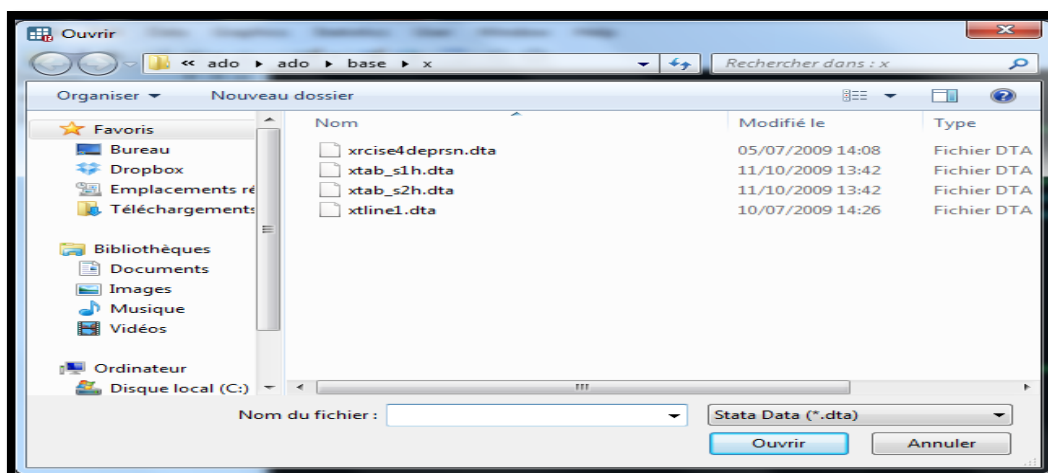
1- Importation de données

Les données doivent être chargées dans le logiciel avant tout traitement. Ce chargement de données se fait de plusieurs manières suivant le type de fichiers de données et l'organisation préalable des données. On distinguera ici deux cas : le cas des fichiers formats Stata et le cas des autres types de fichiers.

a- Fichier de type dta

Les fichiers de type Stata ont pour extension « .dta ». Ils sont au préalable obtenus soit suite à une exportation de données issues du logiciel de saisie Cspiro, soit suite à une conversion de fichiers de données à partir d'un logiciel de transfert de données statistiques (tel que Stat Transfert) ou d'un logiciel le permettant (tel que SPSS), soit suite à un travail précédent effectué sous Stata et enregistré sous format dta.

L'ouverture de fichier dta sous Stata est aisée. Il faut aller dans le menu de Stata et cliquer sur « file » et « open ». On obtient la fenêtre suivante :



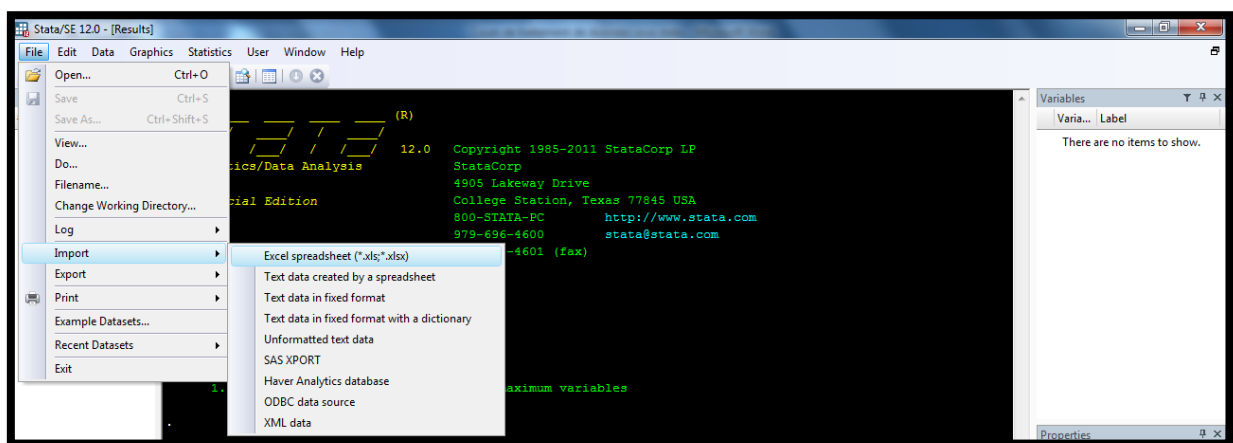
C'est à ce niveau qu'on va choisir le fichier et cliquer sur ouvrir.

La syntaxe est la suivante : lorsqu'aucune base de données n'est ouverte et que l'on souhaite ouvrir une base use "**chemin_d'accès\nom_fichier.dta**" ou lorsqu'une base est déjà ouverte que l'on souhaite en ouvrir une autre use "**chemin_d'accès\nom_fichier.dta**", **clear** ou encore use "**chemin_d'accès\nom_fichier.dta**", **replace**

b- Autres fichiers

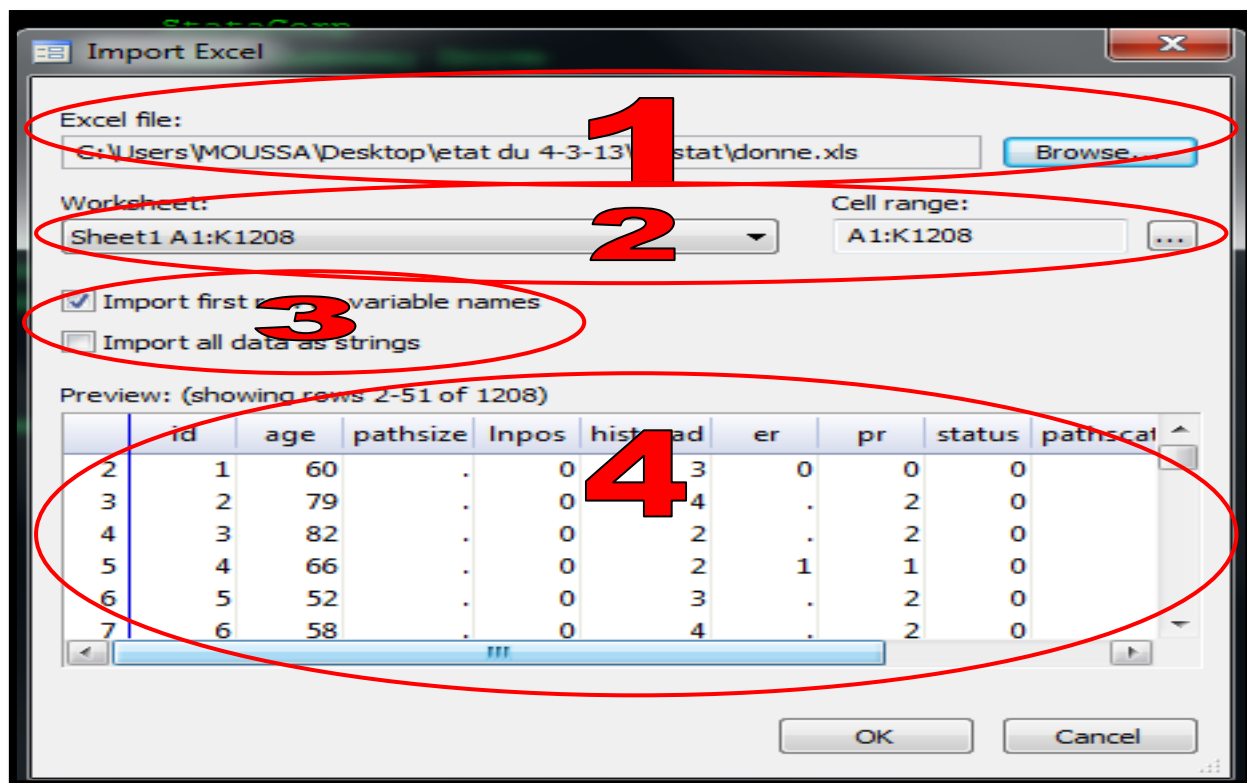
Les autres types de fichier de données accessibles sous Stata sont les fichiers Excel, text (txt) simple ou avec dictionnaire de variable (issu d'un logiciel de saisie tel que Cspiro), SAS export et Access. Pour faire l'importation de ces types de données, il suffit d'aller dans le menu de Stata et de cliquer sur « file », ensuite sur « import » et choisir le type de fichier que l'on souhaite importer.

Dans le cas spécifique des fichiers Excel, on choisit dans le menu décrit ci-dessus, le type « Excel spreadsheet » comme suit :



Mais il faut au préalable se rassurer que dans le fichier Excel que l'on souhaite utiliser, l'on a rangé en colonne les variables et en ligne les individus sur lesquels ont été collectés ces informations.

Après donc la manœuvre ci-dessus, on obtient l'écran ci-dessous :



La zone 1 permet, en cliquant sur le bouton « browse » de sélectionner le fichier Excel qu'on souhaite ouvrir.

La zone 2 permet de spécifier la feuille du fichier Excel qu'on souhaite ouvrir ainsi que la plage de données à utiliser. Par défaut, Stata utilise la feuille qui était active quand on enregistrait pour la dernière fois le fichier Excel et sélectionne toutes les données de cette feuille Excel. Mais l'utilisateur peut modifier cette sélection selon ses intérêts.

La zone 3 est la zone des options. Elle permet de dire ou non à Stata de prendre la première ligne de la feuille ouverte comme la ligne qui contient les noms de variables. Il suffit pour cela de cocher « import first row as variable names » pour que Stata considère cette première ligne comme la ligne des noms de variables, ce qui n'est pas le cas par défaut. La seconde option concerne le type de données. Lorsque l'on coche « import all data as strings », Stata considère les données de la feuille comme des données non numériques (alphanumérique ou strings). Par défaut, Stata attribut un type à chaque variable en fonction de l'analyse qu'il en fait.

La zone 4 est la zone de visualisation des données de la feuille.

Une fois que tout est bien spécifié, l'on peut valider en cliquant sur ok, et vous avez les données importées dans Stata.

La syntaxe est la suivante :

import excel "chemin_accès\nom_fichier.xlsx", sheet("nom_feuille") firstrow

c- Saisie directe

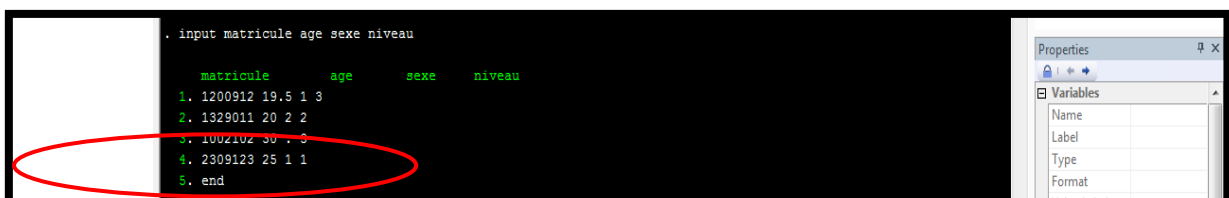
La saisie directe sous Stata est faite sans contrôle et peut s'avérer fastidieuse et très risquée en termes d'erreurs. En effet pour faire une saisie directe sous Stata, il suffit d'écrire « input » dans la barre de commande suivi des noms des variables de la base. On a l'écran ci-dessous :



Il est à noter que Stata interprète chaque mot comme une variable, ainsi il vous est impossible de mettre des espaces dans les noms de variables.

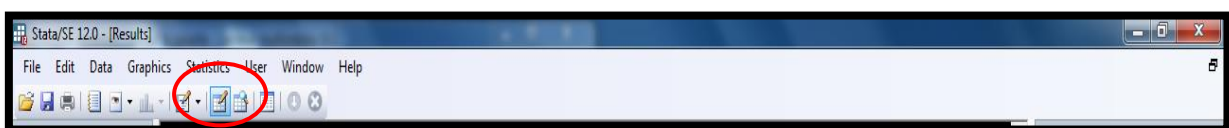
Une fois l'écriture des noms des variables de la base terminée, il faut faire « enter » pour valider. Une fois cette étape validée, l'on peut commencer la saisie des informations individu par individu. Lorsqu'une valeur d'une variable est manquante pour un individu, on met « . » en la place. Il faut toujours séparer les informations par un espace.

Une fois qu'on a terminé de saisir toutes les informations, il faut saisir « end » dans la barre des commandes et faire entrer, ainsi Stata ferme la fenêtre de saisie. Tant que cela n'est pas fait, Stata considère toutes les commandes que vous saisissez comme des tentatives pour rentrer des données.

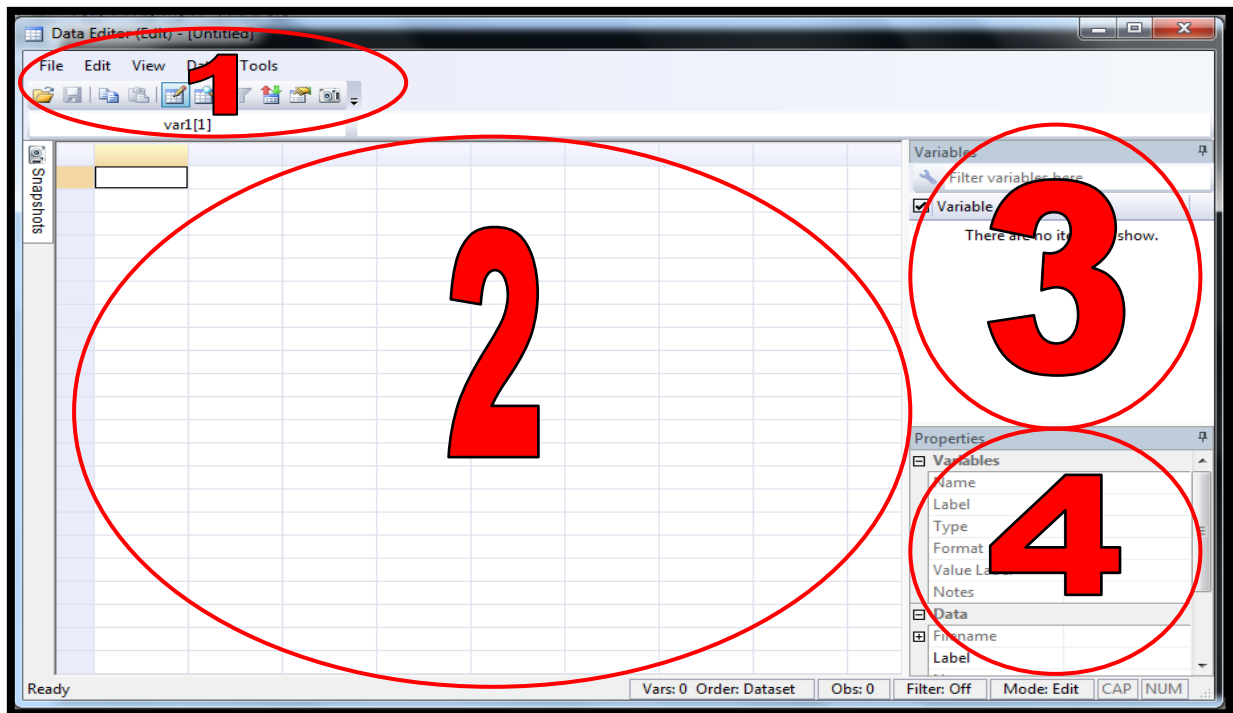


Si l'on souhaite ajouter encore des observations, il suffit de saisir « input » dans la barre de commandes sans ajouter les noms de variables et valider. Puis l'on pourra saisir toutes les informations voulues et terminer par « end ».

Une alternative à cette saisie directe à partir de la barre de commandes est d'ouvrir l'éditeur de données de Stata et y saisir directement les informations. Cette technique permet également de faire des modifications dans une base de données en cours d'utilisation. Pour y arriver, il vous suffit, dans le menu de Stata, de cliquer sur « data », puis sur « data editor » et sur « data editor (edit) » ou encore de cliquer sur l'icône du « data editor » comme sur l'image ci-dessous :



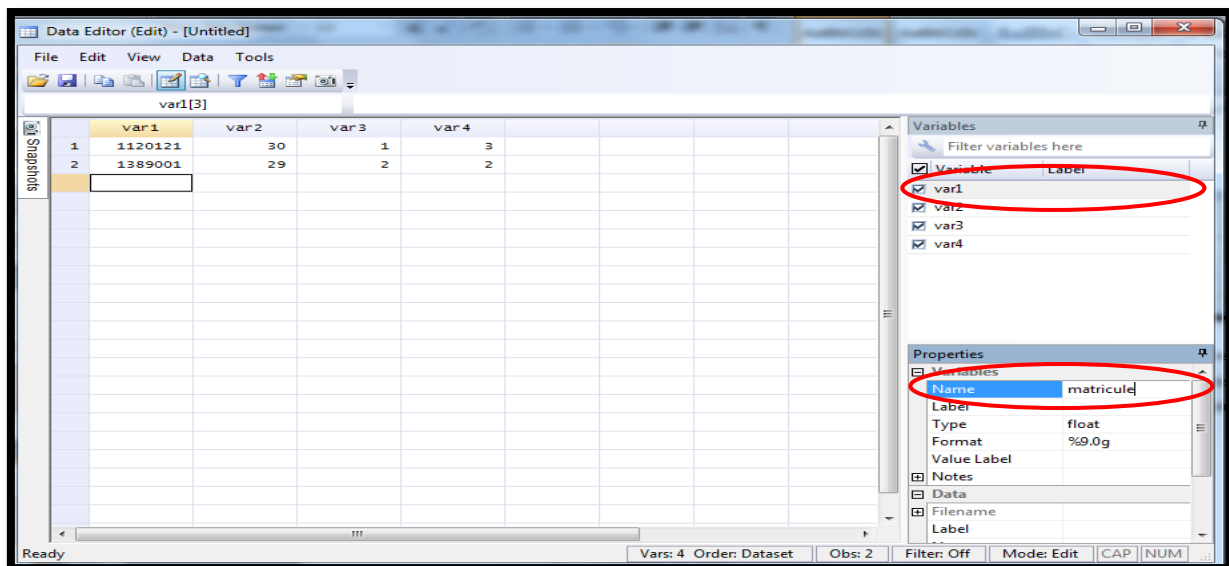
L'on voit alors s'ouvrir la fenêtre ci après :



La zone 1 est le menu du « data editor ». La zone 2 est la base de données, il s'agit d'un tableur. La zone 3 donne toutes les variables de la base de données et la zone 4, pour chaque variable de la base, donne les différentes propriétés.

Ainsi, une fois à cette étape, l'on peut commencer à saisir les données directement dans le tableur suivant un ordre qu'on choisira (soit individu par individu, soit variable par variable). Il faut saisir sans se préoccuper des noms des variables que Stata génère automatiquement (var1, var2, ..., varN).

Une fois la saisie des données individuelles terminées, l'on peut alors modifier les noms et les propriétés des variables. Il suffit pour cela de cliquer sur l'une des variables dans la fenêtre des variables et d'en modifier les propriétés dans la fenêtre dédiée. L'on pourra ainsi modifier le nom par défaut de la variable, ajouter un label à la variable, modifier le type et le format par défaut ainsi qu'ajouter une note d'explication à la variable. Ainsi l'on pourra avoir l'écran ci-dessous dans lequel nous ne modifions que le nom par défaut de la variable :



Remarque : *Stata permet certes la saisie directe mais cette opération doit être rare car ce n'est pas cela son objet.*

2- Fusion

La fusion est une opération qui permet soit d'ajouter des observations à une base de données à partir d'une autre base de données qui contient les mêmes variables que celle en cours d'utilisation, soit d'ajouter des variables à une base de données à partir d'une base de données portant sur les mêmes individus ou des individus liés à ceux-ci.

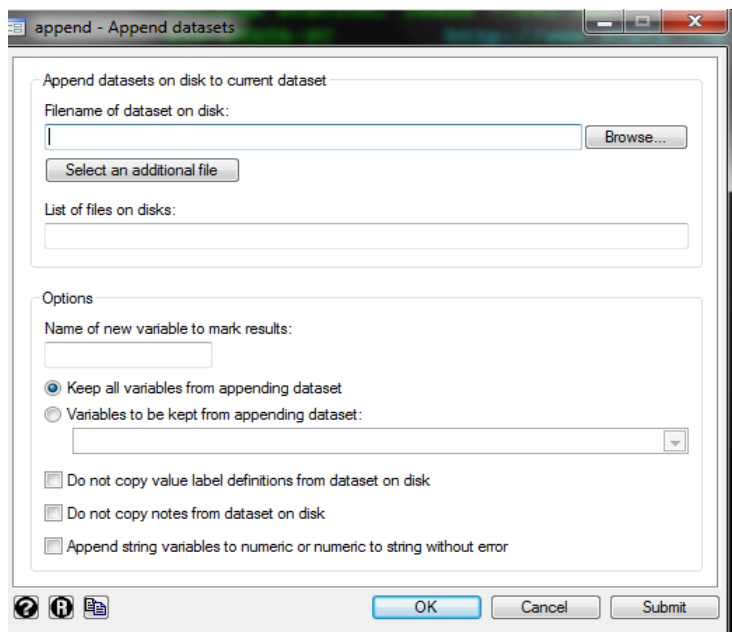
Pour ajouter des variables à une base de données à partir d'une base de données existante, il faut disposer d'une **clé de liaison (un identifiant)** portant le même nom dans les deux bases de données. La clé de liaison permet de retrouver un individu dans les bases afin de procéder à l'association des informations des deux bases.

Par contre, ajouter des individus exige que les variables des deux bases de données soient les mêmes. Pour toute variable qui n'existe pas dans l'une ou l'autre des deux bases, l'ajout engendrera des valeurs manquantes pour les individus de la base où la variable n'existe pas.

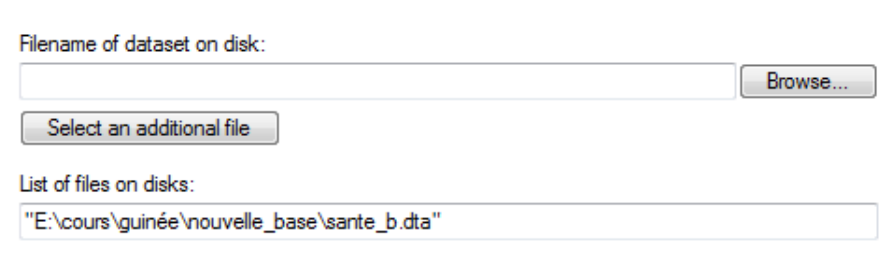
a- Ajout d'individus

Dans l'onglet « Data » du menu, on clique sur « combine datasets » puis « append datasets » et on obtient le menu suivant.

Le premier cadre en haut de ce menu permet de sélectionner les fichiers de données que l'on souhaite ajouter. En cliquant sur « browse », Stata vous ouvre une boîte de dialogue pour le choix du fichier à ajouter. Une fois le premier fichier sélectionné, vous pouvez l'ajouter à la liste des fichiers à fusionner puis sélectionner un autre fichier.

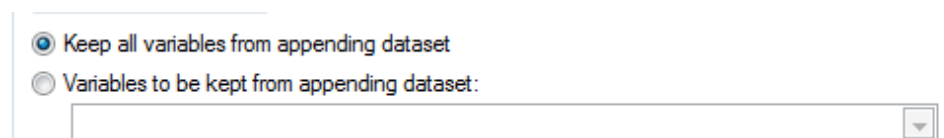


Ainsi dans le text box « list of files on disks », on obtient désormais ceci :



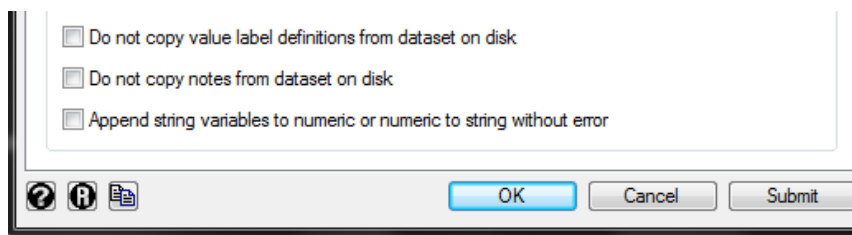
On peut répéter ce processus jusqu'à ce qu'on finisse la sélection des fichiers.

Une fois le processus de sélection des fichiers terminé, on peut renseigner les options de la fusion. La première concerne la création ou non d'une variable contenant une information sur l'origine des différents individus dans la base fusionnée (de quelle base est l'individu). Cela se fait en renseignant le champ « Name of new variable to mark result ». Ensuite il faut choisir les variables à garder dans la nouvelle base fusionnée. A ce niveau, deux options s'offrent à l'utilisateur : sélectionner par défaut toutes les variables ou sélectionner une liste de variables à garder. Cela se fait en cochant l'une des cases comme ci-dessous.



Une fois cette sélection effectuée, il faut maintenant spécifier si Stata copie les labels des bases à fusionner ou pas, si les notes sur la base et les variables de la base devront être maintenues après la fusion. Par défaut Stata conserve les labels et notes des bases à fusionner. Aussi faudrait-il noter que si une variable change de format (de numérique à alphanumérique et vice versa) dans les différents fichiers à fusionner, Stata renvoie un message d'erreur qui empêche la fusion. Ainsi, si l'on souhaite éviter cette erreur, on peut demander à Stata de faire

automatiquement la conversion de la variable concernée en le format le plus approprié des deux (texte alphanumérique ou valeur numérique). Ces différentes opérations sont effectuées en cochant les options dans le cadrant ci-dessous.



Une fois les différentes spécifications effectuées, on peut soit cliquer sur « OK » pour exécuter et fermer la boîte de dialogue, soit cliquer sur « Submit » pour exécuter et laisser la boîte de dialogue toujours ouverte.

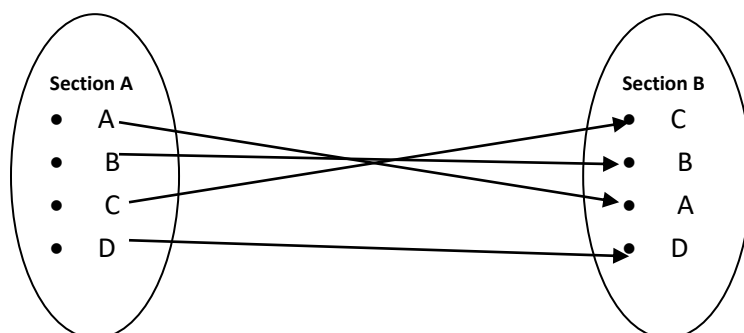
La commande standard est la suivante : **append using "chemin_d'_accès\nom_fichier.dta", keep(liste de variables)**

b- Ajout de variables

La fusion par ajout de variables consiste à ajouter des variables observées sur les mêmes individus que ceux de la base en cours d'utilisation ou sur des individus qui sont liés à ceux-ci à l'aide d'une clé de liaison. Plusieurs types de fusions par ajouts de variables peuvent être distingués. Il s'agit de :

b-1- Fusion un par un

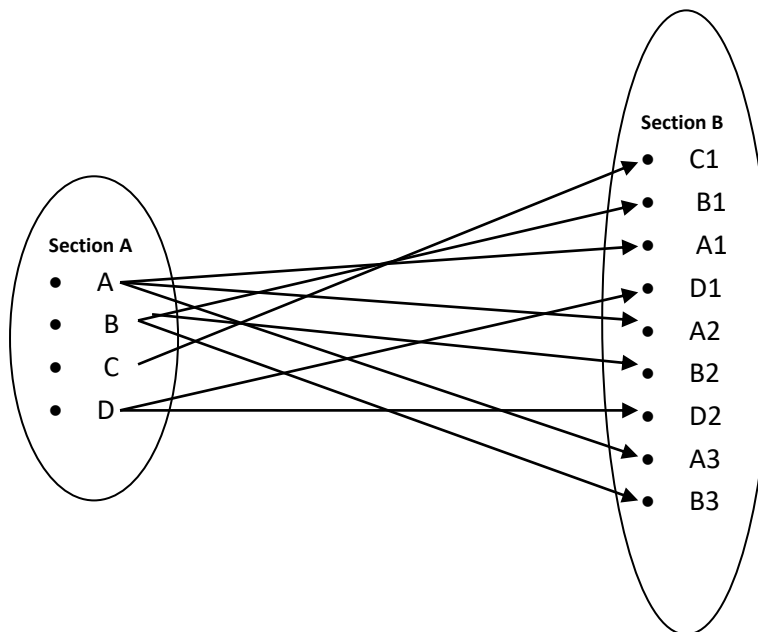
La fusion « un par un » : cela signifie qu'un individu de la base ouverte est associé à un et un seul individu de la base de données que l'on souhaite fusionner. C'est le cas par exemple lorsque l'on a réalisé une enquête à plusieurs volets sur les mêmes individus ou que l'on a saisi séparément les données par section d'une même enquête et que l'on veut fusionner par la suite.



b-2- Fusion un pour plusieurs

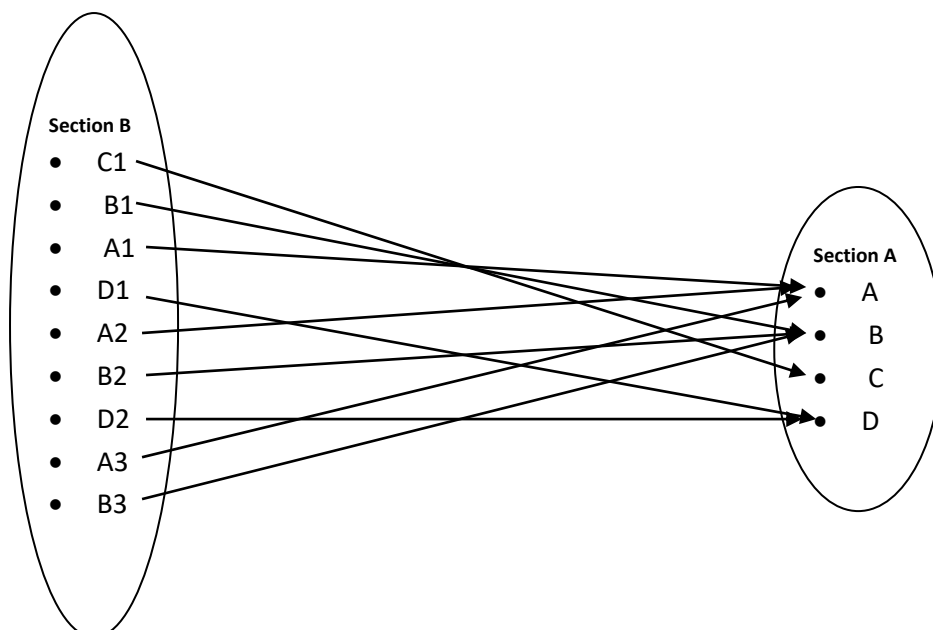
La fusion « un pour plusieurs » : cela signifie qu'un individu de la base ouverte est associé à un ou plusieurs individus de la base de données à fusionner. C'est le cas par exemple lorsqu'on a réalisé une enquête pour laquelle on a une section ménage dans laquelle on prend

des renseignements généraux sur le ménage et on a une section sur les individus du ménage où pour chaque individu, on collecte des informations spécifiques. Si l'on souhaite par la suite fusionner ces deux bases pour n'avoir qu'une seule dans laquelle on a pour chaque individu, les informations sur son ménage d'origine, alors on ouvre la base des ménages et on associe celle des individus.



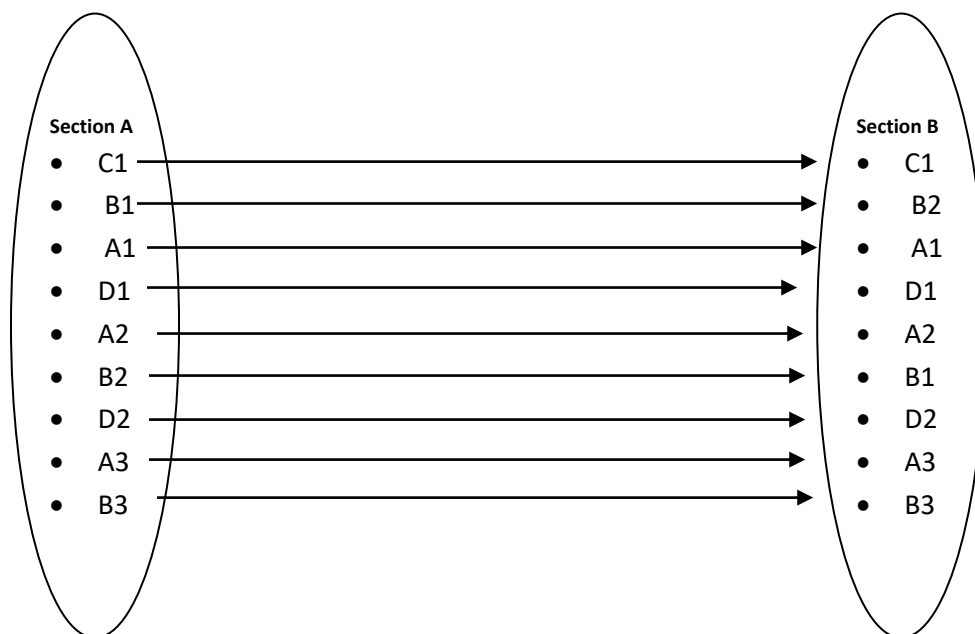
b-3- Fusion plusieurs pour un

La fusion « plusieurs pour un » : cela signifie qu'à plusieurs individus de la base ouverte est associé à un individu de la base de données à fusionner. C'est l'inverse du processus de la fusion « un pour plusieurs »



b-4- Fusion plusieurs pour plusieurs

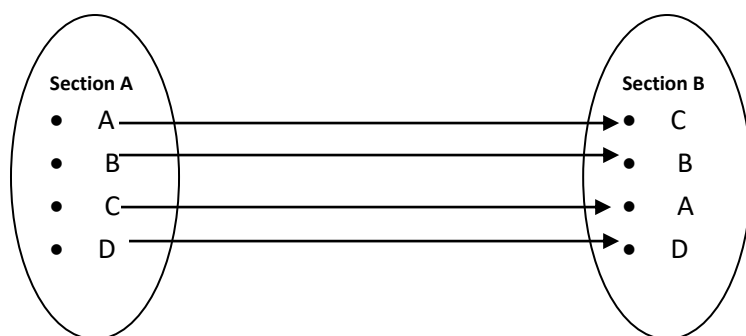
La fusion « plusieurs pour plusieurs » cela signifie qu'à plusieurs individus de la base ouverte, on associe plusieurs individus de la base à fusionner. Ce type de fusion est rarement utilisé. Toutefois, son usage engendre une base fusionnée qui respecte la logique suivante : les individus des deux bases n'étant pas uniques selon l'identifiant, on associe le premier individu d'un groupe ayant le même identifiant au premier individu du même groupe dans la base à fusionner, et ce, jusqu'à ce qu'on ne trouve plus d'individus du même groupe dans l'une des deux bases.



Remarque : Ces quatre types de fusion par ajout de variable nécessitent la présence d'une variable identifiant (la variable déjà créée ou l'ensemble des variables dont la combinaison « concaténation » engendre l'identifiant). Ainsi pour la fusion « un pour un », la variable identifiant doit identifier de manière unique les individus dans les deux bases à fusionner. Pour s'en assurer, on peut utiliser la fonction « isid », voir section IV-1-a. Pour la fusion « un pour plusieurs » ou « plusieurs pour un », l'identifiant doit également identifier de manière unique chaque individu de la base à partir de laquelle on ajoute un individu ou à partir de laquelle on veut ajouter plusieurs individus. Pour la fusion « plusieurs pour plusieurs », l'identifiant n'est pas nécessairement unique dans les deux bases.

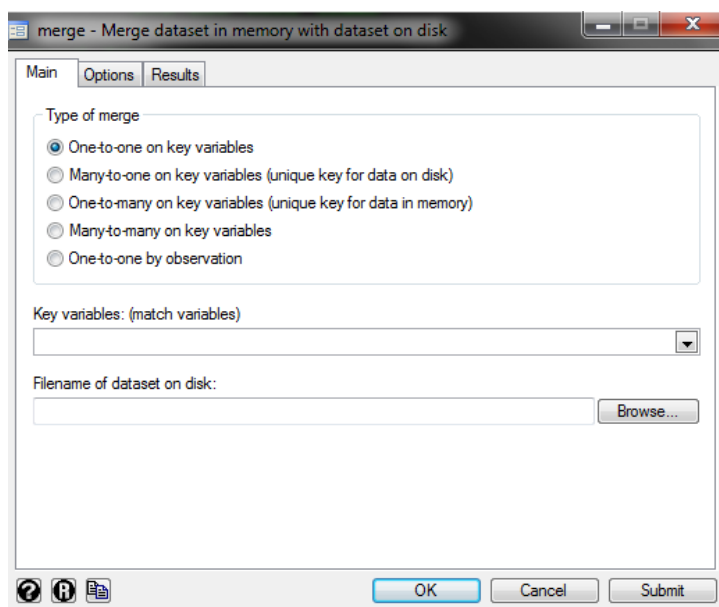
b-5- Fusion un pour un par individu

La fusion « un pour un par individu » : cela consiste à fusionner deux bases de donner sans disposer d'un identifiant. La fusion ainsi réalisée se fait en associant à chaque individu de la base ouverte, un et un seul individu de la base à fusionner. Cette association se fait un selon l'ordre d'apparition des individus dans chaque base. Ainsi, au premier individu de la base ouverte, est associé le premier individu de la base à fusionner, ainsi de suite jusqu'au dernier individu.



b-6- Mise en pratique

Pour réaliser une fusion par ajout de variables, il faut aller dans le menu général et cliquer sur « data » puis sur « combine datasets » puis sur « merge two datasets », on obtient alors la boîte de dialogue suivante :

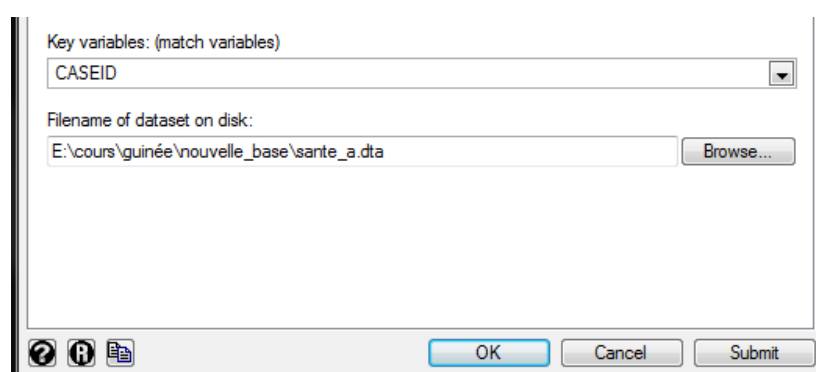


Dans cette boîte de dialogue, il faut commencer par choisir le type de fusion. On a « one-to-one on key variables » pour la fusion « un pour un », « many-to-one on key variables » pour la fusion « plusieurs pour un », « one-to-many on key variables » pour la fusion « un pour plusieurs », « many-to-many on key variables » pour la fusion « plusieurs pour plusieurs » et « one-to-one by observation » pour la fusion « un pour un par observation ».

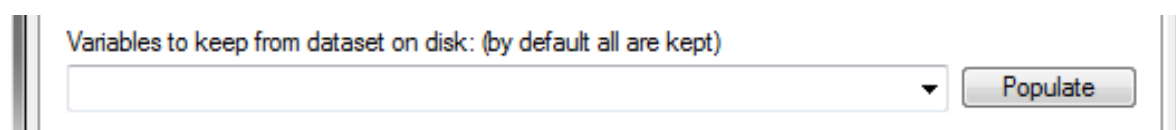
Quand on a choisi le type de fusion, lorsque c'est l'un des quatre premiers types, il faut maintenant choisir l'identifiant ou la liste de variables qui le forment. Cela se fait grâce à la liste de choix nommée « key variables (match variable) ». Lorsqu'il s'agit d'une liste de variables formant l'identifiant, il est nécessaire de les ranger dans l'ordre dans lequel ces variables forment l'identifiant. La concaténation (combinaison) par exemple dans l'ordre suivant « code région, code grappe, numéro ménage, et numéro individu » forme l'identifiant unique pour un individu dans un ménage car un numéro individu est unique dans le ménage, un numéro ménage unique dans une grappe, un code grappe unique dans une région, et les

codes région sont unique pour les régions. Tandis que si l'on intervertit l'une des variables citées ci-dessus, on n'a plus nécessairement un identifiant. En effet, si on met la variable numéro individu avant numéro ménage, on peut avoir deux individus de la même région, même grappe, et qui ont les numéros individu et numéro ménage tels qu'on ait numéro ménage = 12 et numéro individu = 1 pour le premier, numéro ménage = 1 pour le second et numéro individu = 21 pour le second. La concaténation donnera RG121 pour les deux (avec R code région et G code grappe ». On a ainsi deux individus différents mais avec un identifiant identique. Cela est donc à éviter. Il faut prendre soin de ranger les variables dans le bon ordre.

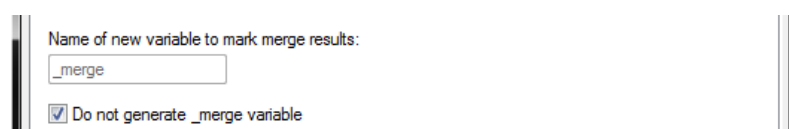
Après avoir sélectionné les variables pour l'identifiant, bien sûr en s'assurant que cette (ces) variable (s) est (sont) présente (s) dans les deux bases, on va maintenant indiquer la base à fusionner. Cela se fait dans la partie « filename of dataset on disk ». Il faut cliquer sur « browse » et aller sélectionner la base à fusionner. Un fois la base sélectionnée, on obtient :



Dans les options (onglet « options »), il faut ensuite choisir les variables de la base à fusionner que l'on souhaite ajouter à la base en cours d'utilisation. Cela se fait grâce à la liste de choix « variables to keep from dataset on disk ». Par défaut, si aucune sélection n'est effectuée, Stata garde toutes les variables de la base à fusionner. On peut également cliquer sur « populate » pour charger uniquement dans la liste de choix, les variables de la base à fusionner uniquement.



Ensuite, il faut donner un nom à la variable qui donnera l'origine de chaque individu dans la base fusionnée. Par défaut, cette variable se nomme « _merge ». Toutefois l'utilisateur peut décider de ne pas créer cette variable. Il suffit pour cela de cocher « do not generate _merge variable ».



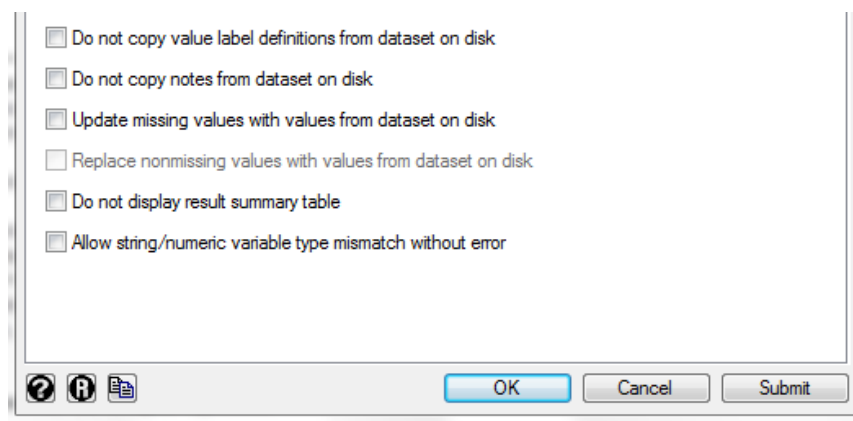
La variable « _merge » générée est faite selon la typologie décrite ci-dessous :

- elle a pour modalité « 1 » si l'individu est uniquement dans la base ouverte au départ ;
- elle a pour modalité « 2 » si l'individu est uniquement dans la base qu'on a utilisée pour la fusion ;
- elle a pour modalité « 3 » si l'individu apparaît dans les deux bases ;

Les autres options vous permettent de ne pas copier les labels des variables de la base à fusionner dans la base fusionnée obtenue, dans ce cas on coche « do not copy value label definitions from dataset on disk » ; par défaut, Stata copie ces labels. Il en est de même pour les notes, cocher dans ce cas « do not copy notes from dataset on disk ».

On peut également combler les données manquantes de la base ouverte avec des variables sans données manquantes de la base à fusionner. Il suffit de cocher dans ce cas « update missing values with values from dataset on disk ». Par défaut, cette mise à jour n'est pas effectuée. Lorsque cette option est choisie, la variable « _merge » peut prendre deux nouvelles modalités que sont « 4 » pour les individus ayant des valeurs manquantes dans les deux bases, et la modalité « 5 » pour les individus qui apparaissent dans les deux bases avec des valeurs non manquantes mais différentes (cas de mise à jour des valeurs). Cette modalité n'apparaîtra que si on demande à Stata de faire la mise à jour des valeurs non manquantes de la base ouverte par les valeurs non manquantes de la base à fusionner. Pour cela, on coche « replace nonmissing values with values from dataset on disk ».

Dans le cas où des variables changent de format (de numérique à alphanumérique ou vice versa), on peut demander à Stata de réaliser la fusion sans erreur, cocher dans ce cas « allow string/numeric variable type mismatch without error ». Enfin, par défaut Stata génère un tableau qui résume les résultats de la fusion effectuée. On pourra au besoin demander à Stata de ne pas générer ce tableau. Il faut simplement cocher « do not display result summary table ».

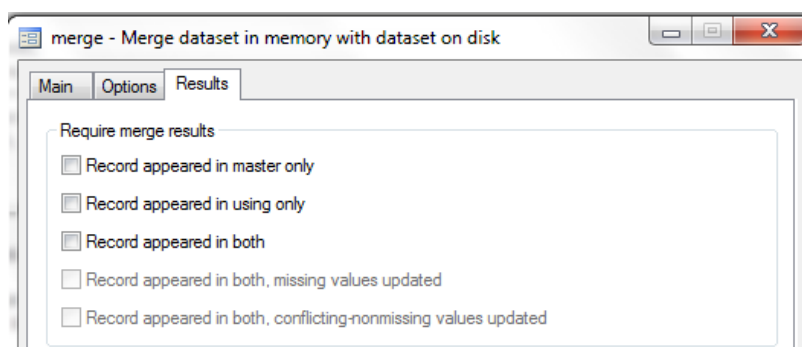


Dans l'onglet « result », on va ensuite spécifier le type d'individus à conserver dans la base de données fusionnées, voir les différentes modalités de la variable « _merge » décrite ci-dessus.

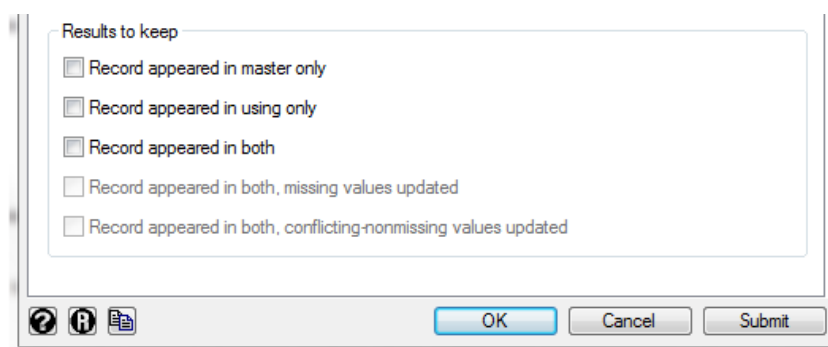
Dans cette première partie « require merge results » (voir schéma ci-dessous), on désire vérifier après la fusion que les individus de la base fusionnée respectent une certaine

condition. Si cette condition n'est pas respectée, alors Stata n'exécute pas la fusion. On distingue les cas suivants :

- si l'utilisateur souhaite faire la fusion uniquement lorsque la base fusionnée ne contient que les individus de la base ouverte alors il coche « record appeared in master only ». Stata affichera un message d'erreur si la fusion doit conduire à un code de la variable « _merge » différent de « 1 », en d'autres termes si on n'a pas uniquement les individus de la base ouverte qui sont fusionnés.
- si on souhaite faire la fusion uniquement lorsque la base fusionnée ne contient que les individus de la base à fusionner (c'est-à-dire variable « _merge » prend uniquement la modalité « 2 ») alors on coche « record appeared in using only ».
- si on souhaite faire la fusion uniquement lorsque la base fusionnée ne contient que les individus qui apparaissent dans les deux bases (c'est-à-dire variable « _merge » prend uniquement la modalité « 3 ») alors on coche « record appeared in both ».
- si on souhaite ne faire la fusion que si la base fusionnée sera mise à jour de ses valeurs manquantes (cas où on a choisi l'option de mise à jour décrite ci-dessus), on cochera « record appeared in both, missing values updated ».
- enfin on cochera « record appeared in both, conflicting nonmissing values updated » si on souhaite ne faire la fusion que si la base fusionnée sera mise à jour de ses valeurs non manquantes mais actualisée (cas où on a choisi l'option de mise à jour décrite ci-dessus).



Dans la seconde partie « results to keep » (voir schéma ci-dessous), on désire que Stata réalise la fusion mais ne retient dans la base de données que les individus qui respectent une certaine condition. Les spécifications sont les mêmes que celles décrites ci-dessus pour le « require merge results ». La seule différence est qu'aucun message d'erreur n'est délivré et la fusion est tout le temps réalisée, à condition bien sûr de respecter les conditions standards.



III- Travaux préliminaires

Dans cette section, il s'agira d'aborder les travaux généralement effectués en premier lieu avant de commencer une quelconque analyse de la base de données. Il s'agit principalement de l'organisation de la base de données, de l'étiquetage des données, de la suppression des données inutiles et de la création de variables que l'on devra analyser mais qui ne sont pas dans la base.

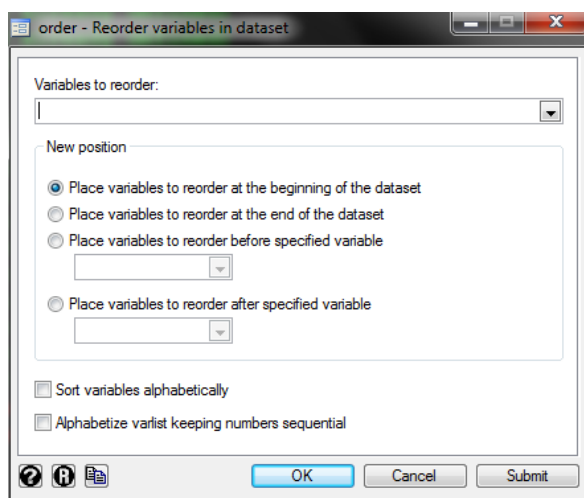
1- Organisation de la base de données

L'organisation de la base de données est une étape très importante avant toute analyse. C'est à ce niveau que vous décidez de l'ordre d'apparition des variables et des individus, et de la suppression ou de la garde de variables ou d'individus inutiles. Aussi, vu que Stata n'affiche pas les données sur lesquelles vous travaillez (comme sous Excel), vous avez la possibilité de visionner votre base de données avant son utilisation.

a- Ordonner la base

Ordonner la base de données consiste à déterminer l'ordre d'apparition des variables ou des individus de la base. C'est une opération dont le but est de rendre plus aisée la manipulation des données au cours du traitement.

Pour ordonner une base de données selon les variables, on peut soit le faire par ordre alphabétique, soit faire apparaître les variables les unes après les autres selon un ordre qu'on établira. Quelle que soit l'option envisagée, pour réaliser un ordre des variables, il faut dans l'onglet « data » du menu, cliquer sur « data utilities » puis sur « change order of variables ». On obtient la boîte de dialogue ci-dessous :

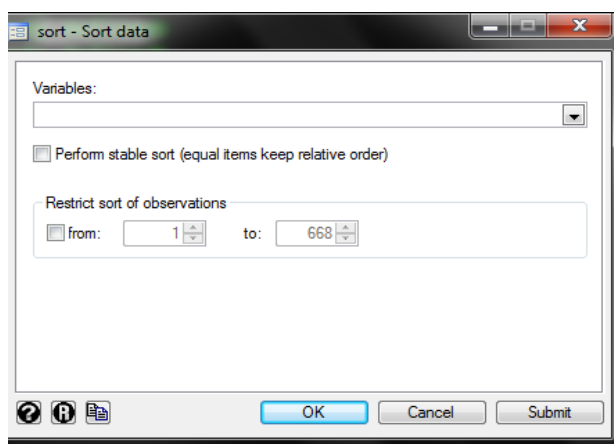


La liste de choix « variables to reorder » vous permet de sélectionner les variables que vous voulez déplacer dans la base de données ». Lorsque cette plage est laissée vide, la commande ne s'exécutera que si vous cochez « sort variables alphabetically », ce qui signifie que toutes les variables de la base de données seront rangées par ordre alphabétique. Lorsqu'une sélection est réalisée, vous devez par la suite indiquer comment vous souhaitez voir ordonnées ces variables. Dans le cadrant « new position », vous cochez au choix « place variables to reorder at the beginning of the dataset » pour indiquer à Stata que les variables que vous avez sélectionnées doivent apparaître dans cet ordre en première position dans la base de données. Vous cochez au choix « place variables to reorder at the end of the dataset » pour indiquer à Stata que les variables que vous avez sélectionnées doivent apparaître dans cet ordre en dernière position dans la base de données. Au-delà de ces deux spécifications génériques, vous pouvez indiquer à Stata que les variables que vous avez sélectionnées doivent apparaître avant une variable donnée (cochez « place variables to reorder before specified variable » et sélectionnez dans la liste de choix juste en bas), ou que ces variables doivent apparaître après une variable donnée (cochez « place variables to reorder after specified variable » et sélectionnez dans la liste de choix juste en bas). Enfin, lorsque certaines variables de votre base ont le même préfixe et ne diffèrent que par un suffixe qui est un numéro, alors vous pouvez cocher « alphabetize varlist keeping number sequential » pour que Stata ordonne ces variable en fonction du numéro.

NB : l'ordre alphabétique simple donne x1, x10, x2 tandis qu'avec la dernière option décrite ci-dessus, vous aurez x1, x2, x10.

La syntaxe est la suivante : si on veut ranger par ordre alphabétique **aorder** et si on veut mettre en premier plan un ensemble de variable **order liste_de_variable**.

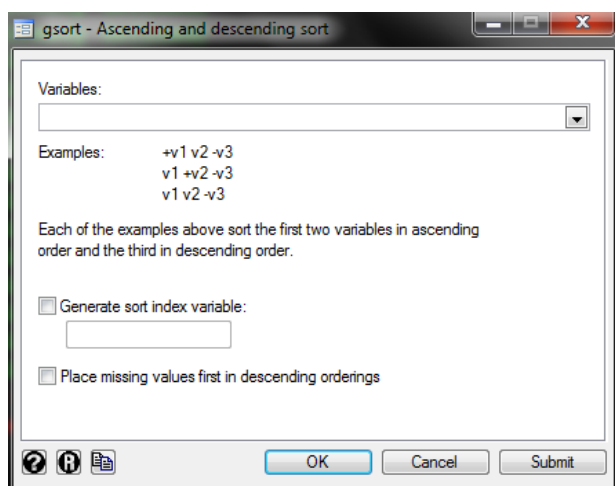
Après avoir ordonné les variables, on peut ordonner ou trier les individus suivant un ou plusieurs critères. Vous pouvez par exemple décider de ranger les individus de la base selon leur âge, ou encore selon la région, la commune puis l'âge. Si l'ordre dans lequel vous voulez faire ce rangement est l'ordre croissant (du plus petit au plus grand), alors vous allez dans l'onglet « data » du menu et vous cliquez sur « sort » puis sur « ascending sort ». Vous obtenez le menu suivant :



Dans la liste de choix « variables », on choisit dans l'ordre la ou les variables suivant(s) la(les)quelle(s) on veut ordonner les observations de la base de données. Ainsi Stata rangera les observations selon les valeurs des variables listées, en mettant dans une position aléatoire les individus d'un groupe (c'est-à-dire si on range par région uniquement, vu que plusieurs individus ont la même région, l'ordre des individus de la région sera aléatoire). On peut cocher l'option « perform stable sort » pour que dans le cas où on a plusieurs individus par groupe, le premier dans la base rangée pour un groupe soit le premier individu observé du groupe. On peut également appliquer le ranger à une partie de la base. Dans ce cas, on coche « restrict sort of observations » pour que par exemple, Stata range les p premiers individus de la base (on mettra from 1 to p) ou les p derniers (on mettra from N-p+1 to N, N étant la taille de la base) ou pour les p individus d'une plage quelconque (on mettra from x to x+p-1).

La syntaxe est la suivante : **sort liste_de_variable**

Lorsque l'ordre que l'on veut mettra dans la base change en fonction des variables suivants lesquelles on range, c'est-à-dire, ordre croissant pour certaine et ordre décroissant pour d'autre, on utilisera dans l'onglet « data » du menu, le « ascending and descending sort ». On obtient le menu ci-dessous. A ce niveau, il faut sélectionner par ordre les variables selon lesquelles on veut faire le rangement. Pour ranger par ordre décroissant selon une variable, il faut précéder le nom de la variable par le signe « - » de la soustraction. Pour les variables selon lesquelles un ordre croissant sera appliqué, on peut précéder le nom de la variable par un signe « + » de l'addition ou ne mettre aucun signe.



On cochera l'option « generate sort index variable » puis on indiquera le nom de la variable à générer si l'on souhaite générer un variable qui numérote les groupes formés par les variables permettant de réaliser le rangement.

La syntaxe est la suivante : **gsort +var1 -var2** où +/- indique le sens du rangement.

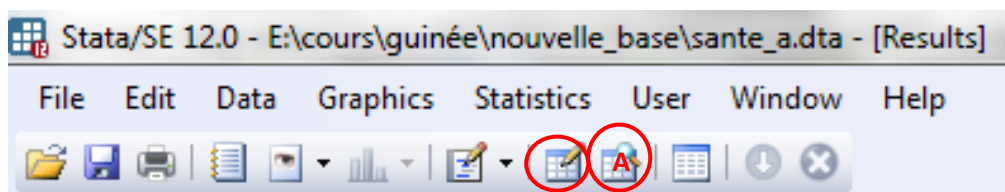
Remarque 1 : Par défaut Stata considère, pour une variable numérique, la valeur manquante (matérialisée par un « . ») comme étant la plus grande possible, et pour une valeur alphanumérique, la valeur manquante (matérialisée par un « ») comme étant la plus petite. Ainsi, si vous voulez changer cet ordre et placer les valeurs manquantes en première position pour un rangement par ordre décroissant, il vous suffit de cocher « place missing values first in descending orderings ».

Remarque 2 : ces actions ne sont pas obligatoires mais utiles pour le traitement de données car elle rendre plus aisée la navigation dans les données.

b- Visualiser la base

La visualisation d'une base de données peut se faire selon deux options. La première consiste à visualiser en se donnant le droit de modifier des valeurs (Edit), et la seconde, on n'a pas de droit de modification (Browse). Pour visualiser la base, il vous suffit de cliquer sur « Data editor » dans l'onglet « data » du menu. Vous choisissez par la suite sur « Edit » pour visualiser avec possibilité de modification et sur « Browse » pour visualiser sans possibilité de modification.

On peut également accéder directement à la visualisation en cliquant sur les icones « edit » ou « browse », icones encadrées en rouge dans la figure ci-dessous.



La syntaxe est la suivante : **edit liste_variable** ou **browse liste_variable**

c- Suppression

La suppression peut concerner une ou des variables ou encore un ou plusieurs individus. Pour la suppression de variables, on clique sur l'onglet « data » puis sur « variables manager » et on obtient l'écran suivant :

#	Variable	Label	Type	Format	Value Label	Notes
	CASEID	Case Identification	str15	%15s		
	MIDX	Index to birth history	double	%10.0g		
	V000	Country code and phase	str3	%3s		
	V001	Cluster number	double	%10.0g		
	V002	Household number	double	%10.0g		
	V003	Respondent's line number	double	%10.0g		

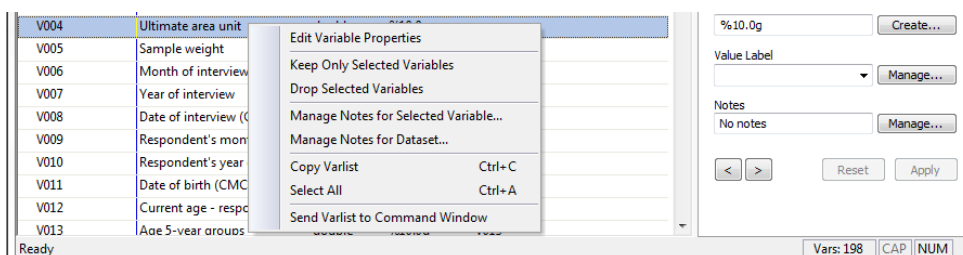
Name
V004

Label
Ultimate area unit

Type
double

Format

Deux options s'offrent à l'utilisateur : il peut sélectionner les variables à supprimer ou les variables à maintenir dans la base. Une fois la sélection effectuée, on peut faire un clic droit et on obtient la boîte suivante :

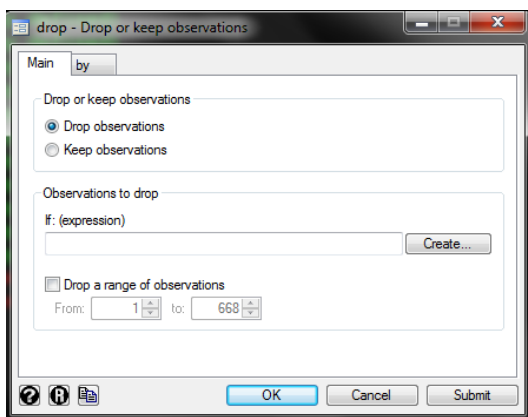


L'utilisateur peut donc cliquer sur « keep only selected variables » s'il souhaite ne retenir que les variables sélectionnées ou sur « drop selected variables » s'il souhaite supprimer les variables sélectionnées.

Il faudrait enfin remarquer que ce menu ne permet pas uniquement de supprimer ou retenir les variables de la base. Il sert également à définir les propriétés des variables. Il suffit dans ce cas de sélectionner la variable à éditer et de cliquer sur « edit variable properties » ou encore de modifier directement certaines propriétés dans le cadran droit du menu ci-dessus mentionné.

La syntaxe est la suivante : **keep liste_variable** ou **drop liste_variable**

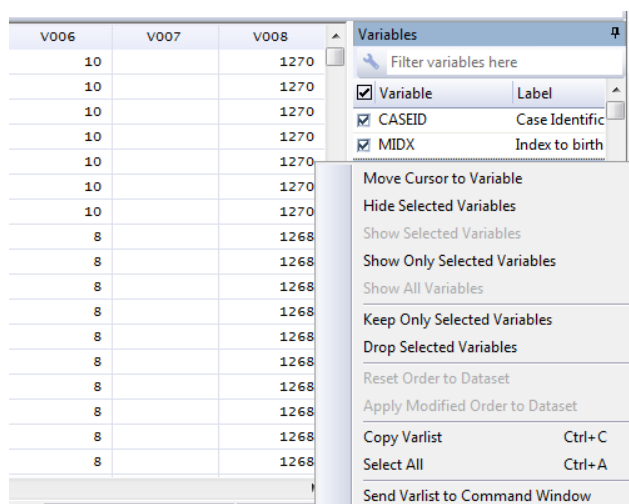
Pour la suppression d'individus, il faut dans l'onglet « data » cliquer sur « create or change data » puis sur « keep or drop observations ». On obtient le menu ci-après :



Ici également, deux options s'offrent à l'utilisateur : sélectionner les individus à supprimer ou les individus à garder dans la base. Il faut commencer par choisir le type d'opération : « drop observations » pour la suppression et « keep observations » pour le maintien d'individus. Ensuite, on peut définir les conditions de suppression ou de maintien. Cela se fait dans la partie « if : (expression) » où on peut écrire directement les conditions ou encore aller dans « create » puis écrire la condition. Le choix des individus peut être également effectué par liste, c'est-à-dire définir d'un numéro d'ordre à un autre. Il suffit de cocher sur « drop a range of observations » et définir le numéro de départ et le numéro de fin.

La suppression des individus peut être répétée dans des groupes d'individus. Il s'agit de dire par exemple dans chaque groupe formé par les modalités d'une variable V1, supprimer les individus qui vérifient une condition. Pour réaliser une telle opération, il suffit de cliquer sur « by » et cocher « repeat command by groups » et sélectionner la ou les variables de groupe.

Une autre approche consiste à entrer dans l'éditeur de variable « data editor » et sélectionner des variables puis faire un clic droit. On obtient le menu suivant :



On peut alors cliquer sur chacune des options en fonction de ce qu'on souhaite réaliser.

La syntaxe est la suivante : **drop if/in condition** ou **keep if/in condition**

2- Etiquettes et noms

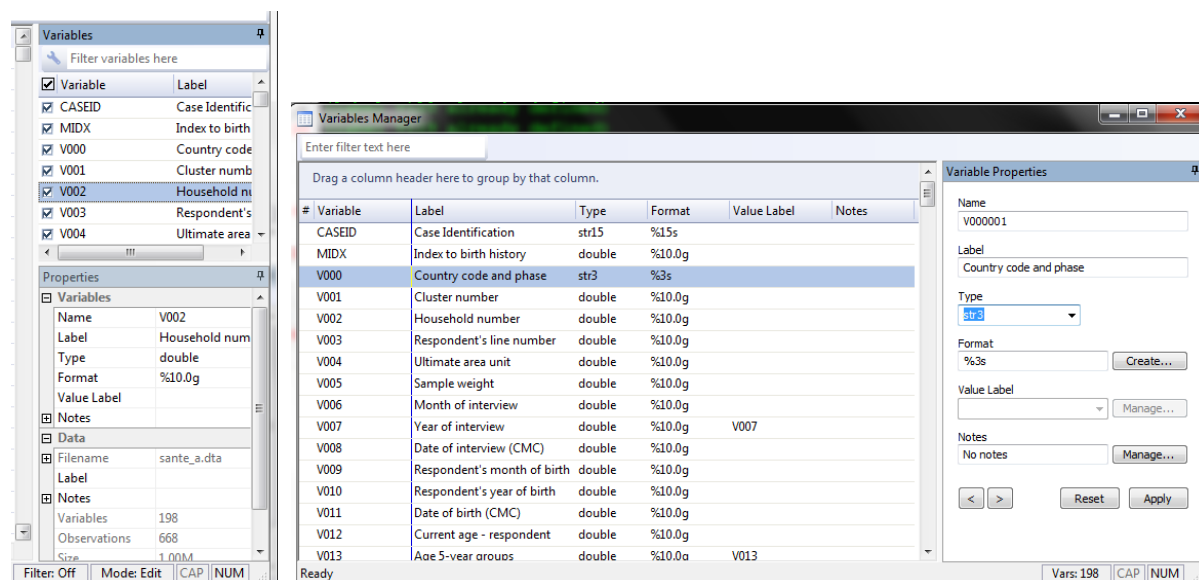
L'étiquetage est une opération très importante dans les travaux préliminaires de préparations des bases de données. Cette opération concerne aussi bien les noms de variables, les labels de variables que les labels des modalités de réponses des variables.

a- Renommer

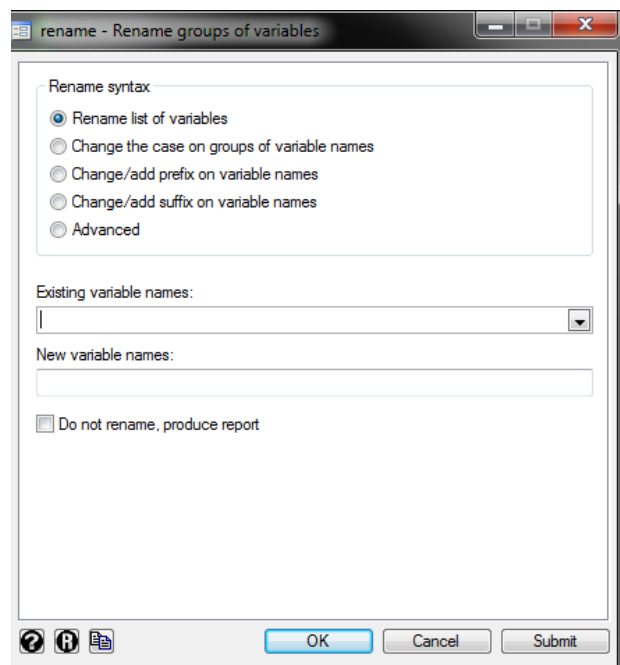
Le nom d'une variable est un ensemble de caractères qui permet d'identifier de manière unique une variable d'une base de données. Ce nom devra être succinct et ne devra pas contenir de caractères spéciaux tels que les espaces, les signes arithmétiques de base (*,-,+,/,...). Quant aux labels, ils peuvent être suffisamment longs (mais ne pas dépasser 225 caractères). Il peut contenir des espaces et des caractères spéciaux.

La première approche consiste à entrer dans le « data editor » ou dans le menu « data » puis dans le « variables manager » et cliquer sur la variable à renommer et dans les propriétés, aller dans « name » et renommer la variable.

Il en est de même pour le label de la variable. Il suffit de cliquer sur la variable et dans les propriétés, aller dans « label » et changer le label de la variable.



La seconde manière de procéder consiste à aller dans « data » puis « data utilities » et cliquer sur « rename groups of variables » et on obtient le menu suivant :



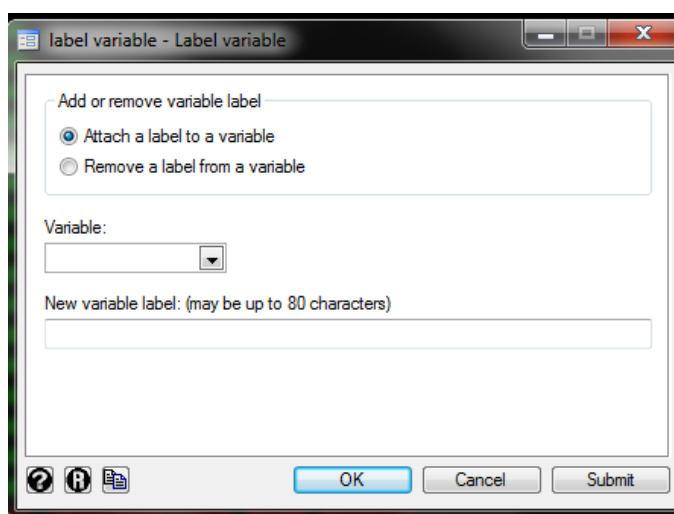
Il faut commencer par choisir le type d'opération à faire. « rename list of variables » pour renommer, « change/add prefix/suffix on variable names » pour précéder ou faire suivre le nom des variables sélectionnées par une chaîne de caractère, « change the case on groups of variable names » pour convertir le nom de la ou des variables en minuscule, majuscule ou

convertir uniquement la première lettre du nom de la variable en majuscule, ou « advanced » pour des options avancées c'est-à-dire soit ajouter un numéro séquentiel aux noms des variables. Une fois le type d'opération sélectionné, il suffit de sélectionner la ou les variables à renommer dans la partie « existing variable names » puis dans la partie « new variables names », écrire dans l'ordre (bien sur selon l'ordre de sélection des variables à renommer) les noms des nouvelles variables.

La commande de base est : **rename ancien_nom nouveau_nom**

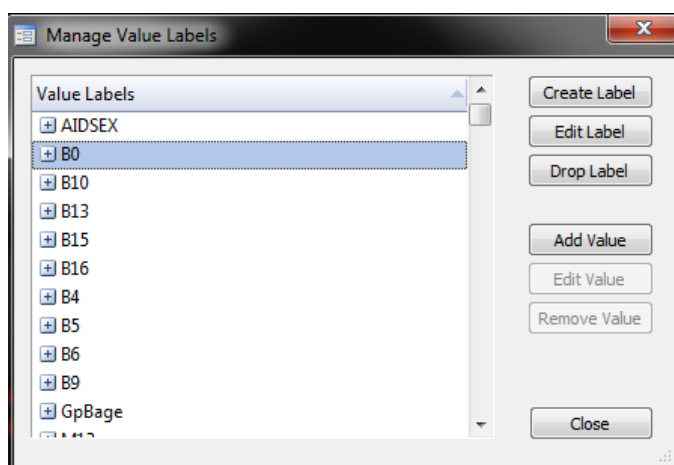
b- Etiquette de variables et de modalités

Pour l'étiquetage des variables, on peut aller dans « data », « data utilities » puis « label utilities » et « label variable ». On obtient le menu suivant :

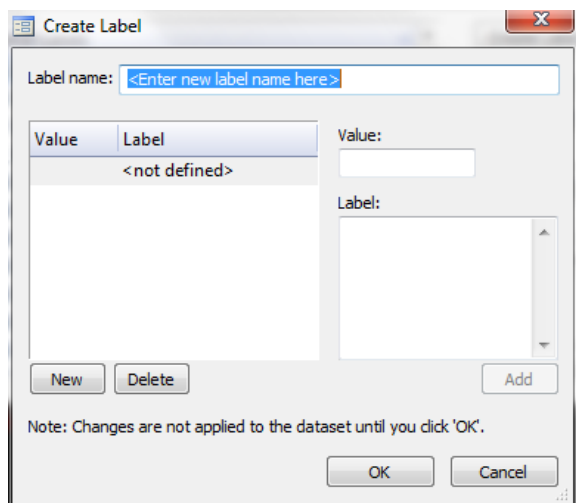


On peut commencer par choisir le type d'opération à réaliser : soit ajouter un label « attach a label to a variable », soit supprimer un label « remove a label from a variable ». Il faut ensuite choisir la variable et dans le cas où on veut ajouter un label, il faut écrire le label à ajouter puis faire OK.

Pour le label des modalités des variables, il faut aller dans « data », « data utilities » puis « label utilities » et « manage value labels ». On obtient le menu ci-dessous :

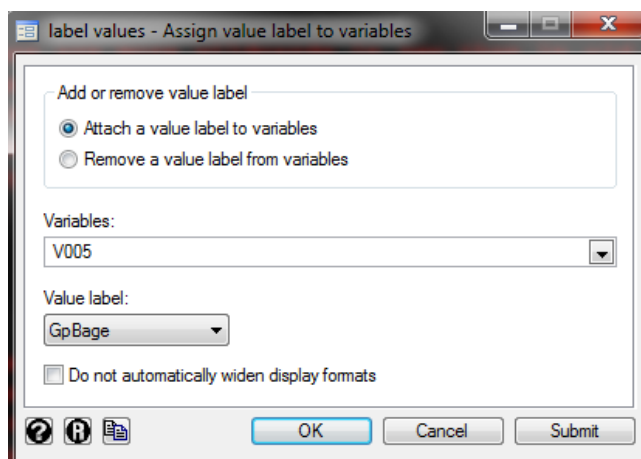


On peut créer un nouveau label « create label », modifier un label « edit label » ou supprimer un label « drop label ». La création d'un label commence par la définition de son nom, puis pour chaque valeur, on définit l'étiquette de la modalité. Une fois la sélection terminée, on peut valider. Tout ceci se fait dans le menu ci-dessous :



Aussi, dans le menu ci-dessus, est-il possible, pour un label déjà défini, de modifier les valeurs. Il suffit de cliquer sur ce label, et sur « add value », « edit value » ou « remove value » selon le cas.

Pour associer le label à la variable, il faut par la suite aller dans « data », « data utilities » puis « label utilities » et « assign value label to variables » et on obtient le menu ci-dessous :



Il faut choisir le type d'opération (ajouter ou supprimer un label) et choisir la ou les variables dans « variables » et dans la liste « value label », choisir le nom du label à ajouter à la variable.

La succession de commandes est la suivante :

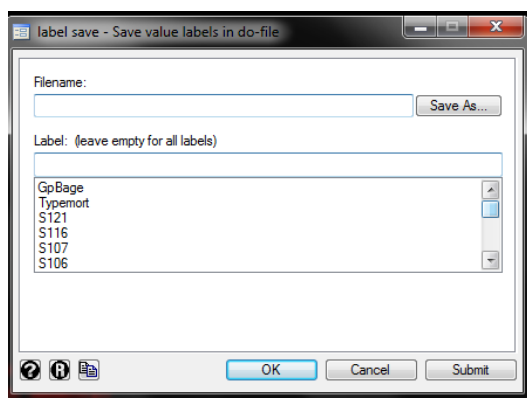
label define nom_label code1 label1 code2 label2 ... puis

label value nom_variable nom_label si l'on souhaite créer et ajouter un label aux modalités d'une variable.

label drop nom_label si l'on veut supprimer un label.

label var nom_variable si l'on veut étiqueter une variable

Remarque : les labels d'une base de données peuvent être exportés sous forme de fichier do en vue d'une réutilisation ultérieure. Il suffit pour cela de faire « data », « data utilities » puis « label utilities » et « save value labels as do-file » et on obtient le menu ci-dessous.



Il suffit enfin de faire sélectionner les labels et donner un nom au fichier do en spécifiant le chemin d'accès.

3- Transformation de données

La transformation de données consiste en la création de variables en se basant sur les variables déjà existantes dans la base de données ou encore de modifier le contenu d'une variable.

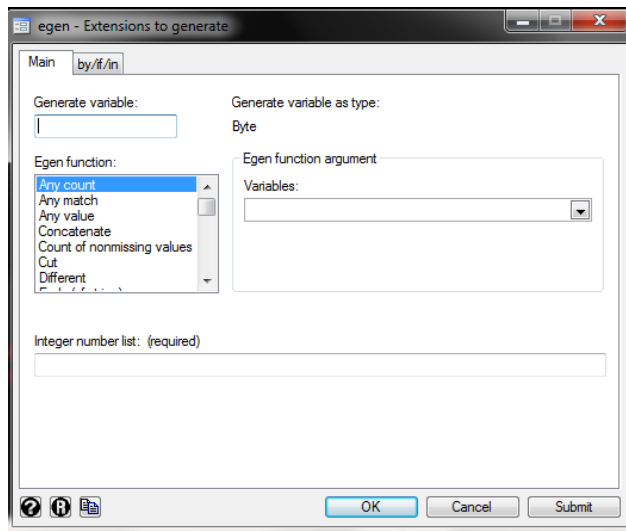
a- Création de variables

La création de variables peut se faire de plusieurs manières et ce, en fonction du type de variables que l'on souhaite obtenir et du type de variables utilisées en input.

La première approche consiste à créer une variable en utilisant les formules mathématiques, arithmétiques ou logiques de base sur un ensemble de variable en input. Cela se fait en cliquant sur « data », puis « create or change data » et sur « create new variable ». Il faut par la suite donner le nom de la variable, définir le format de la variable à créer et définir son expression.

La syntaxe est la suivante : **generate nom_variable = expression**

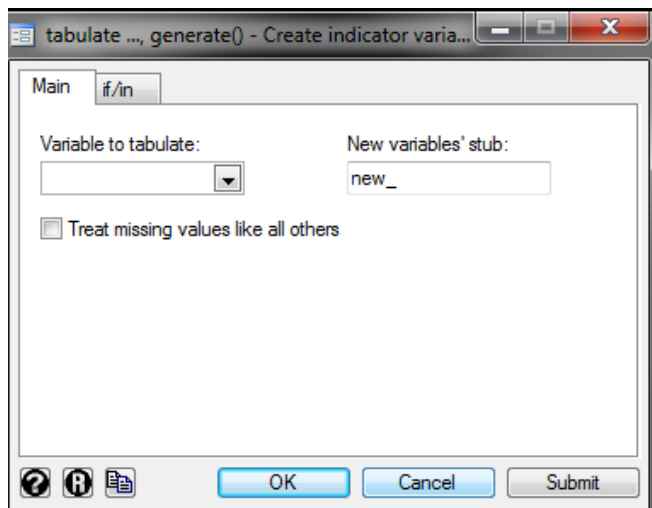
La seconde approche consiste à créer une variable à l'aide de formules plus sophistiquées. Il faut cliquer sur « data », puis « create or change data » et sur « create new variable (extended) » et on obtient le menu ci-après :



Il faut définir le nom de la variable, choisir le format de la variable à générer, choisir la formule à utiliser dans « egen function » et choisir les variables arguments de la fonction sélectionnée.

La syntaxe est la suivante : **egen nom_variable = nom_fonction(liste_variable)**

Il est également possible de créer des variables dichotomiques à partir des modalités d'une variable. Il suffit de cliquer sur « data », puis « create or change data » puis sur « other variable-creation commands » et sur « create indicators variables » et on obtient le menu ci-après :



Il faut sélectionner la variable à dichotomiser dans « variable to tabulate », donner le préfixe des nouvelles variables « new variables stub » et valider.

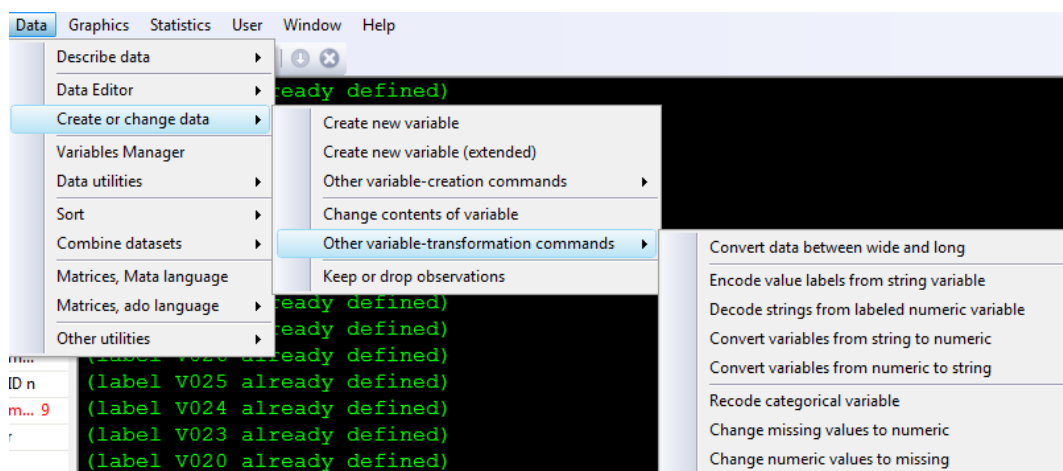
La syntaxe est la suivante :

tabulate nom_variable, generate(prefixe_nom_variable_indicatrice)

D'autres approches sont disponibles dans le menu en cliquant sur « data », puis « create or change data » puis sur « other variable-creation commands ». On peut alors réaliser des transformations telles que celle de Box-Cox ou encore l'orthogonalisation de variables.

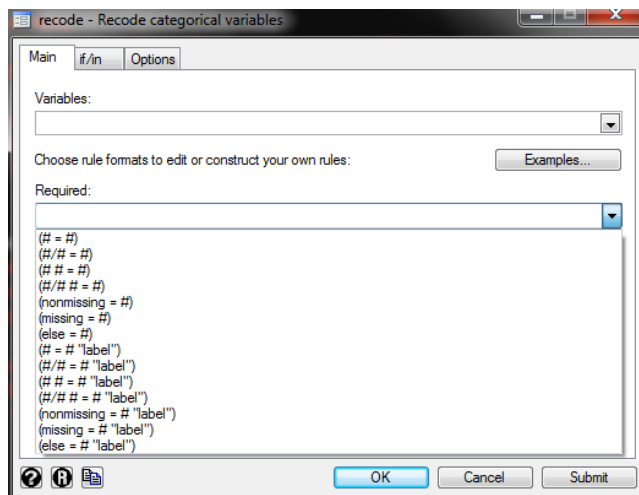
b- Changement de format de variable

Le changement de format consiste à passer d'une variable de format numérique à une variable de format alphanumérique ou vice versa. On parle également de changement de format lorsque pour une variable numérique, on passe d'un format à un autre (c'est-à-dire qu'une variable numérique définie par un type byte, long, wide, double, ..., peut être modifiée vers l'un des formats numériques ci-dessus cité). Pour réaliser cette opération, il suffit de faire « data », puis « create or change data » puis sur « other variable-transformation commands » puis de choisir une option en fonction de ce que l'on souhaite réaliser. Le passage de numérique à alphanumérique ou vice versa se fait en choisissant « convert variables form numeric/string to string/numeric ». Lorsqu'on dispose d'une variable qualitative dont les modalités ont été saisies en alphanumérique et qu'on souhaite convertir en gardant les labels, on choisit « encode value labels from string variable ».



La syntaxe est : pour convertir de alphanumérique à numérique en gardant les labels **encode nom_variable, generate(nom_nouvelle_variable)** ou pour le passage de numérique à alphanumérique **destring nom_variable, generate(nom_nouvelle_variable)** ou pour le passage de alphanumérique à numérique **tostring nom_variable, generate(nom_nouvelle_variable)**

On peut également décider de catégoriser une variable (créer des classes pour une variable quantitative par exemple) ou encore recoder les modalités d'une variables qualitatives en de nouvelles modalités de réponses basées sur les premières. Cela se fait en choisissant « recode categorical variable » dans le menu obtenu en faisant « data », puis « create or change data » puis sur « other variable-transformation commands ». On obtient alors le menu ci-après :



Il faut sélectionner la variable à recoder « variables », définir dans « required » les éléments à recoder puis dans les options, choisir s'il convient de créer une nouvelle variable ou d'écraser le contenu de l'ancienne. La précision des éléments à recoder se fait de la manière suivante :

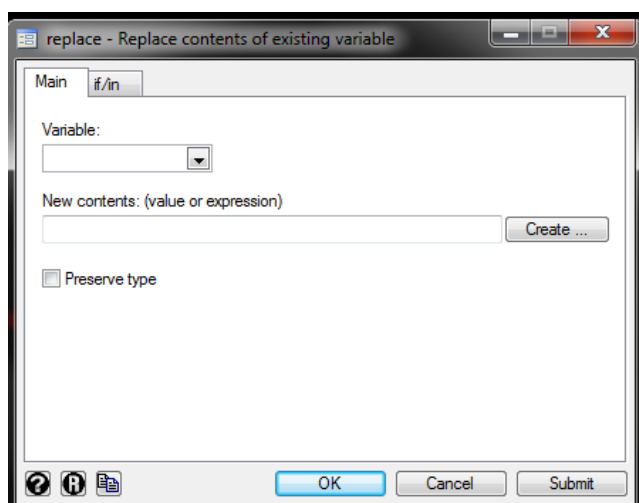
(anciennes_valeurs = nouvelle_valeur "label") où anciennes_valeurs peut être une plage que l'on construit en précisant valeur_de_départ/valeur_d'arrivée (ou borne inf et borne sup).

La syntaxe est : **recode nom_variable (anciennes_valeurs = nouvelle_valeur "label") ..., generate(nom_nouvelle_variable)**

c- Transformation des observations

La transformation des observations d'une variable consiste à remplacer une observation par une autre (cette autre étant une observation, une expression fonction d'une ou plusieurs observations) tout en précisant les individus pour lesquels on souhaite faire cette transformation.

Pour réaliser cette opération, il faut « data », puis « create or change data » puis sur « change contents of variable » et on obtient le menu ci-après :



Il suffit de préciser la variable pour laquelle la transformation est voulue dans « variable », donner la valeur dans « new contents » et définir la condition dans « if/in ». On peut également demander à Stata de ne pas changer le type de la variable de base en cochant « preserve type ».

La syntaxe est : **replace nom_variable = valeur**

IV- Contrôles et apurement

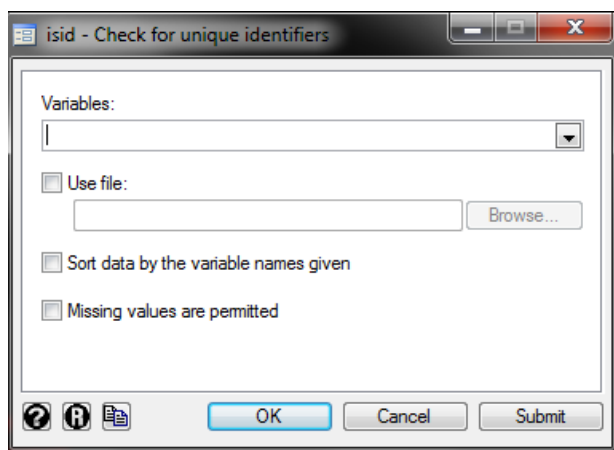
Il s'agit d'une étape cruciale dans tout processus d'analyse statistique. Il consiste à identifier les incohérences et à apporter des corrections. Mais avant on peut s'intéresser à l'identification des doublons.

1- Traitement de doublons

Les doublons dans une base de données peuvent être légitime ou non. Toutefois il convient de s'assurer de la non existence de doublons pour éviter des analyses faussées aux cas où les doublons ne sont pas autorisés.

a- Unicité de l'identifiant

L'identifiant est la ou les variables dont la combinaison ou la concaténation permet de retrouver de manière unique un individu de la base de données. Lorsqu'il s'agit d'une liste de variables, il n'est pas nécessaire de créer la variable identifiant avant de vérifier s'il existe ou non de doublon selon l'identifiant. Pour faire cela, il suffit d'aller dans « data » puis « data utilities » et cliquer sur « check for unique identifiers » et on obtient le menu ci-dessous :



Il faut lister dans « variables », la ou les variables qui constituent l'identifiant et valider. Lorsqu'aucun message ne s'affiche, alors les variables sélectionnées sont un identifiant pour la base.

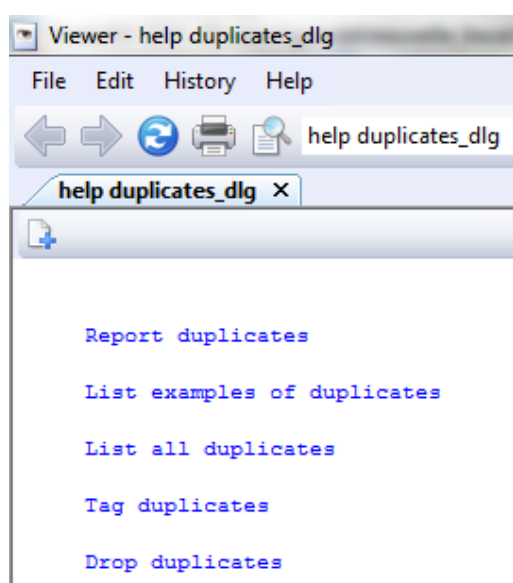
On peut également faire ce test sans même ouvrir la base de données. Il suffit de préciser le nom et le chemin d'accès de la base de données en cochant sur « use file » et sur « browse ». La syntaxe est : **isid liste_variable** ou **isid liste_variable using "chemin_accès\nom_fichier.dta"**

On peut également profiter de cette opération pour ordonner la base selon les variables de l'identifiant en cochant « sort data by the variable names given ».

Il est également possible d'autoriser les valeurs manquantes (car par défaut aucune valeur manquante n'est tolérée dans un identifiant). On peut avoir à utiliser cette option si l'une des variables constituant l'identifiant peut avoir des valeurs manquantes (extrêmement rare).

b- Recherche de doublons suivant des critères

Les doublons sont des individus de la base qui présentent exactement les mêmes caractéristiques au regard d'une ou plusieurs variables. On peut les identifier, les lister ou les supprimer. Pour le faire, il suffit de cliquer sur « data », puis sur « data utilities » et sur « manage duplicate observations » pour obtenir le menu suivant :



A ce niveau, il faut choisir le type d'opération à effectuer. « report duplicates » permet de faire un rapport sur les doublons (c'est-à-dire générer un tableau qui donne le nombre d'individus selon le nombre de fois où il est dupliqué). « list all duplicates » pour lister tous les individus en double ou triple, « tag duplicates » pour créer une variable qui contient « 0 » si l'individu est unique et « i » si l'individu est dupliqué i fois. « drop duplicate » qui permet de supprimer les individus dupliqués en ne conservant que le premier d'un groupe de doublon.

Une fois le choix effectué, il faut lister les variables selon lesquels on recherche les doublons et réaliser l'opération sélectionnée.

La syntaxe est la suivante : **duplicates list liste_variable** ou **duplicates tag liste_variable, generate(nom_nouvelle_variable)** ou **duplicates report liste_variable** ou **duplicates drop liste_variable, force**

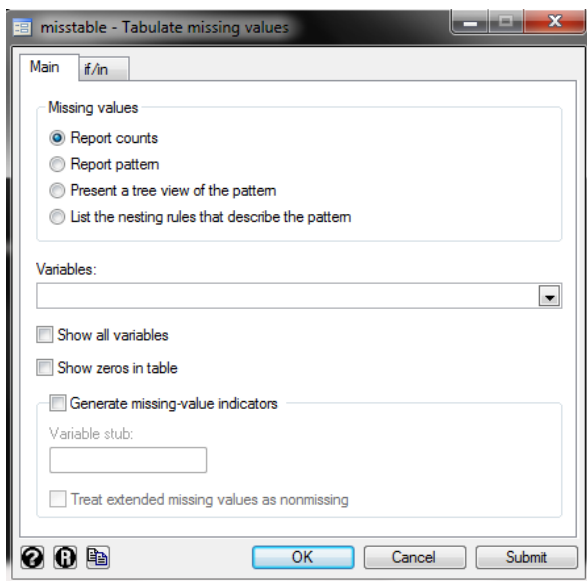
2- Traitement de données manquantes

Les données manquantes sont des non réponses issues de la phase de collecte ou de la phase de numérisation des données. En statistique, il convient de traiter ces valeurs manquantes (les identifier puis les corriger) dans la mesure du possible.

a- Recherche de données manquantes

La recherche de valeurs manquantes peut être réalisée sur toute la base de données ou sur un groupe de variables ou encore sur une variable de la base de données. Toutefois avant de choisir comment réaliser cette recherche, il convient de s'appropriier la logique que doivent respecter ces données afin d'identifier si une valeur manquante est légale ou non. Une valeur manquante sera légale si elle est le fait d'un saut (donc fait intentionnellement). C'est par exemple le cas si tout individu de moins de 5 ans n'a pas été pris en compte pour une section de la base de données portant sur l'activité économique ou encore sur l'éducation.

Une fois cette connaissance de la base acquise, pour rechercher les valeurs manquantes, il suffit d'aller dans « statistics », puis cliquer sur « summaries, tables and tests » puis sur « tables » et cliquer sur « tabulate missing values » et on obtient le menu ci-après :



Il faut choisir le type d'opération à réaliser. Vous cochez « report counts » pour que Stata fasse un rapport sur le nombre de données manquantes par variable. Dans « variables », il faut indiquer les noms des variables pour lesquelles les vérifications sont souhaitées. Lorsque la liste de variable est vide, alors Stata exécute la recherche pour toutes les variables de la base de données. En cochant « generate missing-value indicator », il est possible de générer des variables (pour chaque variable de la base) qui prennent « 1 » si l'individu a une valeur manquante pour la variable considérée.

Dans l'onglet « if/in » du menu ci-dessus, il est possible d'indiquer pour quel groupe d'individu effectuer la recherche de données manquantes. La définition du groupe se fait selon un ou plusieurs critères.

En cochant « report pattern », Stata vous fait un tableau comparatif pour les valeurs manquantes des différentes variables indiquées dans « Variables ». Pour le cas de deux variables par exemple, Stata vous compilera un tableau comme ceci :

Variable 1	Variable 2	Proportion en (%)
renseigné	renseigné	P1
renseigné	non renseigné	P2
non renseigné	renseigné	P3
non renseigné	non renseigné	100-P1-P2-P3

Cocher « present a tree view of the pattern » permet simplement de changer la présentation des résultats. Ainsi, pour deux variables on aura :

Variable 1	Variable 2
Effectif ou pourcentage de non renseigné	Effectif ou pourcentage de non renseigné
	Effectif ou pourcentage de renseigné
Effectif ou pourcentage de renseigné	Effectif ou pourcentage de non renseigné
	Effectif ou pourcentage de renseigné

La syntaxe est la suivante : **misstable summarize liste_variable** ou **misstable nested liste_variables** ou **misstable patterns liste_variables** ou **misstable tree liste_variables**

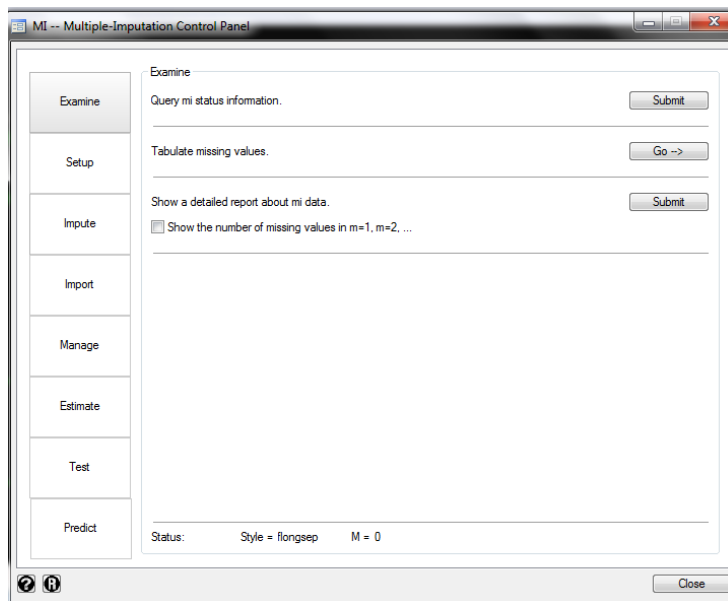
b- Imputation des données manquantes

L'imputation de données manquantes se fait selon plusieurs critères. Le premier réflexe est de faire un retour au questionnaire et vérifier la valeur exacte pour l'insérer dans la base de données. Mais au cas où la donnée n'existe pas dans le questionnaire, il est possible d'imputer en étant le plus vraisemblable possible. Si par exemple on remarque un chef de ménage qui a une épouse et pour qui le sexe n'est pas renseigné, il est vraisemblablement un « homme ». Il est également possible d'imputer par le mode, la moyenne ou la médiane en s'assurant de l'homogénéité dans des individus considérés ; il s'agira alors d'utiliser les techniques de création de variable utilisant les fonction de variables pour générer les variables de moyenne, mode ou médiane.

Pour l'imputation, on peut alors utiliser les techniques de transformation de données (section III-3).

La syntaxe est la suivante : **replace nom_variable = valeur if nom_variable == . & conditions**

D'autres techniques d'imputation existent. Il s'agit notamment des techniques d'imputation multiple (utilisant les régressions). Pour le faire il suffit d'aller dans « statistics », et de cliquer sur « multiple imputation » et on obtient le menu ci-dessus :



A ce niveau, il faut spécifier dans « setup » le type d'opération (c'est-à-dire imputation) et la variable à imputer. Cela se fait dans « add registered variables » où on choisit « type » et « variable » et on fait « submit ». Après cela on va dans « impute » et on choisit le type de modèle, on clique sur « Go » puis on peut spécifier le modèle et faire « OK ».

La syntaxe est la suivante : **mi register imputed nom_variable** puis

mi impute type_regression nom_variable variables_explicatives

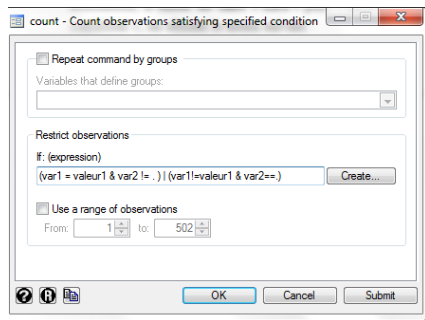
3- Contrôle de cohérence

Les contrôles de cohérences sont d'une importance capitale en statistique. Ils consistent en la vérification des données avant production des statistiques. Ces vérifications font donc partie du processus d'assurance qualité dont le but est de s'assurer que les résultats produits sont d'une très bonne qualité. Il faudra par exemple vérifier que tous les âges déclarés sont positifs, ou que tous les prix déclarés respectent une certaine condition, ou encore que tous ceux qui répondent à une question données en ont le droit ou que tous qui n'ont pas répondu en ont également le droit. Le contrôle de cohérence est immédiatement suivi par la correction des incohérences relevées.

a- Sauts de variables

La vérification pour les sauts se fait généralement par contraposée, c'est-à-dire qu'on s'intéresse aux individus qui ne remplissent pas les différentes conditions et qui ont fait le saut ou aux individus qui remplissent les conditions et qui n'ont pas fait le saut.

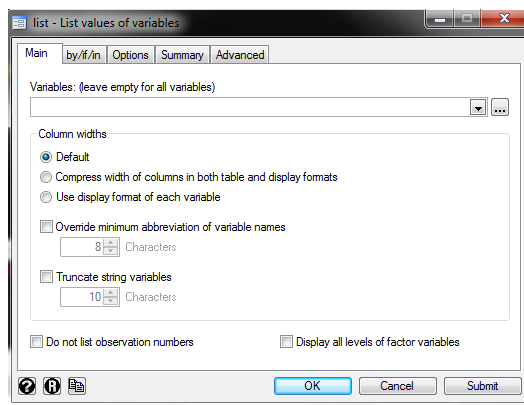
Pour le traitement, on peut dans un premier temps les dénombrer puis les lister. Pour les dénombrer, il suffit de faire « data » puis « data utilities » et « count observations satisfying condition », on obtient le menu suivant :



Il faut préciser la condition dans « if » puis faire « ok ».

La syntaxe est la suivante : **count if (var1 = valeur1 & var2 != .) | (var1!=valeur1 & var2==.)**

Pour lister, il suffit de faire « data » puis « describe data » et faire « list data ». On obtient le menu ci-dessous



Il suffit de lister les variables dont on veut voir apparaître les valeurs puis dans « by/if/in » spécifier les conditions comme précédemment.

La syntaxe est : **liste liste_variable if (var1 = valeur1 & var2 != .) | (var1!=valeur1 & var2==.)** ou **liste liste_variable if (var1 = valeur1 & var2 != " ") | (var1!=valeur1 & var2== " ")**

La correction se fait grâce aux techniques de transformation de données vues en III-3. Il est très important de reprendre exactement les conditions spécifiées pour l'identification des sauts non respectés lors de la correction pour éviter d'écraser des données correctement renseignées. Généralement ces corrections se font soit par retour au questionnaire (pour regarder la vraie information et la renseigner dans ce cas, la syntaxe est **replace nom_variable = vraie_valeur if identifiant == id_individu** ou **replace nom_variable = vraie_valeur in numéro_ligne_individu**) ou par imputation selon la vraisemblance ou par imputation multiple (voir IV-2). Lorsque par exemple on a un individu qui déclare être malade sans dire la maladie, si le questionnaire donne la même information, on peut vérifier la vraisemblance de chaque réponse. Cela consiste à vérifier par exemple s'il a consulté un médecin, vu un pharmacien ou réalisé toute opération que font les malades. Au cas où il n'a fait aucun de ces actes, on dira qu'il n'a pas été malade donc on corrigera la réponse à la

question concernant l'état de santé. Sinon, s'il a réalisé une de ces actions, on devra imputer la maladie soit par le mode, ou par une imputation multiple.

b- Cohérence des données

Lorsque par exemple on se trouve devant un individu de 5 ans qui déclare être médecin, il est dès lors évident qu'il s'agit d'une incohérence. La détection des incohérences se fait par test logique. On peut procéder par dénombrement ou par liste comme dans la sous-section précédente. La condition est alors la suivante : **if (var1 == valeur1 & var2 == valeur_indésirée1) | (var1 == valeur2 & var2 == valeur_indésirée2)**

On parle aussi d'incohérence lorsque les déclarations pour une variable sortent de son ensemble de définition c'est-à-dire pour le sexe par exemple on observe une troisième modalité ou pour une variable définie positive, on observe des valeurs négative. Dans ce cas, la détection se fait par l'ajout à la liste d'une condition qui s'écrit comme suit : **if !inlist(var,modalité1,...,modalitéN)** ou **if !inrange(var,valeur_min,...,valeur_max)**

Une fois l'incohérence identifiée, la correction se fait prioritairement par retour au questionnaire. Lorsque le retour au questionnaire confirme cette l'incohérence, il peut procéder à la correction à l'aide de la vraisemblance. Lorsque par exemple un individu a 5 ans et est médecin, on a une incohérence. Le principe de vraisemblance consiste à rechercher la vraie information (entre l'âge et la profession) et à corriger la fausse information. Si par exemple le niveau d'étude est supérieur, alors il y a de fortes chances c'est l'âge qui est incorrect. Par contre si le niveau d'étude est aucun ou si l'individu n'est jamais allé à l'école, c'est vraisemblablement la profession qui est incorrecte. La correction se fait alors par imputation (voir section IV-2).

V- Statistique descriptive

Dès lors qu'on dispose d'une base de données corrigée, on peut procéder à la compilation de statistiques. Les éléments d'illustrations en statistique sont soit des tableaux soit des graphiques.

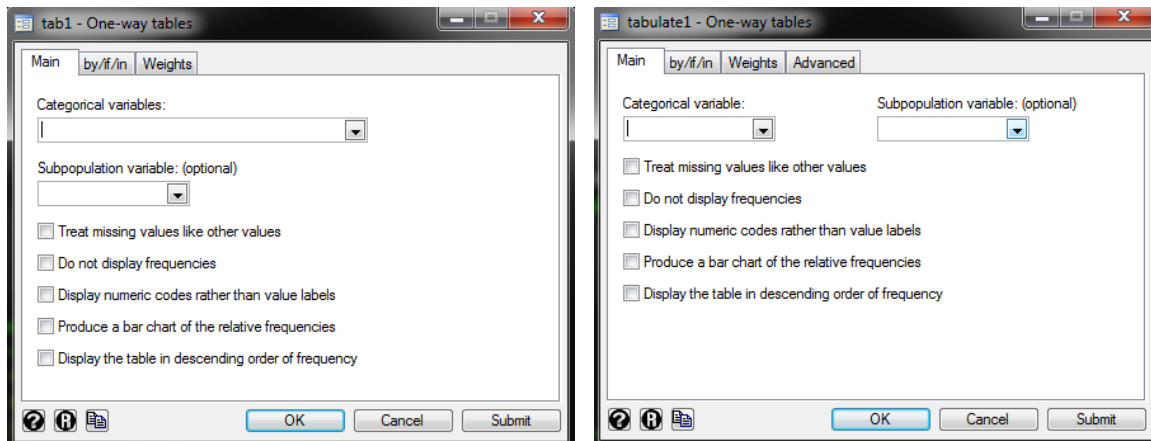
1- Tableaux à une ou plusieurs entrées

Les types de tableaux sont multiples. La présentation ainsi que le remplissage des tableaux dépendent fortement de l'usage ou du but visé. On distingue les tableaux de fréquences simples à une ou plusieurs entrées et les tableaux de statistiques à une ou plusieurs entrées.

a- Tableaux de fréquences

Les tableaux de fréquences peuvent être à une ou plusieurs entrées. En fonction du type d'analyse à réaliser, l'utilisateur pourra choisir entre ces différents types de tableaux. La réalisation d'un tableau à une entre se fait en entrant dans l'onglet « statistics », puis on clique sur « summaries, tables and tests » puis sur « tables » et on choisit « one-way tables » si on

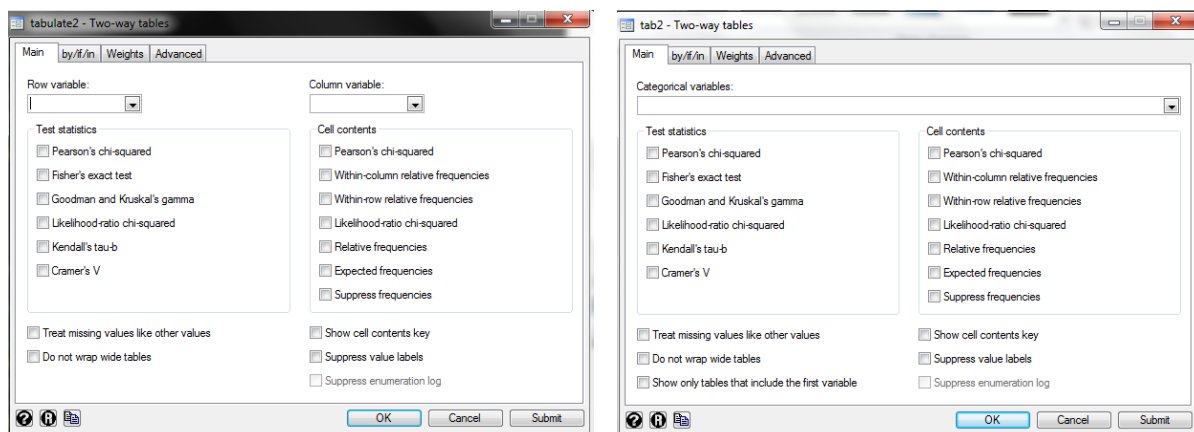
souhaite tabuler une seule variable (schéma à droite) ou « multiple one-way tables » si on souhaite tabuler plusieurs variables (schéma à gauche).



Il suffit ensuite de renseigner la partie « categorical variables » selon le cas avec une ou plusieurs variables, préciser les conditions dans « by/if/in », préciser la pondération dans « weights » au cas où la pondération est non uniforme. Cette pondération devra être entière (valeur dans l'ensemble des entiers naturels). Pour les options, il faut préciser que par défaut, Stata ne compile pas les statistiques sur les données manquantes. Ainsi, dans le tableau construit, l'utilisateur ne sait pas l'existence des valeurs manquantes. Pour afficher les fréquences de ces valeurs manquantes, l'utilisateur peut cocher l'option « treat missing values like others values ». Il est également possible de ranger les modalités en fonction de leurs effectifs en cochant « display the table in descending order of frequency ». Il est également possible d'exclure de la tabulation les individus qui ont une valeur nulle pour une variable donnée. Il faut simplement indiquer le nom de cette variable en « subpopulation ».

Les syntaxes sont : **tab var1, sort plot missing** ou **tab1 var1, sort plot missing**

Pour réaliser un tableau croisé, il suffit d'entrer dans l'onglet « statistics », puis on clique sur « summaries, tables and tests » puis sur « tables » et on choisit « two-way tables with measure of association » si l'on souhaite réaliser un tableau à double entrée (schéma de gauche) en précisant la variable en ligne « row variable » et la variable en colonne « column variable », ou plusieurs tableaux à double entrée (schéma de droite) en précisant dans « categorical variables » la liste de toutes les variables qu'on souhaiterait croiser deux à deux.



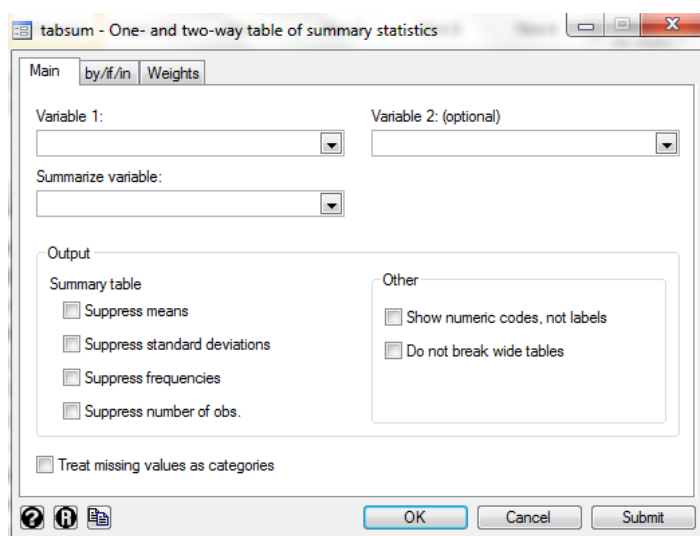
Pour les options, on peut réaliser des tests d'indépendance (voir la partie test statistics) et cocher le test ou la statistique qu'on veut compiler. Dans la partie « cell contents », on peut préciser comment remplir le tableau à afficher (profil ligne ou pourcentage en ligne, cocher « within-row relative frequencies », profil colonne ou pourcentage en colonne, cocher « within-column relative frequencies », fréquence relative par cellule, cocher « relative frequencies », les effectifs théoriques, cocher « expected frequencies ». Ici également, on peut afficher les statistiques sur les valeurs manquantes.

Les syntaxes sont les suivantes : **tab var1 var2, missing row col cel expected** ou encore **tab2 var1 var2 var3 ... varN, missing row col cel expected**.

b- Tableaux de statistiques

Les tableaux de statistiques sont de plusieurs types. Il peut s'agir de réaliser un tableau simple d'une variable qualitative ou croisé de deux ou plusieurs variables qualitatives dont les cellules sont remplies avec des statistiques descriptives (moyenne, écart type, mode, médiane, quantile, erreur type, ...) d'une ou plusieurs variables quantitatives selon les modalités de la variable qualitative, ou encore de réaliser un tableau compilant les statistiques descriptives d'une variable quantitative.

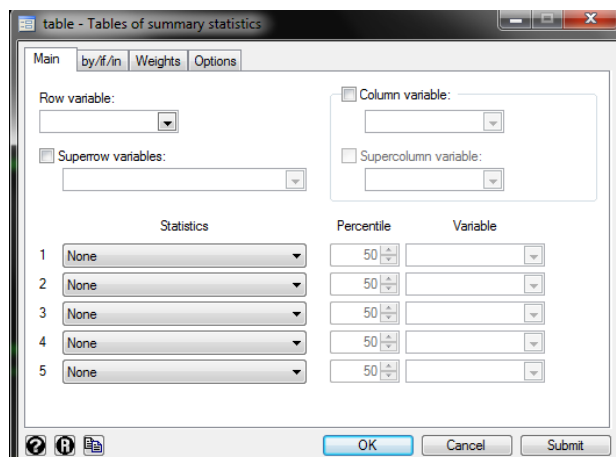
Le cas le plus simple consiste à réaliser un tableau de fréquence simple d'une variable qualitative ou un tableau croisé de deux variables qualitatives dont les cellules donnent la moyenne, l'écart-type et les effectifs de la variable quantitative choisie selon les modalités de la ou des variables qualitatives. Il suffit d'aller dans l'onglet « statistics », puis on clique sur « summaries, tables and tests » puis sur « tables » et on choisit « one/two way tables of summary statistics ». On obtient le menu suivant :



Il faut indiquer dans « summarize variable » la variable quantitative, dans variable 1 la première variable qualitative et dans variable 2 la seconde variable qualitative si on souhaite réaliser un tableau croisé.

La syntaxe est la suivante : **tab var1 var2, summarize(var_quanti)**

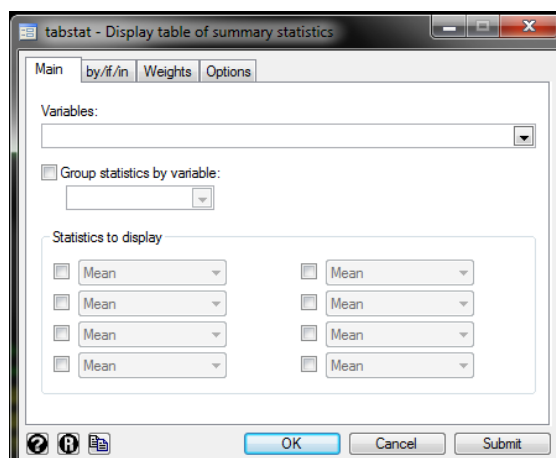
Lorsqu'on veut réaliser un tableau à plusieurs niveaux (plusieurs variables qualitatives à croiser simultanément) ou compiler des statistiques pour plusieurs variables quantitatives selon les modalités d'une ou plusieurs variables, on entre dans l'onglet « statistics », puis on clique sur « summaries, tables and tests » puis sur « tables » et on choisit « table of summary statistics (table) » puis on obtient le menu ci-dessous :



Il suffit de sélectionner les variables qualitatives à positionner selon le cas (row « var1 », column « var2 », supercolumn « var3 », superrow « var4 » var4 pouvant être une liste de variables qualitatives) et de sélectionner une statistique dans la partie « statistics » et définir dans la partie « variable » la variable quantitative à laquelle s'applique la statistique choisie.

La syntaxe est la suivante : **table var1 var2 var3, contents(nom_stat1 var_quant1 nom_stat2 var_quant2 ... nom_stat5 var_quant5) by(var4)**

La dernière option consiste à créer un tableau de statistiques pour une ou plusieurs variables quantitatives. Toutefois, on peut toujours afficher ces statistiques selon les modalités d'une ou plusieurs variables qualitatives. Pour ce faire, il suffit d'aller dans l'onglet « statistics », puis on clique sur « summaries, tables and tests » puis sur « tables » et on choisit « table of summary statistics (table) » et on obtient le menu ci-dessous :



Il suffit de sélectionner dans « variables » la ou les variables quantitatives pour lesquelles on souhaite calculer les statistiques, et dans « statistics to display », les statistiques à compiler.

En cochant « group statistics by variable », on peut afficher les statistiques compilées selon les modalités d'une variable qualitative. Dans les options, on peut définir l'affichage en ligne ou en colonne des statistiques compilées en choisissant dans « use as column » soit variable soit statistics.

La syntaxe est la suivante : **tabstat var1 var2 ... varN, statistics(nom_stat1 ... nom_stat8) by(varM) columns(statistics)**

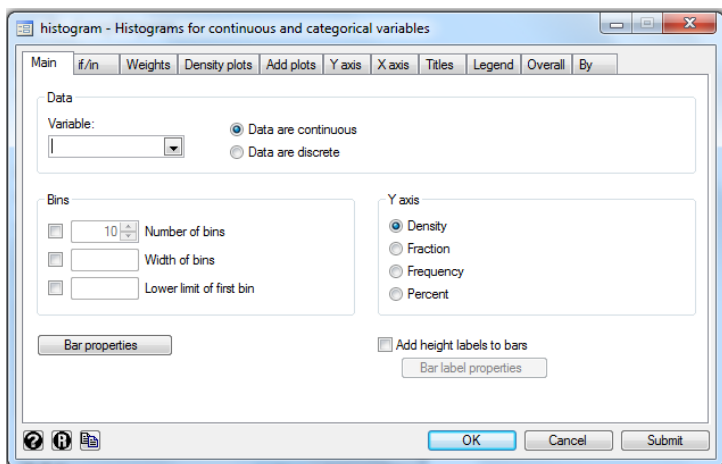
2- Graphiques

Les graphiques illustratifs en statistique sont multiples. Nous nous efforceront dans ce manuel de ne présenter que l'essentiel.

a- Histogrammes

Les histogrammes (variables quantitatives continues) et les diagrammes à bâton (variables quantitatives discrètes) permettent de représenter la distribution d'une variable quantitative. On a la possibilité de représenter la densité empirique de la distribution afin de faire des comparaisons.

Pour réaliser un tel graphique, il suffit d'aller dans le menu et de cliquer sur « Graphics » puis sur « Histogram » et on obtient le menu ci-dessous :



Dans la partie « Variable », on choisit la variable quantitative à représenter puis on coche « Data are continuous » si la variable est continue (histogramme) ou « Data are discrete » si la variable est discrète (diagramme à bâton). Dans la partie « Y axis », on choisit la statistique à afficher en ordonnée (**Frequency** pour les effectifs, **Percent** pour les pourcentages, **Fraction** pour les fréquences en fraction, ou **Density** pour les fréquences corrigées par l'amplitude ; cette option étant l'option par défaut). Ensuite dans la partie « By », on choisit la ou les variable(s) qualitative(s) dont les modalités ou les croisements de modalités permettent de constituer les groupes selon lesquelles on veut dupliquer l'histogramme ou le diagramme à bâton construit. Dans la partie « Density plots », on peut ajouter une estimation de la densité par la méthode des noyaux (kernel density) en cochant « Add kernel density plots »; l'option par défaut étant un noyau d'Epanechnikov. On peut également ajouter la densité de la loi

normale en cochant « Add normal-density plot ». On peut aussi ajouter un nuage de point ou une courbe de régression en cliquant sur « Add plots », le menu obtenu est celui permettant de faire des courbes ; menu décrit dans le paragraphe suivant.

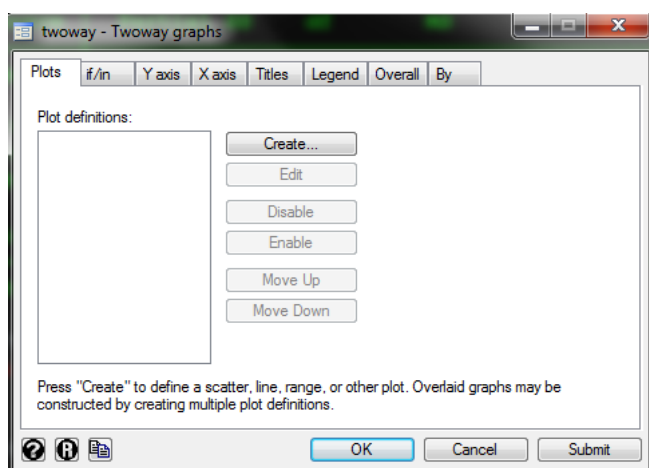
La syntaxe est la suivante :

histogram var_quant, percent normal kdensity kdenopts(parzen) addplot(lfit var1 var2)). L'option « **percent** » peut être remplacé par « **frequency** », « **fraction** », « **density** » (option par défaut) selon ce qu'on souhaite représenté en ordonnée. L'option « **normal** » ajoute la densité d'une loi normale, l'option « **kdensity** » ajoute une estimation de la densité par la méthode des noyaux, l'option « **kdenopts(parzen)** » permet de choisir le noyau pour l'estimation de la densité (on peut choisir epanechnikov, epan2, rectangle, triangle, cosine, biweight, gaussian) et l'option « **addplot(lfit var1 var2)** » permet d'ajouter des courbes ou nuages de points.

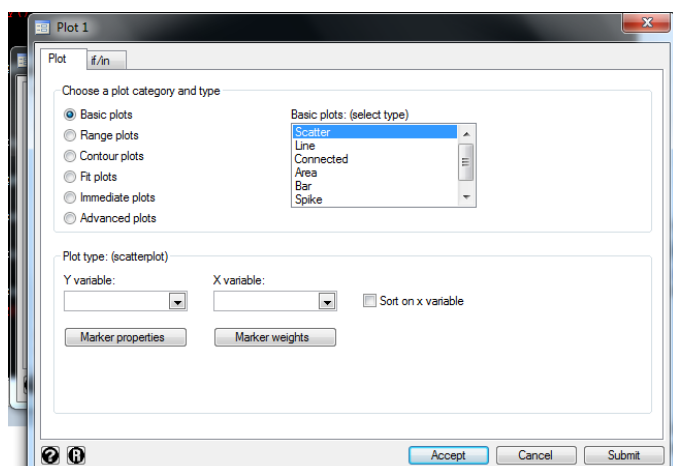
b- Courbes

Les courbes permettent de représenter l'évolution d'une variable quantitative dans le temps ou sur les individus ou encore en fonction d'une autre variable quantitative. Cette représentation peut également autoriser le tracé d'une courbe de régression permettant de mieux approcher la distribution analysée. Les courbes peuvent également être réalisées selon les modalités d'une variable qualitative.

Pour faire une telle représentation, il suffit d'aller dans « Graphics » puis dans « Twoway graph » et on obtient le menu ci-dessous :



Pour réaliser la spécification, il faut ensuite cliquer sur « Create » dans le menu ci-dessus, et on obtient la boîte de dialogue ci-dessous :



En choisissant « Scatter », on construit un nuage de points. On peut également cocher « fit plots » pour choisir des ajustements « Linear prediction », « Quadratic prediction » ou « Fractional polynomial » ou encore l'un de ces ajustements avec un intervalle de confiance. Il faut également noter qu'on peut mettre plusieurs graphiques sur la même représentation. Il suffit pour cela, après la spécification, de cliquer sur « Accept » et de cliquer à nouveau sur « Create » dans la boîte de dialogue initiale. Aussi, a-t-on la possibilité de suspendre un graphique (sélectionner le graphique en question et cliquer sur Disable). Dans l'option « By », on peut choisir la ou les variable(s) qualitative(s) dont les modalités ou les croisements de modalités permettent de constituer les groupes selon lesquelles on veut dupliquer la courbe construite.

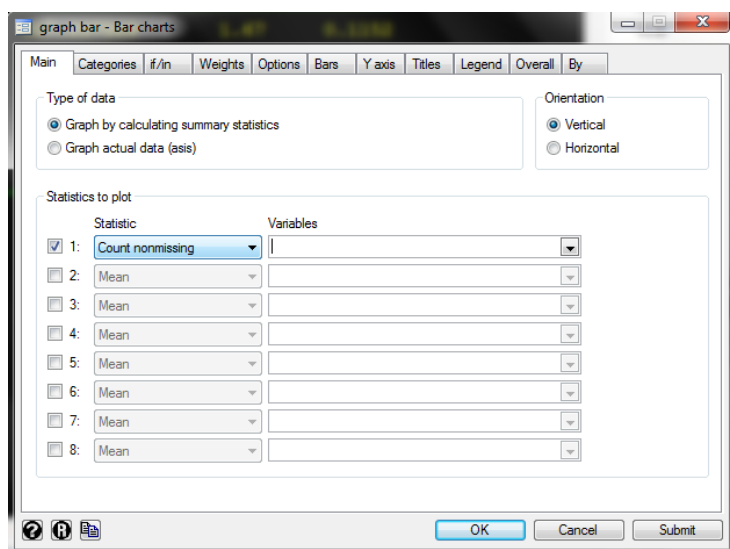
La syntaxe est la suivante :

twoway (qfit var1 var2) (scatter var1 var2) , by(var_qual1 ... var_qualN, total) pour un nuage de point avec un tracé d'une courbe quadratique, pour un tracé linéaire, on mettra « **lfit** » au lieu de « **qfit** » et pour un tracé polynomial, on mettra « **fpfit** ». Pour avoir les intervalles de confiance à 95%, il suffit de faire « **lfitci** » ou « **qfitci** » ou encore « **fpfitci** ». L'option « total » dans « **by(var_qual1 ... var_qualN, total)** » permet d'ajouter un graphique global après les graphiques par groupes.

c- Graphiques à bandes

Les diagrammes à bandes permettent de donner une idée de la distribution d'un ou plusieurs caractères qualitatifs à travers une population donnée. Le cas le plus commun consiste à représenter les effectifs (ou les fréquences en pourcentage) de chaque catégorie de la variable qualitative comme une portion du disque permettant la représentation. Toutefois, il est possible de faire la répartition du disque en fonction d'une caractéristique d'une variable quantitative (moyenne ou médiane par exemple).

La réalisation d'un diagramme à bandes se fait en suivant le menu, dans la partie « Graphics » puis « Bar chart » et on obtient l'écran ci-dessous :



On peut commencer par choisir l'orientation du graphique (vertical/horizontal) dans la partie « Orientation » ; vertical étant l'option par défaut. Ensuite dans la partie « Statistics to plot », il faut choisir la statistique permettant de déterminer la hauteur des bandes à tracer ; l'option « Count nonmissing » donnant les effectifs. Ensuite dans la partie « Categories », on choisit la ou les variables qualitative(s) dont les croisements forment les groupes ; le « Grouping variable 1 » désignant la variable de niveau hiérarchique le plus bas et le « Grouping variable 3 », celle avec le niveau hiérarchique le plus élevé.

La syntaxe est la suivante :

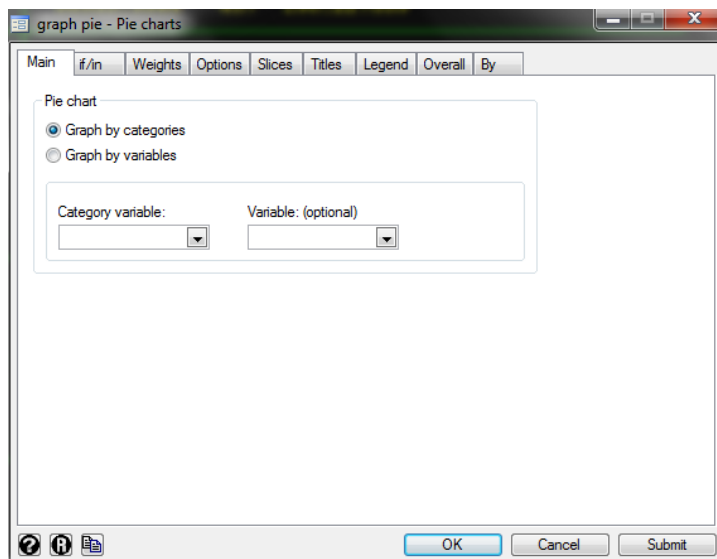
graph bar (statistic1) var1 ... (statisticN) varN, over(var_qual1) ou

graph hbar (statistic1) var1 ... (statisticN) varN, over(var_qual1) pour une orientation horizontale des graphiques construits.

d- Diagrammes circulaires

Les diagrammes circulaires permettent de donner une idée de la distribution d'un caractère qualitatif à travers une population donnée. Le cas le plus commun consiste à représenter les effectifs (ou les fréquences en pourcentage) de chaque catégorie de la variable qualitative comme une portion du disque permettant la représentation. Toutefois, il est possible de faire la répartition du disque en fonction d'une caractéristique d'une variable quantitative (total par exemple).

La réalisation d'un diagramme circulaire se fait en suivant le menu, dans la partie « Graphics » puis « Pie chart » et on obtient l'écran ci-dessous :



Le graph par défaut concerne les catégories d'une variable qualitative qu'on devra choisir dans « Category variable ». Au cas où on souhaiterait faire la représentation selon le total d'une variable quantitative, on choisit la variable en question dans « Variable : (optional) ». Dans le menu « By », on peut choisir une variable qualitative qui permettrait de dupliquer le diagramme circulaire selon les modalités de celle-ci.

La syntaxe est la suivante :

graph pie , over(var_qual) ou

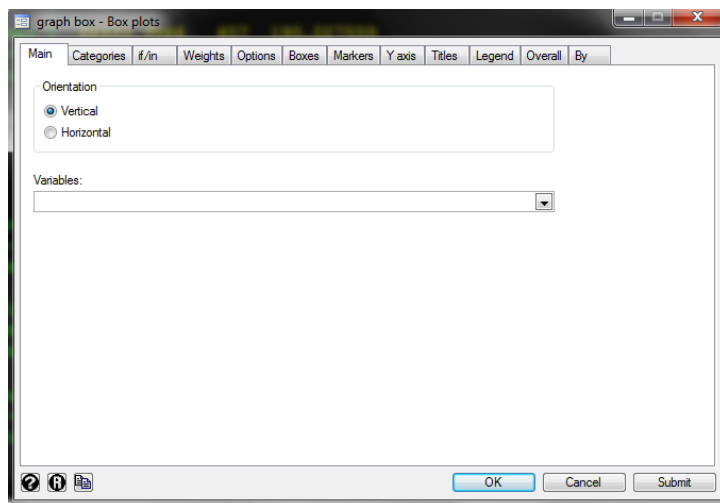
graph pie var_quant, over(var_qual) lorsqu'on souhaite faire la représentation selon le total d'une variable quantitative.

Remarque : On pourrait ajouter après l'option « **over(var_qual)** » dans la ligne de commande ci-dessus, l'option « **by(var_qual1)** » au cas où l'on souhaite dupliquer le graphique selon les modalités de la variable qualitative « **var_qual1** ».

e- Boîtes à moustaches

Les boîtes à moustaches sont des graphiques utiles pour la représentation des chiffres clés portant sur la distribution d'une série ou variable quantitative (exemple : les cours boursiers, l'âge, le revenu, ...). La boîte à moustaches permet également une comparaison aisée de plusieurs groupes au regard des caractéristiques clés d'une distribution quantitative. Elle permet également de comparer plusieurs distributions.

La réalisation de la boîte à moustaches se fait en suivant le menu, dans la partie « Graphics » puis « Box plot » et on obtient l'écran ci-dessous :



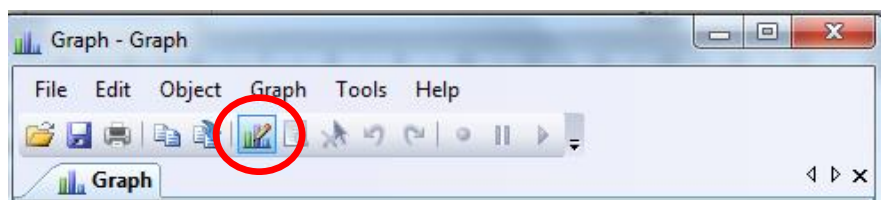
Une fois affichée, on choisit dans « Orientation » le sens des boîtes à moustaches à produire (vertical/horizontal), vertical étant l'orientation par défaut. Puis dans la partie « Variables », on choisit la ou les variable(s) pour lesquelles on souhaite réaliser les boîtes à moustaches. Ensuite dans la partie « Categories », on peut choisir la variable qualitative distinguant les groupes qu'on souhaite comparer. Lorsque les groupes sont formés par croisement de deux variables, il suffit de choisir dans « Grouping variable 1 » et « Grouping variable 2 », les variables qualitatives dont les croisements forment les groupes (On peut également former des groupes par croisement de trois variables. Il suffit de cocher « Grouping variable 3 » puis choisir la troisième variable qualitative). Il faut également remarquer que le regroupement se fait en commençant par la troisième, puis par catégorie de la troisième variable, la deuxième puis la première « Grouping variable ».

La syntaxe est la suivante :

graph box var1 var2 ... varN, over(var_qual1) over(var_qual2) over(var_qual3) ou

graph hbox var1 var2 ... varN, over(var_qual1) over(var_qual2) over(var_qual3) pour les boîtes à moustaches en horizontal.

Remarque : une fois un graphique réalisé, on peut modifier le graphique en ouvrant l'éditeur de graphique de Stata.



Une fois cet éditeur ouvert, on peut ajouter ou modifier les libellés des axes, les couleurs et les épaisseurs des tracés.

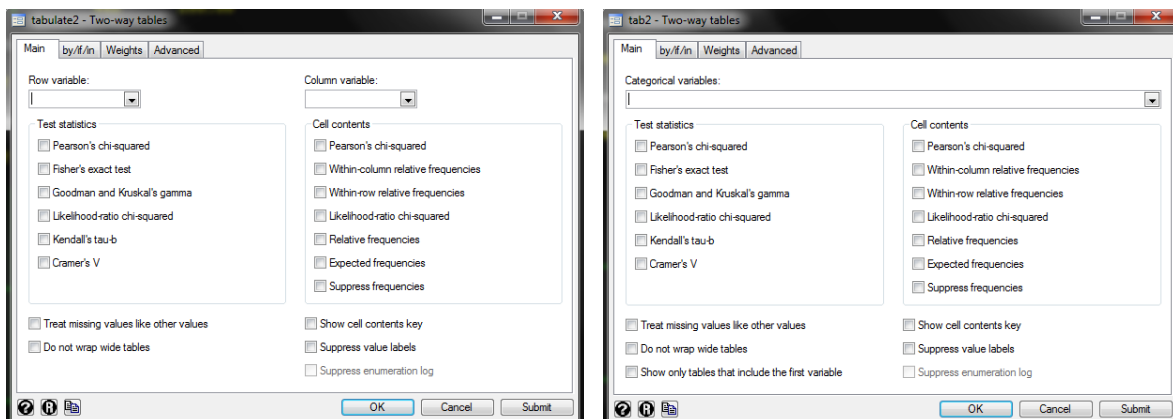
3- Analyse de liaisons entre variables

L'analyse descriptive univariée est certes importante mais ne permet pas à elle seule de cerner tous les contours de l'étude d'un phénomène. Les interactions entre variables permettant de caractériser un phénomène permettent d'affiner l'étude du phénomène. L'analyse de ces interactions se fait selon plusieurs méthodes en fonction du type de variable à analyser. Nous nous intéresserons dans cette section à l'analyse bivariée. L'analyse dans le cas de deux variables de même nature (quantitative ou qualitative) sera exposée dans un premier temps puis on s'intéressera à l'analyse dans le cas de deux variables de natures différentes.

a- Test de khi 2

Le test de khi2 (ou encore chi2) est appliqué lorsque l'utilisateur souhaite tester l'indépendance entre deux variables qualitatives (l'hypothèse nulle étant l'indépendance).

La réalisation de ce test sous Stata passe par la réalisation d'un tableau croisé entre les deux variables (voir section V-1-a de ce manuel). Une fois qu'on a affiché l'un des menus de tableaux croisés, on précise les deux variables (l'une dans row et l'autre dans column) puis parmi les options, on coche « pearson's chi-squared » et on peut faire « OK ». Aussi si l'utilisateur ne souhaite pas faire afficher le tableau croisé, il peut cocher l'option « suppress frequencies » dans les options « cells contents ».



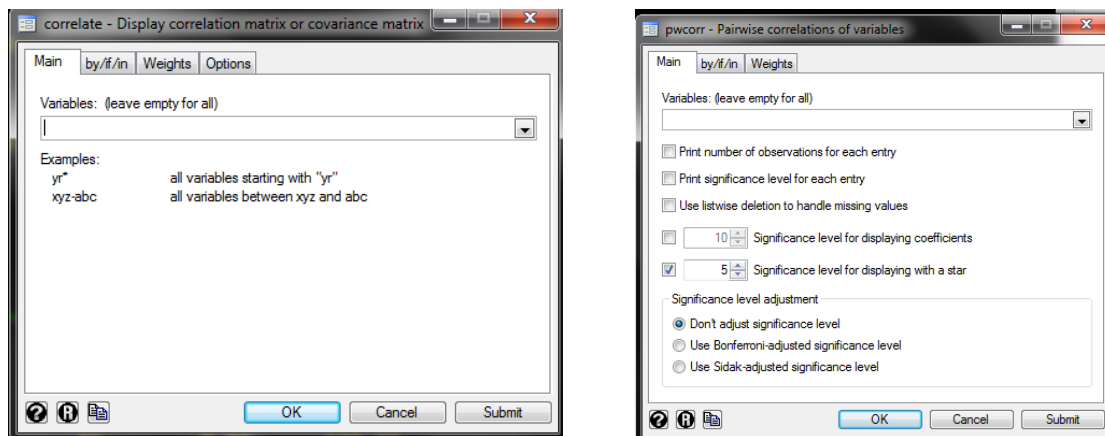
La syntaxe est la suivante : **tab var1 var2, chi2 nofreq** (le « nofreq » après la virgule permet de ne pas afficher le tableau croisé) ou **tab2 var1 var2... varN, chi2 nofreq**

b- Test de corrélation

Le test de corrélation est effectué lorsque nous sommes en présence de deux variables quantitative. On calcule un coefficient de corrélation (qui est compris entre -1 et 1) et qui lorsqu'il est proche de 0, on parle d'indépendance. On peut également tester la nullité de ce coefficient avant de prendre la décision.

Pour réaliser une analyse de la corrélation sous Stata, il suffit d'aller dans le menu « statistics », de cliquer sur « summaries, tables, and tests » puis sur « summary and descriptive statistics » et choisir « correlations and covariances » si l'on veut uniquement calculer les coefficients de corrélation (schéma à gauche) ou choisir « pairwise correlation » si

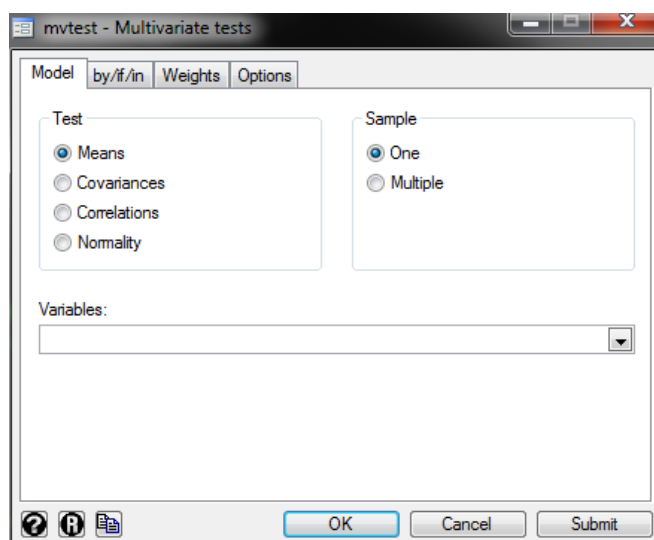
on souhaite en plus des coefficients de corrélation réaliser les tests de nullité de ces coefficients. Dans « variables (leave empty for all) », on dresse la liste des variables pour lesquels on souhaite avoir les corrélations deux à deux puis on fait « OK ».



Dans le second cas (où l'on souhaite réaliser les tests de nullité), il suffit de cocher l'option « significance level for displaying with star » et de choisir le seuil (par défaut de 5%). Stata mettra dans les résultats une étoile à chaque fois que le coefficient de corrélation est significativement différent de 0 au seuil de 5%.

Les syntaxes sont : **corr var1 ... varN** pour le premier cas, ou **pwcorr var1 ... varN, star(5)** pour le second cas.

Il est également possible de tester l'égalité de tous les coefficients de corrélation (entre toutes les variables listées, bien sûr corrélation deux à deux). Il suffit alors d'aller dans le menu « statistics », de cliquer sur « summaries, tables, and tests » et choisir « multivariate test of means, covariances, and normality ». On obtient le menu ci après :



Il suffit alors de lister les variables (dans variables) puis de choisir « correlations » dans « tests » et de faire « OK ».

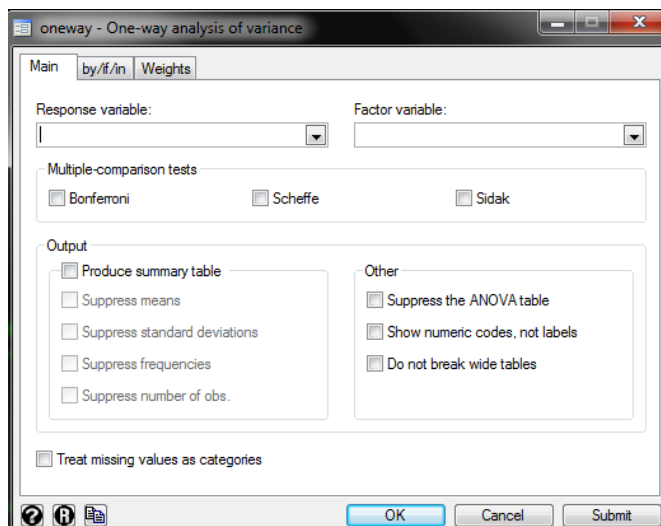
La syntaxe est la suivante : **mvtest correlations var1 ... varN**

c- Analyse de la variance (Anova)

Lorsqu'on est en présence d'une variable quantitative et d'une variable qualitative, on s'intéresse à l'homogénéité de la distribution de la variable quantitative dans les groupes formés par les modalités de la variable qualitative. Cela revient à se demander si quelle que soit la modalité de la variable qualitative, les valeurs moyennes de la variable quantitative sont identiques, dans ce cas on conclut d'une indépendance entre les deux variables (vu que l'appartenance à une classe n'a aucune influence sur la valeur attendue de la variable quantitative). L'hypothèse nulle est l'homogénéité (c'est-à-dire l'indépendance).

Pour réaliser cette analyse dans le cas simple, sous Stata, il suffit d'aller dans le menu « statistics », de cliquer sur « linear models and related » puis sur « ANOVA/MANOVA » puis sur « one-way anova » et on obtient le menu ci-après.

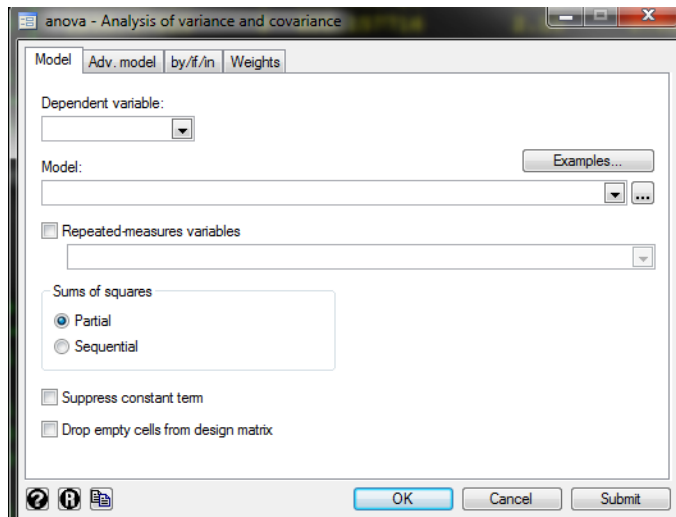
Dans « response variable », on précise la variable quantitative et dans « factor variable », la variable qualitative puis on fait « OK ».



La syntaxe est : **oneway var_quantif var_quali**

L'analyse de la variance (ANOVA) peut être réalisée avec plusieurs variables explicatives (c'est-à-dire qu'au lieu d'une variable qualitative, on peut dresser une liste de variables à la fois qualitatives et/ou quantitatives avec la possibilité d'inclure des interactions entre variables).

Pour réaliser une telle analyse, il suffit d'aller dans le menu « statistics », de cliquer sur « linear models and related » puis sur « ANOVA/MANOVA » puis sur « analysis of variance and covariance » et on obtient le menu ci-après :

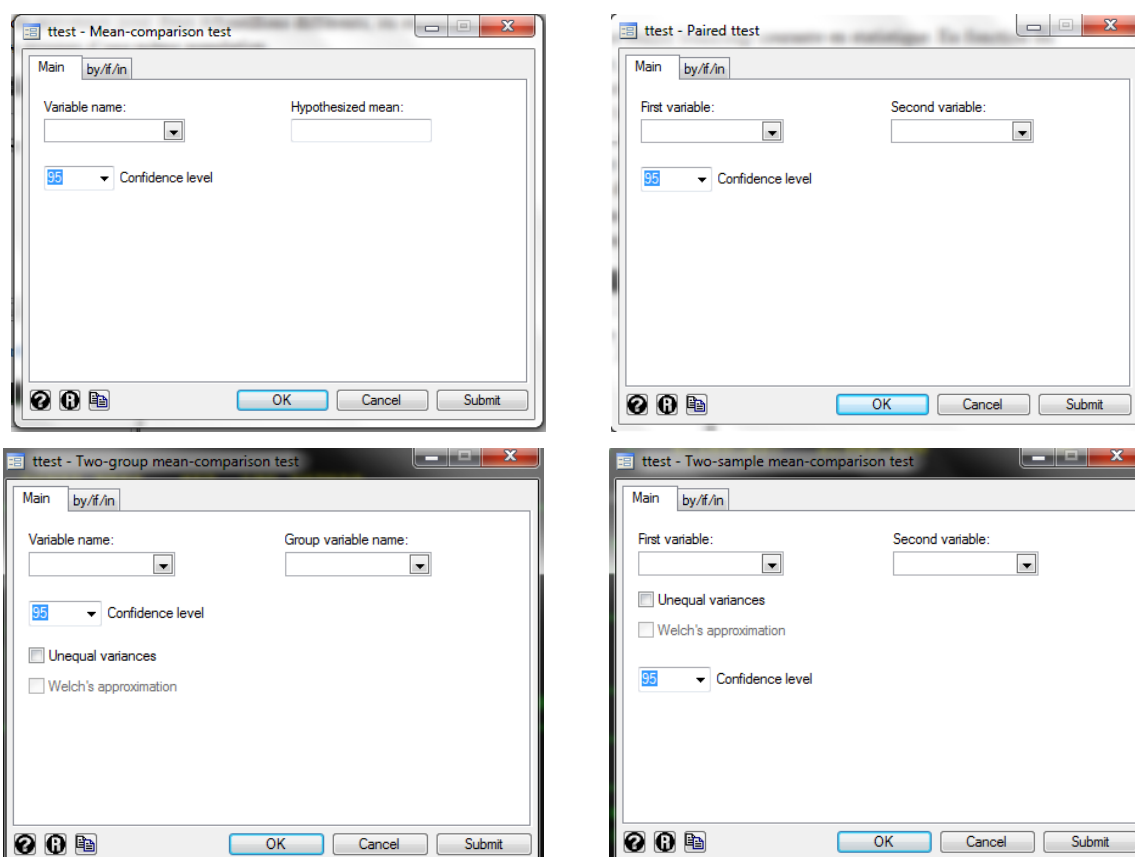


Il faut spécifier la variable quantitative dans « dependent variable » et la liste de variables et d'interactions dans « model ». La définition des interactions se fait cliquant sur les points de suspension devant « model ».

d- Test d'égalité de moyennes

La comparaison d'échantillon est une réalité beaucoup courante en statistique. En fonction du type de variables dont on dispose, l'approche peut changer. Dans le cas d'une variable quantitative, la comparaison d'échantillons se fait en comparant les moyennes. Toutefois, le test de comparaison de moyenne change en fonction de l'hypothèse sur la variance de la variable quantitative dans les différents échantillons.

Pour réaliser un test d'égalité de moyennes, il suffit d'aller dans le menu « statistics », de cliquer sur « summaries, tables, and tests » puis sur « classical test of hypotheses ». Les tests d'égalités de moyenne sont multiples. On peut (i) l'égalité d'une moyenne à une valeur donnée (voir schéma nord-ouest, cliquer sur one-sample mean-comparaison test), (ii) l'égalité de moyennes d'échantillons appariés « échantillon d'une variable observée avec un certain décalage sur les mêmes individus » (voir schéma nord-est, cliquer sur mean-comparaison test, paired data), (iii) l'égalité de moyennes pour deux échantillons différents (voir schéma sud-est, cliquer sur two-sample mean-comparaison test), ou encore (iv) l'égalité de moyennes entre deux groupes d'une même population (voir schéma sud-ouest, cliquer sur two-group mean-comparaison test).



Il faut préciser la ou les variables quantitatives dans (variable name, ou first variable et second variable). Dans le cas de la comparaison de deux groupes d'une même population, il faut préciser le nom de la variable de groupe dans « group variable name » (cette variable de groupe doit avoir deux modalités a priori). Dans le cas de la comparaison de deux groupes ou de deux échantillons, par défaut, l'hypothèse d'égalité des variances est supposée ; toutefois, l'utilisateur peut cocher « unequal variances » pour préciser à Stata que les variances sont différentes (ce qui implique un changement de la statistique de test).

La syntaxe est : **ttest var_quant1 = valeur** (égalité de la moyenne à une valeur) ou **ttest var_quant1 == var_quant2** (égalité de deux moyennes pour échantillons appariés) ou **ttest var_quant1 == var_quant2, unpaired** (égalité de moyennes de deux échantillons différents) ou encore **ttest var_quant1, by(var_quali) unequal** (égalité de moyennes pour deux groupes d'une même population). L'option « unequal » permet de spécifier que les variances sont différentes.

Remarque 1 : dans le cas d'un test d'égalité de moyennes de deux groupes d'un même échantillon, si la variable de groupe a plus de deux modalités, on peut utiliser la condition if pour exclure les modalités pour lesquelles on ne souhaite pas réaliser le test. Dans ce cas, de

sorte à ce qu'il ne reste que deux modalités, on mettra if **!inlist(var_quali, modalité3, ..., modalitép)** dans l'onglet « by/if/in » ou dans la commande avant la virgule.

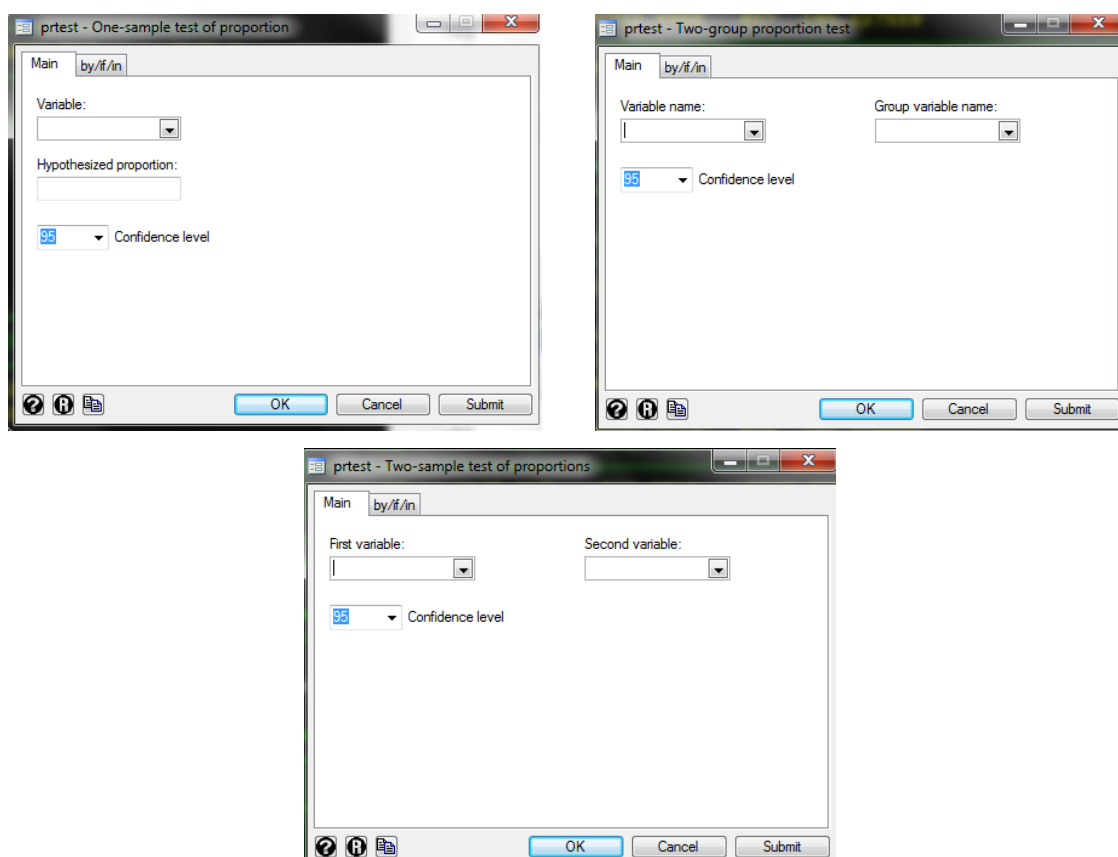
Remarque 2 : on peut réaliser un test d'égalité de variance avant de choisir si l'on doit utiliser l'option « unequal » ou pas. Pour ce faire, il suffit d'aller dans le menu « statistics », de cliquer sur « summaries, tables, and tests » puis sur « classical test of hypotheses » et dans le bloc des variances-tests, on choisit le type de tests d'égalité de variance à réaliser (selon les mêmes descriptions que celles effectuées ci-dessus pour les tests d'égalité de moyenne).

e- Test d'égalité de proportions

Dans le cas d'une comparaison d'échantillon, lorsque la variable d'intérêt est qualitative à deux modalités, on peut comparer les proportions, comme précédemment, soit à une valeur, soit entre elles.

Pour réaliser un test d'égalité de proportions, il suffit d'aller dans le menu « statistics », de cliquer sur « summaries, tables, and tests » puis sur « classical test of hypotheses ». Ici, on dispose de trois possibilités : (i) la comparaison à une valeur (cliquer sur one-sample proportion test, on obtient le schéma nord-ouest), (ii) la comparaison pour deux groupes d'un même échantillon (cliquer sur two-group proportion test, on obtient le schéma nord-est), et (iii) la comparaison pour deux échantillons (cliquer sur two-sample proportion test, on obtient le schéma sud).

Il faut spécifier la variable qualitative d'intérêt dans « variable ». Il est important de signifier que la variable qualitative d'intérêt doit être recodée de manière binaire (0 ou 1) et la modalité 1 doit être la modalité dont vous souhaitez comparer la proportion. Une fois cette étape réalisée, il faut spécifier la valeur (valeur naturellement comprise entre 0 et 1) à laquelle on compare la proportion dans « hypothesized proportion » ; ou la variable de groupe dans « group variable name » (cette variable de groupe devra avoir deux modalités) ; ou la seconde variable à laquelle on compare la première dans « second variable ».



La syntaxe est : **prtest var_quali == valeur** (pour tester l'égalité à une valeur) ou encore **prtest var_quali, by(var_quali2)** (pour tester l'égalité de proportion pour deux groupes), ou **prtest var_quali1 = var_quali2** (pour tester deux échantillons).

Conclusion

Ce manuel est élaboré pour donner quelques éléments clés, commandes et instructions par click, permettant à tout utilisateur d'améliorer sa pratique du logiciel STATA. Il n'est donc pas exhaustif mais permet un usage de base du logiciel.

Plusieurs sources en ligne et en librairie permettent de couvrir d'autres aspects assez diversifiés d'utilisation du logiciel. Toutefois, l'aide du logiciel est un atout majeur, un document de référence très riche et détaillé permettant d'avoir une excellente maîtrise des commandes du logiciel. Il est donc important de s'y référer pour améliorer sa pratique du logiciel.