Manuel d'initiation au traitement de données sous SPSS

Formateur : Sansan Honkounne **KAMBOU,** *statisticien-Economiste*

Septembre 2021

1 INTRODUCTION

1. SPSS est l'un des logiciels pionniers en matière de traitement d'analyse statistique des données, lancé à la fin des années 60 et commercialisé par la société SPSS Inc. Le logiciel SPSS dont le sigle signifie « Statistical Package for Social Sciences » permet de traiter des données dans plusieurs domaines notamment en économie, en science de la santé, en marketing, etc.

2. En 2009, SPSS change de nom et devient « **PASW**» pour « *Predictive Analytics Software* ». Les versions 17.03 à 18.03 de SPSS ont été commercialisés sous ce nom. En 2010, IBM acquiert la société de commercialisation de SPSS (SPSS Inc) et le logiciel change de nom pour « **IBM SPSS**». La version 28 sortie en 2021 est l'édition la plus récente.

3. Depuis la version 14, SPSS a intégré une extension vers python en remplacement des scripts basés anciennement sur Visual Basic. Les versions récentes de SPSS prennent également en charge des extensions de R.

4. SPSS offre une large gamme de possibilité allant de la préparation des données, au traitement et à l'analyse des données. Il permet de répondre à la plupart des problèmes statistiques et son interface est facile d'utilisation :

- ✓ Analyse descriptive ;
- ✓ Analyse graphique ;
- ✓ Inférence statistique ;
- ✓ Modélisation ;
- ✓ Analyse prédictive ;
- ✓ Traitement et analyse des données d'enquêtes complexes ;
- ✓ Etc.

5. La communauté des utilisateurs de SPSS dispose d'un forum dans lequel Modules spécifiques pour traiter des questions spécifiques

2 OBJECTIFS DE LA FORMATION

6. L'objectif de cette formation est d'offrir aux participants les éléments de base pour la préparation et l'exploitation des données en vue de la production des analyses sur la base du logiciel statistique SPSS.

7. Cette formation met l'accent sur l'utilisation des syntaxes en vue d'une meilleure organisation des travaux sous SPSS. Elle passe en revue un ensemble de commandes usuelles et utiles pour la préparation des données et la production des résultats utiles à l'analyse. Les cas pratiques sont basés sur les données de l'enquête régionale intégrée sur l'emploi et le secteur informel (ERI-ESI) réalisée par les pays membres de l'UEMOA en 2017. Il s'agit d'une enquête de type 1-2. La phase I est une enquête sur l'emploi. Cette phase permet le repérage des chefs d'unité de production informelle (UPI) qui sont alors enquêtés lors de la phase II (Enquête sur le secteur informel).

3 GENERALITES SUR LE LOGICIEL SPSS

3.1 APERÇU DES MENUS PRINCIPAUX DE SPSS

Fichier/File	✓ **Ouvrir les fichiers (données, syntaxes, résultats, script) ;** ✓ **Importer des données ;** ✓ **Enregistrer les fichiers ;** ✓ **Etc.**
Edition/Edit	✓ **Copier les variables ;**
Affichage/View	✓ **Afficher les variables ;** ✓ **Afficher les données ;** ✓ **Personnaliser l'affichage des variables ;** ✓ Etc.
Données/Data	✓ **Définir les rôles des variables ;** ✓ **Trier les données ;** ✓ **Trier les variables ;** ✓ **Sélectionner les données ;** ✓ **Agréger les données ;** ✓ **Pondérer les observations** ✓ Etc.
Transformer/Transform	✓ **Transformer les variables ;** ✓ **Créer de nouvelles variables ;** ✓ **Recoder les variables ;** ✓ **Etc.**
Analyse/Analyze	✓ **Réaliser les tableaux ;** ✓ **Modéliser les données ;** ✓ **Analyser les données d'enquêtes ;** ✓ Etc.
Graphiques/Graphs	✓ **Créer plusieurs types de graphiques** ○ **histogrammes,** ○ **boîtes à moustaches,** ○ **courbes,** ○ **etc.**
Utilitaires/Utilities	✓ **Afficher les informations sur les variables du fichier de données ;** ✓ **Configurer les identificateurs OMS ;** ✓ **Etc.**
Extensions/Extensions	
Fenêtre/Window	✓ **Figer les volets (découper la fenêtre en plusieurs cadrants ;** ✓
Aide/Help	✓ **Accéder** : ○ **Au support et forum de SPSS ;** ○ **A la documentation en ligne** ○ **au manuel de référence des commandes SPSS (en local)** ✓

NB : Le menu « Aide » en l'occurrence le « manuel de référence » contient les informations utiles sur l'utilisation de chaque commande de SPSS. Il est conseillé de s'y référer

3.2 LES TYPES DE FICHIERS

Le logiciel SPSS gère quatre principaux fichiers. Il s'agit des :

- **Fichiers de données/ Data** : Les formats de données SPSS sous sauvegardées avec l'extension « .**sav** »
- **Fichiers de syntaxe/ Syntax** : Les listes d'instructions (commandes) sont gérées en SPSS par des fichiers d'extension « **.sps** » ;
- **Fichiers de résultats/Output :** Les sorties de SPSS peuvent être exportées ou copiées vers divers format (MS Excel, MS Word, etc.). Toutefois, pour exploitation de ces résultats sous SPSS, ils devront être enregistrés avec l'extension appropriée «**. spo** »
- Scripts : Depuis la version 14 l'intégration du plug in Python permet d'exécuter des codes en codes sous SPSS. Les scripts sont des programmes en Python d'extension « **.py** » ; « **.pyc** » ou « **.pyo** ».

8. Pour créer l'un des quatre types de fichiers sous SPSS, il faut aller au menu « Fichier » puis « Nouveau » et sélectionner le type de fichier à créer. **Exemple : Fichier/Nouveau/Données va ouvrir une fenêtre de données. Tandis que Fichier/Nouveau/Syntaxe ouvrira une nouvelle de l'éditeur de syntaxe.**

9. Dans la suite de cette formation, nous ferons constamment recours aux fichiers de type « Données » et « Syntaxe » et quelquefois aux fichiers de « Sortie ». L'écriture des scripts va au-delà de la présente formation et nécessite en plus de la maîtrise de SPSS celle de Python.

4 OUVRIR, IMPORTER ET EXPORTER DES FICHIERS DE DONNEES

10. Pour un travail organisé, nous allons créer des répertoires de travail. Pour ce faire, c***réez un répertoire par exemple « C:\Formation SPSS» et copiez-y les dossiers « PrimaryData, SecondaryData ; Program, TempData et OutPut » .***

- **PrimaryData** : Ce répertoire devrait, dans vos pratiques, contenir les données brutes.
- **TempData *:*** Il devra contenir les données temporaires. Il s'agit des fichiers issus de vos manipulations des données primaires mais qui ne vous serviront pas aux analyses. Ce dossier peut être vidé à tout moment.
- **Secondary Data :** Vous veillerez à y sauvegarder les données destinées à vos analyses.
- **Output :** Tous les résultats seront enregistrés dans ce répertoire.
- **Program :** Ce répertoire contient vos fichiers syntaxes

4.1 OUVRIR UN FICHIER DE DONNEES SOUS SPSS

11. Un fichier de données sous SPSS est un fichier d'extension «.**sav** ». Pour ouvrir un fichier de données sous SPSS, aller dans le menu « **Fichier** » puis « **Ouvrir/Données** » et indiquer le nom du fichier ainsi que le chemin d'accès.

12. Après avoir créé le répertoire de travail, ouvrez via l'interface le fichier « individu.sav » se trouvant dans le dossier « PrimaryData ».

13. Le logiciel SPSS offre deux fenêtres pour la visualisation des données (« Vue des données » et « Vue des variables »). La « Vue des variables » donne un aperçu des attributs de chaque variable du fichier de données

- **Nom/Name** : Il s'agit du nom de la variable
- **Type** : Numérique ou Chaîne.
- **Largeur :** largeur de la variable
- **Décimales** : Le nombre de décimales si la variable en comporte ;
- **Libellé** : Il s'agit de l'étiquette de la variable ;
- **Valeurs** : Ce sont les étiquettes des valeurs
- **Manquant** : Permet de définir pour une variable donnée les valeurs qui doivent être traitées comme valeurs manquantes. Par défaut, la valeur manquante pour SPSS est le point pour les variables de « Type » numérique ou « Date ».
- **Colonne** : largeur de la colonne d'affichage de la variable dans la vue des données. Permet donc d'élargir ou réduire la colonne d'affichage de la variable sans aucune incidence sur le format de celle-ci.
- **Alignement** : Alignement à gauche (Left), au centre (Center) ou à droite (Right).
- **Mesure :**
  - **Nominale/Nominal :** Lorsque les catégories de la variable n'ont pas d'ordre défini. **Exemple :** Variable « Sexe » avec les catégories 1 Homme 2 Femme
  - **Ordinale/Ordinal** : Lorsque les catégories de la variable sont établies dans un certain ordre. **Exemple :** Variable « Niveau d'instruction » 1 Aucun 2 Primaire 3 Secondaire 4 Supérieur
  - **Echelle/Scale**: Les catégories de la variable sont dans un ordre avec une métrique de telle sorte que leurs différences ont une signification. **Exemple** : Variable « Age en années révolues »
- **Rôle :** Permet de définir les rôles des variables pour les analyses entre les variables qui jouent le rôle de :
  - « **entrée/Input**» dans les modèles. Comme des variables indépendantes.
  - **« Cible/Target »** ce sont les variables dépendantes
  - **« les deux/Both »** lorsqu'elles jouent à la fois le rôle d'entrée et de cible ;
  - **« aucun/None »** la variable n'a pas de rôle prédéfinie
  - **« partition/Partition »** variable de partition des données pour tester, entraîner ou valider le modèle
  - **« scindée/Split » :** Pour les versions de « IBM SPSS » les variables avec ce rôle ne sont pas utilisées comme telles . ce rôle a été introduit pour « IBM SPSS Modeler ».

14. Par défaut toutes les variables jouent le rôle d'« **entrée** ».

4.2 SAUVEGARDER UN FICHIER SOUS SPSS

15. Pour enregistrer un fichier de données sous SPSS, aller au menu « Fichier » puis choisir « Enregistrer » si l'on souhaite sauvegarder sous le nom déjà existant ou « Enregistrer Sous » si l'on veut donner un nouveau nom au fichier de données ou changer de répertoire. Ce qui permet de laisser ce dernier intouché. Cette dernière commande est très importante dans la mesure où elle laisse le fichier initial intouché. La première par contre altère le fichier initial de façon permanente, ce qui est en général un désastre.

16. En outre la procédure « Enregistrer sous » permet de sélectionner les variables à sauvegarder ou à supprimer selon les besoins de l'analyste.

Exemple 4-1 : Enregistrer le fichier « individu.sav » ouvert dans le répertoire « Temp Data » sous le nom « IndivTemp.sav » en ne gardant que les variables region, hh1, hh2 et celles dont les noms comment par « M ».

4.3 IMPORTER DES DONNEES SOUS SPSS

17. SPSS permet d'ouvrir/importer des données de plusieurs formats en l'occurrence les données (Stata, SAS, Excel etc.). Pour ce faire, aller dans le menu « **Fichier** » puis « **Ouvrir** » et sélectionner le format du fichier à ouvrir dans le menu déroulant « **Fichiers de type** ». La même procédure peut être exécutée à partir du menu « **Fichier** » puis « **Importer des données** ».

Exemple 4-2 : Importez dans SPSS les fichiers suivants se trouvant dans le sous-répertoire « Importation » répertoire « TempData »

a) Le fichier de données Stata « IndivStata.dta » ;
b) Le fichier de données en MS Excel « IndivExcel.xls. ;
c) Le fichier de données en CSV (Données tabulaires séparées par des virgules) « IndivCSV.csv ».

18. Il peut cependant arriver que vous receviez des données dans un autre format fixe (ASCII[1] fixe (.csv), ou texte (.txt ou .dat)) comme dans l'**Exemple 4-3** ci-après.

Exemple 4-3 : Essayez l'importation dans SPSS du fichier « IndivTexte.dat » se trouvant dans le sous-répertoire « Importation » répertoire « TempData ». Quelles conclusions pouvez-vous en tirer ? (Utilisez un éditeur de texte pour visualiser les données).

19. Un tel fichier sans autres informations sur les positions des variables ne peut être exploitée. Il doit être accompagné d'une description de la structure des données. Nous reviendrons plus tard sur la façon de lire ces types de données sous SPSS à l'aide de commande.

4.4 EXPORTER DES FICHIERS

20. Pour exporter des données sous SPSS vers d'autres formats, il faut aller au menu « Fichier », puis « Exporter » et sélectionner le format d'exportation souhaitée (Excel, CSV, Délimité, Texte fixe, stata, etc.). Les fenêtres, qui suivent, permettent de définir ce qui doit être exporté (Noms ou libellés,

[1] American Standard Code for Information Interchange

Valeurs ou étiquettes, etc.) pour certains formats d'exportation comme MS Excel ou ASCII qui ne peuvent conserver à la fois les valeurs et les étiquettes.

21. La procédure d'exportation peut être également réalisée à travers le menu « Fichier » puis « Enregistrer sous » et en choisissant le type de format d'exportation dans le menu déroulant « Enregistrer sous le type ».

Exemple 4-4 : Ouvrez le fichier « Individu.sav » et faire des exportations dans le dossier « Exportation » du répertoire « TempData » suivant les formats ci-dessous :

a) Format de données Stata « IndivStata.dta » ;
b) Format MS Excel « IndivExcel.xls ;
c) Format CSV « IndivCSV.csv ».

22. Lorsque les données contiennent des étiquettes, et les formats d'exportations sont de type ASCII fixe, il vaut mieux exporter les valeurs que les étiquettes et prendre soin de copier ces étiquettes dans un fichier.

Exemple 4-5 : Exportez en ASCII fixe (.dat) le fichier «Individu.sav » dans le dossier « Exportation » du répertoire « TempData » sous le nom « IndivTexte.dat ». Commentez.

4.5 EDITEUR DE SYNTAXE

23. Le logiciel SPSS offre une interface conviviale. Toutefois pour un travail organisé, la liste des instructions peut et devrait être structurée dans un format que le logiciel peut lire et interpréter. En outre, travailler avec les lignes de commandes permet d'économiser un temps précieux, revisiter son travail pour apporter des corrections éventuelles et surtout de gérer des travaux collaboratifs.

24. SPSS à l'instar des autres logiciels offre un éditeur de commandes. Pour l'ouvrir, il faut aller à « Fichier » puis « Nouveau » et choisir « Syntaxe ». Une page blanche s'ouvre et est prête à recevoir vos instructions.

4.5.1 Structure générale des commandes en SPSS

25. Une instruction en SPSS commence par un mot/expression clé. Elle se termine toujours par un point. Au contraire, si nous oublions d'écrire le point de fin, SPSS renvoie lors du traitement un message d'erreur et n'exécute pas la commande. Les versions récentes de SPSS gèrent quelques aspects visuels (couleurs) pour la plupart des commandes usuelles.

Exemple 4-6 : Ouvrez la base de données « Individu.sav » du répertoire « Primary Data » et saisir dans votre éditeur de syntaxe les deux instructions ci-dessous. Quels constats faites-vous ?

```
COMPUTE Region2=Region.
RECODE Region2 (9=2) (ELSE=1) INTO Region3
```

26. Pour une commande ou un ensemble d'instructions, il suffit de sélectionner les lignes correspondantes et avec un clic droit sélectionner « Exécuter la sélection ». Un raccourci est

généralement[2] disponible sous la forme d'un triangle  sur lequel il faut juste cliquer après avoir fait la sélection.

Exemple 4-7 :

a) Sélectionnez les deux commandes précédentes et puis les exécutez. Examinez la vue des données. Quels constats faites-vous ?

b) Ajoutez la commande « EXECUTE. » à la suite des deux autres, puis réexécuter la sélection prenant en compte cette dernière ligne. Commentez et proposez une écriture correcte pour la structure de ces lignes de commandes.

27. Il est conseillé de précéder le point «. » de fin d'une commande ou d'une série de commandes par l'instruction « **EXECUTE** » qui permet à SPSS d'exécuter les transformations pendantes. Certaines instructions comme la suppression de variable par exemple ne s'exécuteront pas tant que des transformations en cours ne sont pas exécutées.

Exemple 4-8 :

c) Exécutez les lignes de commandes suivantes et commentez.

```
COMPUTE Region4=Region.
DELETE VARIABLES Region2 Region3.
EXECUTE.
```

d) Proposez une écriture correcte de ces lignes de commandes

28. Contrairement à certains logiciels statistiques (Stata), SPSS ne fait pas de différence des mots au niveau syntaxique. Autrement dit, SPSS traitera de façon les instructions avec l'une des commandes suivantes :

COMPUTE, compute , Compute, comPute

29. Nous opterons conventionnellement d'écrire nos commandes en majuscules. Nous écrirons donc « COMPUTE » au lieu de « compute », « Compute » ou tout autre forme.

30. Si l'écriture d'une commande doit s'étendre sur plus d'une ligne, seule la dernière portera le point « . » de fin de commande. En outre, il ne doit exister de ligne vide entre les différentes lignes constituant les éléments de la commande.

Exemple 4-9 :

a) La commande suivante s'exécutera sans problème sous SPSS.

```
RECODE hh1 (1 THRU 19=1) (20 THRU 29=2) (30 THRU 39=3) (40 THRU 49=4)
        (50 THRU 59=5)  (60 THRU 69=6) (70 THRU 79=7) (80 THRU 89=8)  (90 THRU 99=9)
        (100 THRU 199=10) (200 THRU 299=11) (300 THRU 399=12) (400 THRU 499=13)
        (500 THRU 599=14)  (600 THRU 699=15) (ELSE=16) INTO hh1Rec1.
```

b) L'exécution de la commande ci-dessus renverra par contre une erreur.

```
RECODE hh1 (1 THRU 19=1) (20 THRU 29=2) (30 THRU 39=3) (40 THRU 49=4)
```

[2] La barre des tâches des personnalisable, il peut arriver que par mauvaise manipulation cette icône soit supprimée. Vous n'avez qu'à la réinitialiser

(50 THRU 59=5) (60 THRU 69=6) (70 THRU 79=7) (80 THRU 89=8) (90 THRU 99=9)

(100 THRU 199=10) (200 THRU 299=11) (300 THRU 399=12) (400 THRU 499=13)
(500 THRU 599=14) (600 THRU 699=15) (ELSE=16) INTO hh1Rec2.

31. Lors de l'écriture de la syntaxe, il est raisonnable de faire attention à la clarté et à la lisibilité (ce qu'on appelle la structure syntaxique). Pour rendre une syntaxe plus claire, il est recommandé de lui adjoindre des commentaires (des notes ou des en-têtes) qui expliquent ce que fait la syntaxe. Les commentaires peuvent être séparées d'une commande ou se mettre à la fin d'une commande.

Exemple 4-10 :

a) Commentaire comme ligne de commande.

*Je fais une copie de la variable Region vers Region5.

COMPUTE Region5=Region.
COMPUTE Region5=Region. /* Region5 est une copie de la variable Region.

32. En tant que ligne de commande séparée, le commentaire débute par * et se termine par le point de fin « . » comme pour toute ligne de commande. Un commentaire peut s'étendre sur plusieurs lignes, il doit cependant se soumettre aux règles du §30 .

33. C'est également une bonne idée de faire des indentations des instructions pour une bonne lisibilité, c'est-à-dire mettre en retrait les commandes subordonnées à une autre commande.

34. Les lignes de commandes sous SPSS sont enregistrées avec l'extension « .sps ». Toutes les instructions contenues dans un fichier d'extension « .sps » peuvent être réexécutées ce qui n'est pas le cas si ces instructions avaient été exécutées à travers l'interface du logiciel. Pour enregistrer le contenu de l'éditeur de syntaxe, aller à « Fichier » puis « Enregistrer » ou « Enregistrer sous » et choisir le répertoire dans lequel l'on souhaite sauvegarder le fichier et saisir un nom. On peut également utiliser le raccourci « CTRL+S » ou cliquer sur l'icône dans la barre des icônes.

4.5.2 Définition des répertoires

35. Un répertoire est défini par le chemin qui permet d'accéder à son contenu. Sous SPSS, on peut définir le travail par défaut avec la commande CD (Change Directory). La commande définit le dossier « Primary Data» se trouvant dans le répertoire « Formation SPSS » qui se trouve sur le bureau de l'utilisateur « KSHC ».

CD "C:\Users\KSHC\Desktop\Formation SPSS\PrimaryData".

36. On peut alors enregistrer ou ouvrir se trouvant dans ce répertoire sans avoir à indiquer le chemin complet d'accès. La commande ci-dessous ouvrira le fichier « Individu.sav » se trouvant dans PrimaryData c'est ce répertoire qui est défini comme répertoire de travail par défaut.

GET FILE="Individu.sav".

37. Toute instruction faisant référence à un fichier se trouvant hors du répertoire par défaut doit explicitement indiquer le chemin d'accès du fichier en question. La commande « **FILE HANDLE** » permet de définir et nommer les répertoires.

Exercice 4.1 : Description des répertoires

a) Ouvrez une nouvelle page de l'éditeur de syntaxe, et en utilisant la commande « FILE HANDLE », définissez les répertoires suivants :

WorD : Formation SPSS

PrimD : PrimaryData

SecD : SecondaryData

TempD : TempData

ProgD : Program

ResultD : Output

b) Ajoutez les commentaires suivants à vos lignes de commandes respectives.

Répertoire de formation SPSS

Données primaires

Fichiers d'analyse

Fichiers temporaires

Fichiers programmes

Fichiers résultats

c) Enregistrez le contenu de l'éditeur dans le répertoire « Program » Sous le nom « **Prg01, Aperçu General sur SPSS** »

d) En utilisant la commande « **GET FILE** », ouvrir le fichier de données « Individu.sav » du répertoire « PrimaryData » (se référer à l'aide pour mieux comprendre la commande et les options disponibles). Reprendre l'**Exemple 4-1** en lignes de commandes à l'aide de « **SAVE OUTFILE** »

e) Avec les commandes « **GET STATA** » et « **GET DATA** » reprendre les points a) et b) de l'Exemple 4-2.

f) Réaliser les exportations de l'Exemple 4-4 en utilisant la commande « **SAVE TRANSLATE** ». Reprendre également l'**Exemple 4-5** en utilisant commande « **WRITE** ».

g) Reprendre l'**Exemple 4-3** avec les informations complémentaires sur la structure du fichier du Tableau 4-1 en utilisant la commande « **DATA LIST** ». Auriez-vous pu importer ces données avec la commande « **GET DATA** » ?

Tableau 4-1 : Structure du fichier « individu.dat »

Variables	Position (N° Ordre)	Libellé	Longueur	Type
region	1.	Régions	2	Numérique
hh1	2.	Grappe (Séquentiel)	4	Numérique
hh2	3.	N° ménage	3	Numérique
m1	4.	M1. N° d'ordre	2	Numérique
m2	5.	M2. Lien avec le chef de ménage	1	Numérique
m3	6.	M3. Sexe	1	Numérique
m4	7.	M4. Quel âge aviez-vous lors de votre dernier anniversaire ?	2	Numérique
m5	8.	M5. Lieu de naissance	2	Numérique
m6	9.	M6. Nationalité	2	Numérique

Variables	Position (N° Ordre)	Libellé	Longueur	Type
M7	10.	M7. Présence d'un ou de plusieurs handicaps	4	Alphanumérique
m8a	11.	M8a. Au cours des 12 derniers mois, durant combien de mois avez-vous vécu dans	1	Numérique
m8b	12.	M8b. Si moins de 6 mois, pendant combien de mois comptez-vous rester dans le ménage	1	Numérique
m9	13.	M9. Avez-vous passé la nuit dernière dans le ménage ?	1	Numérique
m10	14.	M10. Depuis combien d'années vivez-vous dans cette région de manière continue ?	2	Numérique
m11	15.	M11. Dans quelle région habitiez-vous avant de venir ou de revenir dans cette ré	2	Numérique
m12	16.	M12. Pourquoi êtes-vous venu ou revenu dans cette région ?	1	Numérique
m13	17.	M13. Avez-vous déjà été à l'école ?	1	Numérique
m14	18.	M14. Quel type d'école avez-vous fréquenté pour la dernière fois ?	1	Numérique
m15	19.	M15. Allez-vous actuellement à l'école (année scolaire 2016-2017) ?	1	Numérique
m16a	20.	M16a. Quel est votre niveau d'étude actuel ?	1	Numérique
m16b	21.	M16b. En quelle classe êtes-vous actuellement ? (Convertir en nombre d'année d'	2	Numérique
m17	22.	M17. Pourquoi avez-vous arrêté vos études ?	2	Numérique
m18	23.	M18. Aviez-vous fréquenté l'école au cours de l'année scolaire 2015-2016 ?	1	Numérique
m19a	24.	M19a. Quel était votre niveau d'études ?	1	Numérique
m19b	25.	M19b. En quelle classe étiez-vous ?	2	Numérique
m19c	26.	M19c. Avez-vous été admis en classe supérieure ?	1	Numérique
m20a	27.	M20a. Quel niveau d'enseignement avez-vous atteint (en rapport avec la dernière	1	Numérique
m20b	28.	M20b. Quelle est votre dernière classe suivie avec succès ?	2	Numérique
m21	29.	M21. Diplôme le plus élevé obtenu ?	2	Numérique
m22	30.	M22. Pourquoi n'avez-vous pas été à l'école ?	2	Numérique
m23	31.	M23. Savez-vous lire et écrire une phrase dans l'une de ces langues suivantes ?	6	Alphanumérique
m25	32.	M25. Situation matrimoniale	1	Numérique
m24	33.	M24. Quelle langue parlez-vous principalement à la maison ?	1	Numérique
fp1	34.	FP1. Avez-vous suivi une formation ?	1	Numérique
fp2	35.	FP2. Qui vous a formé ?	2	Numérique
fp3	36.	FP3. Quel est le dernier type de formation avez-vous suivi ?	3	Numérique
fp4	37.	FP4. Combien d'années a duré ou aura duré votre apprentissage (9 pour plus de 9	1	Numérique
fp5	38.	FP5. Quel type d'apprentissage avez-vous suivi pendant votre formation ?	1	Numérique
fp6	39.	FP6. Votre formation est-elle en cours ou déjà achevée ?	1	Numérique
fp7	40.	FP7. Exercez-vous actuellement le métier que vous avez appris ?	1	Numérique
fp8	41.	FP8. Pourquoi avez-vous changé de métier ?	1	Numérique
fp9	42.	FP9. Pourquoi n'avez-vous pas encore ou pas du tout exercé le métier de base ?	1	Numérique

5 GERER ET TRANSFORMER LES VARIABLES

38. Les données ne sont pas très souvent disposées selon les besoins immédiats de l'analyste. Très souvent, l'analyste est appelé à réorganiser les données pour qu'elles correspondent à ses besoins. Parmi ces travaux préliminaires, on peut noter entre autres l'étiquetage des variables, le recodage, la génération de nouvelles variables à partir de celles existantes etc.

5.1 Etiquetage des variables

39. Il est plus aisé, en matière de gestion des données et particulièrement en ce qui concerne les données d'enquêtes statistiques, de travailler avec des codes que du texte. C'est ainsi dès la collecte, nous préférions codifier les réponses aux différentes questions. C'est ainsi par exemple que pour enregistrer l'information sur le sexe de l'individu, nous allons créer des codes pour les différentes modalités possibles (Homme, Femme). Dans le fichier « individu.sav » la variable M3 renseigne sur le sexe de l'enquêté. Ses valeurs sont pourtant 1 ou 2.

40. Pour permettre à la variable M3 d'être plus explicite c'est-à-dire d'être plus parlant, il faut lui adjoindre les étiquettes. Pour une variable donnée, on distingue deux types d'étiquettes à savoir l'étiquette de la variable qui est le libellé de la variable et les étiquettes de valeurs qui représentent les libellés des modalités de la variable.

5.1.1 Etiquette de variable

41. Dans le cas précis le libellé de la variable M3 est « Sexe de l'enquêté ». SPSS permet avec une grande flexibilité de décrire les variables. Pour ce faire, aller dans la **« vue des variables »,** se positionner dans la colonne **« Libellé »** puis saisir le l'étiquette de la variable sur la ligne correspondante à son nom dans la colonne « **Nom** ».

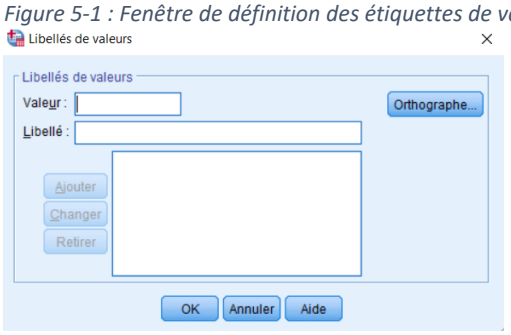
42. Les étiquettes peuvent également être ajoutées aux variables sous SPSS à l'aide de lignes de commandes « **VARIABLE LABELS** »

```
VARIABLE LABELS M3 "Sexe de l'enquêté ".
EXECUTE.
```

5.1.2 Ajout d'étiquettes de valeur

43. Pour les étiquettes de valeurs, il faut plutôt se positionner la colonne « **Valeur** ». Une fenêtre apparaît comme ci-après dans la Figure 5-1. Saisir les valeurs et les étiquettes correspondantes à chaque valeur en cliquer sur le bouton « **Ajouter** ». Lorsque les étiquettes auront été définies pour toutes les valeurs, valider le processus en cliquant sur le bouton « OK »

Figure 5-1 : Fenêtre de définition des étiquettes de valeurs




```
VALUE LABELS M3 1"Homme " 2"Femme".
EXECUTE.
```

Exemple 5-1 : Etiquetage des variables et des valeurs

a) Ouvrir la base de données « IndivNoLab.sav » du repertoire « TempData » et ajouter à travers l'interface, les étiquettes à la variable M3. Ajouter également les étiquettes de valeurs. (1 pour Homme et 2 pour Femme).

b) Créer une nouvelle page de syntaxe que vous enregistrerez sous le nom « Prg02, Manipulation des donnees.sps » dans le répertoire indiqué. Reprendre en tête, la définition des répertoires suivant le programme 1.

c) Ajouter les lignes de commandes requises au point a).

d) En vous aidant de MS Excel, ajouter les étiquettes (variable et valeur) au fichier « IndivNoLab.sav »

e) Le statisticien d'enquête vient de s'apercevoir qu'il a oublié les étiquettes pour valeurs de la variable « Region ». Il s'agit des valeurs 8 et 9 qui ont respectivement pour étiquettes « Kidal » et « Bamako ». En utilisant la commande « **ADD VALUE LABELS** », compléter les étiquettes de valeurs pour la variable région. (Pour se rendre compte de cet oubli, Faites une fréquence sur la région. Assurez-vous d'avoir activé « Valeurs et libellés » dans les paramètres de sortie.

Remarque 5-1 : Edition du dictionnaire des variables

44. Les étiquettes de variables et de valeurs peuvent être visualisées à partir de la « vue des variables ». A travers des lignes de commandes utilisant « **DISPLAY DICTIONNARY** »

```
DISPLAY DICTIONNARY
        /VARIABLES=Region.
```

45. L'exécution de la commande ci-dessus permet de renvoyer les informations sur la variable « Region » y compris ses étiquettes (variable et valeur).

46. Notez également que « **DISPLAY DICTIONNARY** » renverra le dictionnaire pour toutes les variables du fichier actif.

```
DISPLAY DICTIONNARY.
```

5.2 Autres propriétés et attributs des variables

47. Tout comme les étiquettes de variable et de valeurs, les autres propriétés des variables présentes dans la « vue des variables » peuvent être définies soit directement via l'interface soit à travers des lignes de commandes.

5.2.1 Attributs « Type, Largeur, Décimales » :

48. La commande « **FORMATS** » permet de définir à la fois le « **Type** », la « **Largeur** » et les « **Décimales** » des variables.

- ✓ Le « Type » des variables numériques est défini dans « **Formats** » par (**F**) tandis que celui des variables alphanumériques est (**A**).
- ✓ La largeur de la variable est définie par un nombre entier. Il suit immédiatement le type de la variable. **Fw** ou **Aw (w=Width/largeur)**

- ✓ Lorsque la variable est numérique, le format permet de préciser le nombre de décimales au cas échéant. **Fw.d (d=Decimal/décimales)**

FORMATS
/REGION (F1.0)
/M7 (A4).

- ✓ La variable « REGION » est de type numérique et de largeur 1, sans décimale
- ✓ La variable « M7 » (Type de Handicap) est alphanumérique avec une largeur de 4. Naturellement, une variable alphanumérique ne peut avoir de décimale.
- ✓ La largeur d'une variable décimale comprend à la fois celle de la partie entière et le nombre de décimales.

**********Nous n'avons pas de variables décimales dans notre base**********
***La Taille des individus soit mesurée en cm au millimètre près.

FORMATS
/Taille (F3.1).

- ✓ Ce programme n'est pas correct, dans la mesure il attribue une largeur 3 à la taille. SPSS va réserver 2 positions pour la taille en cm et 1 position pour les décimales. Il faut plutôt écrire :

FORMATS
/Taille (F4.1).

Remarque 5-2 : Autres formats

49. En dehors des formats standards (numérique et alphanumérique), il y a des formats spécifiques comme :

- ✓ Les dates : ADATE, SDATE, DATETIME ;
- ✓ Les valeurs monétaires : DOLLAR ;
- ✓ Les Pourcentages : PCT ;
- ✓ Les séparateurs : COMMA (pour la virgule) et DOT (pour le point) ;
- ✓ Custom Currency Formats (CCA).

5.2.2 Attribut « Mesure/Measure »

50. La commande « **VARIABLE LEVEL** » permet de définir la mesure des variables (SCALE, Ordinal ou Nominal).

VARIABLE LEVEL
/REGION (NOMINAL)
/M4 (SCALE)
/M21 (ORDINAL).

- ✓ Il n'y a aucun ordre de classement ou de mesure pour ce qui est des régions. Le classement peut être purement alphabétique ou politique.
- ✓ M4(Age en années révolues). La différence d'âge a bien un sens.
- ✓ M21(Diplôme le plus élevé) : La Maîtrise est un diplôme supérieur à la Licence qui est également supérieur au BAC etc. Mais la différence n'est pas de sens.

b) VARIABLE ALIGNMENT : Alignement des variables dans la « vue des données » trois positions possibles (LEFT, CENTER, RIGHT)

VARIABLE ALIGNMENT
/REGION (LEFT)
/M4 (CENTER)
/M21 (RIGHT).

La variable « REGION » sera alignée à gauche, la variable « M4 » au centre et la variable « M21 » à gauche.

5.3 Transformer les variables

51. L'analyste des données est amené à créer de nouvelles variables qui correspondent à ses besoins à partir des variables déjà existantes. La création de nouvelles variables se fait soit à travers une transformation fonctionnelle ou simplement par regroupement de certaines modalités d'une variable, et ce, dans le but de répondre à une question d'analyse que les variables présentes ne permettent pas.

5.3.1 Créer des variables

52. Pour créer/générer une nouvelle variable, aller au menu « **Transformer** » puis « **Calculer la variable** ». Une fenêtre apparait comme à la Figure 5-2 permettant de :

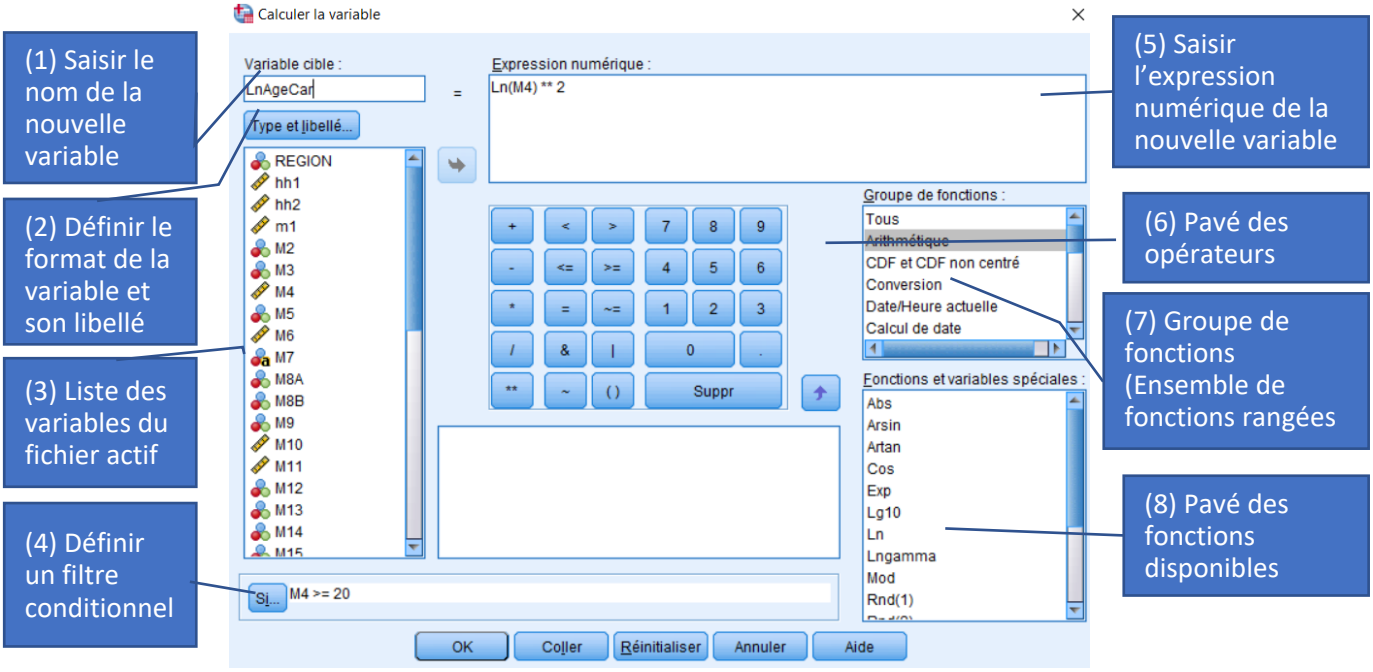
a) saisir le nom de la nouvelle variable (1) ;

b) définir le format (Chaîne ou numérique) de la variable (2) ;

c) définir une condition si la transformation ne s'applique pas à toutes les observations du fichier (3) ;

d) entrer l'expression de la variable comme transformée des variables existantes (4) par combinaison des opérateurs (6) et/ou des fonctions (8).

Figure 5-2 : Fenêtre pour la création de variable



Dans l'exemple qui apparaît sur cette fenêtre, la variable « LnAgeCar » est définie par l'expression Ln(M4)**2 pour les observations dont M4>=20.

53. La création de variable se fait également sous SPSS à l'aide de ligne de commandes. Pour se faire, il faut :

a) Définir le type de variable ainsi que son format

b) Créer la variable à l'aide de la commande « **COMPUTE** » si la transformation s'applique à toutes les observations.

Cas d'une variable de type numérique

```
NUMERIC LnAgeCar (F7.5). /* LnAgeCar est une variable numérique de largeur 7 avec 5 décimales.
COMPUTE LnAgeCar=Ln(M4) **2. /* Création de la variable LnAgeCar.
EXECUTE.
```

Cas d'une variable de type chaîne

```
STRING M7Min (A4). /* M7Min est une variable alphanumérique de largueur 4.
COMPUTE M7Min=LOWER(M7). /* Création de la variable M7Min comme minuscule de M7.
EXECUTE.
```

c) Dans le cas contraire, définir l'expression de la transformation précédée de la condition.

Cas d'une variable de type numérique

```
NUMERIC LnAgeCar (F7.5). /* LnAgeCar est une variable numérique de largeur 7 avec 5 décimales.
IF M4>=20 LnAgeCar=Ln(M4) **2. /* Création de la variable LnAgeCar pour les observations (M4>=20).
EXECUTE.
```

Cas d'une variable de type chaîne

```
STRING M7Min (A4). /* M7Min est une variable alphanumérique de largueur 4.
IF M4>=20 M7Min=LOWER(M7). /* Création de la variable M7Min pour M4>=20.
EXECUTE.
```

Remarque 5-3 : Déclaration de variables

54. La déclaration d'une variable se fait sous SPSS en utilisant les commandes « **STRING** » pour les variables de type chaîne et « **NUMERIC** » pour les variables numériques suivies du format de la variable.

55. Par défaut, une variable non déclarée sera créée comme une variable numérique. En d'autres termes, on ne peut créer une nouvelle variable de type chaîne avec la commande « COMPUTE » sans que celle-ci ne soit préalablement déclarée.

Remarque 5-4 : (Condition IF)

56. Lorsque la condition « IF » est utilisée lors de la création d'une variable, l'expression est évaluée uniquement pour les observations remplissant la condition. Les valeurs des observations ne satisfaisant pas la condition seront « manquant système ».

5.3.2 Quelques opérateurs logiques et relationnels sous SPSS

57. Le pavé (6) fournit un aperçu des opérateurs logiques et relationnels sous SPSS. Le Tableau 5-1 suivant indique le rôle de chacun opérateurs.

Tableau 5-1 : Opérateurs relationnels et logiques sous SPSS

Opérateurs relationnels			Rôle	Opérateurs logiques			Rôle
>	ou	**GT**	Supérieur	**~**	ou	**NOT**	non
<	ou	**LT**	Inférieur	**|**	ou	**OR**	ou
=	ou	**EQ**	Egal	**&**	ou	**AND**	et
>=	ou	**GE**	Supérieur ou égal				
<=	ou	**LE**	Inférieur ou égal				
~=	ou	**NE**	Pas égal ou différent de				

58. Ainsi il est équivalent d'écrire :

Tableau 5-2 : Quelques expressions avec les opérateurs

1.	IF M4**>=20** LnAgeCar=Ln(M4).	IF M4 **GE** 20 LnAgeCar=Ln(M4).
2.	IF M4 **>**10 **&** M4**<**30 LnAgeCar=Ln(M4).	IF M4 **GT** 10 **AND** M4 **LT** 30 LnAgeCar=Ln(M4).
3.	IF M4**~=**20 LnAgeCar=Ln(M4).	IF M4 **NE** 20 LnAgeCar=Ln(M4).
4.	IF M4=10 |M4=20 LnAgeCar=Ln(M4).	IF M4 **EQ** 10 **OR** M4 **EQ** 20 LnAgeCar=Ln(M4).

5.	IF ~ (M4=10 |M4=20) LnAgeCar=Ln(M4).	IF **NOT** (M4 **EQ** 10 **OR** M4 **EQ** 20) LnAgeCar=Ln(M4).

5.3.3 Fonctions sous SPSS

59. SPSS dispose d'une bactérie de fonctions prédéfinies classées par groupe de fonctions dans le quadrant (7). La sélection d'un groupe de fonction au quadrant (7) fait apparaître systématiquement la liste des fonctions disponibles pour ce groupe dans le quadrant (8).

a) **Arithmétique** : Fonctions ABS (Valeur absolue), Ln (Logarithme népérien), Exp (Fonction exponentielle) , les fonctions trigonométriques et réciproques (Cos, Sin, Tan etc.), SQRT (racine) etc.

b) **Statistiques** : Pour le calcul de certaines staistiques. Mean (moyenne), Sum (Somme), Median (médiane), Max(Maximum) etc.

c) **Chaîne** : Il s'agit des fonctions qui s'appliquent aux variables de type alphanumérique. Lower (Minuscule), Upcase (Majuscule), Length (Longueur), Char.index (pour recherche un caractère dans un texte), Char.substr(Extraire des caractères d'un textes à partir d'une position donnée) ; etc.

d) **Conversion :** STRING (pour convertir une expression numérique en texte) ou NUMBER (pour convertir une expression de type chaîne en nombre).

60. Certaines fonctions sont définies par le système. Elles sont précédées par le signe « **$** »

a) **$Sysmis :** valeur manquante du système.

b) **$Casenum** : Renvoie le numéro séquentiel de chaque observation dans la base de données active.

c) **$Date** : Date courante au format international avec l'année deux chiffres (jj-mmm-aa) de largeur 9 (A9). Exemple : **20-SEP-21, pour 20 septembre 2021.**

61. SPSS fournit la définition et le mode d'emploi de chaque fonctionne dans une fenêtre contextuelle c'est-à-dire lorsque celle-ci est sélectionnée.

5.3.4 Recoder des variables

62. Les besoins de l'analyste peuvent l'amener à préférer un regroupement de modalités pour une variable donnée. Par exemple, au lieu de travailler avec l'âge (variable M4), l'on décide d'analyser les classes d'âge.

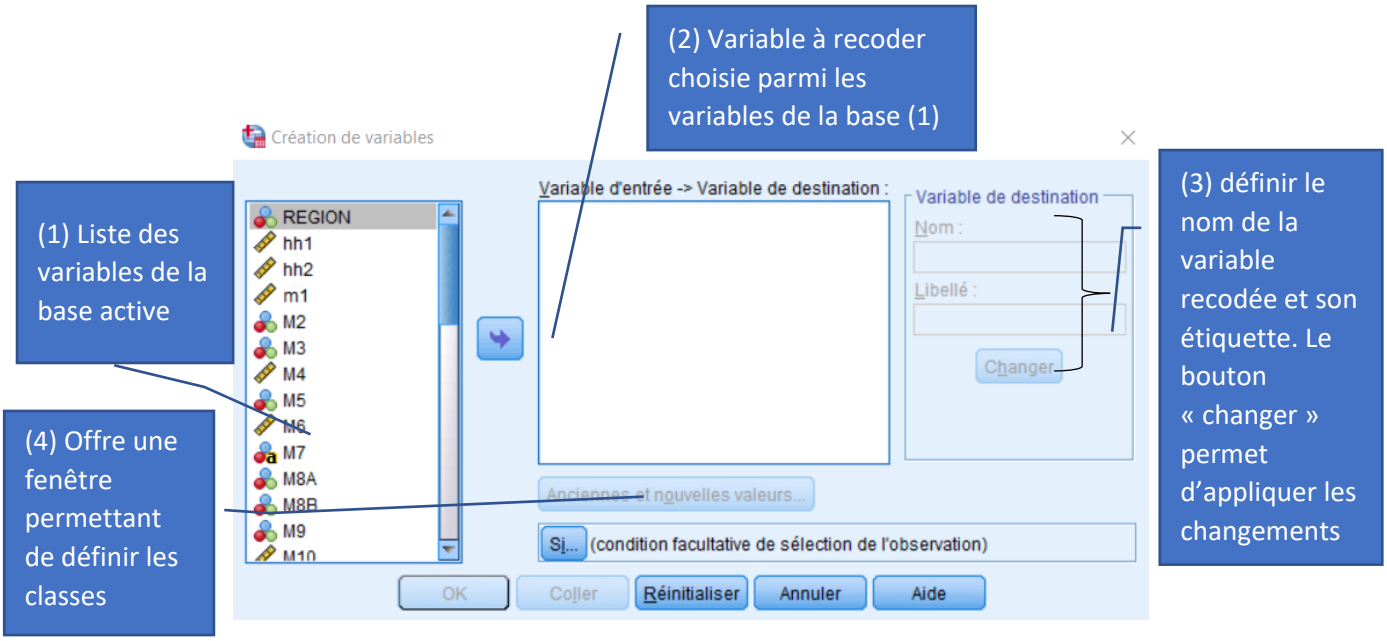
63. Considérons la variable « M4 » du fichier « IndivNoLab.sav» qui correspond à l'âge de l'individu en années révolues. On souhaite étudier certaines caractéristiques des membres de ménages en considérant 7 classes d'âge :

1-Moins de 15 ans
2-Entre 15 et 24 ans
3-Entre 25 et 34 ans
4-Entre 35 et 44 ans
5-Entre 45 et 54 ans
6-Entre 55 et 64 ans
7-65 ans ou plus

64. SPSS permet de recoder une variable sous le même ou créer d'une nouvelle par recodage d'une ancienne.

65. Pour créer une nouvelle variable par recodage d'une autre variable sous SPSS, aller au menu « **Transformer** » puis « **Création de variables** ». Une fenêtre s'ouvre (cf.Figure 5-3)

Figure 5-3 : fenêtre de création de nouvelle variable par recodage



66. Dans cette nouvelle, procéder par les étapes

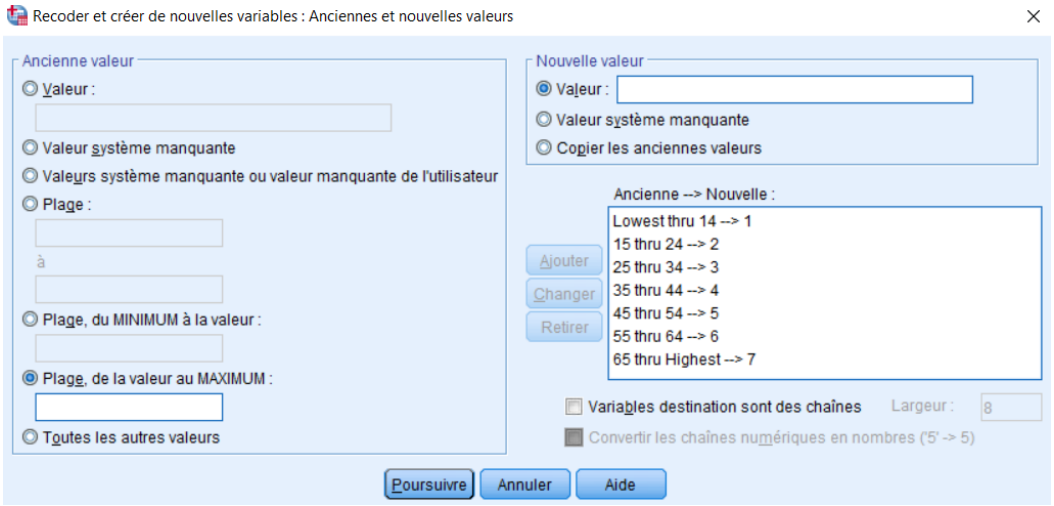
a) Choisir la variable à recoder (variable d'entrée) dans le quadrant (1) vers le quadrant (2);

67. Saisir le nom de la nouvelle variable, son étiquette au cas échéant et valider cette opération en cliquant sur le bouton « **changer** » ;

b) Définir au cas échéant les observations auxquelles s'applique le recodage ;

c) Le bouton « **Anciennes et nouvelles valeurs** » ouvre une nouvelle fenêtre (Cf Figure 5-4) qui permet véritablement de définir les différents codages.

Figure 5-4 : Fenêtre de création des nouvelles valeurs par rapport aux anciennes



68. En ce qui concerne le recodage d'une variable sous le même nom, aller au menu « **Transformer** » puis « **Recodage des variables** ». La fenêtre qui apparaît est de la Figure 5-3, à l'exception du quadrant (3) qui permet de créer la nouvelle variable.

69. Le recodage peut être réalisé à l'aide de la commande « **RECODE** », l'option « **INTO** » permet de recoder une variable sous un autre nom tandis que « **RECODE** » simplement écrase la variable d'entrée.

Exemple 5-2 : Recodage de la variable « M4 » en 7 classes (§63)

Cas de recodage sous un autre nom (création de variables)

```
NUMERIC GrpAge (F1.0).
RECODE M4
        (LOWEST THRU 14=1) (15 THRU 24=2) (25 THRU 34=3) (35 THRU 44=4)
        (45 THRU 54=5) (55 THRU 64=6) (65 THRU HIGHEST=7) INTO GrpAge.
VARIABLE LABELS GrpAge "Classe d'âge "
EXECUTE.
```

Cas de recodage sous le même nom (Cette méthode écrase la variable initiale. A utiliser avec précaution)

*Nous allons créer une copie de la variable « M4 » avec laquelle nous ferons le recodage.

```
COMPUTE M4_Copy=M4.
```

*La variable « Copy_M4 » prendra les valeurs de M4, mais elle n'est pas rigoureusement parlant une copie de M4 dans la mesure où les propriétés (formats, étiquettes, etc.) ne sont pas les identiques.

*Ceux qui disposent de SPSS Python Essentials peuvent installer l'utilitaire de clonage de variables SPSS Clone Variables Tool. La commande ci-après permet de créer une copie parfaite de la variable « M4 ».

```
SPSSTUTORIALS CLONE VARIABLES VARIABLES ='M4' PREFIX ='Copy_'.
*Recodage de « M4 » à partir de la copie « Copy_M4 ».
RECODE Copy_M4
        (LOWEST THRU 14=1) (15 THRU 24=2) (25 THRU 34=3) (35 THRU 44=4)
```

```
(45 THRU 54=5) (55 THRU 64=6) (65 THRU HIGHEST=7).
EXECUTE.
```

5.3.5 *Assistant regroupement visuel*

70. La méthode de recodage utilisée précédemment suppose de connaître préalablement les points de césure (regroupement). Ce qui n'est pas souvent le cas. En effet, supposons que la préoccupation de l'analyste est de créer des classes de déciles, ou quintiles ou tout regroupement autre que ce que nous avons présenté jusque-là. Plutôt que l'analyste détermine les points de césure avant de procéder au recodage, il peut, en utilisant l'assistant « Regroupement visuel/ Visual Binning », laisser le soin à la machine de générer ces points à partir des données pendant le recodage.

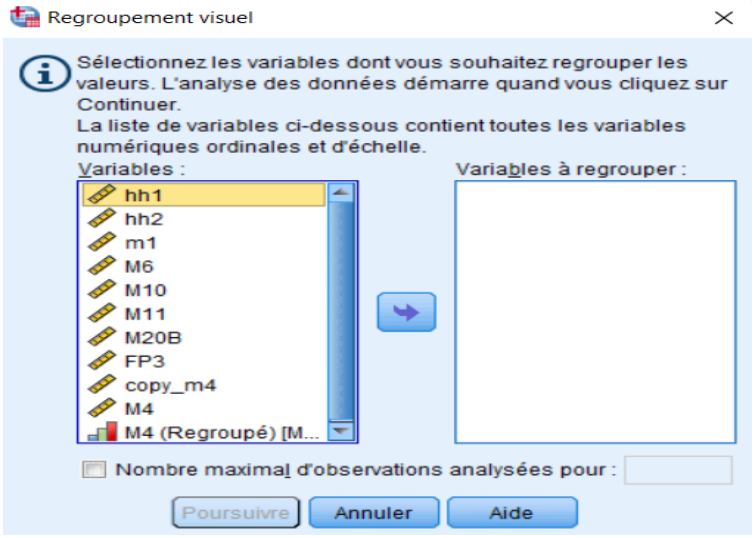
71. Selon la ***Documentation de IBM SPSS Statistics***[3], le regroupement visuel aide à la création de variables basées sur le regroupement des valeurs contiguës de variables dans un nombre distinct de catégories. Le Regroupement visuel peut être utilisé pour :

- Créer des variables catégorielles à partir de variables d'échelle continues. Par exemple, nous pouvons utiliser la variable d'échelle « M4 » (Age) pour créer une variable catégorielle contenant les tranches d'âge.
- Fusionner un grand nombre de catégories ordinales en un jeu de catégories plus petit. Par exemple, vous pouvez fusionner une échelle de classement allant jusqu'à neuf pour obtenir trois catégories qui représenteraient les niveaux faible, moyen et élevé.

72. Pour utiliser l'assistant « Regroupement visuel », aller au menu « Transformer » puis « Regroupement visuel ». La boîte de dialogue de la ci-contre s'affiche et permet de sélectionner parmi les variables d'échelle et ordinales la ou les variables à recoder.

73. Il est donc important de se souvenir que cet assistant n'est utile que pour les variables d'échelle ou ordinale. Le recodage des variables nominales ou de type chaîne doit se faire selon la méthode « RECODE » précédemment expliquée.

Figure 5-5 : 1ère boîte de dialogue de l'assistant « regroupement visuel » (sélection de variables)

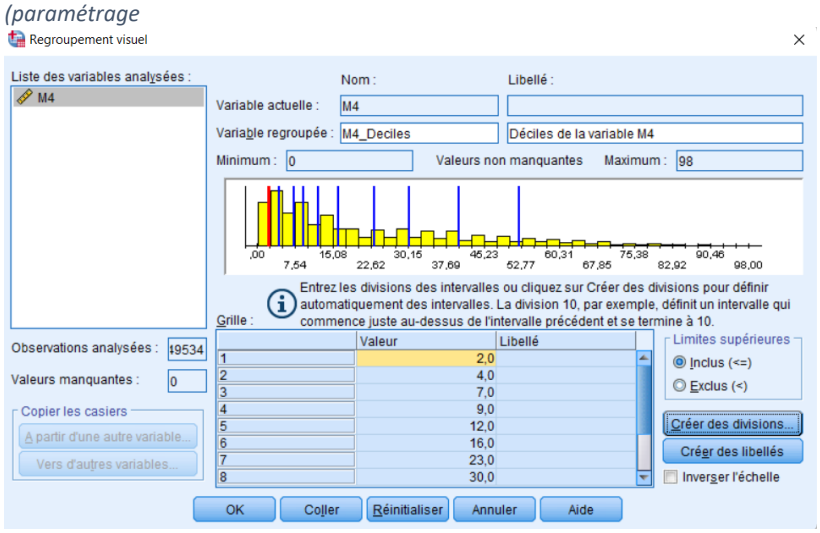


[3] Voir Regroupement visuel - Documentation IBM

74. La sélection d'une variable à regrouper permet d'accéder à la boîte de dialogue de la Figure 5-6 qui permet de paramétrer le mode de regroupement. Le bouton « Créer des divisions » affiche les options disponibles :

- Intervalles de longueur identique
- Percentiles égaux fondés sur les observations analysées
- Divisions au niveau de la moyenne et des écarts types, fondées sur les observations analysées

Figure 5-6 : 2ème boîte de dialogue de l'assistant « regroupement visuel » (paramétrage



75. La deuxième option permet alors en définissant le nombre percentiles de créer une nouvelle variable. Soit M4_Deciles la variable « M4 » recodée suivant les groupes de déciles. Alors M4_Deciles est définie par :

M4_Deciles=1 si (M4<=D1)
M4_Deciles=2 si (D1<M4<=D2)
M4_Deciles=3 si (D2<M4<=D3)
M4_Deciles=4 si (D3<M4<=D4)
M4_Deciles=5 si (D4<M4<=D5)
M4_Deciles=6 si (D5<M4<=D6)
M4_Deciles=7 si (D6<M4<=D7)
M4_Deciles=8 si (D7<M4<=D8)
M4_Deciles=9 si (D8<M4<=D9)
M4_Deciles=10 si (M4>D9)

Où D1, D2, …, D9 sont les déciles de la variable « M4 ».

76. Les lignes de commandes qui suivent constituent la syntaxe créée par l'assistant « regroupement visuel » pour la création des déciles de M4 après paramétrage (9 dividions).

```
* Regroupement visuel.
*M4.
RECODE  M4 (MISSING=COPY) (LO THRU 2=1) (LO THRU 5=2) (LO THRU 7=3)
        (LO THRU 10=4) (LO THRU 14=5) (LO THRU 20=6) (LO THRU 28=7) (LO THRU 38=8)
        (LO THRU 50=9) (LO THRU HI=10) (ELSE=SYSMIS) INTO M4_Deciles.
VARIABLE LABELS M4_Deciles 'Déciles de la variable M4'.
FORMATS M4_Deciles (F5.0).
VALUE LABELS M4_Deciles 1 " 2 " 3 " 4 " 5 " 6 " 7 " 8 " 9 " 10 ".
VARIABLE LEVEL M4_Deciles (ORDINAL).
EXECUTE.
```

77. A travers cette syntaxe nous pouvons déduire les déciles

DECILES	NIVEAU
D1	2
D2	5
D3	7
D4	10
D5	14
D6	20
D7	28
D8	38
D9	50

Remarque 5-5 : Notez que nous n'avons pas eu besoin de connaître les niveaux des déciles. Ils sont déterminés par la machine sur la base des données actives. Il est donc inutile de copier la commande de cet assistant vers d'autre programmes si les jeux de données sont différents.

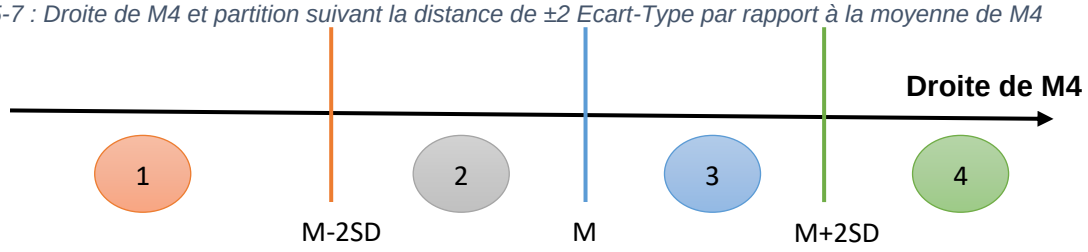
78. Nous pouvons aussi utiliser cet assistant pour créer des groupes basés sur leur distance par rapport à la moyenne où les distances sont exprimées en termes d'écart-type.

```
* Regroupement visuel.
*M4.
RECODE M4 (MISSING=COPY) (LO THRU -17.2401798851367=1) (LO THRU 21.0941373601970=2)
    (LO THRU 59.4284546055308=3) (LO THRU HI=4) (ELSE=SYSMIS) INTO M4_2SD.
VARIABLE LABELS M4_2SD 'Groupes M4 (2 SD)'.
FORMATS M4_2SD (F5.0).
VALUE LABELS M4_2SD 1 '' 2 '' 3 '' 4 ''.
VARIABLE LEVEL M4_2SD (ORDINAL).
EXECUTE.
```

Soit SD l'écart-type (Standard Deviation) de la variable « M4 » et M la moyenne (Mean), alors la variable « M4_SD » est créée telle que :

M4_SD=1 si (M4 <= M- 2*SD)
M4_SD=2 si (M- 2*SD <M4<=M)
M4_SD=3 si (M<M4<=M+2*SD)
M4_SD=4 si (M4>M+2*SD)

Figure 5-7 : Droite de M4 et partition suivant la distance de ±2 Ecart-Type par rapport à la moyenne de M4



5.3.6 Recodage automatique (AUTORECODE)

79. La procédure « AUTORECODE » permet de recoder des variables de type chaîne ou nominales en entiers consécutifs vers une nouvelle variable appelée variable cible (Target variable). Lorsque la

variable d'entrée possède des étiquettes de valeurs, ces étiquettes sont également recopiées dans la nouvelle variable.

80. Pour illustrer cette procédure, nous allons faire quelques transformations sur la variable « M7 » à travers le programme ci-après.

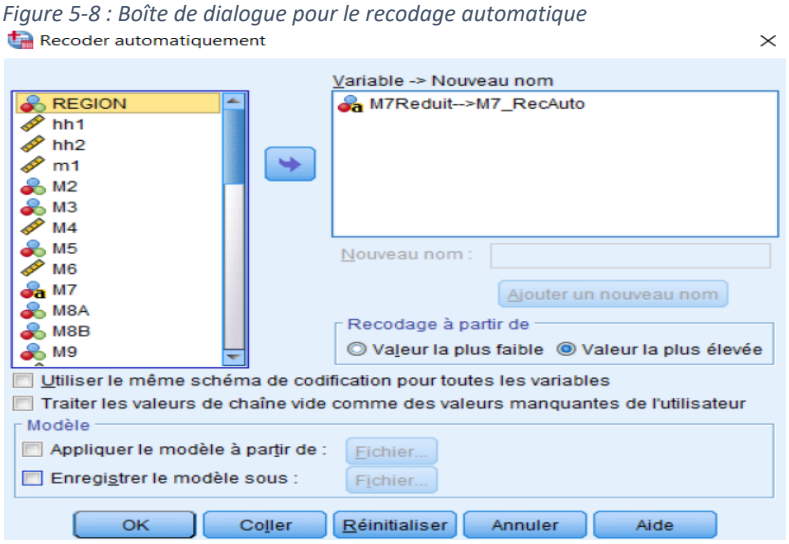
```
STRING M7Reduit (A1).
IF CHAR.LENGTH(M7) =1 M7Reduit=M7.
IF CHAR.LENGTH(M7) =0 M7Reduit=M7.
IF CHAR.LENGTH(M7)>1 M7Reduit='Z'.
EXECUTE.

VALUE LABELS M7Reduit
  'A'  "Sans handicap"
  'B'  "Aveugle/malvoyant"
  'C'  "Sourd/malentendant"
  'D'  "Muet"
  'E'  "Infirme membres inférieurs"
  'F'  "Infirme membres supérieurs"
  'G'  "Malade mental"
  'H'  "Lépreux"
  'X'  "Autre handicap non précisé"
  'Z'  "Plus d'un handicap".
```

81. La variable « M7Reduit » est créée à partir de la variable « M7 » en regroupant les individus qui cumulent plus d'un handicap sous la modalité 'Z'. Ainsi construite, la variable « M7Reduit » est une variable chaîne avec 10 catégories.

82. Pour recoder automatiquement cette variable, aller au menu « Transformer » puis « Recoder automatiquement ». La boîte de dialogue de la Figure 5-8 apparaît permettant de sélectionner la ou les variables à recoder. Il faut absolument indiquer le nom de la variable cible. On peut faire un recodage ascendant ou descendant.

Figure 5-8 : Boîte de dialogue pour le recodage automatique



83. Copier la syntaxe de création de la variable « M7Reduit » à la suite de votre programme. Puis exécuter pour créer la variable « M7Reduit ». Et en suivant la procédure, recoder automatiquement la variable « M7Reduit » en « M7_RecAuto ». Par défaut, la procédure utilise le recodage ascendant.

84. Cette procédure également exécutée à l'aide la commande « AUTORECODE ».

```
AUTORECODE VARIABLES=M7Reduit
  /INTO M7_RecAuto
  /PRINT.
```

Exercice 5.1 : La base de données à utiliser est « IndivNoLab.sav »

85. Les outils de transformation des variables sont nombreux en SPSS à l'instar des autres logiciels. Le présent manuel n'a pas l'ambition de tous les aborder. Cet exercice introduit quelques procédures et fonctions. Le but est d'amener ceux qui souhaitent progresser dans l'utilisation de SPSS à mieux exploiter le document de référence des commandes (**Command Syntax Reference)** de « **IBM SPSS Statistics »** qui accompagne chaque édition du logiciel et disponible dans le menu « **Aide** ». Plus vous vous en référez, plus vous en découvrez et sans doute que les nouvelles découvertes vous feront avancer une bonne maîtrise du logiciel.

86. Plus spécifiquement, cet exercice vous invite à faire un saut dans les structures conditionnelles (**DO IF…END IF**), les structures itératives (**DO REPEAT…END REPEAT** ; **LOOP…END LOOP**) qui n'ont pas été abordées comme sections à part entière.

Remarque 5-6 : Latitude vous est donnée de proposer d'autres solutions n'utilisant pas les procédures/fonction qui vous indiquées mais elles ne devraient intervenir qu'à titre comparatif.

a) Structure conditionnelle « DO IF… END IF » et Fonction « REPLACE »

87. Le responsable de collecte vous signale qu'il y a une mauvaise codification des réponses à la question sur le handicap (variable M7) dans certaines zones. En effet, dans la région 1, les enquêteurs ont interverti les modalités « C : Sourd/Malentendant » et « D : Muet ».

a) Proposez un programme pour corriger cette erreur. Vous prendrez soin de faire une copie de la variable M7 vers Corr_M7 sur laquelle porteront vos corrections.

b) Vérifiez que vos corrections proposées sont exactes. Pour ce faire, vous utiliserez la procédure « **TEMPORARY** » et la commande « **LIST** ». Votre listing sera restreint aux seules observations qui ont fait l'objet de correction.

b) Structure itérative « DO REPEAT… END REPEAT » et Fonction « CHAR.INDEX »

88. Les variables m7 et m23 sont à réponses multiples. En exemple, les réponses à m7 portant sur le handicap se présente comme une combinaison des codes du Tableau 5-3, à l'exception du code « A ». Pour un individu donné qui déclare à la fois être sourd/malentendant et malade mental, m7 aura comme réponse : « CG ». Tandis que pour celui qui n'a aucun handicap, m7 devrait être « A » (soit on n'a aucun handicap « A », soit on a un handicap et le code « A » ne devrait pas apparaitre dans m7).

89. Il est vous demandé de créer des variables dichotomiques m7a, m7b, …, m7x pour chaque handicap et m23a, m23b, …, m23o pour chaque langue d'alphabétisation avec les modalités oui=1/non=0. Vous déclarez avant tout ces variables en numérique au format de largeur 1.

Tableau 5-3 : Codification de la variable m7 et m23

Modalités (m7)	Codes (m7)	Modalités (m23)	Codes (m23)
Sans handicap	A	National	A
Aveugle / malvoyant	B	Français	B
Sourd / malentendant	C	Arabe	C

Modalités (m7)	Codes (m7)	Modalités (m23)	Codes (m23)
Muet	D	Anglais	D
Infirme membres inférieurs	E	Autre langue	E
Infirme membres supérieurs	F	Aucune langue	O
Malade mental	G		
Lépreux	H		
Autre Handicap	X		

c) Fonction « CHAR.SUBSTR »

90. Le programme d'une ONG de la place s'intéresse aux personnes qui cumulent au moins deux handicaps parmi les sept (07) handicaps définis. Elle considère en effet que ces personnes sont plus vulnérables. Elle travaille avec une hypothèse très forte que l'ordre dans lequel les handicaps sont cités déterminent le degré d'invalidité du handicap. Ainsi si une personne est muette et lépreux « DH » alors pour cette dernière son mutisme lui serait plus invalidante que sa lèpre.

a) Pour les personnes souffrant d'au moins deux handicaps, créer trois variables qui indiquent la nature trois premiers handicaps par ordre.

b) Structure itérative « LOOP… END LOOP » (Cf. **Exercice 6.2**, Page 33)

d) Procédure « RANK »

91. Créer les groupes de déciles de la variable « M4 » en utilisant la procédure « RANK ». Comparer les résultats avec ceux obtenus avec l'assistant « Regroupement visuel ».

Exercice 5.2 : Manipulation des dates

92. Pour cet exercice, nous utiliserons la base de données « **DateEnq**.sav » du répertoire « **TempData** ». La variable « DATE_ENQ » est la date de l'interview du ménage format numérique (jour/mois/année).

a) En utilisant la fonction « STRING », convertir la variable « DATE_ENQ » en variable de type chaîne.

b) Déclarer trois variables JJENQ ; MMENQ ; AAENQ numériques de largeurs respectives 2, 2, et 4.

c) En utilisant les fonctions « NUMERIC » ; « CHAR.SUBSTR » et « CHAR.LENGTH », créer les variables précédemment déclarées de telles sorte JJENQ corresponde au jour de l'enquête ; MMENQ au mois de l'enquête et AAENQ à l'année de l'enquête.

d) Supprimer la variable « DateEnq ».

e) En utilisant la fonction « DATE.DMY » reconstruire la date l'enquête sous le nom « DateEnq ». Visualiser les données.

f) Utiliser la procédure « **FORMATS** » pour formater la date « DateEnq » au format (date11), au format (Adate10), au format (Edate10), et au format (Sdate10). Visualiser les données pour voir l'affichage de chaque format.

6 MANIPULER LES FICHIERS DE DONNEES

93. Les manipulations des fichiers qui s'imposent couramment lors de traitement des données (en dehors bien sûr de l'ouverture et de la sauvegarde) sont les opérations de fusion, de sélection, d'agrégation et de transposition.

6.1 OPERATIONS USUELLES SUR LES DONNEES : TRI, SELECTION, ET EDITION DES DONNEES

6.1.1 Tri des observations (SORT CASES BY)

94. Pour ordonner les observations d'un fichier suivant une ou plusieurs variables, on utilise la commande **« SORT CASES BY »** suivie de la variable ou de ces variables.

```
GET
   FILE="PrimD\Individu.sav".
SORT CASES BY hh1 hh2 (A).
```

95. Les données du fichier « Individu.sav » sont triées par ordre croissant suivant les variables hh1 et hh2. L'option **(A)** indique le sens ascendant du tri. Pour trier dans l'ordre décroissant, il suffit de préciser le sens avec l'option **(D).**

6.1.2 Sélection des observations

96. Il arrive souvent que les besoins de l'analyste portent une partie des données et non l'intégralité. M. le Gouverneur de la région de Bamako souhaite avoir des informations sa région, il n'est donc pas nécessaire de travailler avec l'ensemble des données du Mali. On va donc procéder à la sélection des données de la région de Bamako. Sous SPSS, on utilise la commande « **SELECT IF** » pour faire une sélection.

```
GET
   FILE="PrimD\Individu.sav".
SELECT IF (region=9).
EXECUTE.
```

97. Toutes les observations pour lesquelles la variable « region » n'est pas égale à 9 seront supprimées de la base de données « Individu.sav ». Il faut donc faire attention à ne pas écraser le fichier de données initial après une sélection.

98. Plus souvent, c'est l'analyste pour des besoins spécifiques qui souhaite examiner temporairement une partie des données. Il peut alors combiner la sélection avec la commande « **TEMPORARY** ».

```
GET
   FILE="PrimD\Individu.sav".
TEMPORARY.
SELECT IF (region=9).
FREQUENCIES m3. /*Cette instruction sera exécutée pour (region=9).
EXECUTE.
```

**********.

```
GET
   FILE="PrimD\Individu.sav".
TEMPORARY.
SELECT IF (region=9).
```

```
FREQUENCIES m3. /*Cette instruction sera exécutée sur les observations où (region=9).
MEANS m4. /*Cette instruction sera exécutée sur toutes les observations.
```

99. La commande « TEMPORARY » ne s'applique à la première instruction après la sélection. Pour réaliser plusieurs instructions sur la sélection, on peut utiliser un filtre. Sous SPSS, la procédure **« FILTER BY »… « FILTER OFF »** permet de réaliser certaines opérations sur une sélection temporaire.

```
GET
   FILE="PrimD\Individu.sav".
COMPUTE FiltreReg9=(region=9). /*Crée une variable indicatrice (1 si region=9 et 0 ailleurs).
FILTER BY FiltreReg9. /*Début de l'instruction de filtre.
TITLE "Statistiques de la région de Bamako".
FREQUENCIES m3. /*Instruction exécutée sur les observations où (FiltreReg9=1).
MEANS m4. /*Instruction exécutée sur les observations où FiltreReg9=1.
FILTER OFF. /*Fin de l'action du Filtre.
TITLE "Statistiques au niveau national".
FREQUENCIES m3. /*Cette instruction sera exécutée sur toutes les observations.
MEANS m4. /*Cette instruction sera exécutée sur toutes les observations.
```

6.1.3 Edition des données

100. L'édition des données est indispensable lorsqu'il s'agit d'apurer les données. La commande « **LIST** » permet d'éditer les données sous SPSS. Le programme ci-dessous permet d'éditer les informations concernant le sexe (m3), l'âge (m4) et le lien de parenté avec le chef de ménage (m2) des 10 premiers individus de la base de données.

```
LIST region hh1 hh2 m1 m3 m4 m2
      /CASES=FROM 1 TO 10. /*Les 10 premières observations.
```

Remarque 6-1 :

- L'assistant de tri des observations est disponible dans le menu « Données » (« **Données** » puis « **Trier les observations** ». Tout comme les observations, on peut trier les variables à travers l'assistant « **Trier les variables** » disponible également dans le menu « Données ».
- SPSS dispose aussi d'un assistant pour la sélection des observations. ¨Pour ce faire, aller dans le menu « Données » puis « **Sélectionner les observations** ». Cet assistant gère également la sélection par filtrage ainsi que les sélections par processus aléatoire.

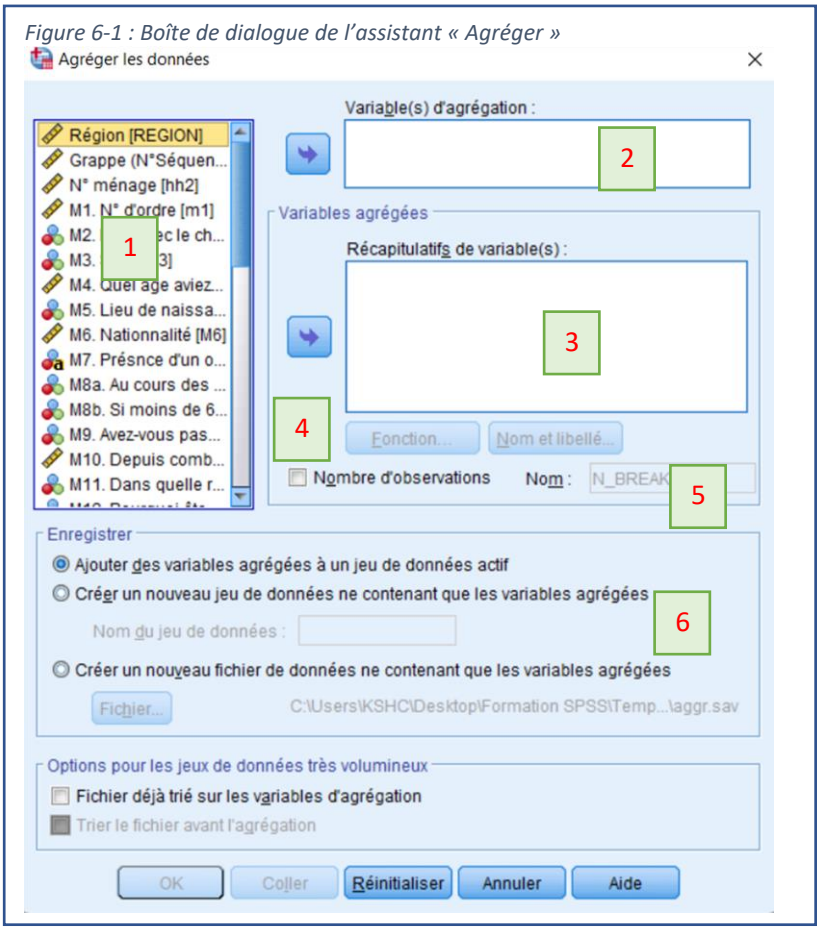
6.1.4 Agrégation des données

101. Considérons le fichier « Individu.sav ». Chaque observation de cette base de données renferme des informations sur les caractéristiques d'un individu bien donné (identifié par les variables « hh1, hh2 et m1 »). L'analyste souhaite constituer de niveau ménage comprenant les données comme fonction des données individuelles de ses membres. Soit par exemple, pour chaque ménage :

- L'âge moyen des membres ;
- L'âge maximum des membres ;

- La taille (nombre de membres) ;
- Le sexe du chef du ménage ;
- Etc.

Figure 6-1 : Boîte de dialogue de l'assistant « Agréger »



102. Chaque observation du nouveau fichier ainsi constitué à renvoie à un groupe d'observations du fichier initial. Le processus permettant de passer des informations individuelles à des données plus groupées est une « **agrégation de données** ».

103. La procédure d'agrégation des données sous SPSS se réalise à l'aide de l'assistant « **Agréger** » disponible dans le menu « **Données** ». L'aperçu de sa boîte de dialogue est donné par la **Erreur ! Source du renvoi introuvable.**.

- **Le quadrant (1)** contient la liste de toutes les variables du fichier.
- **Le quadrant (2)** permet sélectionner les variables d'agrégation à partir de la liste du quadrant (1). Si on veut avoir les informations de niveau ménage, la hh1 hh2 qui identifie chaque ménage de façon unique constitue les variables d'agrégation. Toute autre variable qui n'altère pas cette unicité du ménage peut être ajoutée à la liste des variables d'agrégation. C'est le cas de la variable « region » par exemple.
- **Le quadrant (3)** est réservé aux variables à agréger. Celles-ci sont sélectionnées à partir du quadrant (1).
- **Le quadrant (4)** concerne les boutons « Fonction » et « Libellé ». Le bouton « Fonction » donne accès à une boîte des fonctions prédéfinies.
- **Le quadrant (5)** permet d'ajouter aux données le nombre d'observations du fichier initial ayant servi à constituer chaque observation du fichier final. Il suffit pour cela de cocher la case « Nombre d'observations ». La variable « Nombre d'observations » sera créée dans le fusion agrégée sous le nom renseigné dans la partie « N_BREAK ». Par défaut, cette variable aura pour nom « N_BREAK »
- Le quadrant (6) est très explicite. Il permet de choisir comment sont gérer les nouvelles variables issues de l'agrégation. On peut soit sauvegarder ces variables dans un nouveau fichier, soit les ajouter à l'ensemble de données actif ou dans un nouvel ensemble de données (Cf. 30 pour les ensembles de données).

104. La procédure d'agrégation est également réalisée à travers la commande « **AGGREGATE** ». Le programme ci-dessus va créer un fichier nommé « TempMen.sav » dans le répertoire temporaire contenant les variables : **region hh1 hh2 Agemean AgeMax Hhsize**. Pour chaque observation de ce fichier, AgeMean est la moyenne des âges des membres du ménage en question.

```
GET
    FILE="PrimD\Individu.sav".
AGGREGATE
   /OUTFILE= "TempD\TempMen.sav"
   /BREAK region hh1 hh2
   /AgeMean "Age moyen"=MEAN(M4)
   /AgeMax "Age maximum"=MAX(M4)
   /Hhsize "Taille du ménage"=NU(M1).
```

Exercice 6.1 : Agrégation des données

105. Vous créer deux fichiers nommés Temp6A.sav et Temp6B.sav respectivement pour les questions a) et b) . Ces fichiers seront sauvegardés dans le répertoire temporaire.

a) A partir du fichier « Individu.sav » du répertoire construire un fichier de données contenant pour chaque ménage de chaque region :

- Age du chef de du ménage ;
- Sexe du chef de ménage ;
- Nombre d'enfants de moins de 5 ans ;
- Nombre d'adultes de 18 ans ou plus.

b) Reprendre le point a) en y ajoutant pour chaque région, les variables suivantes :

- Age médian des chefs de ménage
- Nombre moyen d'enfants de moins de 5 ans par ménage
- Nombre total d'adultes de 18 ans ou plus.

6.2 FUSION DEUX OU PLUSIEURS ENSEMBLES DE DONNEES

106. On distingue deux types de fusion de données. Lorsqu'on a une partition de données en deux ou plusieurs fichiers que l'on veut assembler dans un seul fichier de données, la fusion est appelée **ajout de cas**. Il s'agit dans ce cas d'empiler verticalement les enregistrements (observations). En revanche, lorsque pour deux fichiers de données, des informations (en termes de variables) sont disponibles sur l'un et pas dans l'autre, on parle d'**ajout de variables.**

6.2.1 Ajout de cas

107. Supposons que les données sur les caractéristiques des ménages ont été saisies séparément selon le milieu de résidence (urbain/rural). A l'issue de la saisie, le statisticien d'enquête dispose par conséquent de deux fichiers soit « MenageUrbain.sav» et « MenageRural.sav» avec des variables similaires.

108. Pour constituer un fichier de ménage unique avec les deux fichiers, il doit les superposer ou empiler, c'est-à-dire ajouter les observations du fichier « MenageRural.sav» à la suite de celles du fichier « MenageUrbain.sav» (vice versa). Ce type de superposition de fichiers s'appelle **fusion par ajout de cas.** Sa mise en œuvre sous SPSS se fait dans le menu « **Données** » puis « **Fusionner des fichiers** » et « **Ajout des observations** ». Une boîte de dialogue vous permet alors de sélectionner les fichiers à fusionner.

109. Au niveau syntaxique, c'est la commande « **ADD FILES** » qui permet de réaliser une telle fusion. de données les deux ensembles de données en un seul fichier contenant alors l'ensemble des ménages (urbain et rural).

Exemple 6-1 : Fusion (ajout de cas) des fichiers « MenageUrbain.sav » et « MenageRural.sav »

```
GET
   FILE="TempD\MenageUrbain.sav".
ADD FILES
   /FILE= *
   /FILE="TempD\MenageRural.sav".
```

Remarque 6-2 : La procédure « ADD FILES » permet de compiler plusieurs fichiers les uns à la suite des autres.

Exercice *6.2 :* *Fusion de plusieurs fichiers, Structure itérative « LOOP… END LOOP »*

110. Reconstituer le fichier « Menage.sav » à partir des données saisies par opérateur (« MenageOpera1.sav » à « MenageOpera9.sav »). Ces données sont dans le sous-répertoire (RawData) des données primaires (PrimaryData). Créer une variable dans le fichier reconstitué indiquant la provenance de chaque enregistrement. Les données reconstituées seront sauvegardées sous le nom « MenageReconstruit.sav » dans le répertoire temporaire.

6.2.2 Ajout de variables

111. Considérons le fichier « Individu.sav » contenant les caractéristiques des membres de ménages et le fichier « Menage.sav » les caractéristiques telles que le logement des ménages. L'on souhaite produire des statistiques du genre « pourcentage de la population ayant accès à l'électricité ». Pour ce faire, nous avons besoin d'un fichier d'analyse comprenant à la fois les individus et les informations sur leurs logements.

112. Le statisticien d'enquête doit procéder soit à l'ajout des variables portant sur les individus dans le fichier « Menage.sav », soit ajouter les variables portant sur le logement dans la base « Individu.sav ». Ce type de fusion s'appelle « **ajout de variables** ». Il est mis en œuvre sous SPSS, dans en allant dans le menu « Données » puis « **Fusionner des fichiers** » et « **Ajout des variables** ».

113. Cette opération exige que les observations des deux fichiers :

a) aient une clé de fusion. Il s'agit d'une ou plusieurs variables communes aux deux fichiers à fusionner et qui permettent d'identifier les observations de part et d'autre,

b) soient préalablement triées suivant la clé de fusion.

114. La boîte de dialogue Figure 5-2 offre deux onglets « Méthode de fusion » et l'onglet « Variables ». L'onglet « Méthode de fusion » propose trois méthodes de fusion :

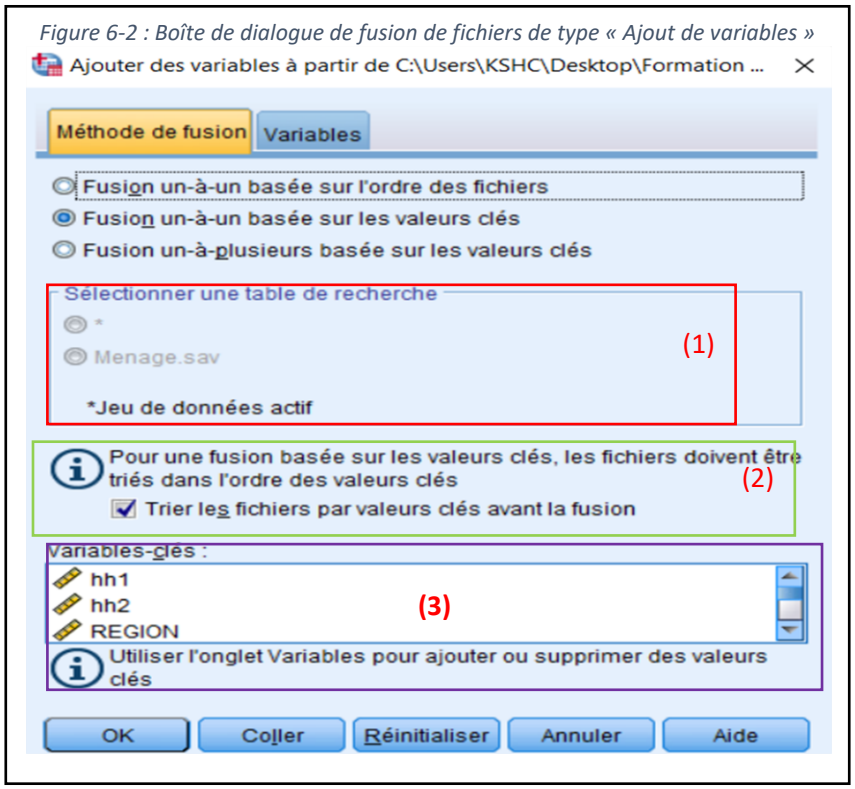
a) **Fusion un-à-un basé sur l'ordre des fichiers :** Ce type de fusion ne nécessite pas une clé de fusion. Les variables du second fichier sont ajoutées à la suite de celles du premier fichier. Ce type de fusion n'est pas adapté aux données tabulaires sur lesquelles nous travaillons.

b) **Fusion un-à-un basée sur les valeurs clés :** Chaque observation de la base de données de départ sera reliée par une seule observation de la seconde base de données (analogie avec la relation de **1 :1** dans les bases de données relationnelles).

c) **Fusion un-à-plusieurs basée sur les valeurs clés :** A une observation de la base de données de départ peut correspondre une ou plusieurs observations de la seconde base de données (relation 1 :m ou m :1).

115. Le quadrant **(1)** invite à la sélection d'une « **table de recherche** ». Il s'agit d'une précision importante à laquelle il faut prêter attention lors de vos travaux. A défaut, SPSS vous renverra une erreur du genre « Clé en double dans un fichier table ». La base de données appelée « **Table de recherche »** ou **« Table de référence »** est une base de données dans laquelle la clé de fusion est unique. Les méthodes de fusion de type **b) Un-à-Un** et **c) Un-à-Plusieurs** font appel absolument à une base de données dans laquelle la clé de fusion est unique pour chaque observation. Ce quadrant permet de sélectionner cette base de données.

Figure 6-2 : Boîte de dialogue de fusion de fichiers de type « Ajout de variables »



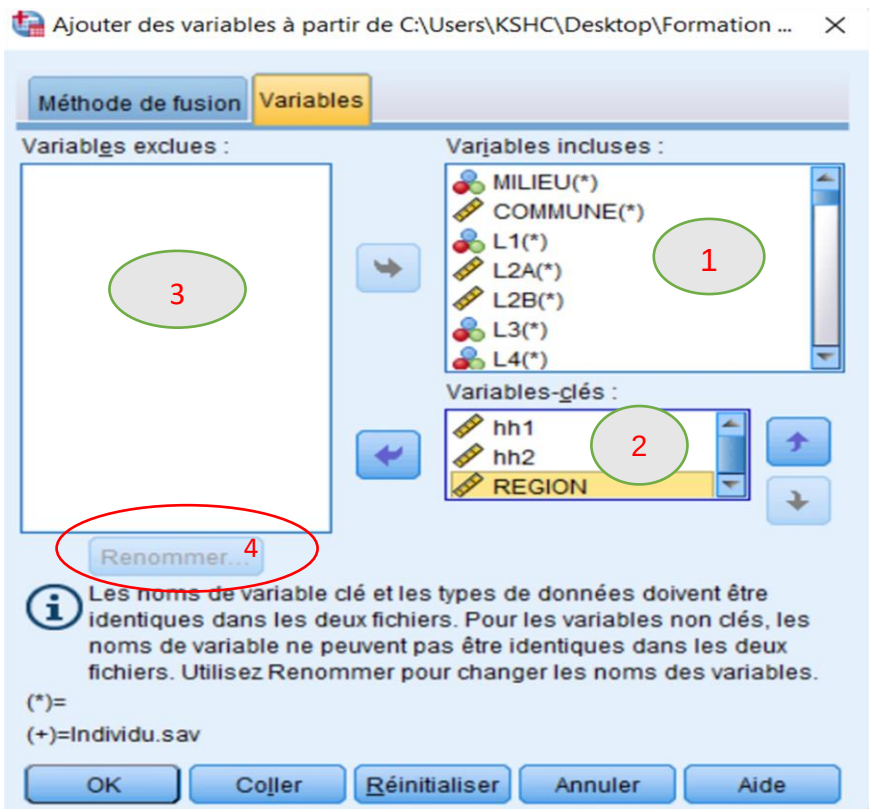
116. Le quadrant (2) permet en cochant la case « Trier les observations par valeurs clés avant la fusion » permet à SPSS de réaliser les tris des données avant de procéder à la fusion des données. Les versions antérieures n'offraient pas cette possibilité et il fallait nécessairement faire les tris avant la fusion.

117. Le quadrant **(3)** présente les variables communes aux deux fichiers comme « variables-clés ». Un avertissement vous renvoie cependant à aller au deuxième onglet « Variables » pour ajouter ou supprimer des valeurs clés.

118. Le deuxième onglet « Variables » se présente comme la Figure 6-3. Par défaut, le logiciel inclut toutes les variables non communes aux deux fichiers dans la liste des variables pour la fusion. Il propose dans le cas spécifique comme variables clés de fusion les variables « hh1 », « hh2 » et « region » qui sont des variables communes aux deux fichiers. Mais il appartient à l'analyste d'indiquer lesquelles des variables sont les variables clés de fusion.

119. En outre, si l'analyste n'a pas besoin de toutes les variables de l'une ou l'autre fichier de fusion, il peut les sélectionner dans le quadrant (1) et les déplacer vers le quadrant (3). De même comme indiquer dans le quadrant (4), si une variable figure dans les deux fichiers sous de noms différents, l'analyste doit la renommer afin que cette dernière ait le même nom dans chaque fichier.

Figure 6-3 : : Aperçu de l'onglet « Variables »



120. Dans notre exemple, il s'agit de fusionner les données sur les caractéristiques des ménages « Menage.sav » avec celles portant sur les caractéristiques des individus « Individu.sav ». Un ménage est identifié par le numéro de la grappe « hh1 » ainsi que son numéro dans la grappe « hh2 ». Cette combinaison est unique dans le fichier « Menage.sav ». En revanche dans le fichier « Individu.sav », pour un ménage donné, la combinaison « hh1 » et « hh2 » est répété autant de fois qu'il y a d'individus dans ce ménage. Ainsi, nous devons choisir comme :

- **Méthode de fusion**, la méthode c) Fusion un-à-plusieurs basée sur les valeurs clés
- **Variables clés de fusion**, les variables hh1 et hh2
- **Table de recherche,** le fichier « Menage.sav ».

121. La variable « region » n'est pas nécessaire parmi les variables clés de fusion dans la mesure où les variables « hh1 » et « hh2 » permettent d'identifier sans ambiguïté chaque ménage. On peut la ressortir, et comme celle-ci est présente dans chacun des deux fichiers, il faut alors choisir laquelle garder. Dans notre exemple, la variable « region » est conservée dans la liste des variables clés de fusion.

122. Les lignes de commandes ci-après permettent de réaliser cette opération lorsque nos deux fichiers sont préalablement triés suivant les clés de fusion (hh1, hh2).

```
GET
   FILE="PrimD\menage.sav".
MATCH FILES /TABLE=*
 /FILE="PrimD\Individu.sav".
```

```
 /BY REGION hh1 hh2.
EXECUTE.
```

123. Dans une fusion de type Un-à-Plusieurs avec la procédure « MATCH FILES », toutefois la différence se situe entre la base de données « TABLE » et la base de données « FILE ». Les deux bases ne sont pas interchangeables, en d'autres termes, mettre la base de données « FILE » à la place de « TABLE » causera un échec de l'opération. Cela reviendrait dans notre exemple à considérer le fichier « Individu.sav » comme la table de recherche.

Exemple 6-2 :

124. Réaliser l'opération de fusion « ajout de variables » des fichiers « Menage.sav » et « individu.sav » du répertoire primaire via l'assistant de fusion de SPSS. Vous supposerez que les deux fichiers de données ne sont pas préalablement triés. Après avoir finalisé le paramétrage de la procédure, coller la syntaxe au lieu de valider par le bouton « **OK** ».

```
DATASET NAME Jeu_de_données1.
GET FILE='C:\Users\KSHC\Desktop\Formation SPSS\PrimaryData\Individu.sav'.
DATASET NAME Jeu_de_données2.
DATASET ACTIVATE Jeu_de_données1.
SORT CASES BY REGION hh1 hh2.
DATASET ACTIVATE Jeu_de_données2.
SORT CASES BY REGION hh1 hh2.
DATASET ACTIVATE Jeu_de_données1.
MATCH FILES /TABLE=*
 /FILE='Jeu_de_données2'
 /BY REGION hh1 hh2.
EXECUTE.
```

125. Le programme ci-dessus est l'opération de fusion réalisée par SPSS avec l'option « Trier les fichiers par valeurs clés avant fusion ».

Remarque 6-3 :

a) Le seul fichier de données ouvert est le fichier « menage.sav ». Dans le programme ci-dessus, SPSS nomme cet ensemble de données « Jeu_de_données1 » à travers la commande **« DATASET NAME ».**

b) SPSS peut gérer plusieurs ensembles de données ouverts à la fois. Chaque ensemble de données est nommé grâce à la procédure « DATASET NAME ».

c) Pour indiquer l'ensemble de données sur lequel, on exécute une procédure, on utilise la commande « DATASET ACTIVATE » suivie du nom de cet ensemble de données. Ainsi, l'instruction « SORT CASES BY REGION hh1 hh2 » ci-dessous s'exécute sur l'ensemble de données appelé « Jeu_de_données2 » car c'est cet ensemble qui est activé avant que cette dernière ne s'exécute.

```
DATASET ACTIVATE Jeu_de_données2.
SORT CASES BY REGION hh1 hh2.
```

d) Dans la procédure « MATCH FILES /TABLE=* », le fichier actif matérialisé par « * » est l'ensemble de données « Jeu_de_données1 » qui est le fichier « Menage.sav ». C'est

l'action « DATASET ACTIVATE Jeu_de_données1 » juste au-dessus qui lui permet d'être activé.

e) Pour fermer un ensemble de données, on utilise la commande « DATASET CLOSE » suivie du nom de l'ensemble de données. « **DATSET CLOSE Jeu_de_données1. »** fermera l'ensemble « Jeu_de_données1 » c'est-à-dire le fichier « Menage.sav ».

6.2.3 Fusion des données avec la procédure « STAR JOIN »

126. En plus la commande usuelle « MATCH FILES », les versions récentes de SPSS comporte une commande de fusion dérivée de SQL. Il s'agit de la commande « **STAR JOIN** ». Elle requiert tout comme la commande « MATCH FILES » une clé de fusion.

127. Considérons le programme ci-dessous (copier et coller à la suite de votre programme) :

```
DATASET CLOSE ALL. /*Ferme les ensembles de données ouverts.
NEW FILE. /*Ouvre un ensemble de données vide.
GET
   FILE="PrimD\menage.sav".
DATASET NAME Menage. /*Le fichier «Menage.sav » ouvert est nommé « Menage ».
 GET
   FILE="PrimD\Individu.sav".
DATASET NAME Individu. /*Le fichier «Individu.sav » est nommé « Individu ».
STAR JOIN
  /SELECT t0.m1, t0.m2, t0.m3, t0.m4, t1.L6, t1.L7, t1.L8, t1.L9, t1.L10, t1.L11
    /FROM "Individu" AS t0
    /JOIN "Menage" AS t1
          ON t0.hh1=t1.hh1
          AND t0.hh2=t1.hh2
  /OUTFILE FILE= "TempD\Acces_Service_Base.sav".
```

- Dans le programme ci-dessus, **/SELECT** permet de sélectionner les variables à conserver dans le fichier fusionné. Dans le cas précis, les variables « m1, m2, m3, m4, L6, L7, L8, L9, L10 et L11 » sont sélectionnées.
- **"Individu" AS t0 : t0** est l'**alias** de l'ensemble de données **"Individu"** qui n'est autre le fichier de données « Individu.sav ». Ainsi, **t0.m1** sépecifie que la variable **« m1 »** est sélectionnée dans la base de données **« t0 »,** c'est-à-dire le jeu de données **"Individu". De même, "Menage" AS t1** indique le jeu de données se nomme également **t1.** En tant qu'alias, t0 et t1 peuvent prendre des noms quelconques (Exemple : AS data1, AS data2 etc.).
- **/FROM** permet de spécifier la table de gauche. C'est la base de données dont toutes les observations seront listées dans le fichier après fusion peu importe qu'elles aient ou pas de correspondance dans le second fichier.
- **/JOIN** indique par analogie avec la commande « MATCH FILES », le fichier « TABLE ». La clé de fusion ne peut être dupliquée dans ce fichier.
- Les variables clés de fusion sont définies par leurs correspondances dans chacun des deux fichiers à la suite « **ON** » et avec « **AND** » pour séparer les correspondances.

- Cette commande ne nécessite pas que les données soient triées préalablement, contrairement à la commande « MATCH FILES ». Il n'est pas non plus nécessaire de renommer les variables clés de fusion lorsqu'elles n'ont pas les mêmes noms dans les deux fichiers.

6.2.4 Création de variables par comptage (COUNT)

7 ANALYSES DES DONNEES

128. Cette section s'intéresse à l'analyse statistique sous SPSS, nous traiterons essentiellement des statistiques simples et des graphiques.

129. L'ensemble des outils d'analyses statistiques sont disponibles dans le menu « Analyse/Analyze » de SPSS.

130. Dans cette section, nous travaillerons dans un nouveau programme que vous nommerez **« Prg03, Analyse des donnees.sps ».** Les données sont essentiellement dans le répertoire d'analyse (répertoire secondaire). Pour démarrer, le programme « **Prg03, Analyse des donnees.sps »** avec la définition des répertoires. Ajoutez à la suite une ligne de commande pour ouvrir le fichier « Indiv.sav » du répertoire secondaire.

7.1 LES STATISTIQUES SIMPLES (UNIVARIEES, BIVARIEES)

131. SPSS offre de nombreuses commandes pour l'analyse des données, autant sous forme de tableaux que de graphiques. La forme des tableaux dépend du nombre de variables croisées. Lorsqu'une seule variable est décrite, on parle de distribution univariée, ou de fréquence simple. Pour deux variables, on parle de distribution bivariée, et, au-delà de deux variables, de distribution multivariée. Plus le nombre de variables est élevé, plus la lecture sera complexe : un tableau croisant plus de quatre variables est souvent incompréhensible pour une personne normalement constituée (y compris un statisticien).

7.1.1 Résumé des variables

132. La première chose à faire avant de travailler sur un fichier est d'examiner l'ensemble des variables, pour détecter d'éventuelles erreurs à la saisie ou lors du transfert des données, et surtout pour se familiariser avec les données.

133. Dans le menu « Analyse/Analyze », la boîte d'outils « Rapports/Report » permet de générer des informations résumées sur une ou plusieurs variables.

Exemple 7-1 : Rapports sur les variables

a) Examinez les résultats des commandes suivantes et commentez.

```
CODEBOOK M4.
CODEBOOK nivinstM.
CODEBOOK Region Milieu.
```

b) Réexécuter les trois instructions ci-dessus en utilisant l'interface de SPSS accessible à travers : Analyse ⟶ Rapports⟶ Livre de codes

7.1.2 Statistiques univariées

134. SPSS dispose d'une boîte d'outils variés pour la production des statistiques descriptives (fréquences, tableaux croisés, etc.). Pour accéder à cette boîte d'outils aller dans le menu « Analyse » puis « Statistiques descriptives ».

135. Pour réaliser les fréquences d'une variable, on utilise la commande « **FREQUENCIES** » ou son assistant accessible via : Analyse ⟶ Statistiques Descriptives⟶ Fréquences.

Exemple 7-2 : Fréquences des variables

```
FREQUENCIES VARIABLES= M4.
FREQUENCIES VARIABLES = nivinstM.
```

136. Pour les variables de mesure « Echelle/Scale » autrement dit les variables continues, on peut produire les paramètres de tendance centrale et dispersion sous SPSS à l'aide de la commande « FREQUENCIES » ou de son assistant :

Exemple 7-3 : Quelques paramètres de position et de dispersion

```
FREQUENCIES VARIABLES=M4
  /FORMAT=NOTABLE
  /NTILES=4
  /STATISTICS=STDDEV MINIMUM MAXIMUM MEAN MEDIAN MODE
  /ORDER=ANALYSIS.
```

7.1.3 Statistiques bivariées

Variables qualitatives

137. Pour l'analyse de deux variables de mesure nominale on utilise généralement un tableau croisé. Le test d'indépendance de Khi-deux permet par ailleurs de se prononcer sur la présence éventuelle d'une liaison entre les deux variables.

138. Sous SPSS, les tableaux croisés sont réalisés grâce à la commande « **CROSSTABS** » dans la boîte d'outils « Statistiques descriptives » à travers l'assistant « Tableaux croisés » accessible via :

Analyse ⟶ Statistiques Descriptives⟶ Tableaux croisés

Exemple 7-4 : Liaison entre deux variables qualitatives

a) Croisement la variable « region » avec la variable « NivinstM » (niveau d'instruction de l'individu)

```
CROSSTABS
  /TABLES=region BY NivinstM.
```

b) Y a-t-il un lien entre le lieu de résidence et le niveau d'instruction ?

```
CROSSTABS
  /TABLES=region BY NivinstM
  /STATISTICS=CHISQ.
```

c) Reprendre le croisement en a) en affichant les pourcentage ligne puis les pourcentage colonne.

Variables quantitatives

139. Lorsqu'on a à faire deux variables continues, les outils d'analyse descriptives sont en général le coefficient de corrélation qui détermine s'il y a une liaison linéaire entre les deux variables. Sous SPSS, pour calculer le coefficient de corrélation entre deux variables quantitatives on utilise la commande « **CORRELATIONS** » ou l'assistant disponible dans la boîte à outils « Corrélation » accessible : Analyse ⟶ Corrélation ⟶ Bivariée

Exemple 7-5 : Liaison entre deux variables quantitatives

140. Pour cet exemple, nous allons charger la base de données « emploi.sav ». Elle contient quelques caractéristiques de l'emploi des individus âgés de 15 ans ou plus.

a) Etudier la liaison entre le nombre d'années d'étude « annee_etud » et le salaire horaire « sal_horair »

```
CORRELATIONS
  /VARIABLES=annee_etud sal_horair.
```

b) Reprendre la procédure avec l'assistant pour marquer d'une étoile au cas où la corrélation serait significative.

Remarque 7-1 : Il existe plusieurs coefficients de corrélation dans SPSS :

- **Pearson**: Permet d'étudier la linéarité entre deux variables continues ;
- **Spearman** (Pearson basé sur les rangs) : Permet de décrire dans quelle mesure la liaison entre deux variables est monotone. Il est utile pour les variables quantitatives non normales ou les variables qualitatives ordinales.
- **Kendall tau-b** (basé sur le nombre de concordances et discordances des rangs) : pour des variables ordinales

7.1.4 Tableaux personnalisés sous SPSS

141. La boîte d'outils « Tableaux » de SPSS permet de créer des tableaux croisés personnalisés. Il s'agit d'un outil puissant qui offre des présentions agréables. Cet outil est accessible via :

Analyse ⟶ Tableaux ⟶ Tableaux personnalisés

Exemple 7-6 :

a) Produire dans un même tableau en utilisant l'assistant « Tableaux personnalisés », les statistiques (âge moyen, salaire horaire moyen) pour chaque région.

b) Produire le salaire horaire moyen par genre (Homme/femme) suivant les régions.

142. Le logiciel SPSS permet de gérer les variables à réponses multiples à l'aide de la procédure « MRSET ». En vous référant au manuel de référence de SPSS, définir un jeu de réponses multiples pour les deux catégories de variables (m7a à m7i) et (m23a à m23e) en vue des analyses prochaines.

NB : la procédure MRSET se trouve dans le menu « Analyse/Tableaux/Jeux de réponses multiples » de SPSS.

```
* Définir les jeux de réponses multiples.
MRSETS
  /MDGROUP NAME=$Handicap LABEL='Nature du Handicap'
CATEGORYLABELS=VARLABELS VARIABLES=M7A M7B M7C
    M7D M7E M7F M7G M7H M7I VALUE=1
  /DISPLAY NAME=[$Handicap].
```

8 Système de gestion des résultats

Les résultats « OUTPUT » de SPSS peuvent être récupérés pour d'autres usages soit forme de données SPSS (.sav), sous format Word, Excel, PDF, fichier de sortie (.spv), format de rapport Web (.spw), XML, html et tex. La récupération des résultats avec SPSS se fait à l'aide de l'utilitaire dit Système de Gestion des Résultats (Output Management System (OMS)).

1) Spécifier le(s) type(s) de sortie (Table, Log, Graphic, Models, etc.) ;
2) Sélectionner un ou plusieurs identificateurs de commande
3) Sélectionner le(s) sous-type(s) de table ;
4) Spécifier le format du fichier de sortie ;
5) Spécifier la destination du fichier de sortie ;
6) Sélectionner les sorties à exclure ;
7) Ajouter les requêtes d'OMS.

Exemple 8.1 : Récupération des déciles avec la commande « FREQUENCIES »

Considérons la base de données « emploi_indicateurs15.sav » dans le sous répertoire « Emploi » du répertoire « Secondaire ». Pour les actifs occupés, l'enquête a saisi le revenu de l'emploi. La variable « sal_horair » contient le salaire horaire de chaque actif occupé. L'analyste désire connaître les déciles de salaire horaire selon le milieu de résidence et voudrait pour des usages qui lui sont propres disposer ces informations dans son fichier de données.

Sous SPSS, la procédure « AGGREGATE » est celle qui permet de calculer le plus souvent les statistiques, lesquelles peuvent être rajoutées directement dans le fichier de données. Cette procédure ne permet cependant pas de calculer les déciles. En revanche, les procédures « FREQUENCIES » ou « EXAMINE » les produisent pour consultation dans la fenêtre des résultats (Output).

Les résultats avec la procédure « FREQUENCIES » sont récupérées comme il suit et réintroduit dans le fichier de données grâce à l'utilitaire OMS.

```
PRESERVE.
SET TVARS NAMES TNUMBERS VALUES.
DATASET DECLARE DecileMilieu.
SORT CASES BY Milieu.
SPLIT FILE LAYERED BY Milieu.
```

```
OMS
   /SELECT TABLES
   /IF COMMANDS=['FREQUENCIES']
      SUBTYPES =['STATISTICS']
   /DESTINATION FORMAT=SAV
      OUTFILE='DecileMilieu'
   /COLUMNS SEQUENCE=[L1 R2].

FREQUENCIES VARIABLES=sal_horair
  /FORMAT=NOTABLE
  /NTILES=10
  /ORDER=ANALYSIS.
OMSEND.
RESTORE.

MATCH FILES FILE=*
  /TABLE='DecileMilieu'
  /rename (Var1=Milieu)
  /BY Milieu
 /DROP command_ TO label_.
EXECUTE.

DATASET CLOSE DecileMilieu.
```

La Commande « PRESERVE » permet de préserver les paramètres actuels de SPSS. Dans la suite de « PRESERVE » la commande « SET » modifie les paramètres de sortie de résultats. Pour les sorties en tableaux (TABLE) ce sont les noms des variables qui seront affichés, en revanche, ce sont les valeurs qui seront affichées pour les nombres du tableau.

TVARS NAMES : Pour « TABLE VARIABLES » « NAMES »
TNUMBERS VALUES : pour « TABLE NUMBERS » « VALUES »

La commande « RESTORE » à la fin De la Commande « OMSEND » permet de revenir aux paramètres de SPSS avant les changements apportés avec « SET ».
La commande « SPLIT FILE » comme décrit précédemment permet ici de produire séparément les déciles pour le milieu urbain et le milieu rural.
La commande « OMS » nous amène à la sortie des résultats.

- /SELECT permet de spécifier le type de sortie. Dans le cas présent on s'intéresse aux sorties du type tableau d'où : /SELECT TABLES
- /IF COMMANDS permet de spécifier la commande qui fournit la sortie. Les déciles que nous voulons récupérer seront produits à partir de la commande « FREQUENCIES ». Ce qui revient à : /IF COMMANDS=['FREQUENCIES']
- SUBTYPES permet de spécifier quels sous-types de sortie sont concernés ou non concernés. Ce sont les statistiques de « FREQUENCIES » qui nous intéressent, soit : SUBTYPES=['STATISTICS']
- /DESTINATION FORMAT permet de spécifier le format de destination de la récupération (spss (.sav), word, excel , etc.). L'analyste souhaite les avoir dans son fichier de données en SPSS, donc le format qui convient est le format SPSS. /DESTINATION FORMAT=SAV

- OUTFILE permet de spécifier le fichier de destination. Les sorties récupérées seront transférées à l'ensemble de données « DecileMilieu » précédemment déclaré. Soit : OUTFILE='DecileMilieu'. On aurait bien pu enregistrer le fichier. Dans ce cas, on aurait indiqué le chemin complet.
- /COLUMNS SEQUENCE permet de préciser l'ordre de position des éléments dans le tableau. R pour Rows (lignes), C pour Columns (colonnes) et L pour Layers (couches). /COLUMNS SEQUENCE=[L1 R2]