جامـــــــــــــــــعة وهـــــــران 2
كلية العلـــوم الاقتصادية التجارية و علوم التسيير

مطبوعة

إعلام آلي 2
محاضرة مع تمارين محلولة
الجزء الأول
الثانية ماستر إقتصاد نقدي و بنكي
السداسي الأول

مقدمة من طرف :

السيد(ة):..... طالب زوقار سعاد.....
الرتبة:..... أستاذة محاضرة أ.....

السـنـــــــــة:.....2023/2024.....

« Informatique 2 : Analyse de données avec SPSS
1ère Partie»

Description du cours :

Le cours est destiné aux étudiants qui font un parcours d'études qui nécessite de faire des analyses sur des données récoltées de différentes sources par différentes manières. Il est composé de douze chapitres, le premier chapitre est purement théorique ; contient une présentation des

méthodes d'analyses statistiques de façon générale en mettant l'accent sur la méthode descriptive et les onze chapitres restant présentent l'outil informatique d'analyse en détails en faisant des illustrations par exemples

***Objectifs** : Un rappel sur les méthodes d'analyse, Une initiation à l'analyse par des outils et des applications informatiques, Apprendre à réaliser des questionnaires et à les utiliser et les analyser en utilisant un logiciel.*

«Computer science 2: Data analysis with SPSS
1st Part»

Course description:

The course is intended for students who are pursuing a course of study that requires performing analyzes on data collected from different sources in different ways. It is composed of twelve chapters, the first chapter is purely theoretical; contains a presentation of statistical analysis methods in general with emphasis on the descriptive method and the remaining eleven chapters present the computer analysis tool in detail by illustrating examples.

***Goals**: A reminder of analysis methods, An introduction to analysis using computer tools and applications, Learning how to create questionnaires and how to use and analyze them using software.*

علوم الكمبيوتر 2: تحليل البيانات باستخدام"
الجزء ١ *SPSS* "

وصف المحاضرة:

الدورة مخصصة للطلاب الذين يتابعون دورة دراسية تتطلب إجراء تحليلات على البيانات التي تم جمعها من مصادر مختلفة بطرق مختلفة. ويتكون من اثني عشر فصلاً، الفصل الأول نظري بحت؛ يحتوي على عرض لأساليب التحليل الإحصائي بشكل عام مع التركيز على المنهج الوصفي أما الفصول الأحد عشر المتبقية فتعرض أداة التحليل الحاسوبي بشكل تفصيلي من خلال توضيح الأمثلة

الأهداف :

تذكير بأساليب التحليل، مقدمة للتحليل باستخدام أدوات وتطبيقات الحاسوب، تعلم كيفية إنشاء الاستبيانات وكيفية استخدامها وتحليلها باستخدام البرمجيات

Avant-propos

Ce cours est destiné aux étudiants de deuxième année Master spécialité EMB (Economie Monétaire et Bancaire), il représente une introduction (une première partie) à l'initiation des étudiants à l'utilisation de SPSS pour se familiariser avec des fonctions de la statistique descriptive.

Le présent cours peut être considéré comme un manuel d'utilisation du logiciel SPSS, pour les étudiants débutants ou toute personne désirant commencer à l'utiliser, présentant des exemples pratiques pour chaque fonctionnalité.

Une deuxième partie du cours (en cours de préparation) a pour objectif de se focaliser sur les différentes méthodes d'analyse avec une partie théorique les présentant avec détails et une partie applicative avec le même outil et en utilisant des exemples concrets relevant du domaine de l'économie.

Table des matières

Résumé

Le cours présente dans son premier chapitre, des généralités sur les méthodes d'analyse statistiques en mettant l'accent sur les différentes méthodes d'analyse et en particulier les méthodes d'analyses descriptives. Les chapitres qui suivent constituent une présentation détaillée du logiciel SPSS avec ses différentes composantes ; fenêtres et fonctionnalités, la présentation s'appuie sur différents exemples d'illustration pour les différentes fonctionnalités.

Mots clés. Analyse, Variable, Donnée, Echantillon, Effectif, Statistique descriptive.

Abstract.

In its first chapter, the course presents general information on statistical analysis methods, emphasizing the different analysis methods and in particular descriptive analysis methods. The following chapters constitute a detailed presentation of the SPSS software with its different components; windows and functionalities, the presentation is based on different illustrative examples for the different functionalities.

Key words. Analysis, Variable, Data, Sample, Size, Descriptive statistics.

ملخص

يقدم المقرر في الفصل الأول معلومات عامة عن طرق التحليل الإحصائي، مع التركيز على طرق التحليل المختلفة وخاصة طرق التحليل الوصفي. تشكل الفصول التالية عرضًا تفصيليًا للبرنامج الإحصائي اس ـ بي ـ اس ـ اس بمكوناته المختلفة؛ النوافذ والوظائف، يعتمد العرض التقديمي على أمثلة توضيحية مختلفة للوظائف المختلف

الكلمات المفتاحية. التحليل، المتغير، البيانات، العينة، العدد، الإحصاء الوصفي

Introduction

SPSS, est un logiciel conçu spécialement pour les analyses statistiques en sciences sociales : « S*tatistical Package for Social Sciences* ». Ses premières versions ont vu le jour dans les années soixante, ces versions étaient sous forme de programmes *open source* (ces programmes donnent la possibilité d'ajouter de nouvelles commandes.

A partir des années 80, ce logiciel a cessé d'être en open source et devient une propriété exclusive de SPSS inc. Plusieurs versions du produit ont été commercialisées en passent rapidement d'une version à l'autre, en une période record les concepteurs sont passés de la version 6 à la version 12 et à partir de la version 7 il est devenu un produit pour Windows.

Il existe d'autres produits qui emplissent le même rôle que SPSS (SAS, SYSTAT, STATISTICA, etc.). Tous ces produits sont très différents d'utilisation mais permettent tous de faire des tests statistiques sans devoir connaître les formules par cœur.

D'autres logiciels, tel Excel, peuvent aussi faire quelques tests statistiques, mais sont limités quant au nombre de données permises, et compliqués d'utilisation malgré l'apparence trompeuse.

L'objectif du logiciel SPSS est d'offrir un produit permettant de réaliser la totalité des analyses statistiques habituellement utilisées en sciences humaines.

Ce cours est une initiation au logiciel SPSS, nous présentons les principales fonctions de ce logiciel (création d'un fichier de données, transformation de variables, analyse statistique). Il est composé de 12 chapitres,

- Au niveau du premier chapitre, nous présentons les notions de base de l'analyse de données statistique en présentant avec plus de détail la statistique descriptive qui est au cœur des chapitres qui suivent dans un cadre applicatif.
- Les 11 chapitres restant présentent avec détails toutes les fonctions et les composants du logiciel SPSS, susceptible d'apporter des éléments de base aux étudiants qui commencent à manipuler le logiciel.

Chapitre I : Analyse de données : Notions de base.

I.1 Introduction

Analyser consiste à décomposer, examiner, raisonner, dégager l'essentiel, interpréter, évaluer,…

L'analyse est l'étude, l'examen, l'évaluation pour mieux comprendre les relations entre le tout et les parties.

Une donnée est une information élémentaire connue ou admise comme telle, sur laquelle peut fonder un raisonnement (les données servent de point de départ pour toute recherche), elle peut prendre différentes formes (numérique, texte, son, images, vidéos, ou une combinaison des différents formats).

De nos jours, il y a une explosion de données d'où le terme Big data ; qui représente la tendance de toute nouvelle technologie.

Pour pouvoir l'exploiter, une donnée doit passer par un processus de transformation.

I.2 Analyse de données

Un processus (un ensemble de méthodes statistiques, mathématiques, ou informatiques) permettant de transformer les données en information. C'est le processus pour lequel des données brutes deviennent des connaissances utilisables pouvant être exploitées.

Ce processus consiste à examiner, étudier et à interpréter des données afin d'élaborer des réponses à des questions, de comprendre le comportement et les liaisons entre les caractéristiques étudiées pour comprendre et optimiser le sujet en question.

I.2.1 Objectifs de l'analyse de données

- Permet de tirer des synthèses d'aide à la décision ; ce qui permet d'améliorer considérablement leurs opérations,
- Avec la transition numérique (digitalisation), l'analyse de données devient une valeur ajoutée indispensable aux stratégies économiques des entreprises,
- Aide à l'amélioration des produits et des services,
- Comprendre les résultats des enquêtes,
- Formuler des objectifs en matière de qualité.

I.2.2 Types d'études

On a deux types d'études

- Etude descriptive : Consiste à estimer des mesures agrégées d'une population cible, par exemple calculer les bénéfices moyens d'une entreprise.
- Etude analytique : Expliquer le comportement de caractéristiques ou les relations entre elles, on peut citer comme exemples :

  Dans le domaine médical pour étudier les facteurs d'obésité chez l'enfant,

  Dans le domaine de marketing il s'agit d'analyser les données pour prédire le comportement des consommateurs et placer les produits sur les marchés concernés,

Dans l'environnement de travail des ressources humaines, pour offrir aux employés un bon environnement de travail.

I.2.3 Les principales étapes du processus d'analyse

- Cerner le sujet d'analyse,
- Déterminer la disponibilité des données appropriées,
- Choisir les méthodes à utiliser,
- Appliquer et évaluer les méthodes,
- Résumer et interpréter les résultats.

I.3 Généralités et vocabulaire

L'analyse de données fait partie du domaine de la statistique ; qui est une discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles, on parlera aussi de Data science.

I.3.1 La démarche statistique

Le processus d'analyse est composé de 3 principales étapes :

a)- Recueil de données : Cette étape consiste à rassembler toutes les données relatives à l'étude, c'est la toute première phase qui consiste à déterminer les différents caractères (variables) à étudier, à déterminer la population avec ses différentes problématiques (son choix, sa taille, sa représentativité).

b)- Traitement de données : Une fois les données collectées lors de la phase précédente, vient la phase de traitement, ce dernier contient toutes les opérations qu'on peut appliquer sur les données ; calcul, classement, résumé visuel en numérique, compression (permet de réduire la dimensionnalité des données), etc.

c)- Interprétation et analyse : l'analyse consiste à identifier la variété de données les plus significatives, après la phase d'analyse vient la phase d'interprétation qui permet la réflexion qu'il faut faire sur les résultats à partir de la problématique.

Cette démarche statistique est utilisée dans différents domaines ; ingénierie, économie, management, biologie, informatique, physique, etc.

En bref, la statistique utilise des règles et des méthodes sur la collecte des données pour que celles-ci soient correctement interprétées, elle est utilisée comme composante à l'aide à la décision.

I.3.1 La démarche statistique

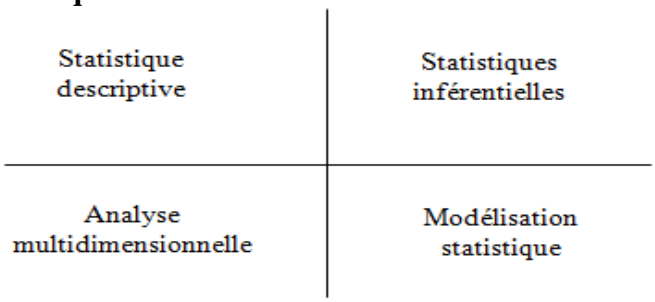


Figure I.1. Les sous domaines de la statistique.

a) Statistique descriptive : C'est une branche statistique qui regroupe des techniques utilisées pour décrire, résumer, représenter un ensemble relativement important de données à l'aide de graphiques et de mesures (moyenne, écart type, variance, etc.).
b) Statistique inférentielle : Consiste à analyser les données d'un sous ensemble d'une population pour en décrire les statistiques globales (estimateur, tests statistiques, etc.).
c) Analyse multidimensionnelle : Appelée aussi analyse exploratoire ; elle représente le prolongement des statistiques descriptives mais avec l'étude de relations entre trois variables et plus (Baccini A., 2003).
d) Modélisation statistique : Son principe consiste à formuler des observations par des règles mathématiques appelées « modèle probabiliste », dans ce cas il est question de faire de la prédiction et non de la prévision.

I.4 Statistique descriptive

C'est un domaine qui vise à étudier les caractéristiques d'un ensemble d'observations appelées aussi individus ; ces individus peuvent être des objets, des mesures, des personnes, des animaux, etc. en résumé un individu représente une unité d'observation.

Les individus ont des caractéristiques appelées variables ou caractères, l'ensemble des individus est appelé population notée N (qui représente le nombre s'individus de la population).

La sélection d'un sous ensemble de la population permet d'obtenir un échantillon de taille noté n (n : représente la taille de l'échantillon).

On appelle jeu de données (data set), l'échantillon qui représente les données collectées. Un échantillon est représenté sous forme d'un tableau ; où chaque ligne correspond à un individu et chaque colonne correspond à une variable.

On considère l'exemple suivant résumant les opérations de débit et de crédit sur un compte avec la nature de l'opération effectuée.

N	Date opération	Libellé	Montant	Solde_avant	catégorie
01	01/01/2023	Achat	20000	4500	Courses
02	02/02/2023	Virement	30000	1500	Autre
03	15/03/2023	Retrait	25000	2000	Loyer
….	….	….	….	….	…

Tableau I.1 : Echantillon mouvement de compte.

- Population : On appelle population l'ensemble sur lequel porte notre étude statistique, par exemple pour l'ensemble des patients dans un hôpital, on s'intéresse aux personnes diabétiques. La population correspond à l'ensemble des personnes diabétiques.
- Echantillon : Un échantillon de taille n est un sous ensemble formé de n individus de la population.
- Individus : On appelle individu ou unité statistique tout élément de la population, pour l'exemple de l'hôpital chaque patient diabétique est un individu.

I.5 Les variables (CARICANO M. et al., 2009)

Les caractéristiques étudiées sur les individus d'une population sont appelées des variables ou des caractères. Sur les unités stratégiques c'est-à-dire les individus on mesure un caractère ou une variable. Les valeurs possibles d'une variable sont appelées des modalités, par exemple pour la variable couleur on a les modalités {vert, rouge, blanc}.

I.5.1 Types de variables

Représente le domaine de variation de la variable, pour l'exemple du Tableau I.1, on a différents types de variables : le type nombre pour les variables montant et solde, type date pour la variable date, et caractère pour les variables catégorie et libellé. On distingue deux grands types de variables (quantitative et qualitative) :

- Les variables quantitatives, prennent des valeurs numériques ; leurs modalités sont mesurables, elles sont à leur tour composées de deux autres sous catégories :
  - Quantitatives discrètes, représente un ensemble de modalités fini ou dénombrable (nombre enfants par famille, nombre de places dans un cinéma, etc.).
  - Quantitatives continues, l'ensemble des modalités n'est pas dénombrable, elles sont représentées par des nombres en écriture décimale elles correspondent au résultat d'une mesure (poids, distance, moyenne, etc.).
- Les variables qualitatives, sont des variables non quantitatives, les valeurs qu'elles prennent sont appelées catégories ou modalités, à leur tour elles peuvent être :
  - Nominales, elles ne présentent pas d'ordre particulier, par exemple la couleur des yeux est une variable quantitative nominale.
  - Ordinale, les modalités sont organisées selon un ordre hiérarchique, par exemple la mention du bac est une variable qualitative ordinale car les mentions sont ordonnées selon la moyenne obtenue.

On peut aussi citer des types qui n'appartiennent à aucun des types su cités, les types date, monétaire, booléen, etc.

I.5.2 Distribution empirique d'une variable

Un échantillon est représenté sous la forme d'un tableau où chaque ligne représente un individu et chaque colonne représente une variable. On considère la variable catégorie de l'exemple du Tableau I.1, si cette variable prend les modalités suivantes pour 8 individus (transactions) {courses, courses, loyer, autre, courses, transport, facture}, on remarque que c'est beaucoup de modalités difficilement analysables juste pour 8 individus et avec l'augmentation du nombre d'individus ce sera encore plus compliqué.

Cependant, il y a une solution bien meilleure qui consiste à dire que « il y a 30 fois la valeur courses, 18 fois la valeur autre, 15 fois la valeur transport, etc. », cette formulation est appelée distribution empirique.

Ceci consiste à associer à chaque modalité un effectif, par exemple l'effectif associé à la modalité courses noté $n_{coureses}$=18. Donc la distribution empirique d'une variable représente l'ensemble des modalités prises par la variable avec les effectifs ou les fréquences associées.

Modalité	Effectif	Fréquences
Courses	30	0.6
Transport	15	0.3
Autre	2	0.04
…	….	…
Total	50	1

Tableau I.2 : Distribution empirique de la variable catégorie.

I.5.3 Représentation graphique d'une variable (Tillé Y., 2023)

On a deux configurations différentes qui sont liées aux types de variables,

a) Cas de variables qualitatives, on a la possibilité d'utiliser deux types de diagrammes :
   - Diagramme en secteur (appelé aussi camembert), représenté dans un cercle chaque modalité est représentée par une portion définie par un angle qui est proportionnel à l'effectif de chaque modalité (chaque fréquence est multipliée par 360° et plus la surface est importante plus la modalité est importante dans la distribution de la population.
   - Diagramme en tuyau d'argue (Bar chart) composé d'un ensemble de bâtons qui correspondent aux différents effectifs ou fréquences (plus le bâton est long plus la variable est importante dans la population)

b) Cas de variables quantitatives, on a la possibilité d'utiliser deux types de diagrammes :

- Dans le cas où la variable quantitative est discrète :
   - On utilise un diagramme en bâtons,
   - Courbes cumulatives, c'est une courbe en escalier qui représente les fréquences ou les effectifs cumulés.

- Dans le cas où la variable quantitative est continue, par exemple la taille T d'une personne T=1.678m et une autre T=1.679m, ceux sont 2 tailles différentes mais on les considère pas séparément car elles sont quasiment les mêmes. Pour cela la solution consiste à regrouper ou agréger les valeurs en classe (des intervalles qui peuvent être de tailles régulières ou irrégulière) c'est un processus de décomposition appelé discrétisation.

- On utilise l'histogramme, pour représenter les classes on n'a plus de fins bâtons mais des rectangles dont la largeur correspond à la largeur de la classe et l'effectif ne sera plus représenté par la hauteur du rectangle mais par sa surface *f*, les classes n'ont pas forcément la même hauteur.

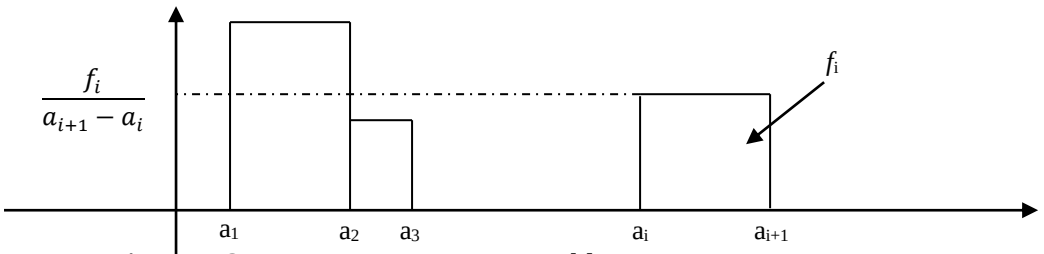


Figure I.2. Histogramme pour variable quantitative continue.

Dans ce premier chapitre, nous avons présenté des notions générales sur la statistique et en particulier sur la statistique descriptive, les différentes définitions et éléments présentées seront illustrées de façon applicative dans les chapitres qui suivent avec le logiciel SPSS.

Chapitre II : Présentation de SPSS : structure et composants

II.1 Lancement de SPSS

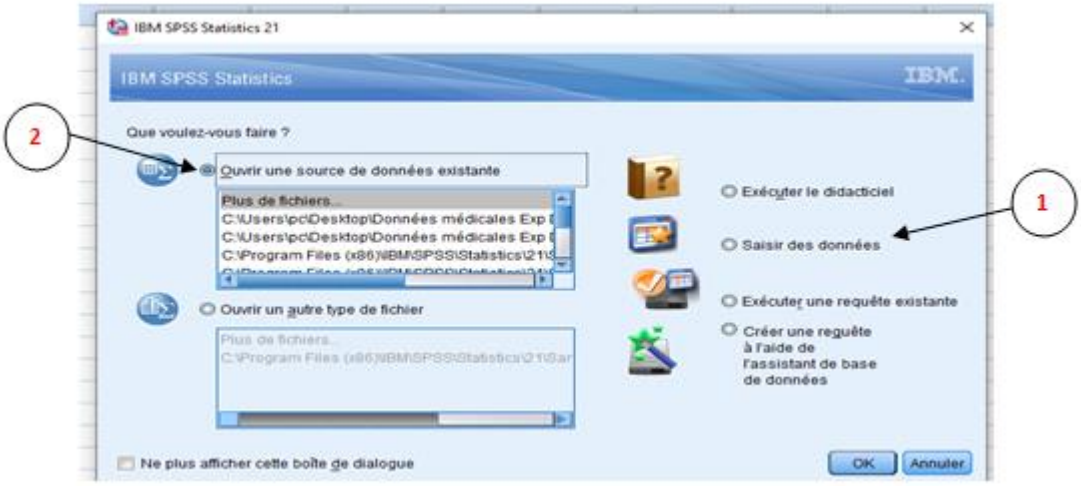


Figure II.1. Lancement de SPSS.

Au démarrage de SPSS on obtient cette boite de dialogue qui nous donne le moyen de choisir entre plusieurs possibilités :

1 : Permet d'obtenir un fichier SPSS vide,

2 : Permet d'ouvrir un fichier SPSS existant et récemment utilisé,

II.2 Les fenêtres SPSS (Jalby V., 2015)

SPSS possède Trois principales fenêtres : SPSS data Editor (éditeur de données), SPSS Viewer (fenêtre des résultats) et SPSS Syntax Editor (fenêtre de syntaxe). Pour chacune des fenêtres un fichier est généré.

II.2.1 La fenêtre éditeur de données

La fenêtre d'éditeur de données possède deux onglets qui contiennent deux pages, une pour les données (Data view) et une pour les variables (Variable view).

a. *La page affichage de données :* Cette page est un tableau de type individus-variables (il ressemble au tableau Excel). On trouve les individus en lignes et les variables en colonnes. Une ligne permet de donner le profile d'un individu, l'intersection d'une ligne et d'une colonne donne une case qui contient un score ou une modalité (représente la valeur que prend un cas donné pour une variable donnée). Dans ce tableau on peut directement introduire nos données. Le fichier, lorsque mis sous l'option affichage des données, se présente sous la forme suivante :

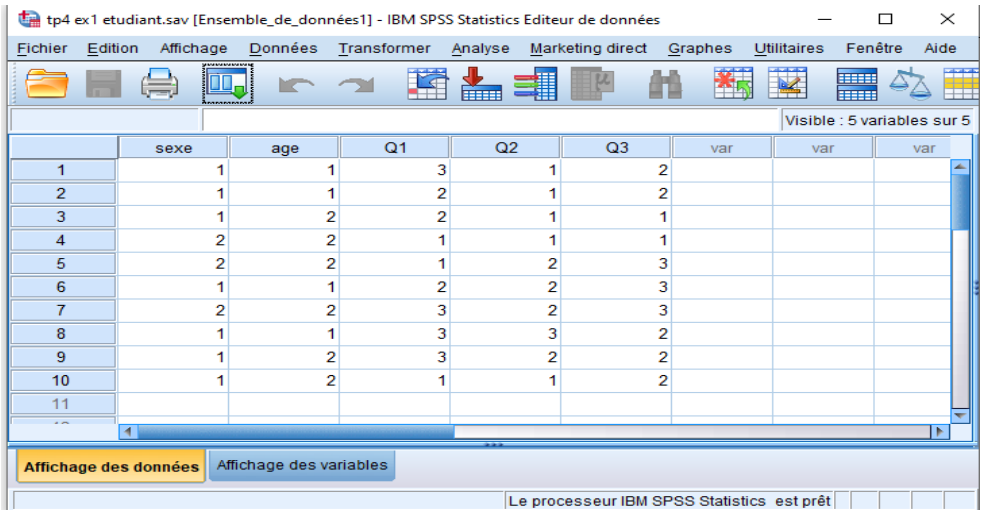


Figure II.2. Fenêtre Affichage des données (Data View).

- Chaque ligne représente un cas, par exemple un sujet (*case*),
- Chaque colonne représente une variable (*variable*),
- Chaque cellule contient une valeur d'un cas sur une variable.

b. *La page affichage des variables :* Permet de voir toutes les variables présentes du fichier de données, leurs noms, ce qu'elles représentent, les valeurs manquantes si elles existent, les valeurs possibles, les étiquettes qui les désignent. Le fichier, lorsque mis sous l'option affichage des variables, se présente sous la forme suivante :

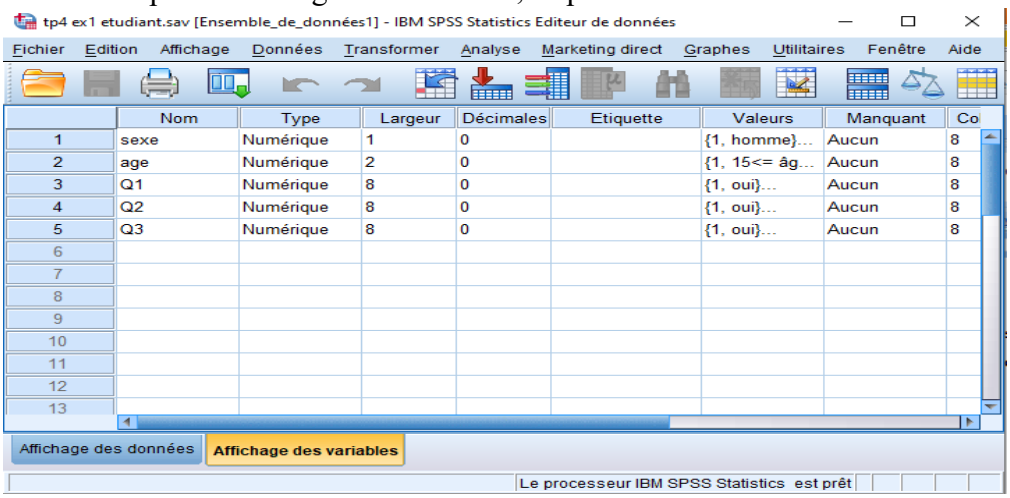


Figure II.3. Fenêtre Affichage des variables (Variables View).

Dans cette interface, chaque ligne représente une variable et les colonnes décrivent les caractéristiques des variables.

II.2.1.1 Caractéristiques des variables (Gilles I. et al, 2008)

Chacune des variables possède un ensemble de caractéristiques :

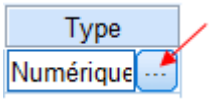
Le Nom:

Représente le nom de la variable, qui doit être unique, débute par une lettre, contient au maximum 64 caractères. Ces derniers peuvent être des lettres majuscules ou minuscules, des chiffres, un point et les symboles @ _ # $ (tous les autres symboles sont interdits).

Le nom ne peut pas se terminer par un point, les espaces vides ne sont pas possibles, majuscules et minuscules ne sont pas différenciées dans l'appellation d'une variable ou dans celui d'un fichier. Les mots clé de SPSS ne peuvent pas être utilisés (all, and, by, eq, ge, gt, le, lt, ne, not, or, to, with) et enfin ne pas mettre de lettre accentuée même si c'est possible.

Le Type :

Représente la nature de la variable (numérique, date, dollar, etc.), par défaut, SPSS considérera la variable de type numérique qui peut être changé dans la boîte de dialogue qui s'ouvre en appuyant sur le bouton qui apparaît quand on sélectionne la cellule.



Dès qu'on appui sur le bouton en question, on obtient la panoplie de type disponibles, qui peuvent être de type Date, monétaire, chaine, etc.

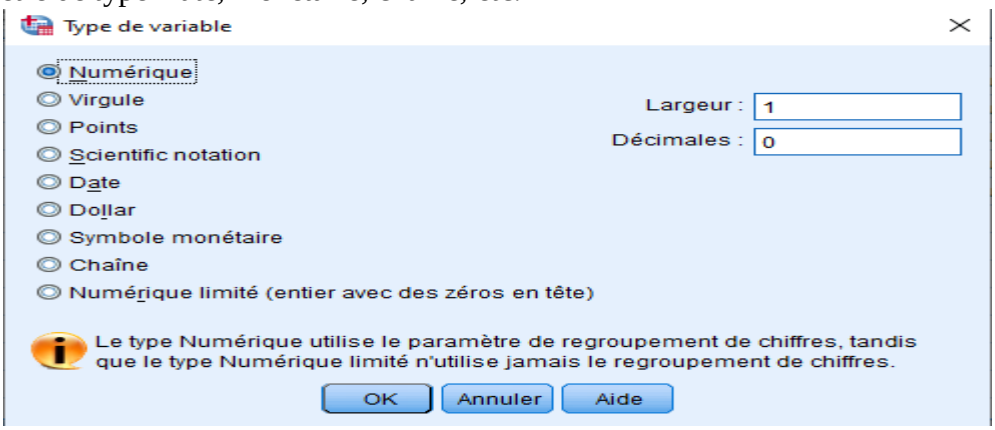


Figure II.4. Les différents types de données.

Largeur : Désigne le nombre de chiffres accordés à la donnée décimaux inclus ; les décimaux représentent les chiffres après la virgule.

Décimales : nombre de décimaux

Etiquette : Etiquette ou description de la variable max. 256 caractères

Valeurs : Valeurs définies et leur description (exp. 1 = Femme, 2 = Homme)

Manquant: Permet l'attribution de certaines valeurs comme codes pour valeurs manquantes

Colonnes: Largeur des colonnes dans la vue de données (en augmentant/diminuant le défaut « 8 », plus/moins de caractères de la colonne seront visibles dans la vue de données)

Align : Alignement des valeurs des variables dans les cellules de la grille de données (à droite, à gauche, centrées)

Measure : Description de l'échelle de mesure (données numériques sur un intervalle ou une échelle de rapport), ordinal ou nominal. Les données nominales et ordinales peuvent être des chaînes de caractères (alphanumériques) ou numériques.

- **Nominale** : Une variable est considérée comme nominale si ses valeurs représentent des catégories non ordonnées. Par exemple, les départements d'une société, la région géographique, le domaine d'activité.

- **Ordinale :** Une variable peut être considérée comme ordinale lorsque ses valeurs représentent des catégories ordonnées ; par exemple, les niveaux d'un indice de satisfaction, variant de très insatisfait à très satisfait
- **Échelle :** Une variable peut être considérée comme d'échelle lorsque ses valeurs sont ordonnées à partir d'une métrique spécifique, et que les distances entre les valeurs ont un sens. Par exemple, l'âge est mesuré en années, un salaire en Dinar.

II.2.2 Fenêtre des résultats (Output Editor)

Cette fenêtre apparaît après qu'une commande d'analyse a été effectuée, et contient les résultats de cette analyse. Les résultats apparaissent à droite dans la fenêtre. A gauche, figure une table des matières des résultats générés par SPSS.

La fenêtre résultats enregistre les résultats des opérations effectuées : tableaux, statistiques et diagrammes obtenus tout au long de votre session de travail. SPSS ouvre automatiquement cette fenêtre et y inscrit l'ensemble des résultats ainsi que le détail des opérations effectuées. La fenêtre de résultats Viewer se présente sous la forme suivante :

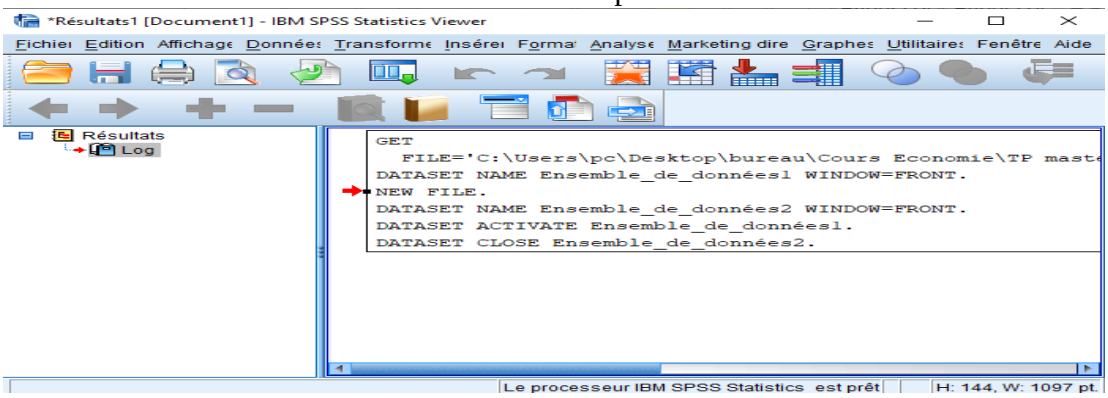


Figure II.5. La fenêtre des résultats.

Les résultats peuvent être imprimés tels quels (mais avec le risque de voir un tableau s'imprimer sur plusieurs pages. Il est également possible de copier les tableaux qui nous intéressent pour les coller ensuite dans Word, Excel ou dans un autre logiciel. Il est possible de copier un tableau de deux manières. En cliquant sur le tableau en appuyant sur le bouton de droite de la souris, SPSS vous propose de copier ou de copier spécial.

Copier correspond à copier les valeurs, mais lorsqu'il est collé il peut perdre son format (utile pour copier les résultats dans une feuille Excel par exemple). Copier les objets corresponds à copier les valeurs et le format du tableau : une fois collé, impossible de modifier les cellules du tableau (utile pour copier les résultats dans Word).

II.2.3 Fenêtre de syntaxe (Syntax Editor)

Jusqu'à maintenant, nous avons vu comment travailler avec les menus déroulant. Il existe une autre manière de lancer des analyses : passer par la fenêtre de syntaxe. Cette fenêtre permet d'écrire les commandes d'analyses statistiques. Elle fonctionne comme un traitement de texte simple.

C'est aussi utile pour faire la même analyse sur plusieurs fichiers de données.

Ainsi, une fois la syntaxe faite pour une opération, il est facile d'enregistrer les commandes et de les réutiliser pour différents fichiers de données.

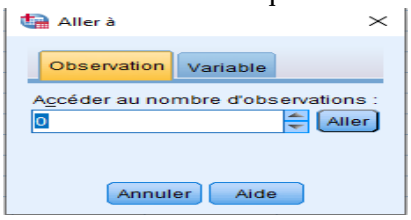
Voici les règles générales pour écrire des commandes dans SPSS :

- Chaque nouvelle commande se trouve en tête de ligne, précédée d'aucun espace.
- Les options qui suivent une commande débutent sur la ligne suivante et sont précédées d'au moins un espace et d'une barre oblique (/).
- Chaque commande doit ABSOLUMENT se terminer par un point.
- Lorsqu'on spécifie un nom de fichier, il doit être "entre guillemets".
- SPSS ne fait pas de différence entre les lettres majuscules et minuscules. Vous pouvez taper les commandes autant d'une manière ou de l'autre.
- De plus, entre les commandes, vous pouvez insérer des lignes vides. SPSS les ignore, mais elles peuvent améliorer la lisibilité des commandes quand il y en a plusieurs dans une fenêtre.

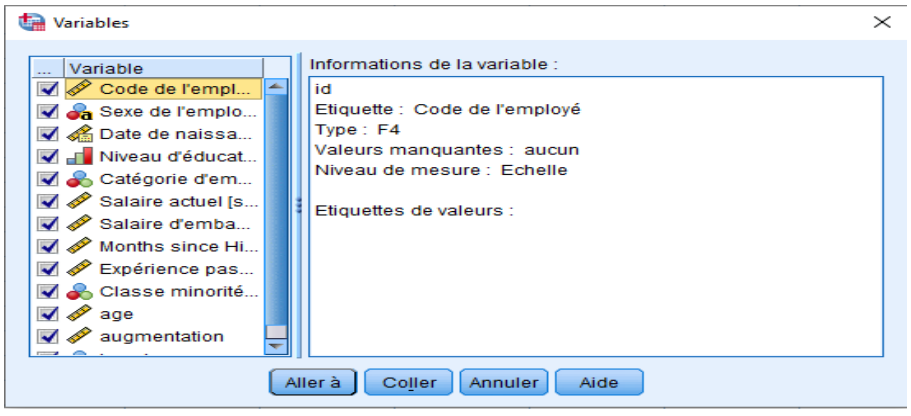
II.3 Quelques composants de la barre d'outils

La barre d'outils SPSS est composée de plusieurs boutons qui sont des raccourcis disponibles au niveau du menu.

-  Aller à observation: Permet d'accéder à une observation en spécifiant le numéro de ligne, permet aussi d'accéder à une variable en spécifiant son nom, cette dernière fonctionnalité peut aussi être exécutée en cliquant sur le bouton  .



-  Variables : Permet de donner la liste de toutes les variables.



Dans la boite de dialogue sont spécifiées des informations sur la variable sélectionnée, à savoir :

Son nom,

Etiquette si elle existe,

Type : qui prend une des valeurs suivantes {F : pour le type numérique, A : pour le type chaîne, ADATE : pour le type Date, DOLLAR8 : pour le type monétaire}, ces types sont suivis par la longueur de la variable.

Valeurs manquantes, la mesure, et les étiquettes de valeurs si elles existent.

- Descriptive : statistiques rapides, des tableaux dans le fichier output.

Effectifs

[Ensemble_de_données1] C:\Users\pc\Desktop\Employee data.sav

Statistiques

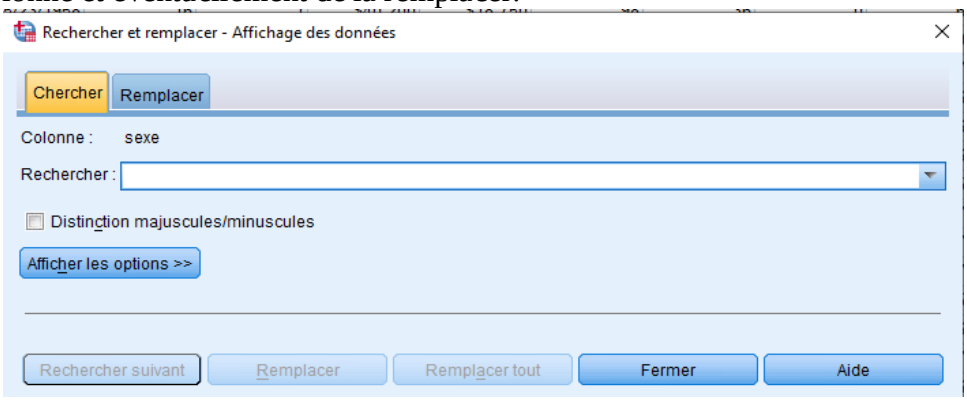
Sexe de l'employé

N	Valide	474
	Manquante	0

Sexe de l'employé

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Féminin	216	45,6	45,6	45,6
	Masculin	258	54,4	54,4	100,0
	Total	474	100,0	100,0	

- Rechercher valeur dans une case, permet de rechercher une valeur dans une colonne et éventuellement de la remplacer.



- Insérer un cas, permet d'insérer une ligne à la position où l'on se trouve.
- Insérer variable, permet d'insérer une variable.
- Décomposer fichier, décompose le fichier selon les valeurs d'une variable.

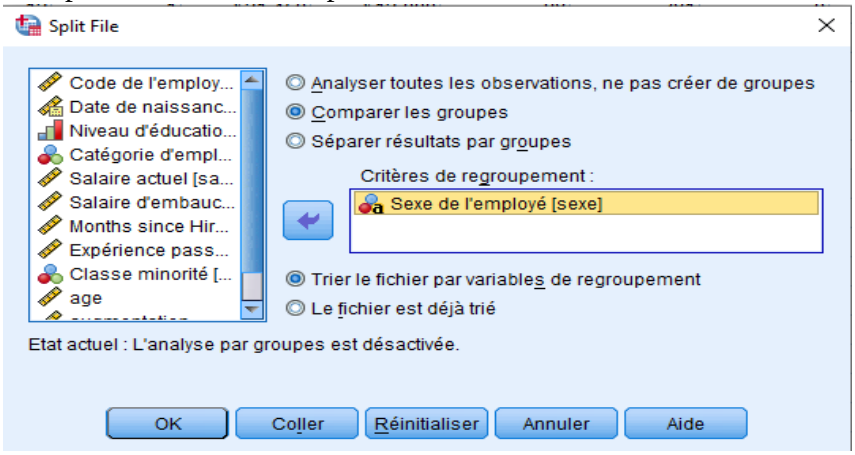


Figure II.6. Scinder un fichier.

Le choix de l'option « Comparer les groupes, permet de trier le fichier de données selon la variable choisie.

```
SAVE OUTFILE='C:\Users\pc\Desktop\Employee data.sav'
  /COMPRESSED.
SORT CASES  BY sexe.
SPLIT FILE LAYERED BY sexe.
```

Le choix de l'option « Séparer les résultats par groupes », permet de faire une séparation selon la variable choisie par effectifs.

- Affecter des poids, permet d'affecter des pondérations à des observations, nous allons revenir plus en détails sur le problème de pondération dans les sections précédentes.

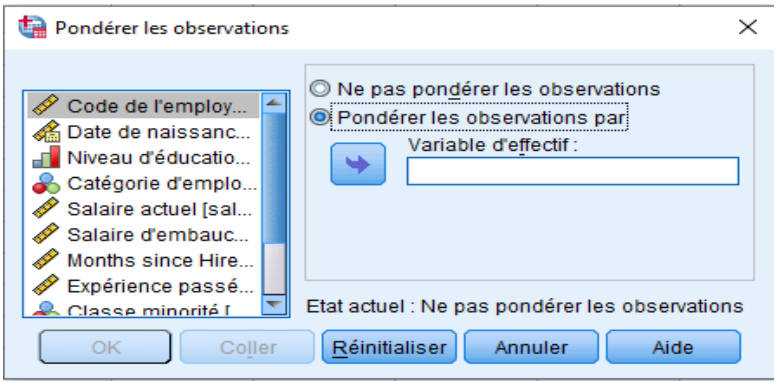
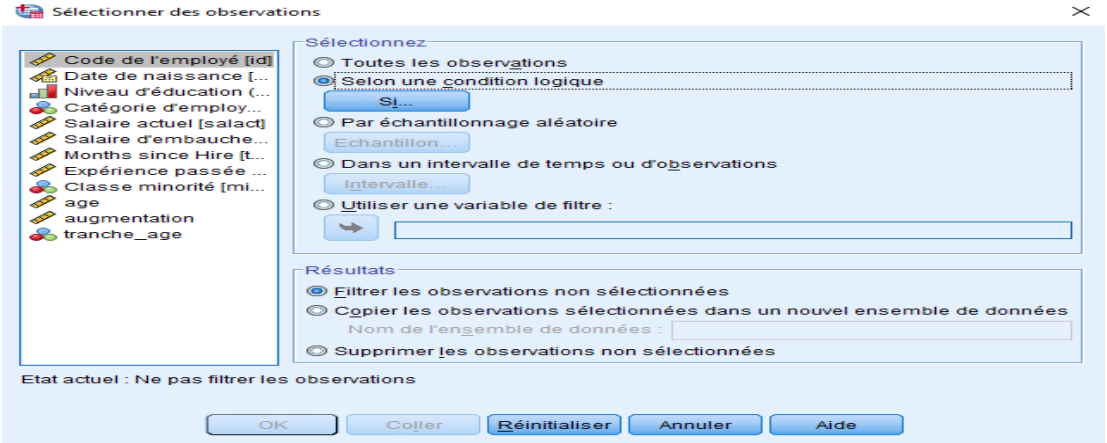


Figure II.7. Pondérer des observations.

- Sélectionner des cas (selon condition), permet de choisir des cas selon une condition.



- Bascule entre affichage numérique et caractères.
- Choix sous ensemble de vars.

- Utiliser toutes les variables, permet d'afficher toutes les variables de départ après avoir sélectionné un sous ensemble en cliquant sur le bouton précédent.

Chapitre III : Manipulation de fichiers

III.1 Types de fichiers de données

Il existe 3 types de fichiers de données générés par SPSS ayant différentes extensions, ces fichiers correspondent aux éditeurs à partir desquels ils sont générés.

- *Les fichiers de données :*

Les fichiers de données qui apparaissent dans la fenêtre Éditeur de données possèdent l'extension .sav suivant le nom.

Ces fichiers de données contiennent les données avec lesquelles on travaille et se présentent sous la forme d'une feuille quadrillée remplie de chiffres ou de lettres répartis en colonnes et en lignes.

Pour ouvrir un fichier de données dans SPSS on sélectionne à partir du menu : **Fichier > Ouvrir > Données …** et rechercher le fichier par son nom.

Pour créer un nouveau fichier de données, on sélectionne : **Fichier > Ouvrir > Données …** et on commence à saisir les données (comme nous allons le voir plus loin dans ce cours).

- *Les fichiers résultats :*

Ces fichiers possèdent l'extension .spv et sont générés par la fenêtre résultat (Output), Ils contiennent les traces de tous les traitements effectués depuis l'ouverture du fichier .sav.

Tant que le fichier n'est pas sauvegardé, il est nommé RESULTAT1 et à partir du moment ou est il sauvegardé il peut être manipulé comme tout autre fichier (on peut l'ouvrir, l'éditer, le sauvegardé à nouveau, etc.).

Pour ouvrir un fichier résultat dans SPSS on sélectionne à partir du menu : **Fichier > Ouvrir > Résultat …** et rechercher le fichier par son nom.

- *Les fichiers syntaxe :*

Ce troisième type de fichier est un éditeur de texte (comme le fichier précédent à la différence que celui-ci n'est pas généré automatiquement à l'ouverture d'un fichier de données. Il possède l'extension .sps, il représente une autre manière d'exécuter des commandes sur les données (sans utiliser les menus déroulants).

L'utilisation de la syntaxe permet de pouvoir suivre la progression des travaux statistiques et celui de pouvoir garder des traces de toutes les commandes et sous-commandes que vous aurez demandé un logiciel d'exécuter. Il est possible aussi d'inscrire des notes directement dans le fichier syntaxe, comme par exemple, *analyses bivariée en date du 1er septembre avec les données initiales*. Ces notes se retrouveront dans vos fichiers-résultats et vous permettront de mieux situer les tableaux.

Pour ouvrir un fichier syntaxe dans SPSS on sélectionne à partir du menu : **Fichier > Ouvrir > Syntaxe …** et rechercher le fichier par son nom.

III.2 Exportation de fichiers de données

Les trois types de fichier générés par SPSS peuvent être exportés vers d'autres logiciels et peuvent par conséquent s'afficher avec une autre extension.

- Les fichiers .sav (générés par l'éditeur) peuvent être enregistrés dans les formats de sauvegarde présents dans la boite de dialogue, ceci est possible en cliquant sur : **Fichier > Enregistrer sous >…**

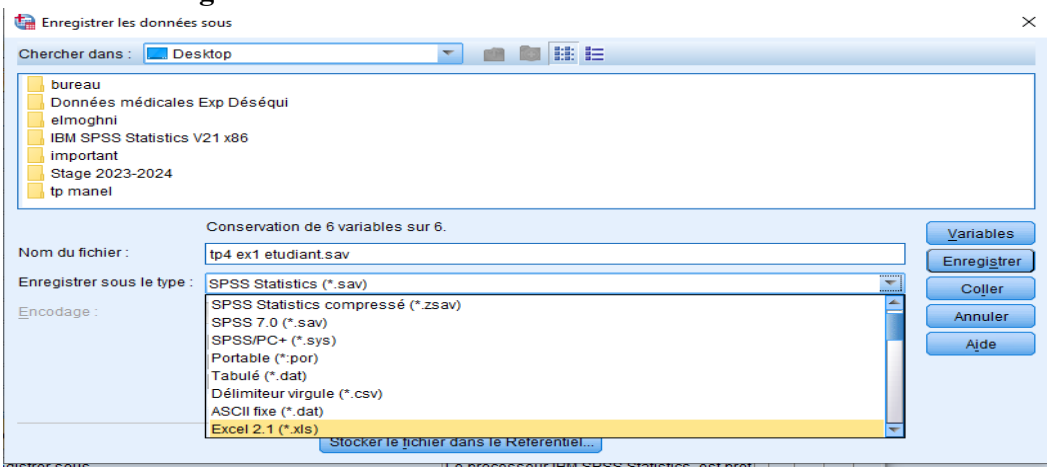


Figure III.1. Exportation d'un fichier .sav vers Excel.

- Les fichiers .spv (générés par la fenêtre des résultats) peuvent aussi être exportés vers d'autres formats, dans le but de faire d'autres traitements ou pour une meilleure visualisation, dans le menu de la page résultats on clique sur : **Fichier > Exporter >…**

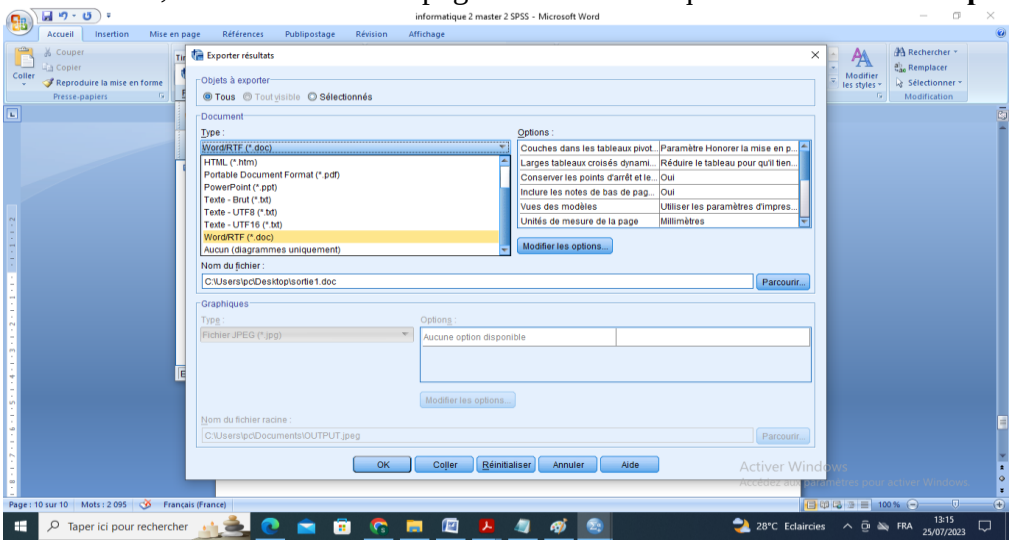


Figure III.2. Exportation d'un fichier .spv vers un autre format.

Par défaut, l'exportation se fait vers un document word (qui peut être changée), on doit aussi choisir un emplacement pour le fichier généré et lui donner un nom.

III.3 Importation de fichiers de données (Kambou H.K., 2021)

On peut aussi ouvrir des fichiers avec différents formats et issus de différentes sources dans SPSS :

III.3.1 Lecture des données Excel

Des quantités importantes de données peuvent se trouver dans un autre format que celui de SPSS. Au lieu de les saisir dans l'éditeur de données, elles peuvent être transférées directement à partir du logiciel en question. Il existe deux manières pour importer les fichiers Excel dans SPSS :

- *1ère méthode :*

- On sélectionne dans le menu **Fichier > Ouvrir > Données > …**
- Aller vers le dossier Samples\French et sélectionner demo.xlsx

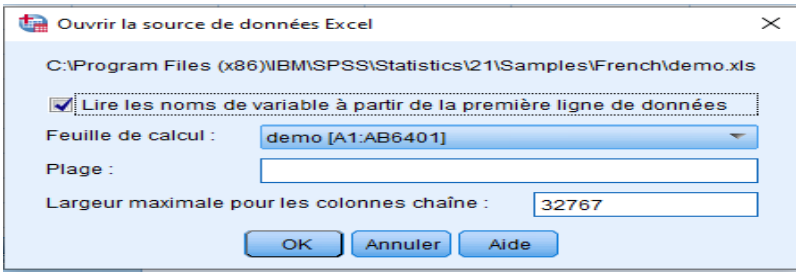
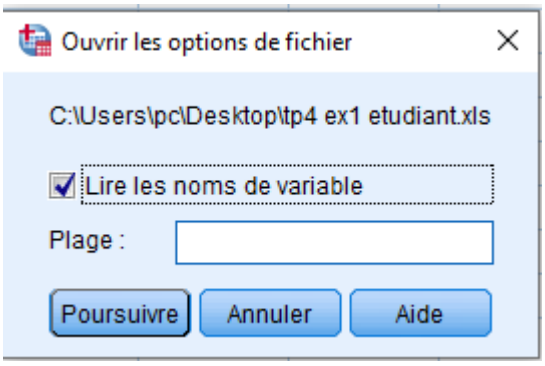


Figure III.3. Boite de dialogue lire fichier Excel.

- Vérifiez que l'option Lire les noms de variable à partir de la première ligne de données est sélectionnée. Les en-têtes de colonne qui ne sont pas conformes aux règles de dénomination de variables sont convertis en noms de variable valides. Les en-têtes de colonne d'origine sont enregistrés en tant que libellés de variable.
- Cliquer sur OK

- *2ème méthode :*

- Glisser le fichier Excel dans une fenêtre SPSS,



- S'assurer que l'option lire les noms de variable est cochée,
- Cliquer sur poursuivre.

III.3.2 Lecture des données texte

Les fichiers texte représentent une autre source commune de données. De nombreux tableurs et bases de données peuvent enregistrer leur contenu dans l'un des nombreux formats de fichier texte. Comme pour les fichiers Excel, il existe aussi deux méthodes pour importer des données texte dans SPSS.

- *1ère méthode :*

- On sélectionne dans le menu **Fichier > Lire les Données Texte> …** on peut aussi utiliser le menu **Fichier > Ouvrir > Données > …**
- Choisir le chemin du fichier texte et le sélectionner

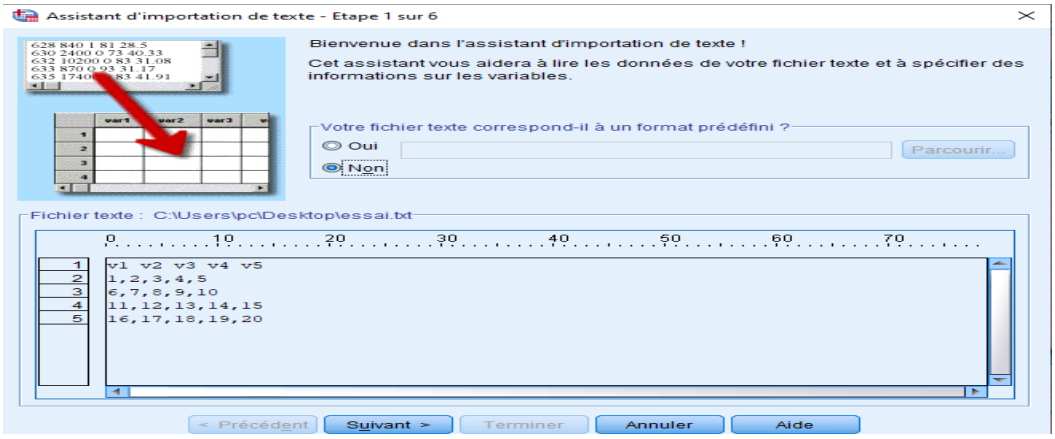


Figure III.4. Assistant d'importation du texte (étape 1).

- Cliquer sur Suivant,

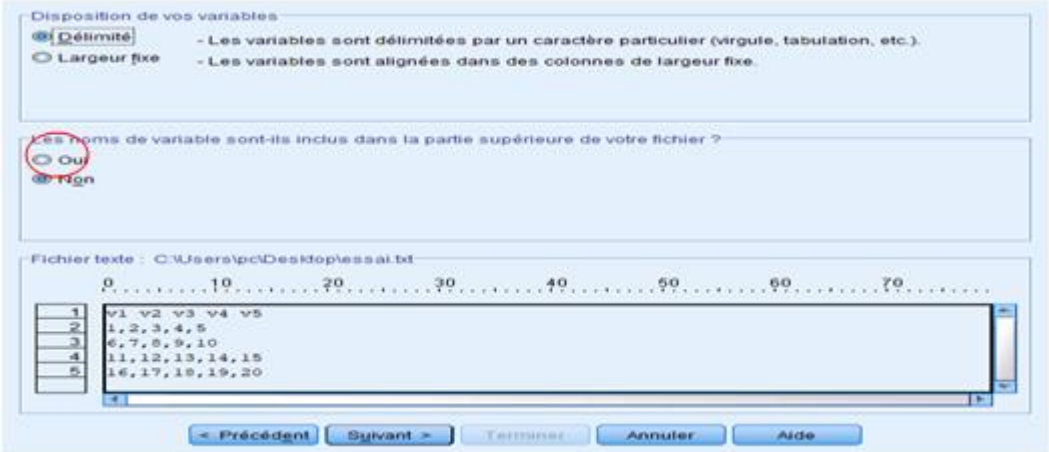


Figure III.5. Assistant d'importation du texte (étape 2).

- Cocher Oui pour la question « Les noms de variables sont-ils inclus dans la partie supérieure de votre fichier ? »,
- Cliquer sur Suivant,

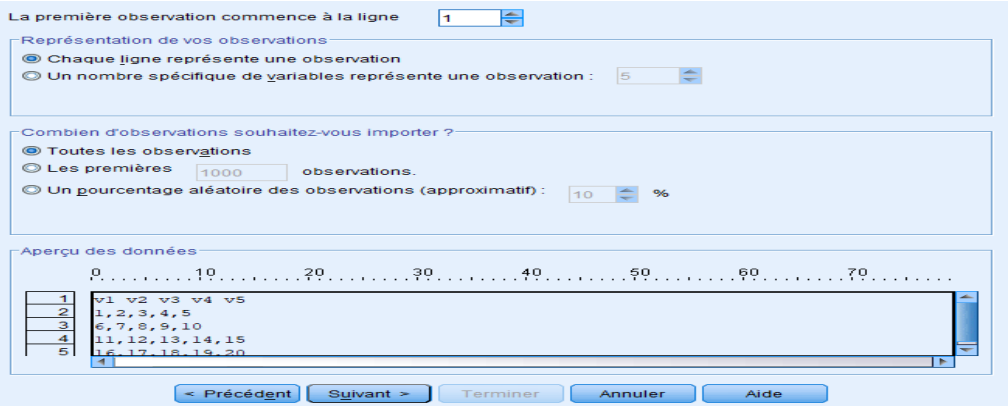


Figure III.6. Assistant d'importation du texte (étape 3).

- Cliquer sur Suivant,

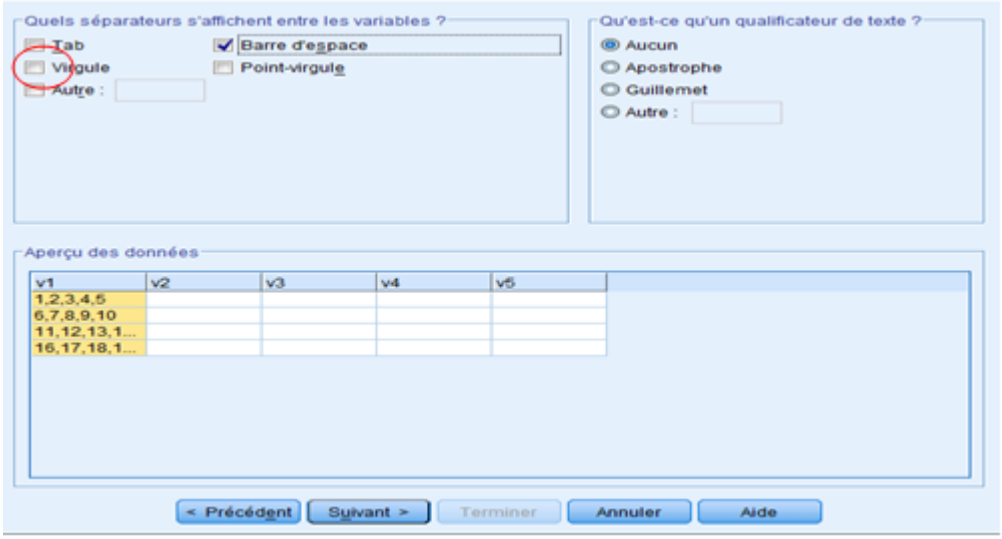


Figure III.7. Assistant d'importation du texte (étape 4).

- On coche virgule pour la question « Quels séparateurs s'affichent entre les variables ? », cela peut changer d'un fichier à un autre selon le séparateur utilisé qui peut être simplement un espace,
- Cliquer sur Suivant,
- Enfin cliquer sur Terminer.

Pour la deuxième méthode, on glisse le fichier texte dans l'éditeur de données SPSS, et on recommence les étapes de la première méthode.

Chapitre IV : Saisie des données

Les données peuvent être saisies dans l'éditeur de données, qui peut s'avérer utile pour traiter les fichiers de données peu volumineux ou pour apporter de légères modifications à des fichiers de données plus volumineux.

Exemple IV.1 :

On considère l'exemple du Tableau 1.

Epargne	Salaire	Education
4000	35000	1
2000	40000	1
3000	28000	1
1100	16000	0
800	12000	0
900	24000	0

Tableau IV.1 : Exemple épargne.

- Préparer l'entrée des données en définissant les variables dans la vue des variables. Entrer ces données dans la vue des données :
  - Cliquez sur l'onglet Vue de variable en bas de la fenêtre de l'éditeur de données. Vous devez définir les variables qui seront utilisées, on remarque que toutes les variables sont numériques.

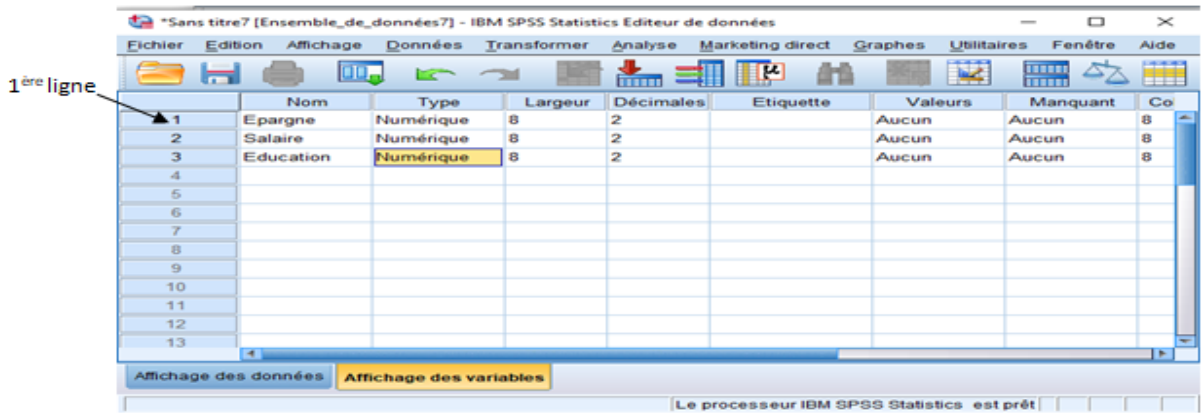


Figure IV.1. Saisie des variables de l'Exemple 1 dans la vue Affichage des variables.

- Dans la 1ère ligne, on saisit Epargne,
- Dans la 2ème, on saisit Salaire,
- Enfin dans la 3ème et dernière ligne, on saisit Education,
- On met Décimales à 0 pour les trois lignes (pour éviter d'avoir des chiffres après la virgule).

Un type de données numérique est automatiquement attribué aux nouvelles variables. Si vous ne saisissez pas de noms de variable, des noms uniques sont automatiquement créés. Cependant, ces noms ne sont pas descriptifs et ne sont pas recommandés pour les fichiers de données volumineux.

- On clique sur l'onglet « vue de données » pour saisir les données (les lignes du Tableau 1 à partir de la deuxième).

Les noms saisis dans la vue de variable sont à présent les en-têtes des trois premières colonnes dans Vue de données. Commencer à saisir des données dans la première ligne, en commençant par la première colonne.

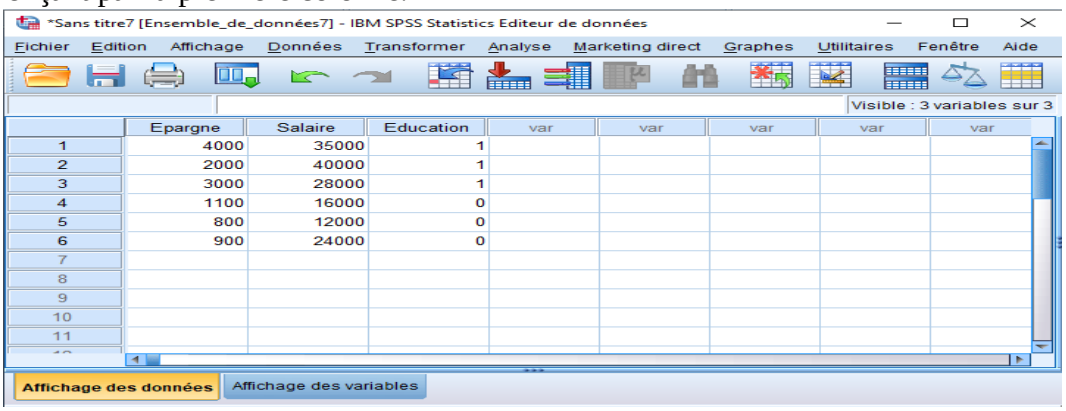


Figure IV.2. Saisie des données l'Exemple 1 dans la vue Affichage des données.

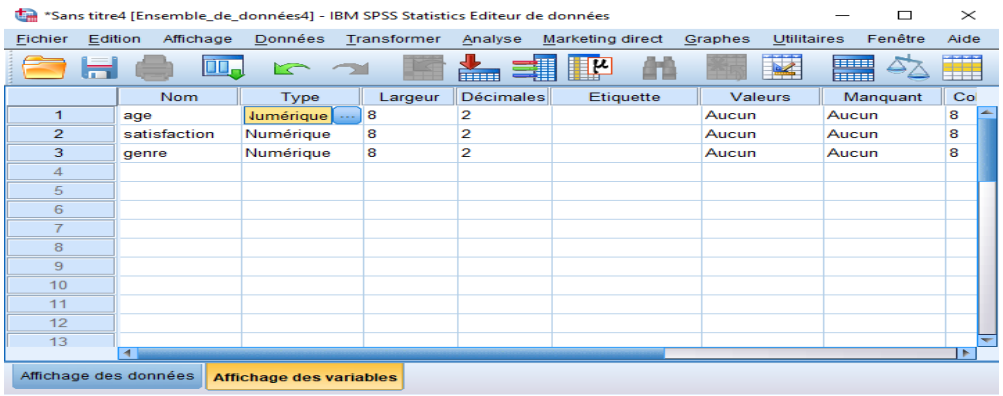
Exemple IV.2 :

Sur un évènement organisé on souhaite savoir ce que les participants pensent de ce dernier, pour cela on collecte 3 informations sur chaque invité (âge, genre, et satisfaction).

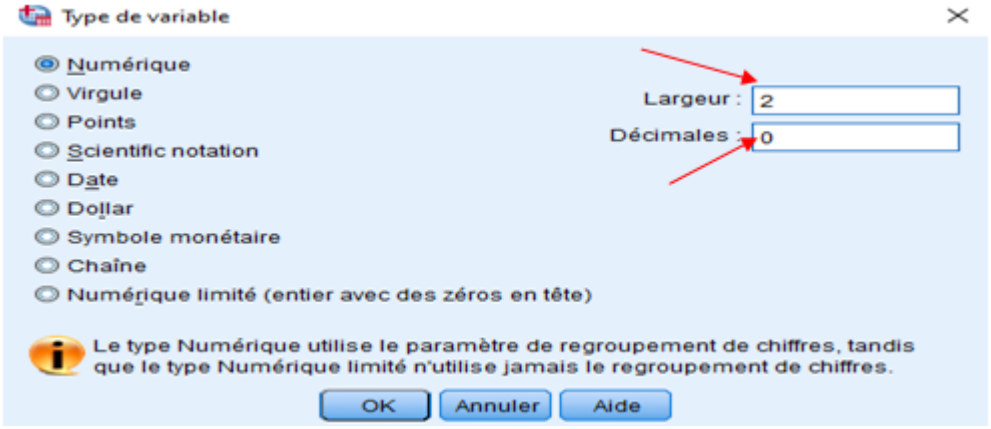
Age	Satisfaction	Genre
21	3	1
17	2	0
16	3	1
17	2	1
13	5	0

Tableau IV.2 : Exemple évènement.

- Créer le jeu de données sous SPSS, sachant que : Satisfaction prend les valeurs : « 1 » : Pas Satisfait ; « 2 » : Pas Très Satisfait ; « 3 » : Moyennement satisfait ; « 4 » : Satisfait ; « 5 » : Très Satisfait. Genre prend les valeurs : « 1 » : Femme ; « 0 » : Homme ;

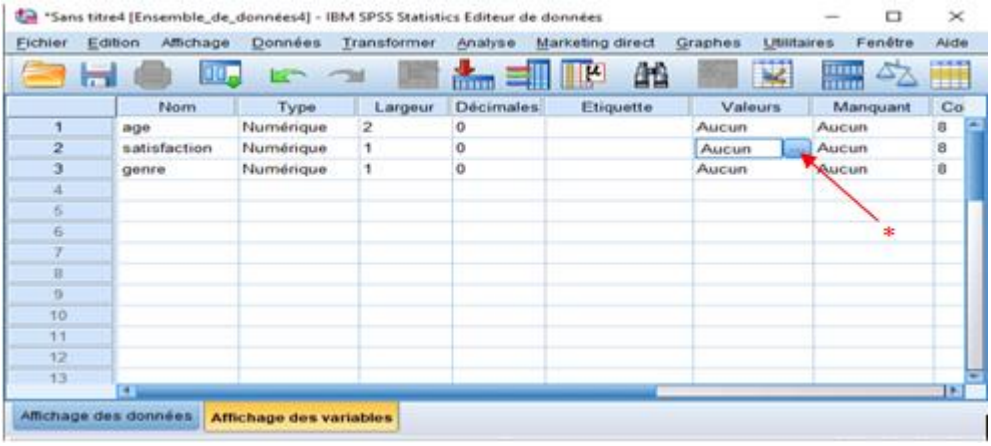


- On écrit les noms de variables dans la zone nom,
- Le Type est par défaut numérique, pour l'âge on a une largeur de 2 max pour les deux autres variables on a largeur 1, dans Type, on modifie la largeur et on met décimales à 0 (on peut le faire directement sur les propriétés Largeur et Décimales.

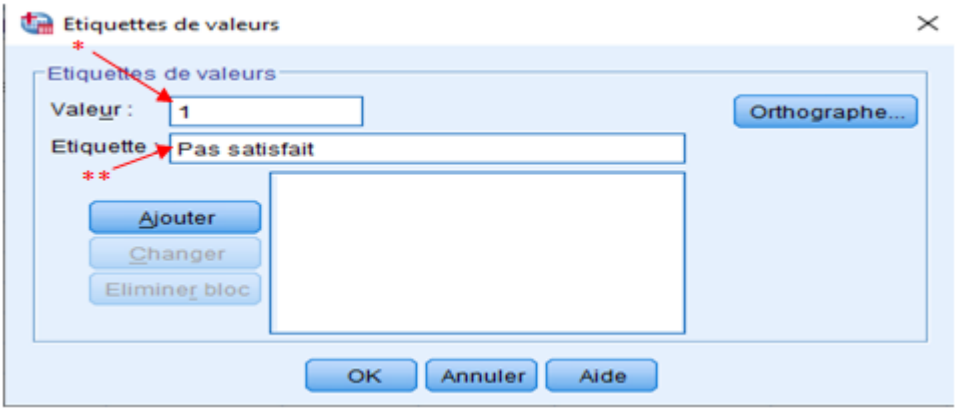


Pour les variables, Satisfaction, on introduit les valeurs suivantes :
« 1 » : Pas Satisfait ; « 2 » : Pas Très Satisfait ; « 3 » : Moyennement satisfait ; « 4 » : Satisfait ; « 5 » : Très Satisfait.

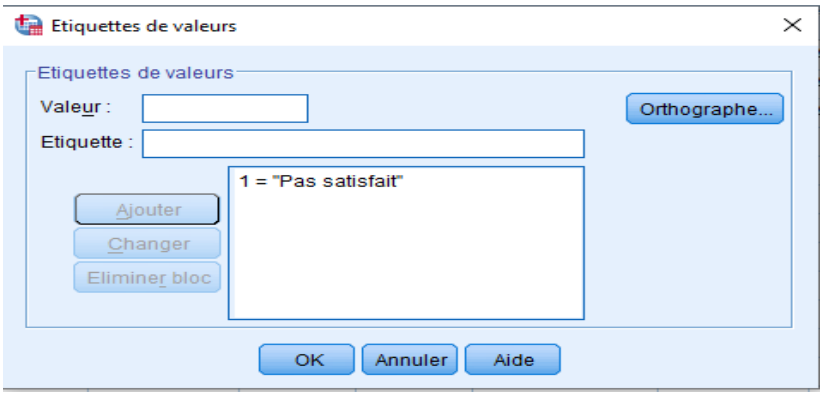
- On clique sur la propriété Valeurs, de la ligne qui correspond à la ligne Satisfaction



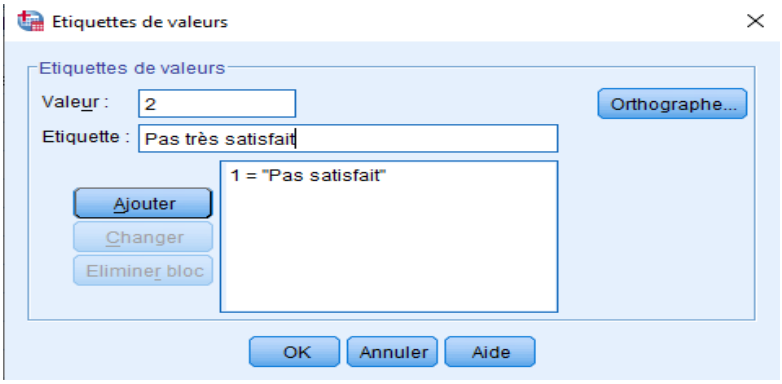
- On clique sur * (comme montré sur la figure ci-dessus),



- On met le code numérique dans * et la chaine de caractères dans **,
- On clique sur le bouton Ajouter,



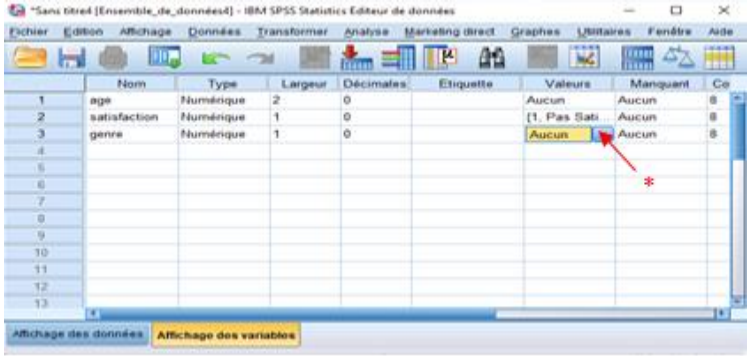
- On continue avec la 2ème valeur et ainsi de suite jusqu'à épuisement de toutes les valeurs.



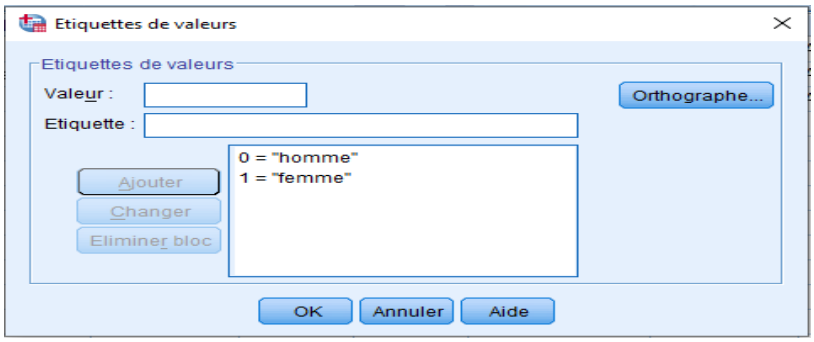
- Enfin on introduit toutes les valeurs

On fait la même chose avec la variable Genre :

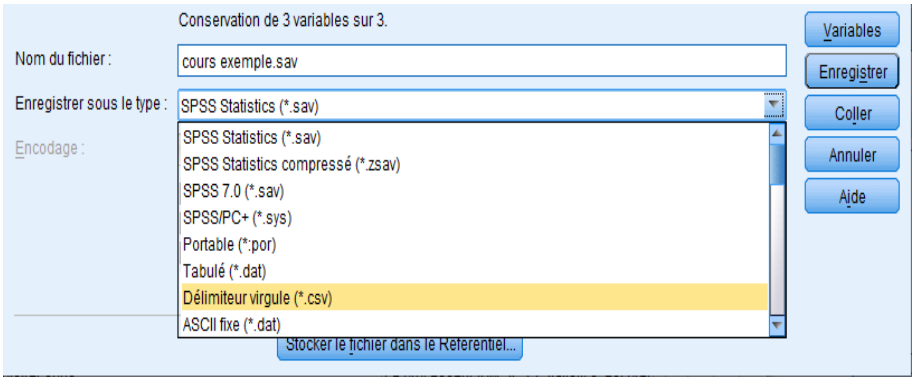
- On clique sur Valeurs qui correspond à la ligne de la variable en question



- On introduit 1 pour Femme et 0 pour Homme,



- Enregistrez vos données au format par défaut (.sav). Changez l'enregistrement au format (.csv) quelle est la différence ?
  - Pour un premier enregistrement on utilise le menu **Fichier > Enregistrer sous> …** choisir le chemin d'enregistrement et donner un nom au fichier.
  - A partir du 2ème enregistrement on utilise le menu **Fichier > Enregistrer> …**
  - Pour enregistrer le fichier au format .csv, on clique sur **Fichier > Enregistrer sous> …** On choisit le format .csv

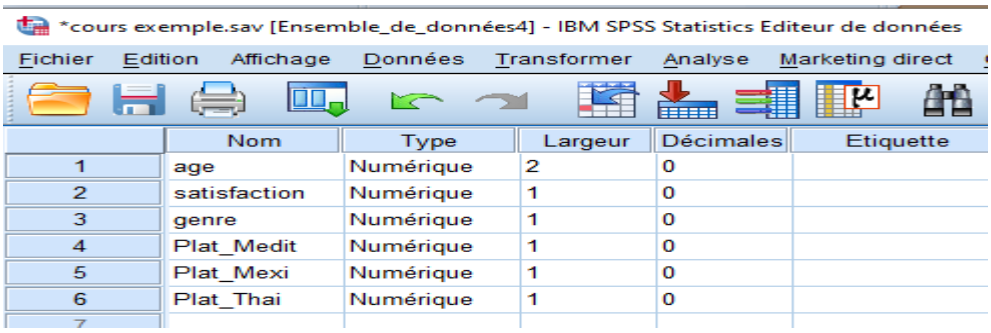


La différence réside dans le faite que le fichier .sav est un fichier de données SPSS et le fichier .csv est un fichier Excel (on n'a plus de vue de variables juste une vue de données).

- Dans la même enquête, on demande aux invités le type de plat qu'ils ont consommé entre : Méditerranéen (1) ; Mexicain (2) ; Thaïlandais (3). Ajouter cette variable 'TypedePlat ' à la matrice de données précédente sachant que : Le premier invités à mangé tous les types de plats, le second a consommé un plat mexicain, le troisième a consommé un plat thaïlandais, le quatrième a consommé un plat mexicain et l'autre thaïlandais, et le dernier a consommé un plat méditerranéen.

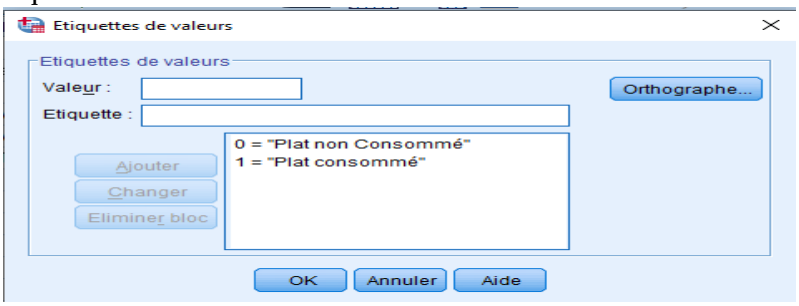
La question est très souvent présente dans les questionnaires car la variable TypedePlat est une variable composée, on l'appelle question à choix multiples pour laquelle les réponses peuvent être une deux ou plusieurs selon le nombre de choix.

La solution pour ce type de question est de créer pour chaque réponse une variable qui lui correspond. Pour note cas, nous avons trois réponses qui correspondent à trois plats => on crée 3 variables : Plat_Medit, Plat_Mexi, Plat_Thai Avec les valeurs {0 : Plat non consommé, 1 : Plat consommé}. On crée les trois variables dans la fenêtre Affichage des variables :

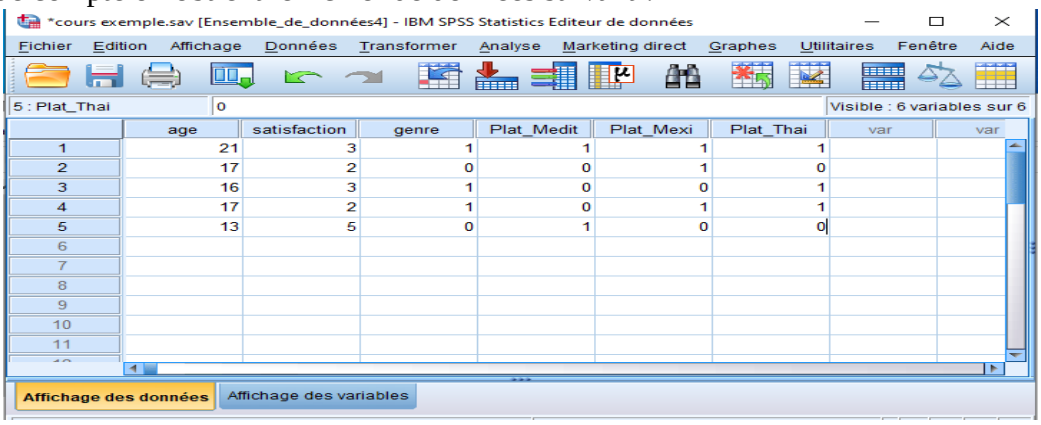


	Nom	Type	Largeur	Décimales	Etiquette
1	age	Numérique	2	0	
2	satisfaction	Numérique	1	0	
3	genre	Numérique	1	0	
4	Plat_Medit	Numérique	1	0	
5	Plat_Mexi	Numérique	1	0	
6	Plat_Thai	Numérique	1	0	

Pour chacune des trois dernières variables, on introduit 1 et 0 pour désigner la consommation ou non du plat en question.



Enfin de compte on obtient le fichier de données suivant :



	age	satisfaction	genre	Plat_Medit	Plat_Mexi	Plat_Thai	var	var
1	21	3	1	1	1	1		
2	17	2	0	0	1	0		
3	16	3	1	0	0	1		
4	17	2	1	0	1	1		
5	13	5	0	1	0	0		
6								
7								
8								
9								
10								
11								

IV.1 Gestion des données manquantes

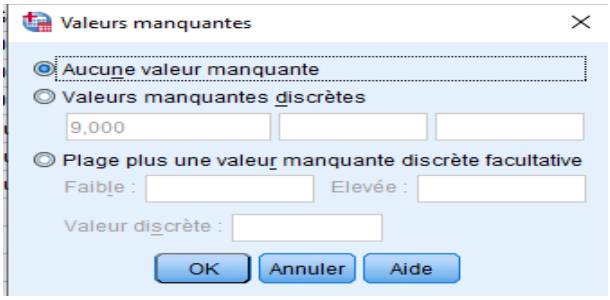
Les données manquantes sont des données non valides et sont beaucoup trop nombreuses pour être ignorées. Elles peuvent être le résultat de refus de réponse à une ou plusieurs questions par les répondants, ou dans le cas où les réponses données sont dans un format inattendu. Si on ne traite pas ces données les résultats de l'analyse peuvent être imprécis.

La raison pour laquelle une valeur est manquante peut être importante pour votre analyse. Par exemple, on peut juger utile de distinguer les personnes qui ont refusé de répondre à une question de celles qui n'ont pas répondu car cette question ne les concernait pas.

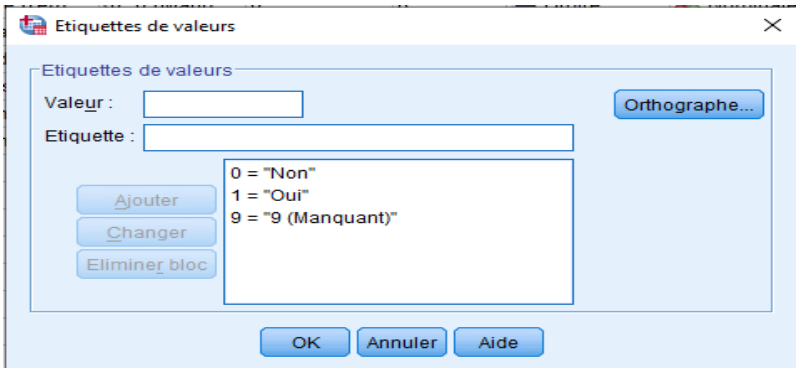
IV.1.1 Valeurs manquantes pour variables numériques

Pour ce type de données, les champs de données vides ou contenant des valeurs non valides sont convertis en valeurs manquantes par défaut, signalés par une virgule (Windows français) ou par un point (Windows anglais).

- Ouvrir le fichier employee data.sav : **Fichier > Ouvrier > …** ,
- Cliquer sur l'onglet « vue de variables » en bas de la fenêtre Editeur de données,
- Cliquer sur la cellule « manquante » de la variable minorité, on peut indiquer dessus jusqu'à trois valeurs manquantes ou une plage de valeurs et une valeur discrète supplémentaire,



- Sélectionner Valeurs manquantes discrètes,
- Saisir 9, sur la première zone de texte, ne rien saisir sur les autres zones,
- Cliquer sur OK pour enregistrer les modifications et revenir à la fenêtre Editeur de données,
- Cliquer sur la cellule Valeurs de la variable minorité et ajouter le libellé « 9 (manquant) » pour l'associer à la valeur manquante,



IV.1.2 Valeurs manquantes d'une variable chaine de caractères

Les valeurs manquantes des variables de type chaîne de caractères sont gérées de la même façon que celles des variables numériques. Mais contrairement aux valeurs numériques, les champs vides dans les variables de chaîne ne sont pas désignés comme données manquantes par défaut. Ils sont pris en compte en tant que des chaînes de caractères vides.

- Cliquer sur l'onglet « Vue de variable » en bas de la fenêtre de l'éditeur de données,
- Cliquer sur la cellule Manquante de la ligne de la variable Sexe, puis sur le bouton à droite de la cellule pour ouvrir la boîte de dialogue Valeurs manquantes,

- Sélectionner Valeurs manquantes discrètes,
- Saisir NR dans la première zone de texte, les valeurs manquantes des variables de chaîne distinguent les majuscules des minuscules. Par conséquent, la valeur nr n'est pas traitée comme une valeur manquante,
- Cliquer sur OK pour enregistrer vos modifications et revenir dans l'éditeur de données.
- Cliquer sur la cellule Valeurs de la ligne sexe, puis sur le bouton à droite de la cellule pour ouvrir la boîte de dialogue Libellés de valeur,
- Saisir NR dans le champ Valeur,
- Saisir Non répondu dans le champ Libellé.

Chapitre V : Fusionner et scinder des fichiers

V.1 Fusionner des fichiers

L'opération de fusion de fichiers peut être réalisée de 2 manières, horizontale ; dans ce cas on ajoute des lignes ou des cas au fichier de données initial, verticale ; dans ce cas des variables sont ajoutées au fichier de données initial.

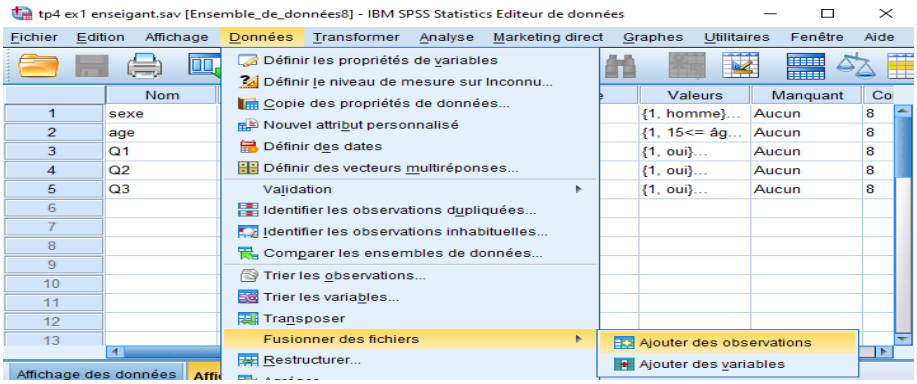
Les fusions sont effectuées sous un certain nombre de conditions :

- a)- La fenêtre du fichier SPSS sur laquelle on ajoute des données doit être active (le fichier à ajouter peut être ouvert ou fermé,
- b)- Les variables à la même position doivent avoir le même type,
- c)- Les deux fichiers peuvent avoir un nombre de variables différent,
- d)- Les variables doivent, de préférence, avoir les mêmes noms.

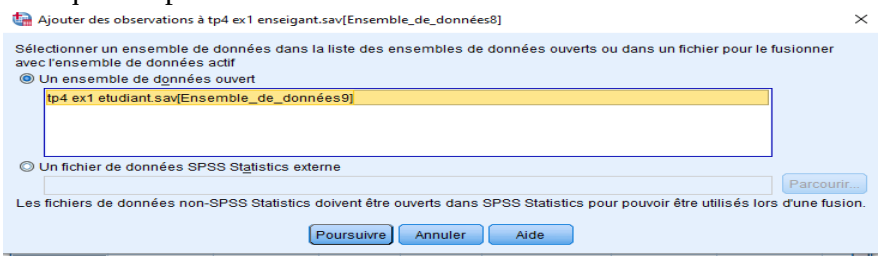
V.1.1 Les étapes de la fusion horizontale

1. Ouvrir le fichier auquel on va ajouter des données (observations ou lignes),

2. Si le fichier à ajouter est ouvert, on clique sur **Données > Fusionner des fichiers > Ajouter des observations>…**

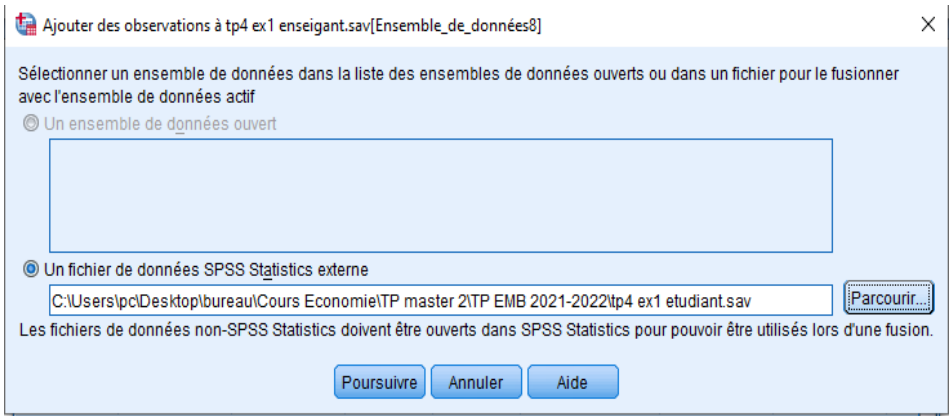


3. Si l'ensemble de données à ajouter est ouvert, on sélectionne son nom dans la boite de dialogue, et on clique sur poursuivre.



4. Si l'ensemble de données à ajouter est fermé,

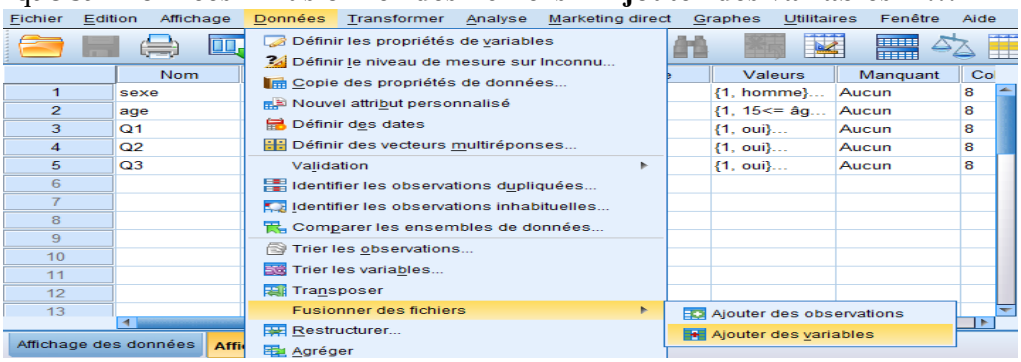
5. on clique sur le bouton parcourir et on sélectionne le fichier à ajouter, ensuite on clique sur poursuivre.



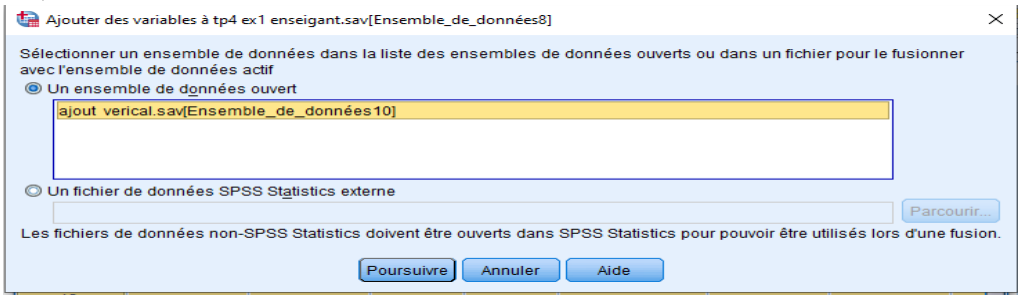
V.1.2 Les étapes de la fusion Verticale

Pour ce ca de figure, les deux fichiers peuvent ne pas avoir le même nombre d'observations mais le fichier à ajouter doit contenir des noms de variables différents du premier.

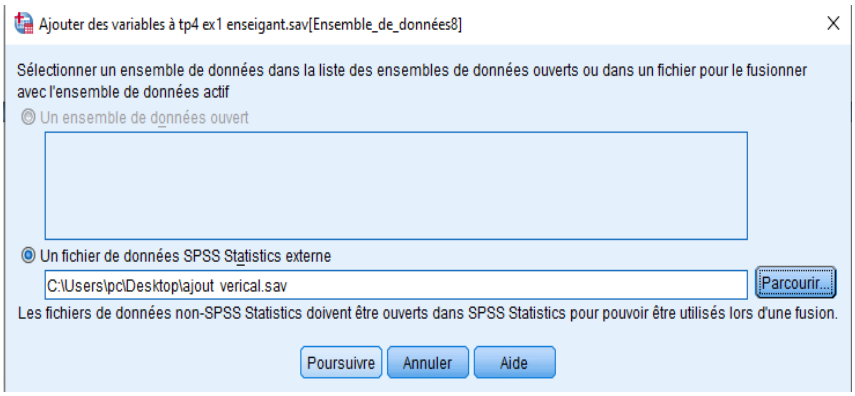
1. Ouvrir le fichier auquel on va ajouter des variables,
2. On clique sur **Données > Fusionner des fichiers > Ajouter des variables > …**



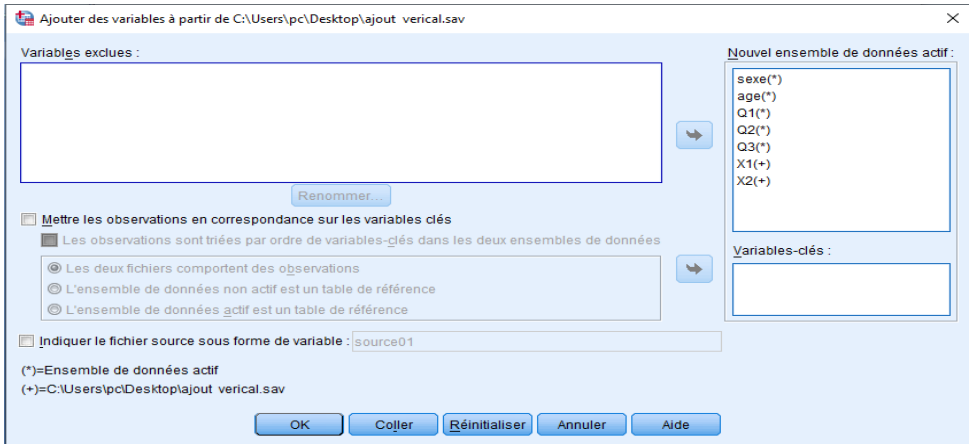
3. Si le fichier à ajouter est ouvert, on le sélectionne dans la boite de dialogue, et on clique sur poursuivre,



4. Si l'ensemble de données à ajouter est fermé,
5. on clique sur le bouton parcourir et on sélectionne le fichier à ajouter, ensuite on clique sur poursuivre.



6. on obtient la fenêtre suivante :



7. les symboles (*) et (+) indiquent respectivement que la variable apparient au premier fichier (auquel on ajoute des données) ou le second (celui qu'on ajoute), ceci est utilise dans le cas où on a les mêmes noms de variables ou des variables avec des types différents pour pouvoir choisir la variable qu'on sélectionne (qui apparait dans le fichier final de fusion).

Exemple V.1 :

- Créer le fichier de données nommé soirée dans SPSS :

Age	Satisfaction	Genre	Cuis_medit	Cuis_mex	Cuis_thai
21	3	1	1	1	1
17	2	0	0	1	0
16	3	1	0	0	1
17	2	1	0	1	1
13	5	0	1	0	0

Tableau V.1 : Exemple évènement.

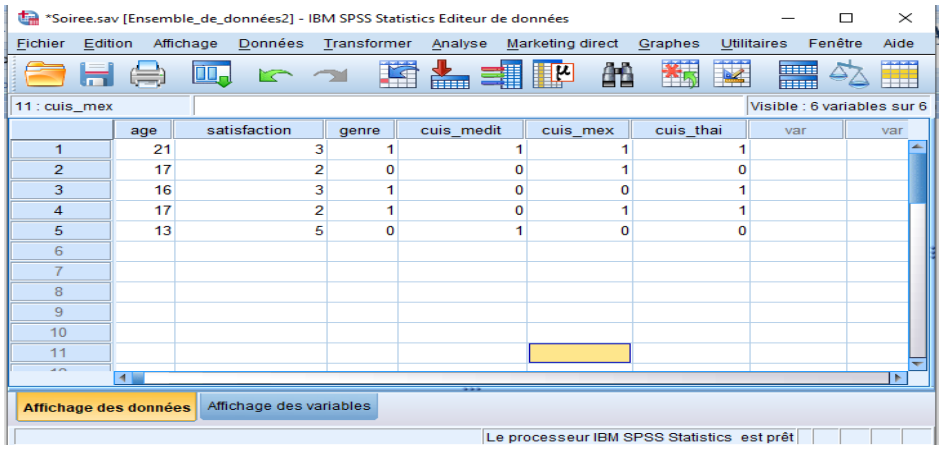
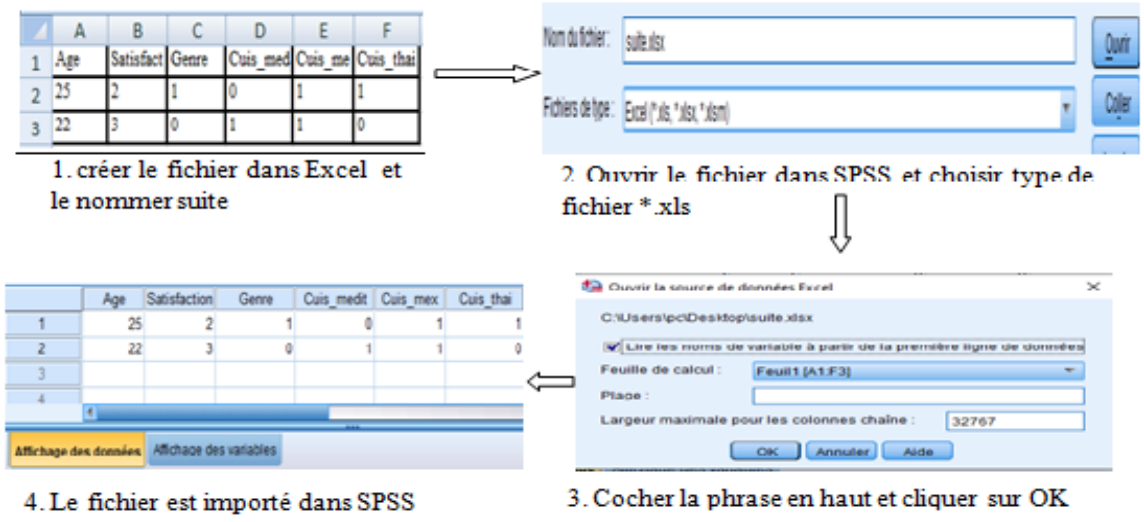


Figure V.1. Fichier soirée.sav de l'exemple évènement.

- Créer au niveau d'Excel le fichier suivant nommé « suite », et importer le dans SPSS :

Age	Satisfaction	Genre	Cuis_medit	Cuis_mex	Cuis_thai
25	2	1	0	1	1
22	3	0	1	1	0

Tableau V.2 : Exemple Évènement suite.



- Fusionner le fichier Excel créé avec le fichier SPSS de la Q1 et vérifier que les valeurs sont correctes.
  - Ouvrir le fichier soirée.sav,
  - Cliquer sur **Données > Fusionner des fichiers > Ajouter des observations>…**
  - On cherche le fichier suite.sav, on sélectionne son nom dans la boite de dialogue, et on clique sur poursuivre.
- Créer le fichier suivant et enregistrer le sous le nom suite2.sav

Poids
85
55

96
56
78
65
77

Tableau V.3 : Exemple Évènement suite2.

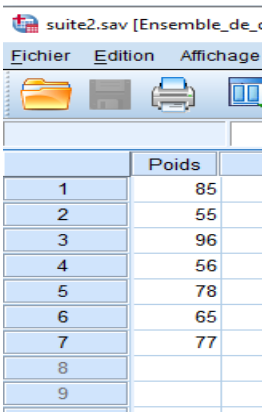


Figure V.2. Fichier suite2.sav.

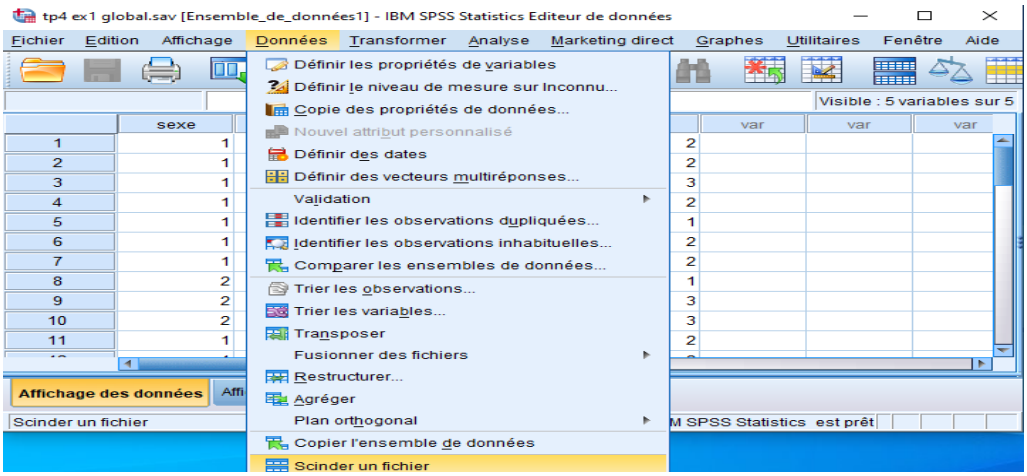
- Fusionner le fichier obtenu précédemment avec ce dernier.
  - Ouvrir le fichier soirée.sav,
  - On clique sur **Données > Fusionner des fichiers > Ajouter des variables > …**
  - On sélectionne le fichier suite2.sav,
  - On obtient la fenêtre dans laquelle la variable poids et ajouté à l'ensemble des variables du fichier soirée.sav :

V.2 Scinder des fichiers

Consiste à décomposer un fichier en deux parties en se basant sur une ou plusieurs propriétés. Des analyses d'effectifs succèdent une décomposition pour pouvoir tirer des conclusions et faire une étude des données.

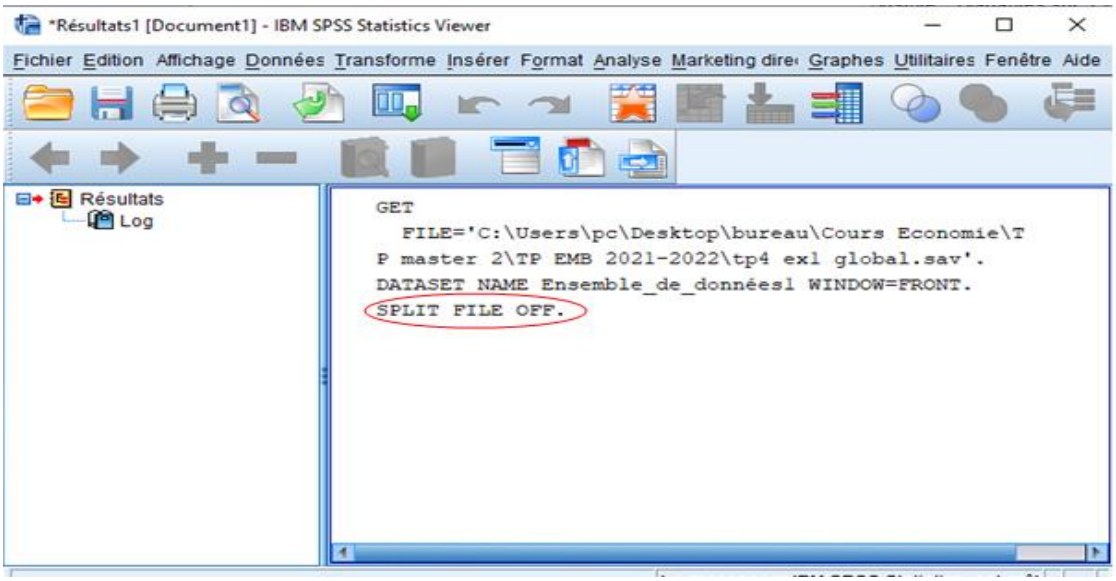
Pour scinder un fichier de données SPSS on choisit dans le menu :

1. **Données > Scinder un fichier > Comparer les groupes > …**



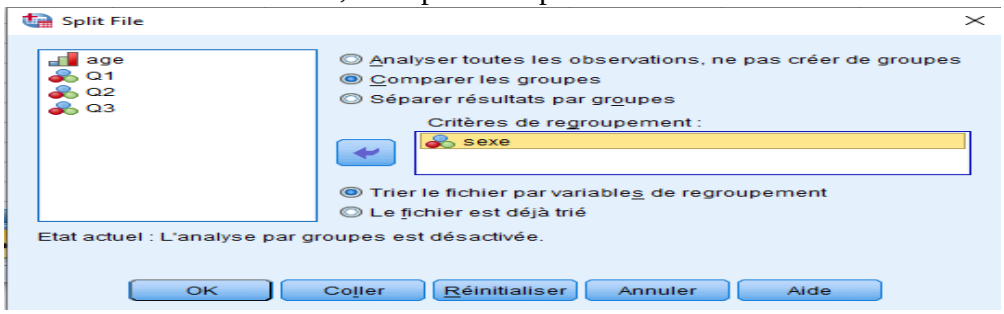
2. On choisit l'une des trois options qui s'affichent,

- La première option « Analyser toutes les observations, ne pas créer de groupes », ne décompose pas le fichier initial. En cliquant sur OK, on obtient le fichier résultat qui indique que le fichier n'a pas été décomposé,

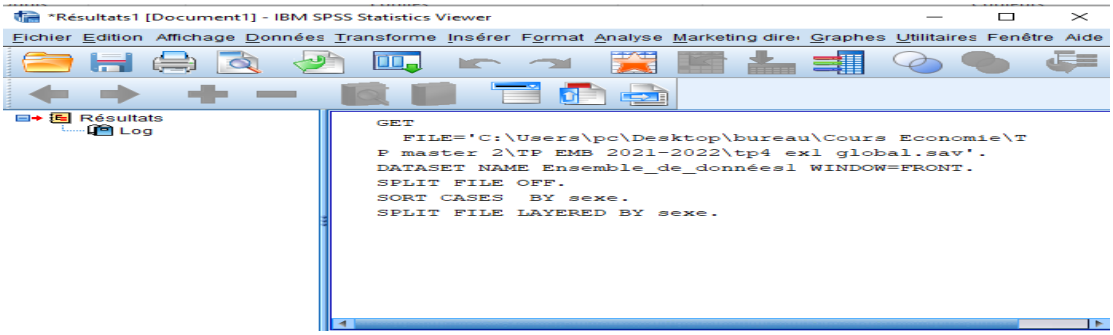


- La deuxième option « comparer les groupes », est associée au choix d'une variable de coupure choisie dans la boite à gauche parmi l'ensemble des variables.

Une fois la variable choisie, sexe par exemple.

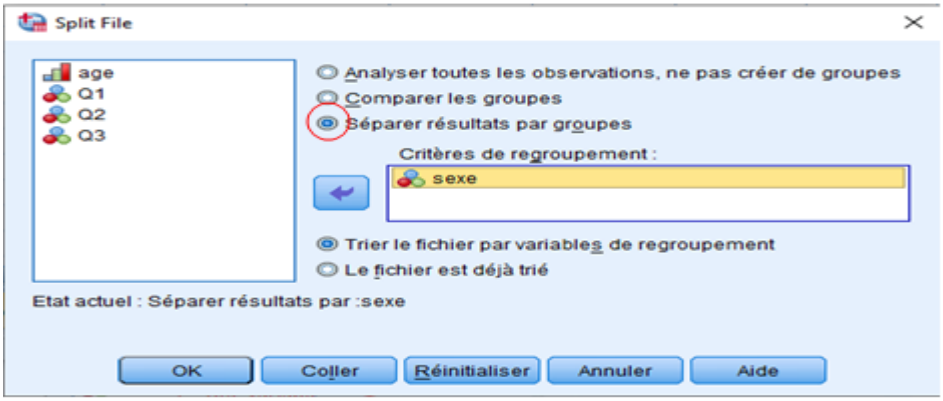


On clique sur OK, on obtient une fenêtre résultat qui indique que le fichier est décomposé selon la variable sexe.

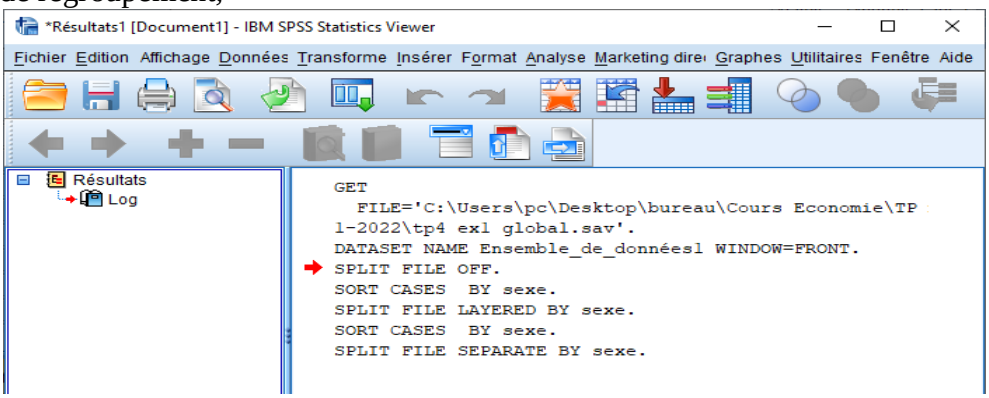


Dans la fenêtre Affichage de données, le fichier de données est ordonné selon la variable sexe.

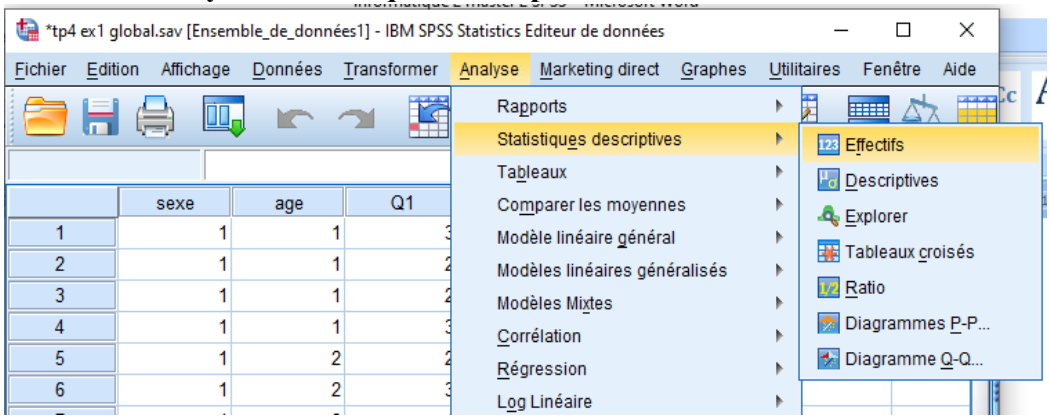
- La troisième option « Séparer résultats par groupes » consiste aussi à choisir un critère de regroupement comme la précédente,



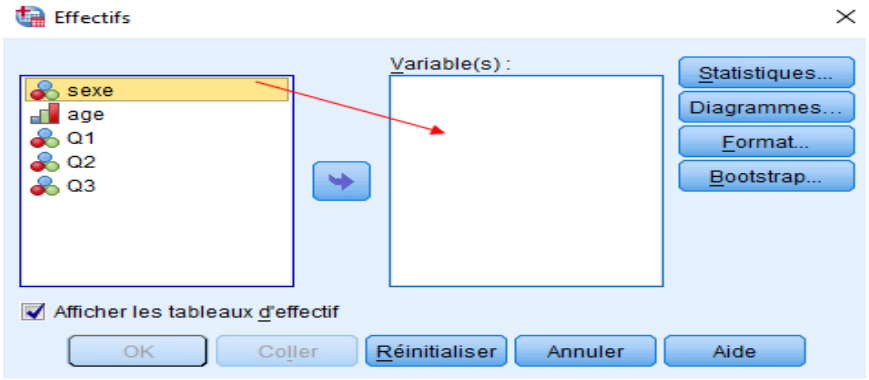
On obtient une fenêtre résultat, qui mentionne le type de décomposition ainsi que la variable de regroupement,



Comme nous l'avons mentionné précédemment, l'intérêt de la décomposition réside dans le fait qu'on puisse analyser les effectifs des groups créés. Pour analyser es effectifs, on choisit le menu : **Analyse > Statistiques descriptive > Effectifs > …**



Si on reconsidère les deux dernières options de la décomposition d'un fichier, et on applique l'analyse des effectifs,



On transfert la variable de regroupement dans la zone droite et on clique sur OK, on obtient des effectifs selon les valeurs de la variable sexe.

Effectifs

[Ensemble_de_données1] C:\Users\pc\Desktop\bureau\Cours Economie\TP master 2\TP EMB 2

Avertissements

Les tableaux de fréquences ne sont pas générés pour les variables suivantes qui sont des variables de division : sexe.

Statistiques

sexe

homme	N	Valide	14
		Manquante	0
femme	N	Valide	6
		Manquante	0

V.2.1 Activation et désactivation du traitement d'un fichier scindé

Après avoir lancé le traitement d'un fichier scindé, ce dernier reste actif pour le reste de la session, il y a toujours possibilité de le désactiver.

- Analyser toutes les observations : Cette option désactive le traitement du fichier scindé.
- Comparer les groupes et Séparer la sortie par groupes : Cette option active le traitement du fichier scindé.

Au cas où le traitement du fichier scindé est activé, le message Séparer fichier actif apparaît sur la barre de statut située au bas de la fenêtre de l'application.

Exemple IV.2 :

- On considère le fichier étudiant.sav suivant :

L'âge possède 4 valeurs :

1 : .15<= âge<20

2 : 20<=âge<30

3 : 30<=âge<40

4 : âge>=40

Sexe : 1 homme et 2 femme

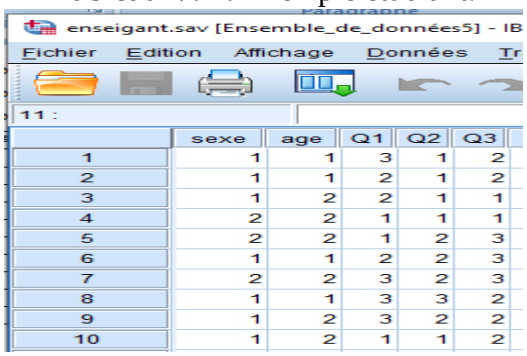
Q1 :1 : oui, 2 : non, 3 : peut être

Q2 :1 : oui, 2 : non, 3 : peut être

Q3 :1 : oui, 2 : non, 3 : peut être

Sexe	Age	Q1	Q2	Q3
1	1	3	1	2
1	1	2	1	2
1	2	2	1	1
2	2	1	1	1
2	2	1	2	3
1	1	2	2	3
2	2	3	2	3
1	1	3	3	2
1	2	3	2	2
1	2	1	1	2

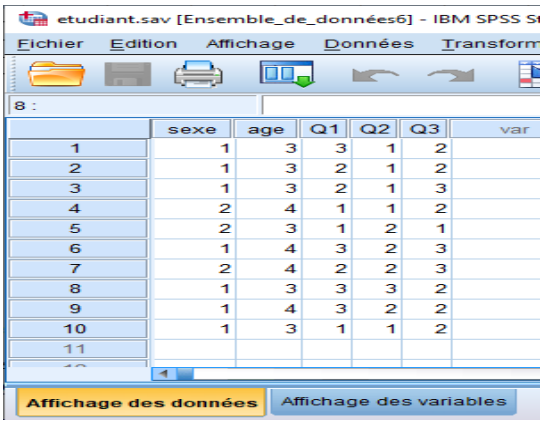
Tableau V.4 : Exemple étudiant.



- On considère le fichier enseignant.sav suivant :

Sexe	Age	Q1	Q2	Q3
1	3	3	1	2
1	3	2	1	2
1	3	2	1	3
2	4	1	1	2
2	3	1	2	1
1	4	3	2	3
2	4	2	2	3
1	3	3	3	2
1	4	3	2	2
1	3	1	1	2

Tableau V.5 : Exemple enseignant.



- Fusionner les 2 fichiers en un seul.
  - Ouvrir le fichier enseignant.sav (composé de 10 lignes),
  - On clique sur **Données > Fusionner des fichiers > Ajouter des observations > …**
  - On sélectionne le fichier etudaint.sav (composé de 10 lignes),
  - On obtient le fichier fusionné composé de 20 lignes (on a les mêmes noms de variables qui posent pas problème) :

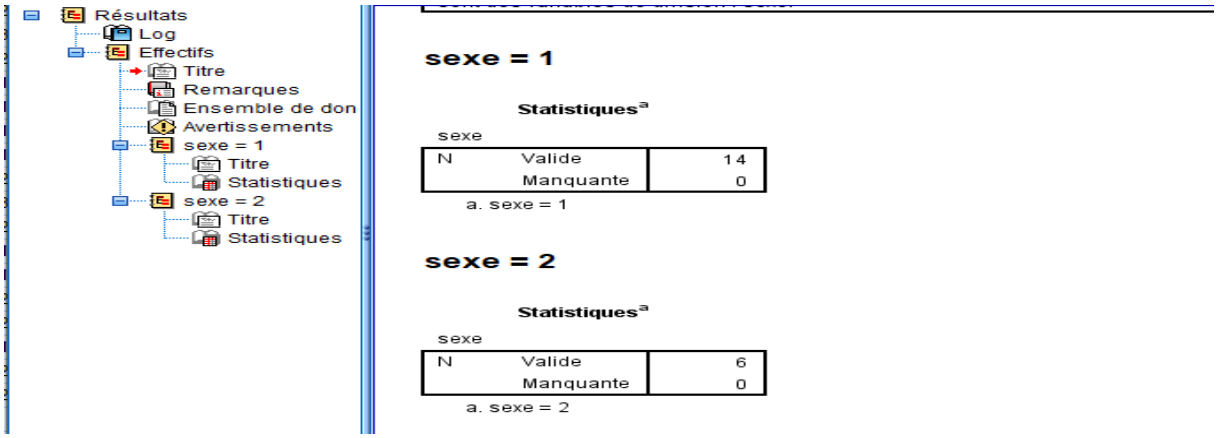
	sexe	age	Q1	Q2	Q3
1	1	1	3	1	2
2	1	1	2	1	2
3	1	2	2	1	1
4	2	2	1	1	1
5	2	2	1	2	3
6	1	1	2	2	3
7	2	2	3	2	3
8	1	1	3	3	2
9	1	2	3	2	2
10	1	2	1	1	2
11	1	3	3	1	2
12	1	3	2	1	2
13	1	3	2	1	3
14	2	4	1	1	2
15	2	3	1	2	1
16	1	4	3	2	3
17	2	4	2	2	3
18	1	3	3	3	2
19	1	4	3	2	2
20	1	3	1	1	2

Figure V.3. Les fichiers étudiant et enseignant fusionnés.

- Scinder le dernier fichier (résultat de fusion) sur la base de la variable sexe. Données→scinder un fichier→comparer les groupes Que remarquez vous ?
  - On remarque que le fichier est ordonné selon la variable sexe,

	sexe	age	Q1	Q2	Q3
1	1	1	3	1	2
2	1	1	2	1	2
3	1	2	2	1	1
4	1	1	2	2	3
5	1	1	3	3	2
6	1	2	3	2	2
7	1	2	1	1	2
8	1	3	3	1	2
9	1	3	2	1	2
10	1	3	2	1	3
11	1	4	3	2	3
12	1	3	3	3	2
13	1	4	3	2	2
14	1	3	1	1	2
15	2	2	1	1	1
16	2	2	1	2	3
17	2	2	3	2	3
18	2	4	1	1	2
19	2	3	1	2	1
20	2	4	2	2	3

- Choisir Données→ scinder un fichier→séparer les résultats par groupe, choisir analyse→statistique descriptive→effectifs et choisir sexe. Qu'est ce que vous remarquez ?
  - On obtient un fichier résultat dans lequel les individus sont séparés en deux sous groupes correspondant aux valeurs de la variable sexe, avec les effectifs de chaque sous groupe.



Chapitre VI : Tri et sélection des données

Pour préparer des données pour une analyse, on peut effectuer un grand nombre de transformations de fichiers. On a le choix entre les opérations suivantes :

- ✓ Trier les données : Consiste à trier les observations en fonction de la valeur de certaines variables.
- ✓ Sélectionner des sous-ensembles d'observations : Limiter l'analyse à un sous-ensemble d'observations ou effectuer des analyses simultanées sur différents sous-ensembles.

VI.1 Tri des données

Il est parfois nécessaire de trier les observations (lignes du fichier de données) pour certains types d'analyse. Pour réorganiser la séquence des observations dans le fichier de données en fonction de la valeur de certaines variables de tri (le tri peut se faire sur une ou plusieurs variables à qui sont affectés des ordres de priorité). On considère le fichier soirée.sav pour réaliser les étapes :

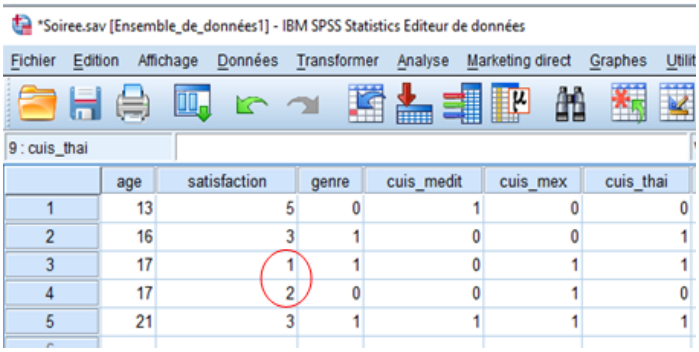
1. On sélectionne dans le menu **Données > Trier les observations > …**,

2. On utilise âge et satisfaction pour le tri donc on les déplace vers la zone « Trier par : »,

3. On choisit l'ordre du tri (croissant ou décroissant) et on clique sur OK,
4. S'affiche le fichier résultat qui indique que le fichier a été trié avec la mention des noms des variables de tri,

```
GET
  FILE='C:\Users\pc\Desktop\exercices cours\Soiree.sav'.
DATASET NAME Ensemble_de_donnéesl WINDOW=FRONT.
SORT CASES BY age(A) satisfaction(A).
```

5. Le fichier de données est trié, on remarque que pour les cas des lignes 3 et 4 , la première variable de tri possède les mêmes valeurs alors on utilise le 2ème critère pour le tri,



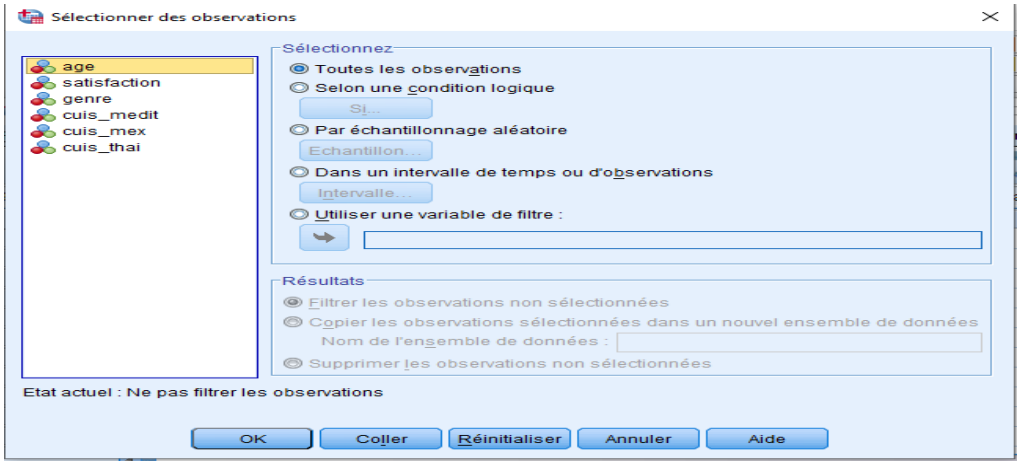
	age	satisfaction	genre	cuis_medit	cuis_mex	cuis_thai
1	13	5	0	1	0	0
2	16	3	1	0	0	1
3	17	1	1	0	1	1
4	17	2	0	0	1	0
5	21	3	1	1	1	1

VI.2 Sélection des données

On peut limiter l'analyse à un sous ensemble donné en fonction de critères contenant des variables et des expressions complexes. On peut également sélectionner un échantillon aléatoire d'observations. Les critères utilisés pour définir un sous-groupe comprennent :

- Plages et valeurs de variables,
- Plages de dates et d'heures,
- Nombres d'observations (lignes),
- Expressions arithmétiques,
- Expressions logiques,
- Fonctions.

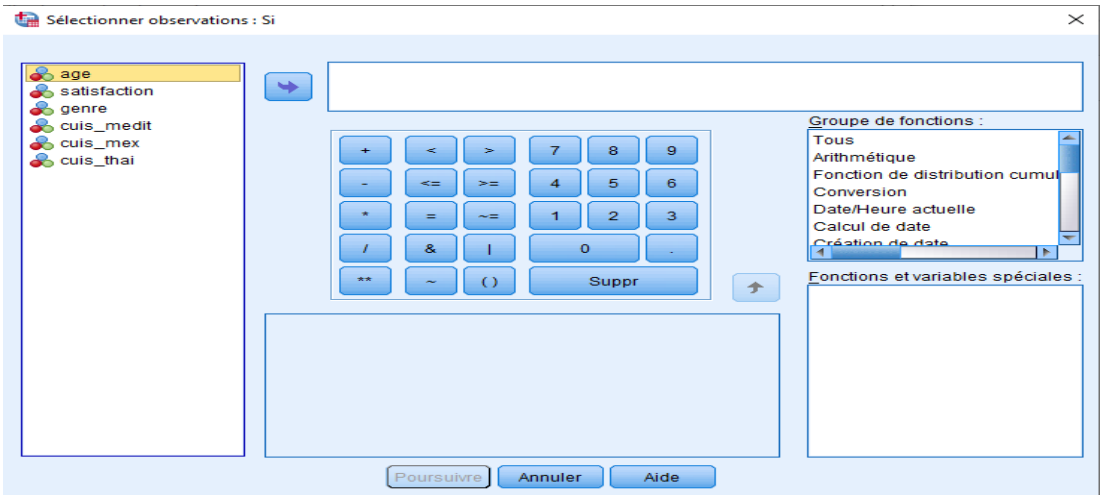
Pour sélectionner un sous-ensemble d'observations à analyser, on sélectionne dans le menu **Données > Sélectionner des observations > ...**



VI.2.1 Sélection à l'aide d'une expression conditionnelle

Pour sélectionner des observations sur la base d'une expression conditionnelle, procédez comme suit :

1. Sélectionnez Selon une condition logique, puis cliquez sur Si dans la boîte de dialogue Sélectionner des observations.



On remarque qu'on a à gauche l'ensemble de toutes les variables du fichier de données, au milieu un pavé numérique pour écrire les formules et à droite un ensemble de fonctions prédéfinies à utiliser directement en cas de besoin.

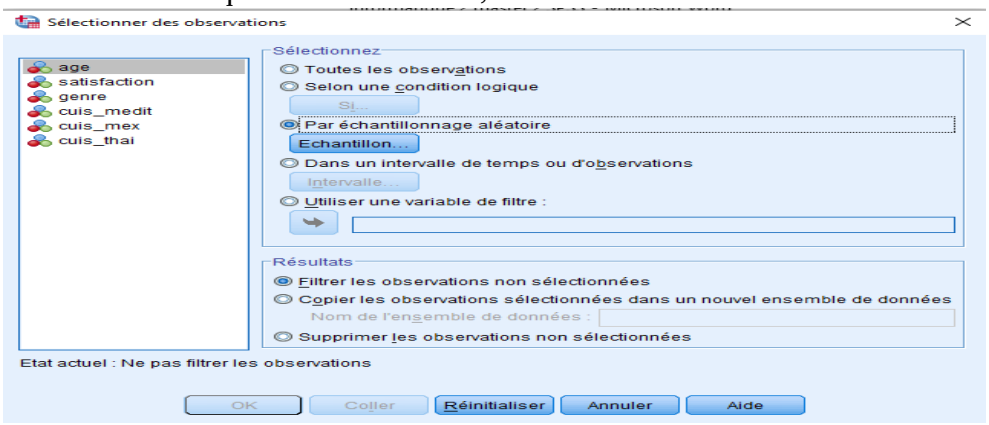
L'expression conditionnelle peut utiliser des noms de variables existantes, des constantes, des opérateurs arithmétiques, des opérateurs logiques, des opérateurs relationnels et des fonctions.

Dans la zone de texte en haut de la boite, on peut saisir et modifier du texte comme n'importe quel texte dans une fenêtre de sortie. On peut également utiliser le pavé numérique de la boîte de dialogue, la liste des variables et la liste des fonctions pour coller des éléments dans l'expression.

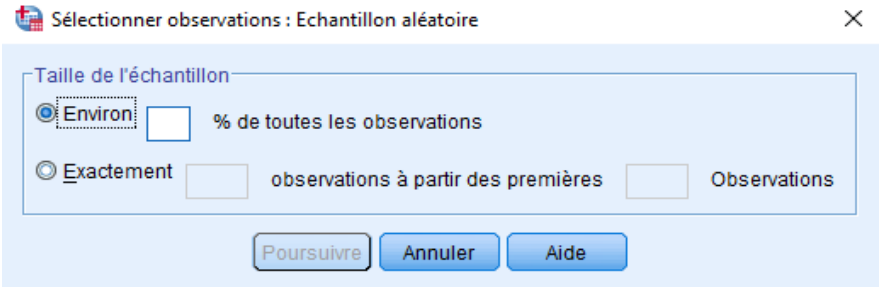
VI.2.2 Sélection d'un échantillon aléatoire

Pour obtenir un échantillon aléatoire, on suit les étapes suivantes:

- Sélectionnez Par échantillonnage aléatoire dans la boîte de dialogue Sélectionner des observations et cliquez sur Echantillon,



- Cette opération permet d'ouvrir la boîte de dialogue Sélectionner observations : Echantillon aléatoire,



Pour la taille de l'échantillon, deux options sont disponibles :

Environ : Pourcentage défini par l'utilisateur. Cette option génère un échantillon aléatoire d'observations dont le nombre correspond approximativement au pourcentage indiqué.

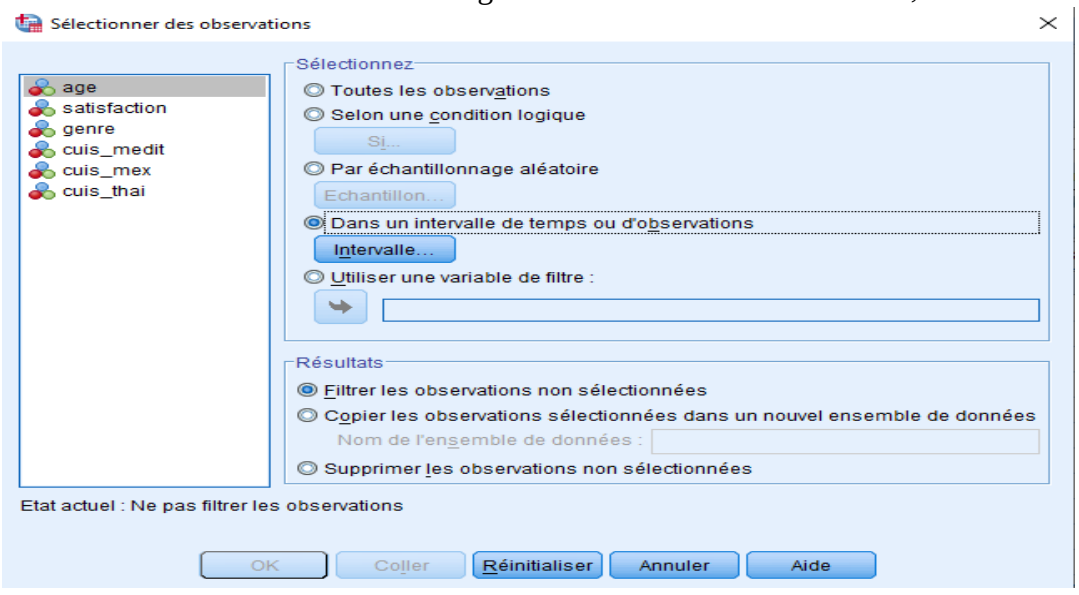
Exactement : Nombre d'observations spécifié par l'utilisateur. On doit également indiquer le nombre d'observations à partir duquel l'échantillon sera généré. Ce deuxième nombre doit être inférieur ou égal au nombre total d'observations dans le fichier de données.

Si ce nombre dépasse le nombre total d'observations dans le fichier de données, l'échantillon contiendra proportionnellement moins d'observations que le nombre demandé.

VI.2.3 Sélection d'une plage de temps

Pour sélectionner une plage d'observations sur la base de dates, d'heures ou de numéros (lignes) d'observations :

- Sélectionner « Dans un intervalle de temps ou d'observations », puis cliquer sur « Intervalle » dans la boîte de dialogue Sélectionner des observations,



- Cette opération ouvre la boîte de dialogue Sélectionner observations : Plage, dans laquelle on peut sélectionner une plage de numéros (lignes) d'observations.

Première observation : Entrer la date de début et/ou les valeurs de temps de la plage. Si aucune variable de date n'est définie alors introduire le numéro d'observation de départ (numéro de ligne dans l'éditeur de données, sauf si l'option Scinder un fichier est activée).

Si aucune valeur n'est indiquée pour la zone Dernière, toutes les observations à partir de la date/l'heure de début jusqu'à la fin de la série chronologique sont sélectionnées.

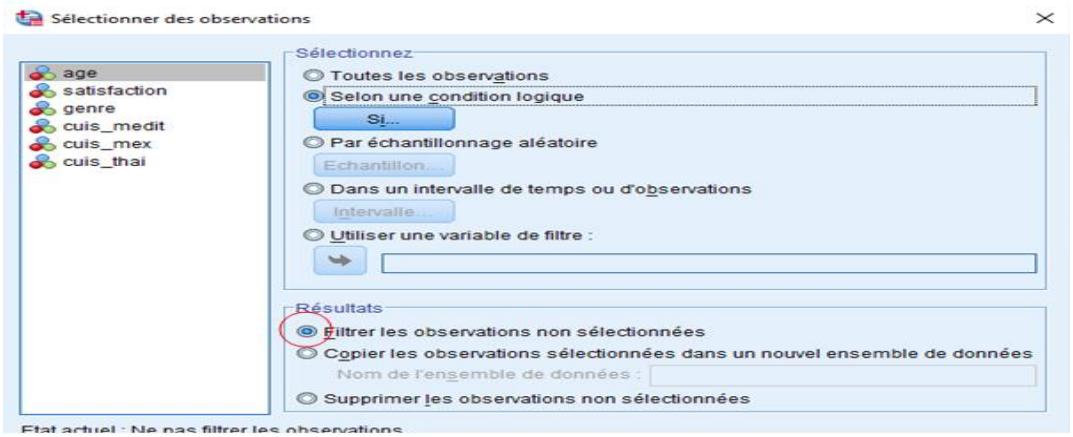
Dernière observation : Entrer la date de fin et/ou les valeurs de temps pour la plage. Si aucune variable de date n'est définie, entrer le numéro d'observation de fin (numéro de ligne dans l'éditeur de données, sauf si l'option Scinder un fichier est activée).

Si aucune valeur d'est indiquée pour la zone Première, toutes les observations à partir du début de la série chronologique jusqu'à la date/l'heure de fin sont sélectionnées.

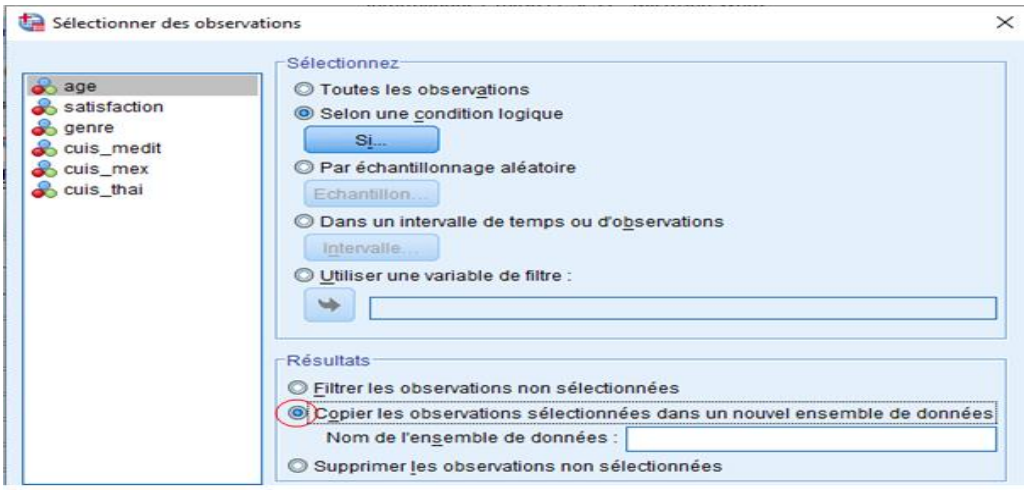
VI.2.4 Traitement des observations exclues

On peut choisir l'une des options suivantes pour traiter les observations exclues :

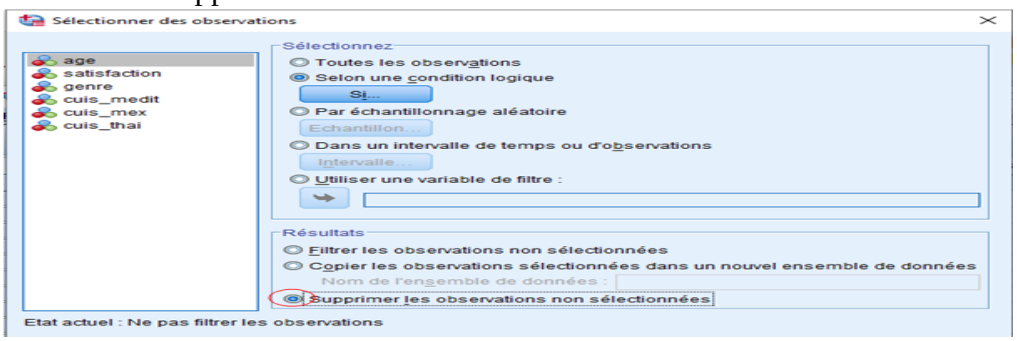
- Filtrer les observations non sélectionnées : Les observations exclues ne sont pas incluses dans l'analyse, mais restent dans le jeu de données. On peut utiliser les observations exclues ultérieurement dans la session si le filtrage est désactivé. Si on sélectionne un échantillon aléatoire ou si on sélectionne des observations sur la base d'une expression conditionnelle, une variable nommée filter_$ est générée ; elle comporte la valeur 1 pour les observations sélectionnées et la valeur 0 pour les observations exclues.



- Copier les observations sélectionnées dans un nouvel ensemble de données : Les observations sélectionnées sont copiées dans un nouveau jeu de données ; le jeu de données d'origine reste inchangé. Les observations exclues ne sont pas incluses dans le nouveau jeu de données et sont conservées dans leur état d'origine dans le jeu de données d'origine.



- Supprimer les observations non sélectionnées : Les observations exclues sont supprimées du jeu de données. On peut récupérer les observations supprimées uniquement en fermant le fichier sans enregistrer les modifications et en l'ouvrant à nouveau. La suppression des observations est définitive si on enregistre les modifications apportées au fichier de données.

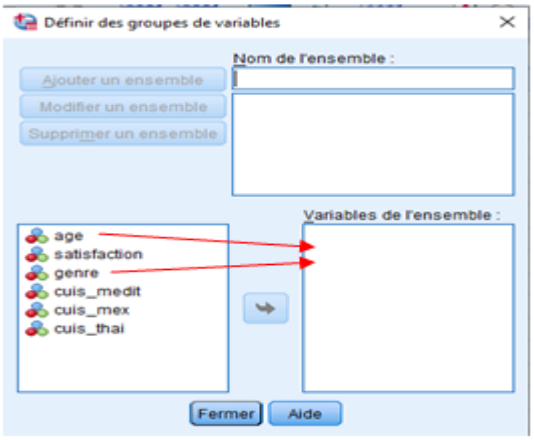


Si un sous-ensemble d'observations est sélectionné mais celles non sélectionnées ne sont pas écartées, celles-ci sont identifiées dans l'éditeur de données par une ligne verticale au niveau du numéro de ligne.

Soit le fichier soirée.sav :

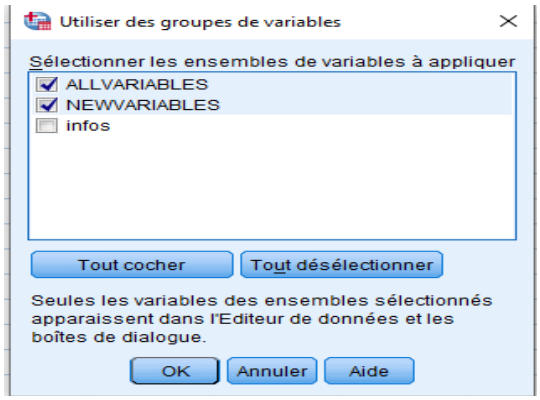
Age	Satisfaction	Genre	Cuis_medit	Cuis_mex	Cuis_thai
21	3	1	1	1	1
17	2	0	0	1	0
16	3	1	0	0	1
17	2	1	0	1	1
13	5	0	1	0	0
25	2	1	0	1	1
22	3	0	1	1	0

- Regrouper les variables âge et genre.
  - On sélectionne dans le menu : **Utilitaires > Définir des groupes de variables > …**

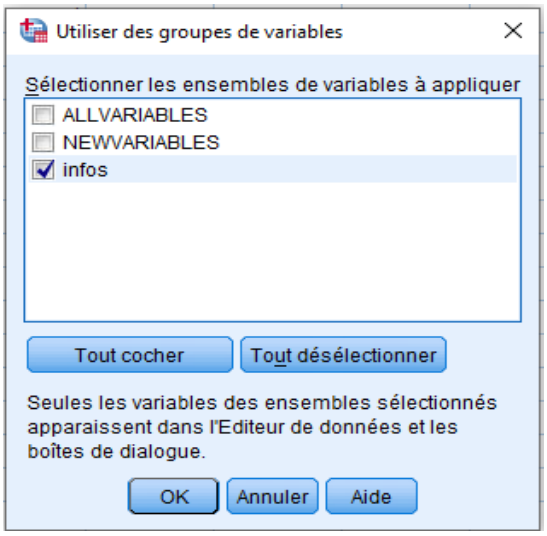


- On introduit un nom pour l'ensemble des variables regroupées (par ex infos), et on sélectionne les variables âge et genre pour les mettre dans la zone variables de l'ensemble.

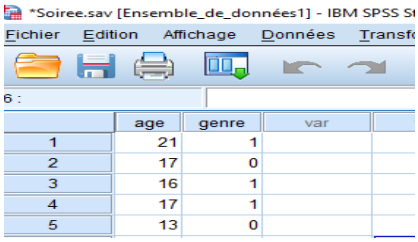
- Afficher uniquement les variables regroupées ensuite afficher toutes les variables.
  - On clique sur **Utilitaires > Utiliser des groupes de variables > …**



- On décoche ALLVARIABLES et NEWVARIABLES et on coche infos, ensuite on clique sur OK,

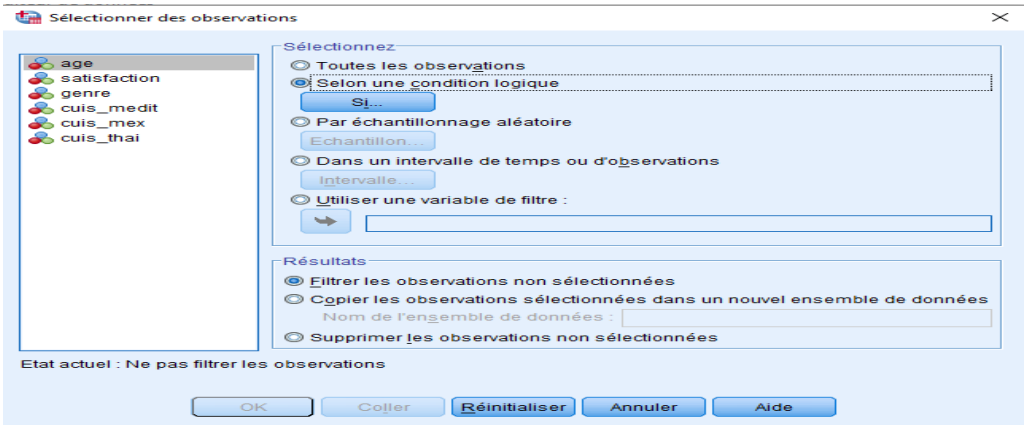


- Dans le fichier de données, est affiché juste les noms des variables déjà regroupé dans l'ensemble infos,

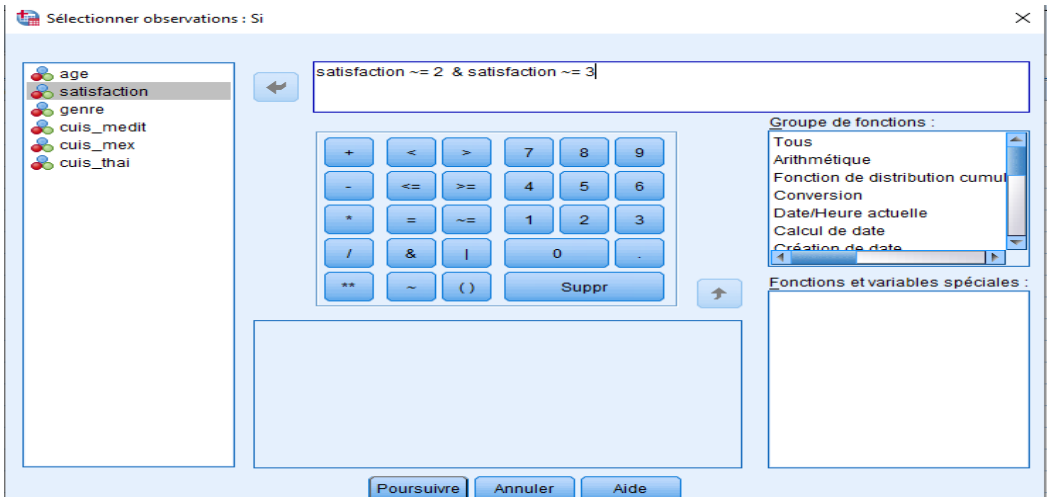


- Pour afficher toutes les variables on choisit dans le menu : **Utilitaires > Afficher toutes les variables > …**

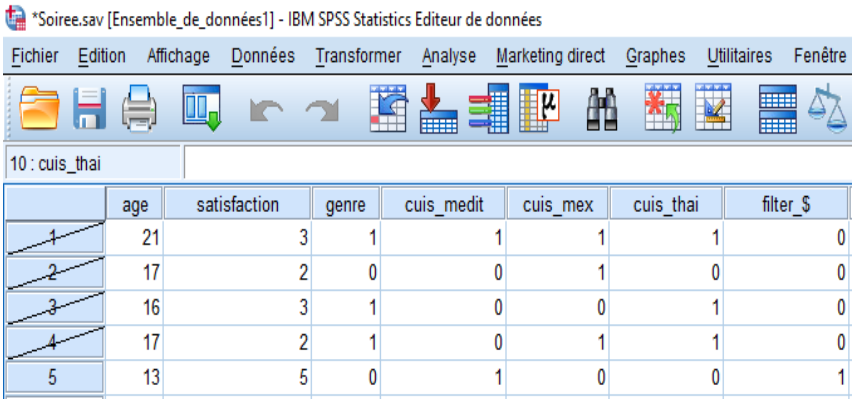
- Filtrer vos données en n'affichant pas les personnes pas satisfaites de la soirée (on élimine les individus avec les valeurs 2 et 3 pour la variable satisfaction).
  - Sélectionner **Données > Sélectionner des observations > …**, on choisit Selon une condition logique et on clique sur le bouton Si,



  - On obtient la fenêtre dans laquelle on introduit la condition, on clique sur poursuivre,

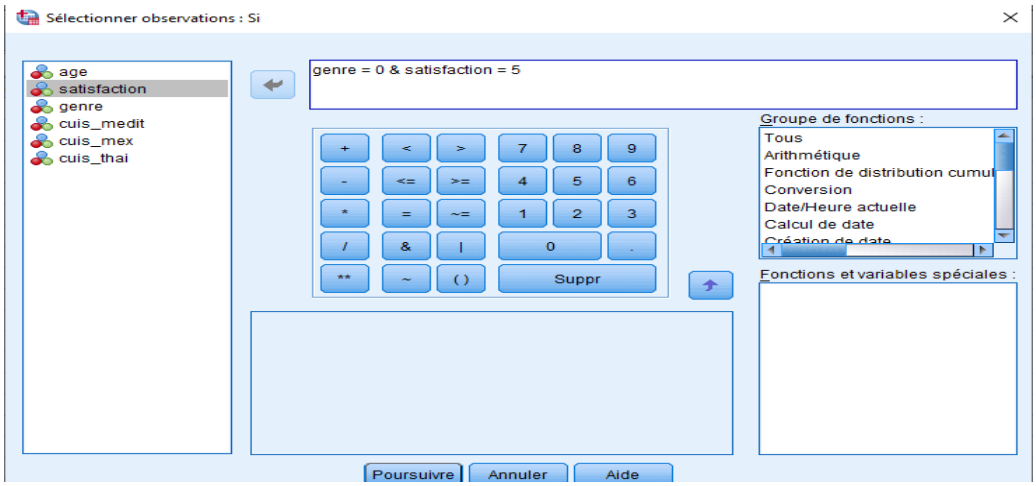


- On obtient le fichier de données avec les lignes qui ne satisfont pas la condition biffées au niveau du numéro de ligne,



	age	satisfaction	genre	cuis_medit	cuis_mex	cuis_thai	filter_$
1	21	3	1	1	1	1	0
2	17	2	0	0	1	0	0
3	16	3	1	0	0	1	0
4	17	2	1	0	1	1	0
5	13	5	0	1	0	0	1

- Sélectionner les hommes ayant appréciées la soirée (satisfaction = 5).
  - S'il y a des filtres dans les questions précédentes il faut les éliminer en supprimant la colonne filter_$,
  - Sélectionner **Données > Sélectionner des observations > …**, on choisit Selon une condition logique et on clique sur le bouton Si,
  - On obtient la boite de condition dans laquelle on écrit la condition (& : représente le ET, ~= : représente le symbole ≠),



- Cliquer sur poursuivre, on obtient un fichier de données dont une seule ligne correspond à la condition,

	age	satisfaction	genre	cuis_medit	cuis_mex	cuis_thai	filter_$
1	21	3	1	1	1	1	0
2	17	2	0	0	1	0	0
3	16	3	1	0	0	1	0
4	17	2	1	0	1	1	0
5	13	5	0	1	0	0	1

- Sélectionner les personnes ayant mangé de la cuisine méditerranéenne ou thaïlandaise.
  - Eliminer les filtres précédents en supprimant la colonne filter_$,
  - Sélectionner **Données > Sélectionner des observations > …**, on choisit Selon une condition logique et on clique sur le bouton Si,
  - On obtient la boite de condition dans laquelle on écrit la condition (| : représente le ou) ,

  - On a le fichier de données suivant :

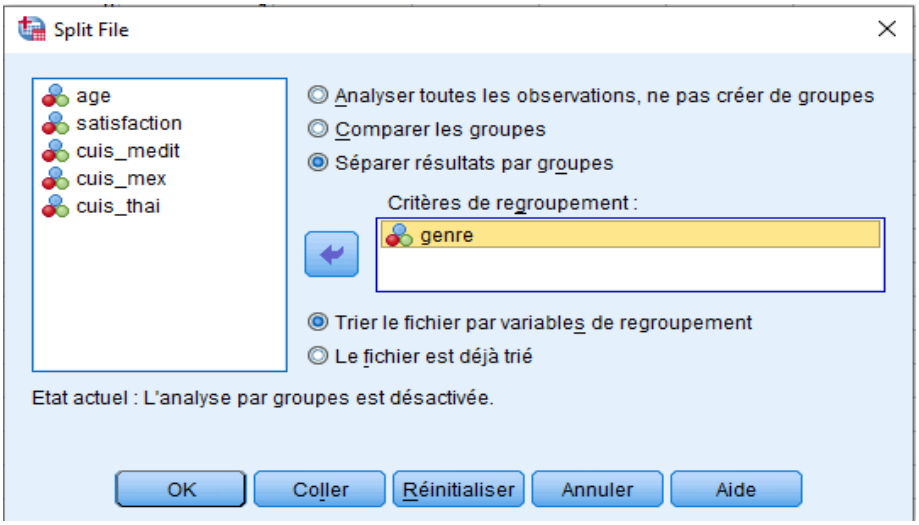
	age	satisfaction	genre	cuis_medit	cuis_mex	cuis_thai	filter_$
1	21	3	1	1	1	1	1
2	17	2	0	0	1	0	0
3	16	3	1	0	0	1	1
4	17	2	1	0	1	1	1
5	13	5	0	1	0	0	1

- Sélectionner les personnes dont l'âge est compris entre 15 et 20.
  - Eliminer les filtres précédents en supprimant la colonne filter_$,
  - Sélectionner **Données > Sélectionner des observations > …**, on choisit Selon une condition logique et on clique sur le bouton Si,
  - On obtient la boite de condition dans laquelle on écrit la condition,

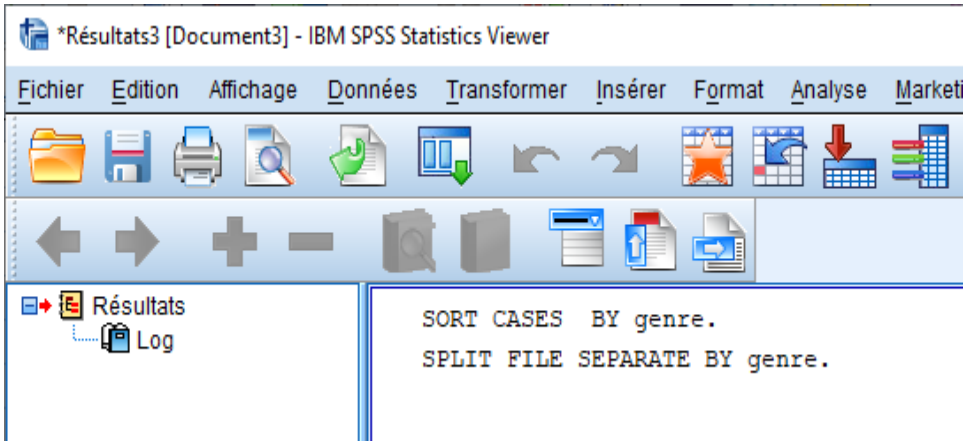
	age	satisfaction	genre	cuis_medit	cuis_mex	cuis_thai	filter_$
1	21	3	1	1	1	1	0
2	17	2	0	0	1	0	0
3	16	3	1	0	0	1	0
4	17	2	1	0	1	1	0
5	13	5	0	1	0	0	0

Aucune ligne ne correspond à la condition, toutes les lignes sont exclues.

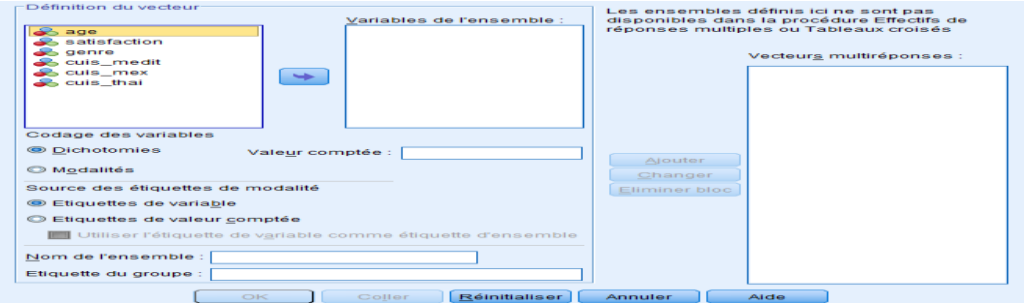
- Scinder les données selon la variable genre.
  - Sélectionner **Données > Scinder un fichier > …**,



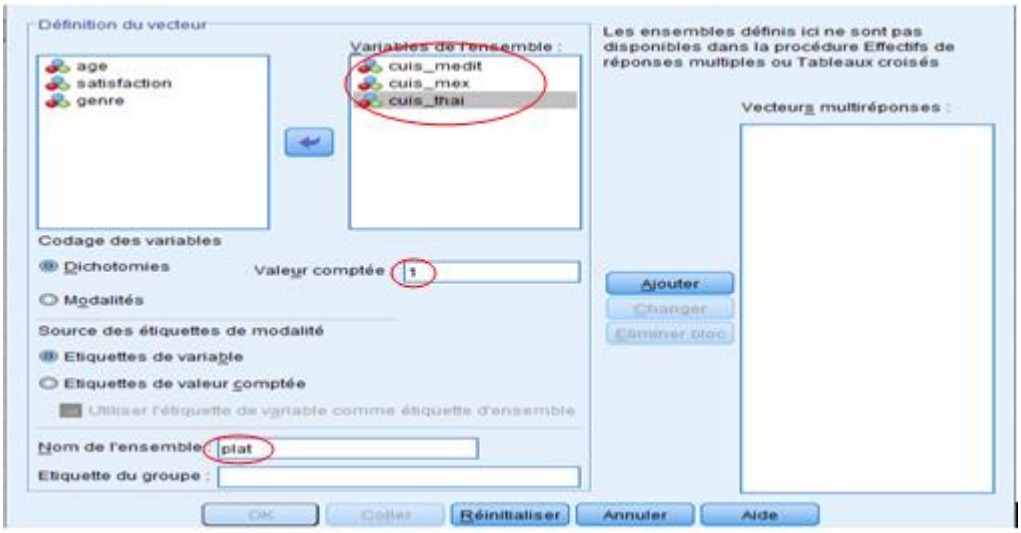
  - On clique sur OK on obtient le fichier résultats suivant :



- Définir un vecteur multi réponses pour le type de cuisine consommé.
  - Sélectionner **Données > Définir des vecteurs multiréponses > …**,



- On transfert les variables du vecteur multiréponses dans la zone Variables de l'ensemble, on met valeur comptée à 1 (indique sur si la valeur est 1 le plat a été consommé), enfin on donne un nom au vecteur,



- On clique sur ajouter pour associer l'ensemble aux variables (le nom de l'ensemble est ajouté dans la zone à droite) ensuite on clique sur OK.

```
* Définir les vecteurs multiréponses.
MRSETS
  /MDGROUP NAME=$plat CATEGORYLABELS=VARLABELS VARIABLES=cuis_medit cuis_mex cuis
_thai VALUE=1
  /DISPLAY NAME=[$plat].
```

Vecteur de réponses multiples

```
[Ensemble_de_donnéesl] C:\Users\pc\Desktop\exercices cours\Soiree.sav
```

Vecteurs de réponses multiples

Nom	Codé comme	Valeur comptée	Type de données	Variables élémentaires
$plat	Dichotomies	1	Numérique	cuis_medit cuis_mex cuis_thai

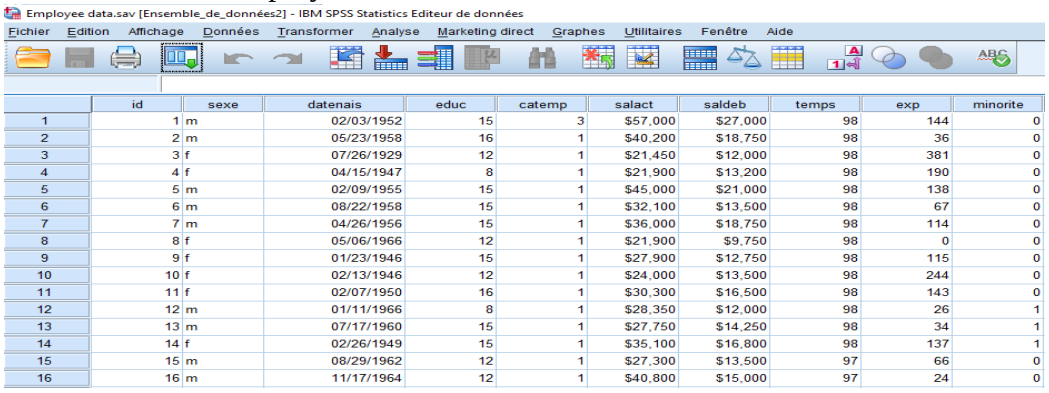
Chapitre VII : Transformation de variables

L'organisation et la codification des données de départ ne répond pas systématiquement aux besoins en matière de création de rapport ou d'analyse. Pour illustrer nous allons utiliser le fichier employee data.sav.

VII.1 Recodage des variables

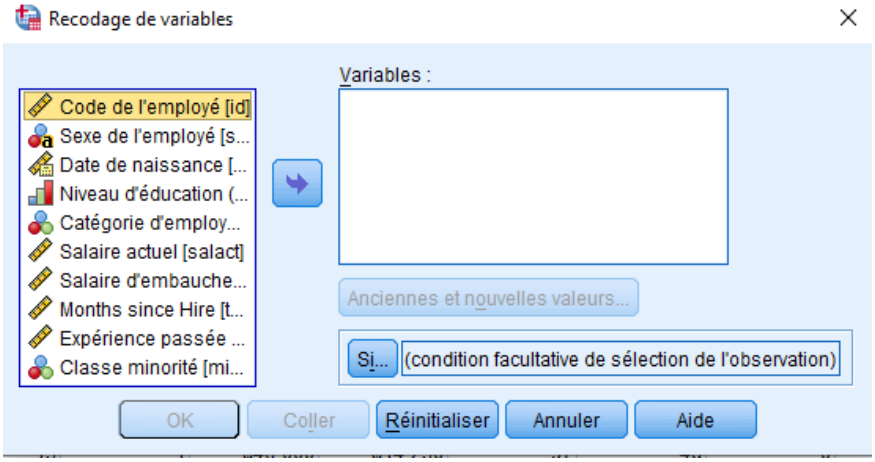
Pour transformer les valeurs d'une variable, on suit les étapes suivantes :

- Ouvrir le fichier employee data.sav : **Fichier > Ouvrier > …**

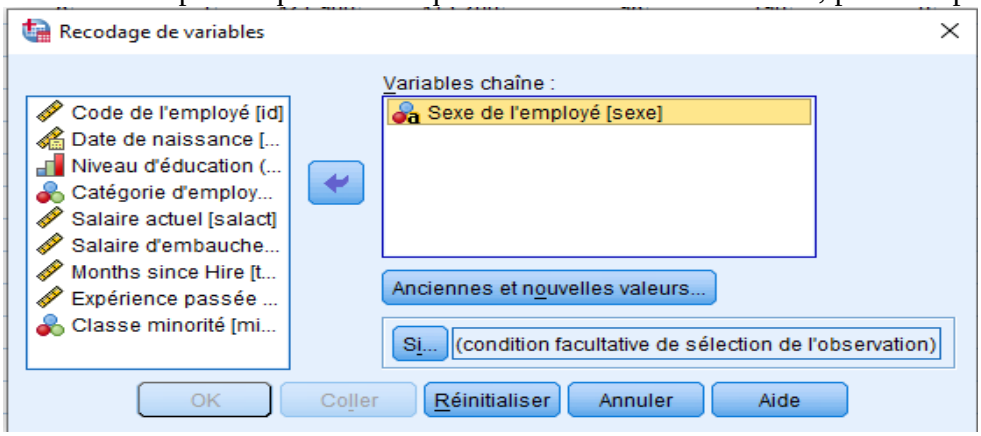


	id	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0
4	4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0
5	5	m	02/09/1955	15	1	$45,000	$21,000	98	138	0
6	6	m	08/22/1958	15	1	$32,100	$13,500	98	67	0
7	7	m	04/26/1956	15	1	$36,000	$18,750	98	114	0
8	8	f	05/06/1966	12	1	$21,900	$9,750	98	0	0
9	9	f	01/23/1946	15	1	$27,900	$12,750	98	115	0
10	10	f	02/13/1946	12	1	$24,000	$13,500	98	244	0
11	11	f	02/07/1950	16	1	$30,300	$16,500	98	143	0
12	12	m	01/11/1966	8	1	$28,350	$12,000	98	26	1
13	13	m	07/17/1960	15	1	$27,750	$14,250	98	34	1
14	14	f	02/26/1949	15	1	$35,100	$16,800	98	137	1
15	15	m	08/29/1962	12	1	$27,300	$13,500	97	66	0
16	16	m	11/17/1964	12	1	$40,800	$15,000	97	24	0

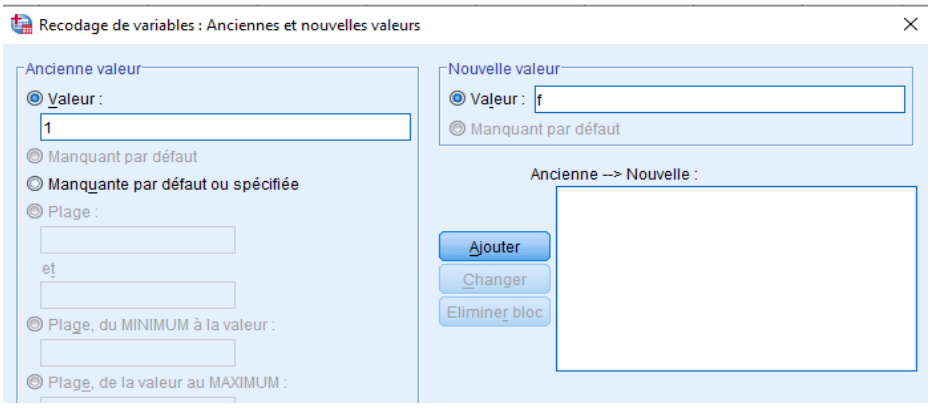
- On sélectionne dans le menu **Transformer > Recoder les variables > …**



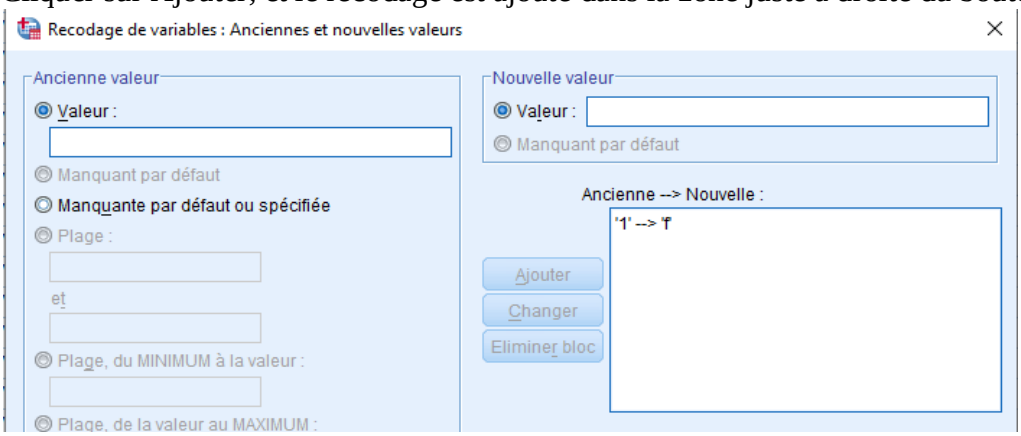
- Choir la variable pour laquelle il est question de recoder les valeurs, par exemple sexe,



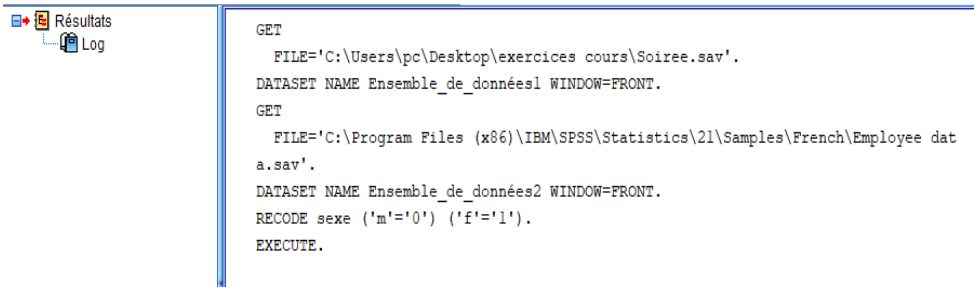
- Cliquer sur « Anciennes et nouvelles valeurs… », on modifie les valeurs de la variable sexe en remplaçant m→0 et f→1, l'ancienne valeur est écrite dans la zone à gauche et la nouvelle dans la zone à droite,



- Cliquer sur Ajouter, et le recodage est ajouté dans la zone juste à droite du bouton,



- Recommencer avec la nouvelle valeur, m→0 de la même façon, cliquer sur Poursuivre ensuite sur OK, on obtient le fichier initial avec la valeur de la variable sexe recodée. La transformation est aussi mentionnée dans le fichier Résultat.



Le fichier de données contient les nouvelles valeurs 0 et 1 pour le sexe.

	id	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite
1	1	0	02/03/1952	15	3	$57,000	$27,000	98	144	0
2	2	0	05/23/1958	16	1	$40,200	$18,750	98	36	0
3	3	1	07/26/1929	12	1	$21,450	$12,000	98	381	0
4	4	1	04/15/1947	8	1	$21,900	$13,200	98	190	0
5	5	0	02/09/1955	15	1	$45,000	$21,000	98	138	0
6	6	0	08/22/1958	15	1	$32,100	$13,500	98	67	0
7	7	0	04/26/1956	15	1	$36,000	$18,750	98	114	0
8	8	1	05/06/1966	12	1	$21,900	$9,750	98	0	0
9	9	1	01/23/1946	15	1	$27,900	$12,750	98	115	0
10	10	1	02/13/1946	12	1	$24,000	$13,500	98	244	0
11	11	1	02/07/1950	16	1	$30,300	$16,500	98	143	0
12	12	0	01/11/1966	8	1	$28,350	$12,000	98	26	1
13	13	0	07/17/1960	15	1	$27,750	$14,250	98	34	1

VII.2 Calcul de nouvelles variables (Cusson F. et Corneau M., 2010)

A l'aide d'une grande variété de fonctions mathématiques, on peut calculer de nouvelles variables en fonction d'équations extrêmement complexes.

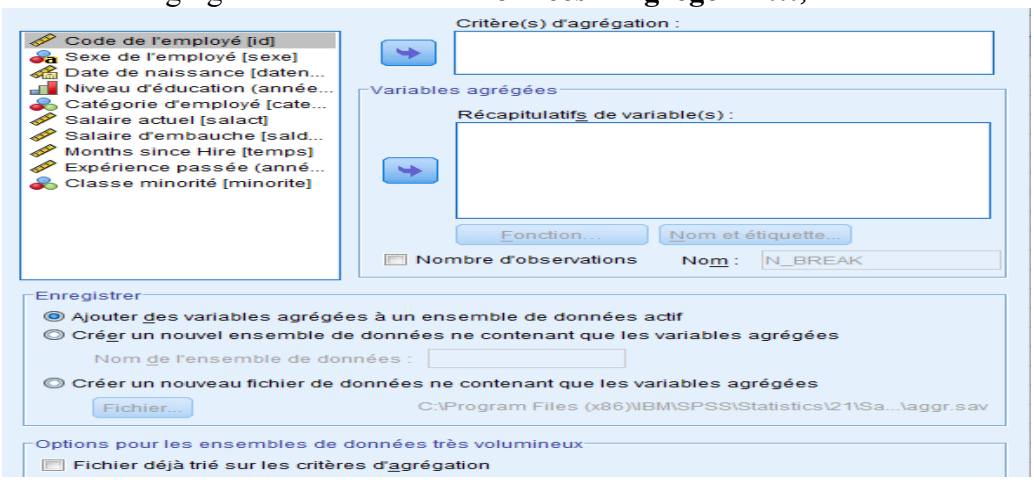
Pour calculer une nouvelle variable (qui sera ajoutée au fichier de donnée) à partir d'une ou plusieurs variables du fichier de données, on a deux possibilités :

VII.2.1 Agrégation de données

L'agrégation est l'opération qui permet de regrouper des données (des lignes) selon un critère. A la suite de l'agrégation des fonctions de calcul sont appliquées) chacun des groupes séparément.

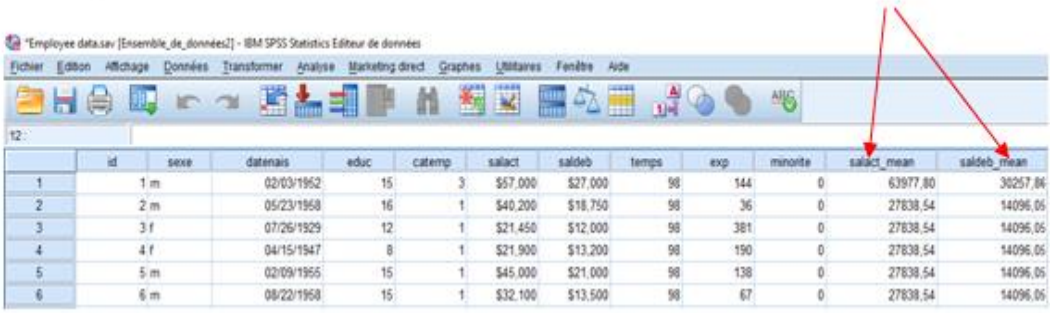
En utilisant le fichier employees data.sav, on veut calculer le saldeb (salaire début) moyen et salact (salaire actuel) moyen pour les différentes catégories. On applique les étapes suivantes :

- Ouvrir le fichier employee data.sav : **Fichier > Ouvrier > …**
- Pour regrouper les variables salact et saldeb selon la catégorie (catemp) de l'employer, on utilise l'agrégation avec le menu **Données > Agréger > …**,



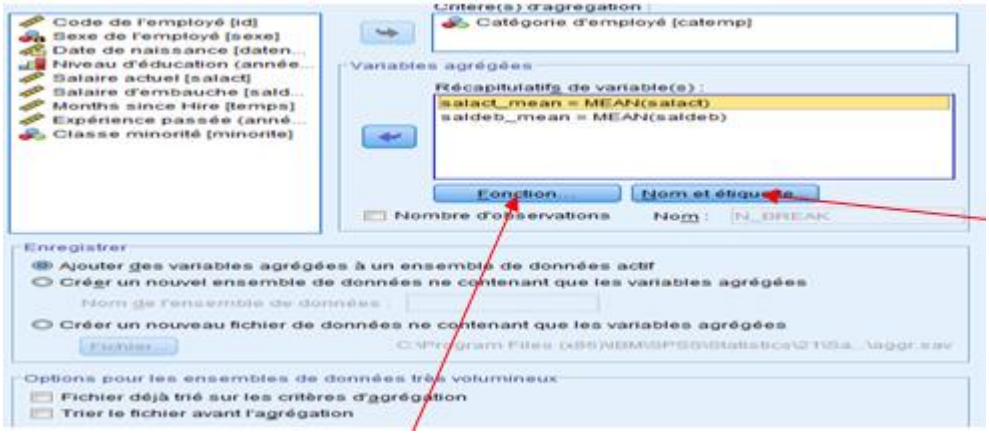
- On met Catégorie d'employé dans la zone de texte Critère d'agrégation et les variables Salaire actuel et Salaire d'embauche dans la zone Récapitulatifs de variables,

- Dans Récapitulatifs de variables, deux variables nommées salact_mean et saldeb_mean qui représentent respectivement les moyennes du salaire actuel et du salaire d'embauche ce sont deux variables qui sont calculées et ajoutées directement au fichier de données car l'option Ajouter des variables agrégées à un ensemble de données actif,

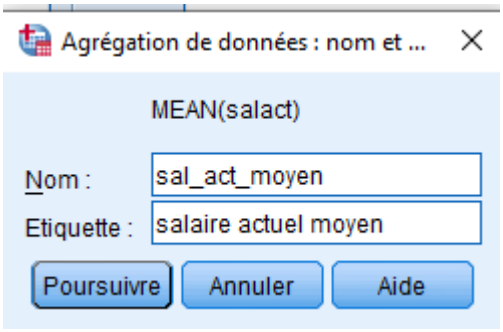


	id	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	salact_mean	saldeb_mean
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0	63977,80	30257,86
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0	27838,54	14096,05
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0	27838,54	14096,05
4	4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0	27838,54	14096,05
5	5	m	02/09/1955	15	1	$45,000	$21,000	98	138	0	27838,54	14096,05
6	6	m	08/22/1958	15	1	$32,100	$13,500	98	67	0	27838,54	14096,05

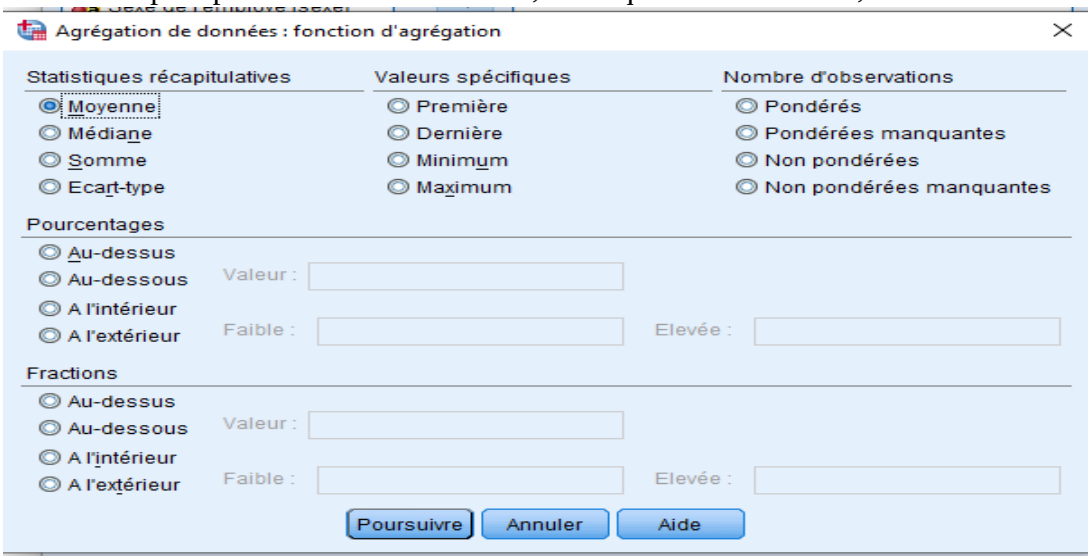
- On peut toujours changer les noms des variables et même les étiquettes en cliquant sur le bouton Nom et étiquette, et changer la fonction de calcul en attribuant une fonction différente à chaque variable,



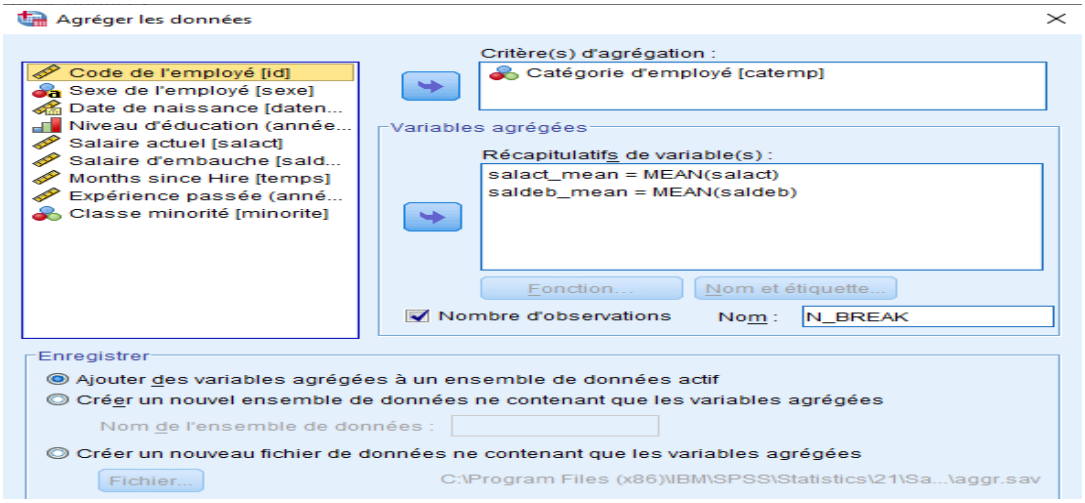
- On change le nom de la variable « salact_mean » en « sal_act_moyen » et on lui affecte une étiquette,



- On peut aussi changer la fonction de calcul qui par défaut utilise la fonction Moyenne, on a une panoplie de fonction à utiliser, en cliquant sur Fonction,



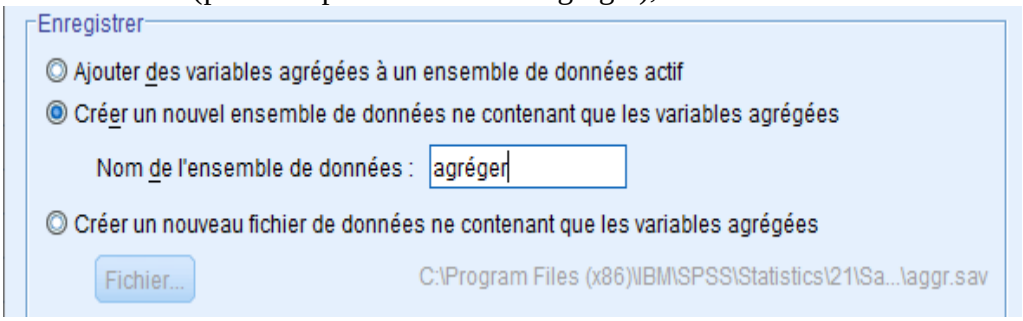
- On peut cocher « Nombre d'observations » automatiquement se crée une variable N_BREAK,



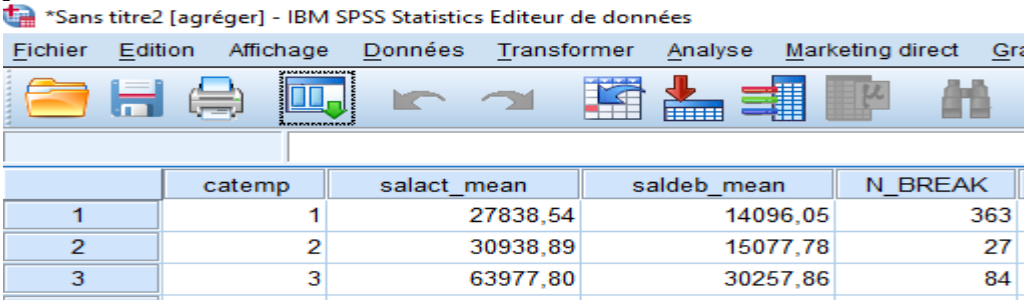
- N_BREAK est ajouté au fichier de données et qui indique le nombre de lignes dans les différents groupes (Pour la catégorie 3 on a 84 lignes et pour la catégorie 1 on a 363 lignes),

	id	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	salact_mean	saldeb_mean	N_BREAK
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0	63977,80	30257,86	84
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0	27838,54	14096,05	363
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0	27838,54	14096,05	363
4	4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0	27838,54	14096,05	363

- On peut aussi changer le fichier dans lequel on ajoute toutes les variables issues d'une opération d'agrégation, la 1ère option permet de les ajouter au fichier de données comme on l'a déjà fait, la seconde permet de créer un fichier à part pour lequel on donne un nom (par exemple on le nomme agréger),

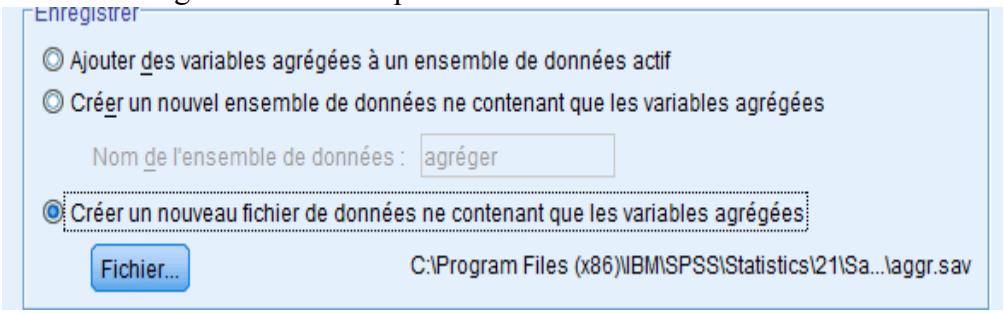


- Le fichier contient 3 lignes car on a 3 catégories (ce fichier doit être enregistré car il ne l'est pas encore), on peut toujours lui donner le nom qu'on a choisit dans l'étape précédente,



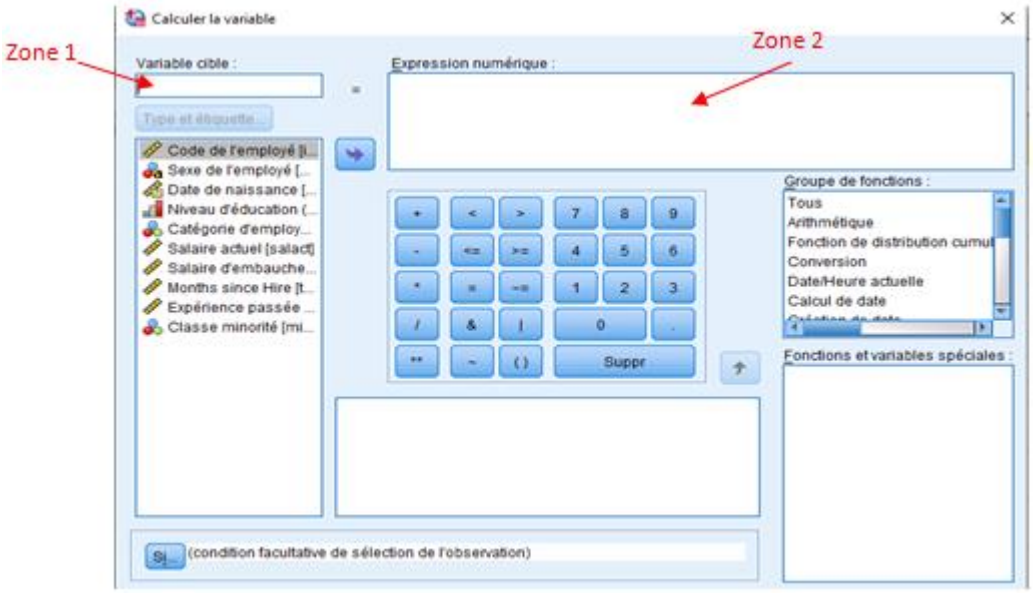
	catemp	salact_mean	saldeb_mean	N_BREAK
1	1	27838,54	14096,05	363
2	2	30938,89	15077,78	27
3	3	63977,80	30257,86	84

- La dernière option permet d'utiliser un fichier externe mais pour lequel on choisit le chemin d'enregistrement ainsi que le nom.



VII.2.2 Utilisation du menu Calculer la variable

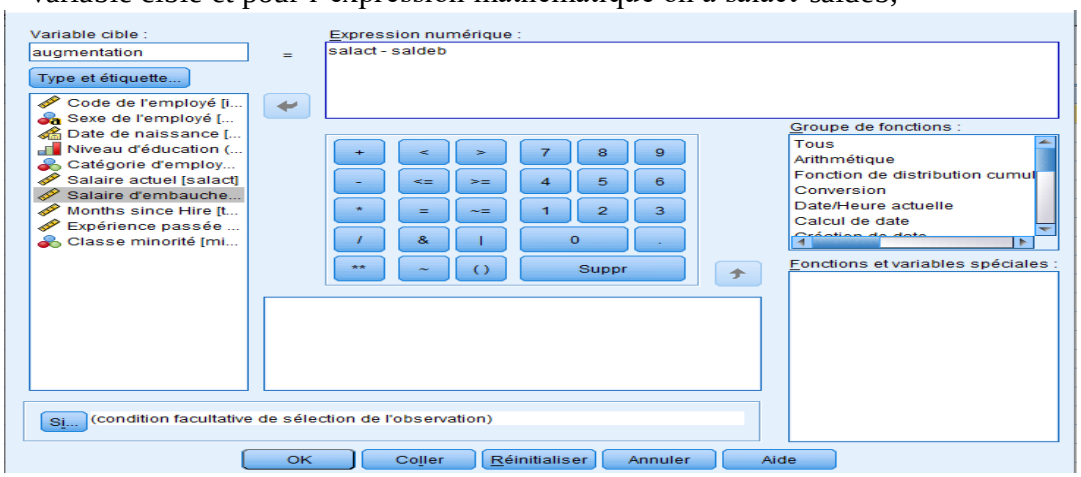
Pour ajouter de nouvelles variables (calculées) au fichier de données, on utilise le menu **Transformer > Calculer la variable …**



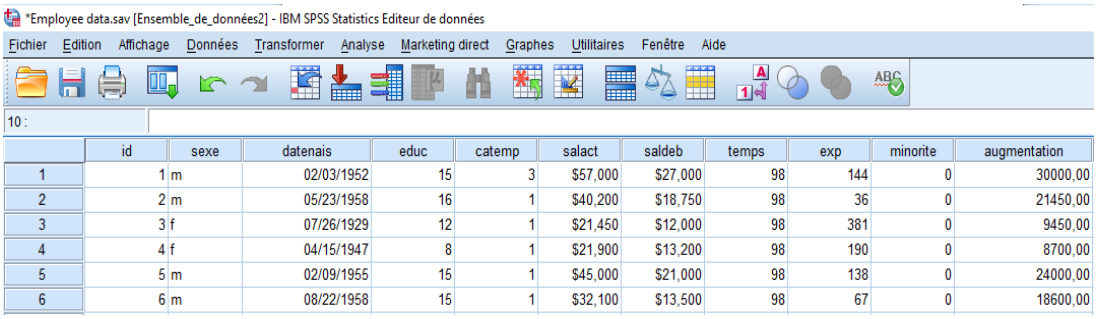
- Dans Zone 1, on écrit le nom de la nouvelle variable qu'on veut calculer et ajouter au fichier de données,
- Dans Zone 2, on introduit l'expression numérique qui permet d'effectuer le calcul en utilisant le pavé numérique et un ensemble de fonctions prédéfinies à droite.

En utilisant le fichier employees data.sav :

- Créer une nouvelle variable *Augmentation* qui calcule la différence de salaire de recrutement et le salaire actuel.
  - Ouvrir le fichier employee data.sav : **Fichier > Ouvrier > …**
  - On clique sur **Transformer > Calculer la variable …**, on écrit augmentation dans la variable cible et pour l'expression mathématique on a salact-saldeb,

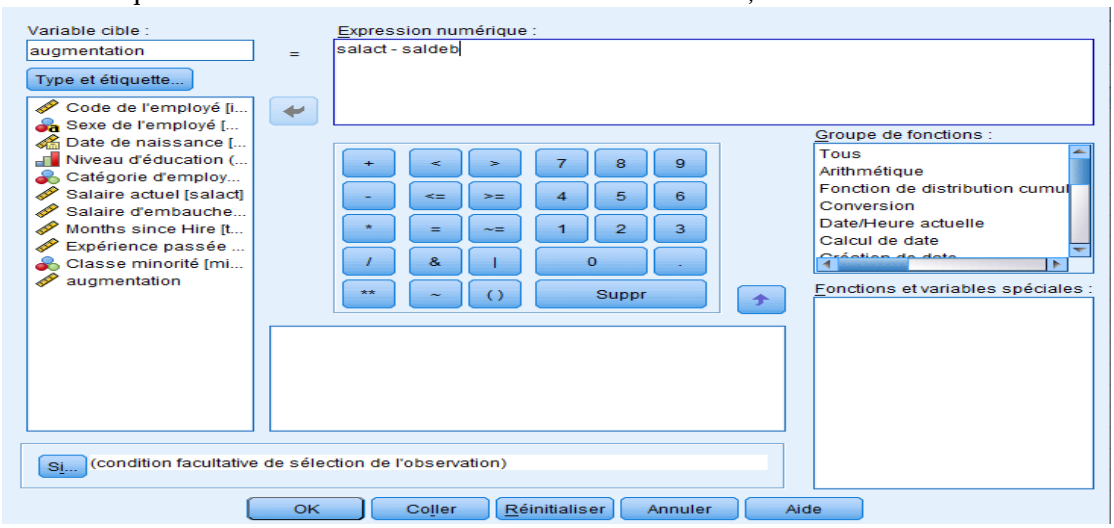


- On obtient le fichier de données initial auquel est ajoutée la variable augmentation,

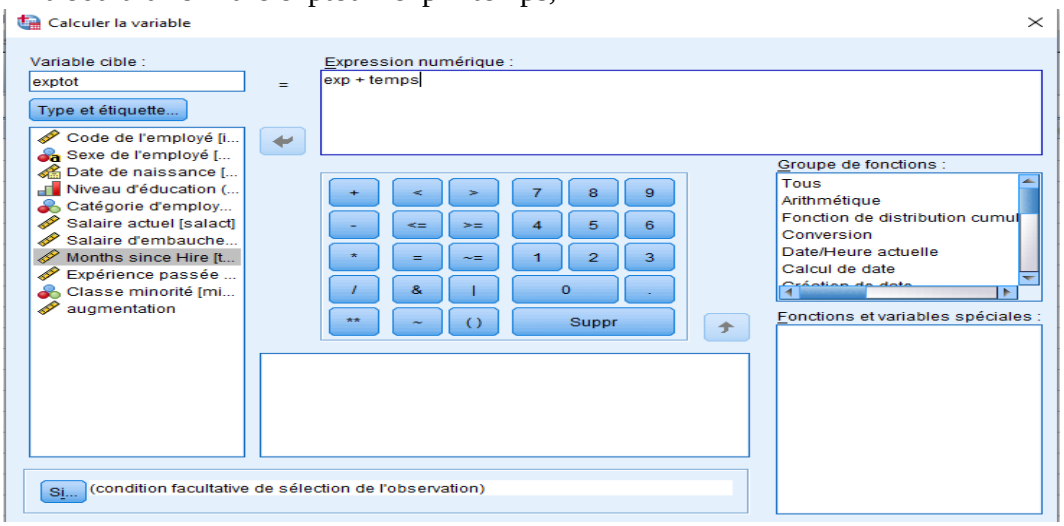


	id	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	augmentation
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0	30000,00
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0	21450,00
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0	9450,00
4	4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0	8700,00
5	5	m	02/09/1955	15	1	$45,000	$21,000	98	138	0	24000,00
6	6	m	08/22/1958	15	1	$32,100	$13,500	98	67	0	18600,00

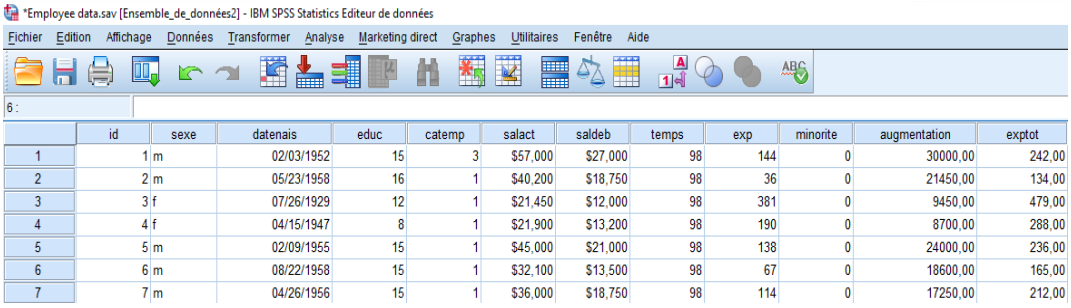
- Calculer la nouvelle variable exptot pour expérience totale soit la somme de expérience passée (exp) et temps :
  - On clique sur **Transformer > Calculer la variable …**,



  - On clique sur le bouton Réinitialiser pour effacer la fonction précédente, et on introduit la formule exptot = exp + temps,



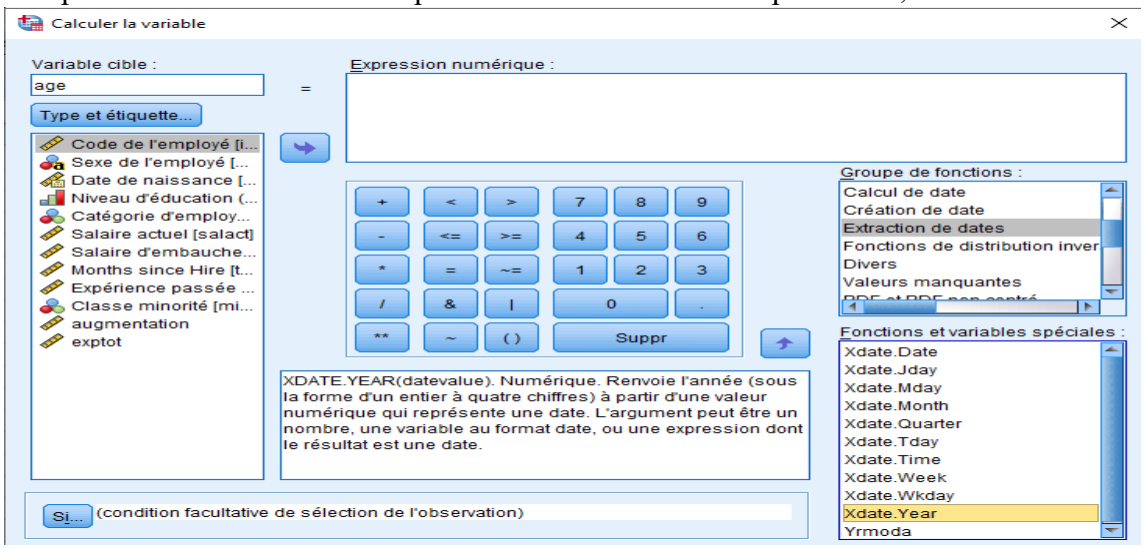
- On obtient le fichier de données avec la nouvelle variable ajoutée,



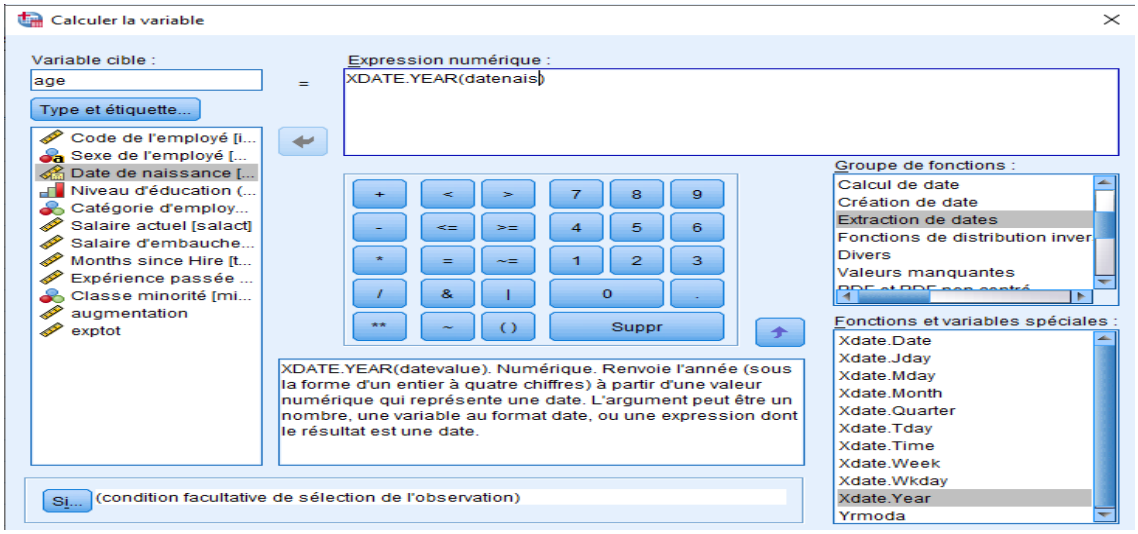
	id	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	augmentation	exptot
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0	30000,00	242,00
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0	21450,00	134,00
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0	9450,00	479,00
4	4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0	8700,00	288,00
5	5	m	02/09/1955	15	1	$45,000	$21,000	98	138	0	24000,00	236,00
6	6	m	08/22/1958	15	1	$32,100	$13,500	98	67	0	18600,00	165,00
7	7	m	04/26/1956	15	1	$36,000	$18,750	98	114	0	17250,00	212,00

- Calculer une nouvelle variable *Age*, qui calcule l'âge des employés en se basant sur leurs dates de naissance :
  - On clique sur **Transformer > Calculer la variable …**

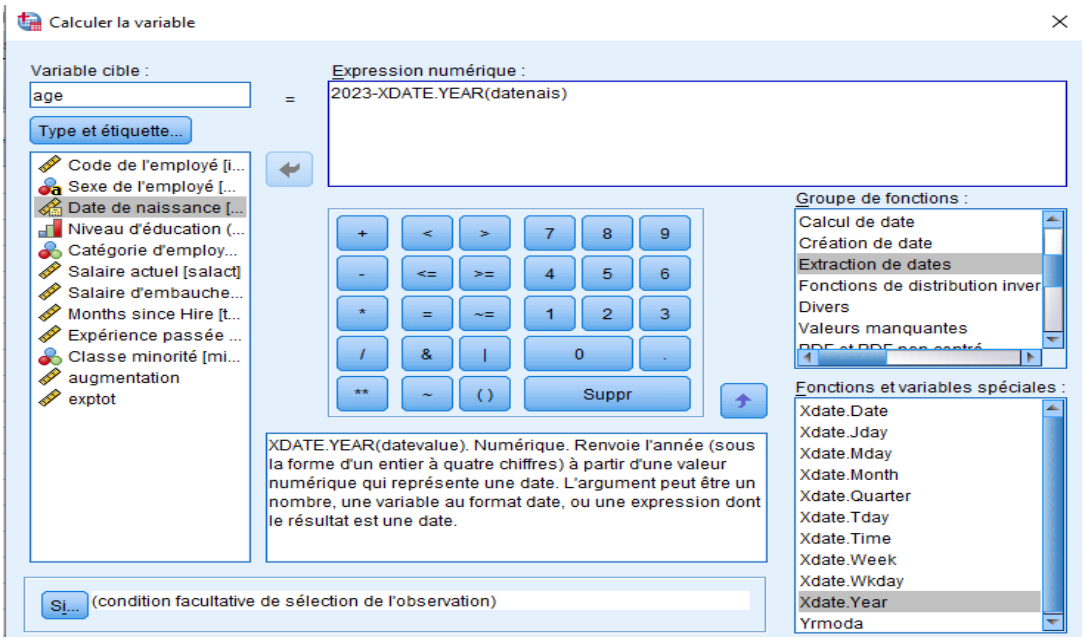
  - On clique sur le bouton Réinitialiser pour effacer la fonction précédente, on introduit le nom de la variable à calculer « âge » et on utilise la fonction Xdate.Year(date), qui permet d'extraire l'année à partir d'une date donnée en paramètre,



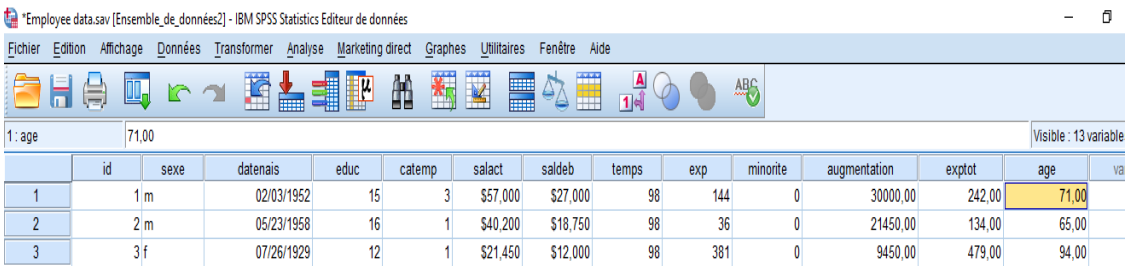
- On clique sur la flèche à coté du pavé numérique pour faire monter la fonction Xdate.Year dans la zone expression numérique, avec comme paramètre la date de naissance,



- On modifie la formule en mettant année actuelle-Xdate.Year(datenais), pour cette année l'année actuelle est 2023,



- La variable calculée Age est ajoutée au fichier de données avec pour chacun des employés son âge calculé.

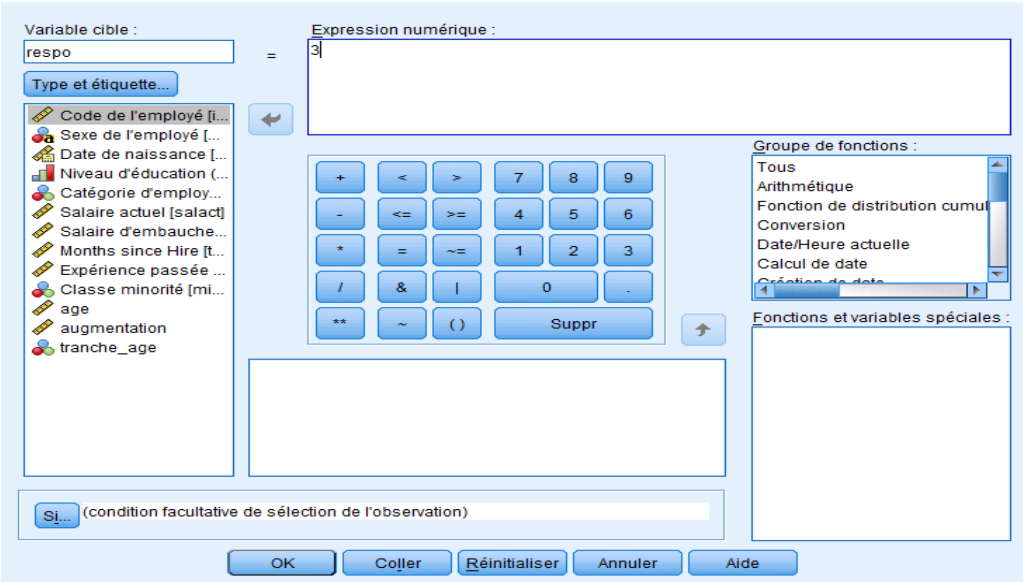


	id	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	augmentation	exptot	age
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0	30000,00	242,00	71,00
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0	21450,00	134,00	65,00
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0	9450,00	479,00	94,00

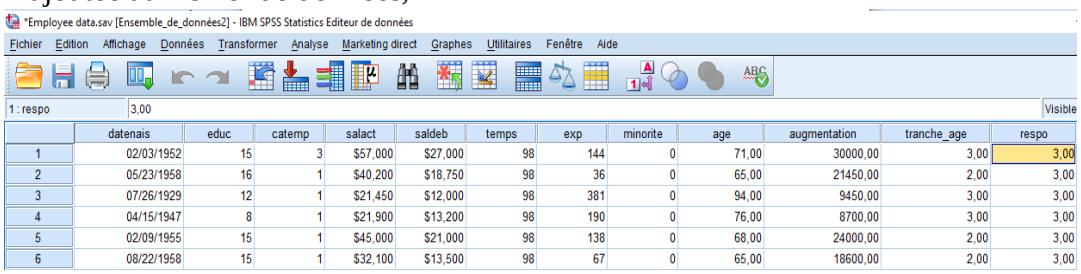
- On désire créer une nouvelle variable "respo" qui se présente sous la forme suivante :
  - "respo = 1" si l'individu est cadre (catemp=2) et possède une expérience (exp) supérieure à 60 mois.
  - "respo = 2" si l'individu est cadre et possède une expérience inférieure à 60 mois.
  - "respo = 3" dans les autres cas.

- On clique sur **Transformer > Calculer la variable …**

- On clique sur le bouton Réinitialiser pour effacer la fonction précédente, la première étape du calcul est un calcul inconditionnel pour tous les individus. On choisit le cas général, ici « respo = 3 ».

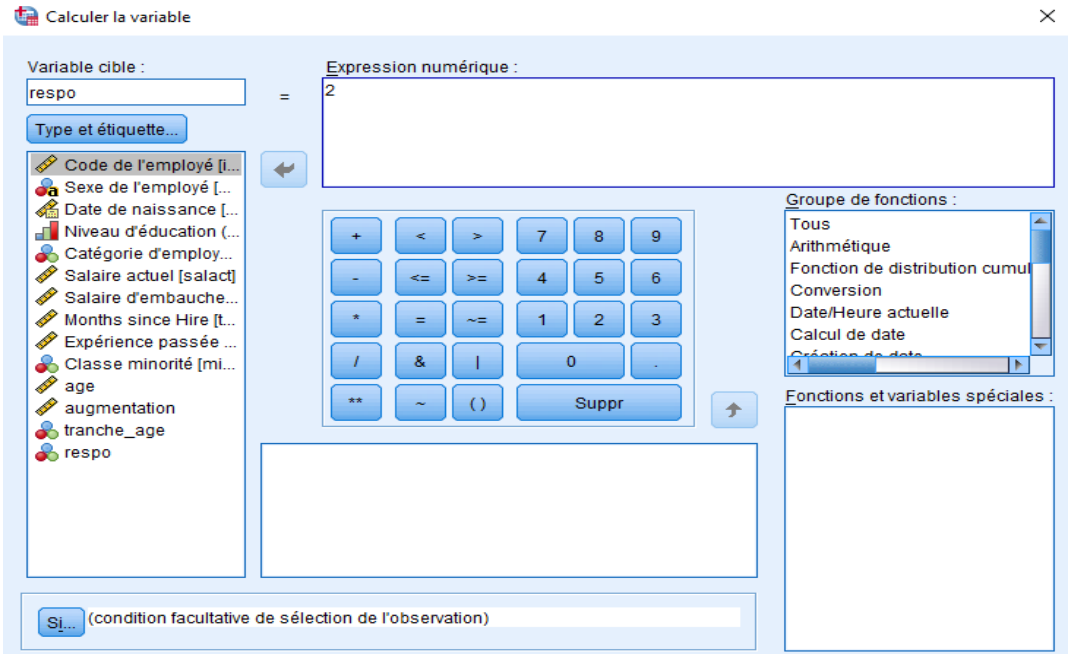


- On clique sur OK, une variable respo dont la valeur est 3 pour toutes les lignes est ajoutée au fichier de données,

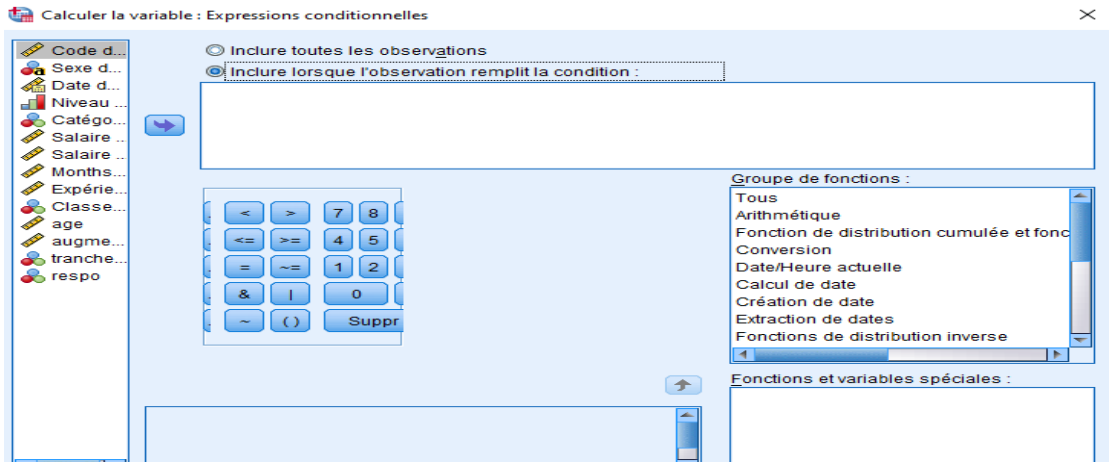


	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	age	augmentation	tranche_age	respo
1	02/03/1952	15	3	$57,000	$27,000	98	144	0	71,00	30000,00	3,00	3,00
2	05/23/1958	16	1	$40,200	$18,750	98	36	0	65,00	21450,00	2,00	3,00
3	07/26/1929	12	1	$21,450	$12,000	98	381	0	94,00	9450,00	3,00	3,00
4	04/15/1947	8	1	$21,900	$13,200	98	190	0	76,00	8700,00	3,00	3,00
5	02/09/1955	15	1	$45,000	$21,000	98	138	0	68,00	24000,00	2,00	3,00
6	08/22/1958	15	1	$32,100	$13,500	98	67	0	65,00	18600,00	2,00	3,00

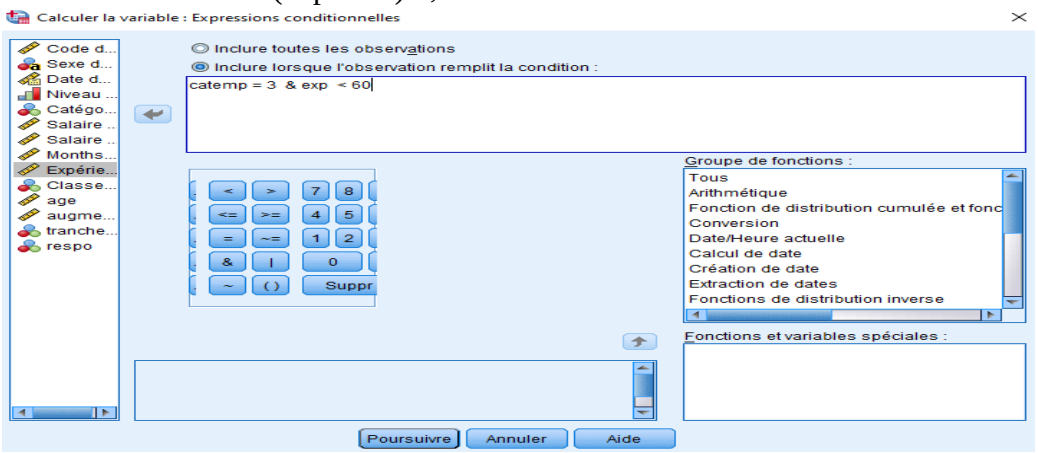
- On clique à nouveau sur **Transformer > Calculer la variable …**, pour introduire la 2ème valeur pour la variable « respo » ; cette valeur est associée à une condition,



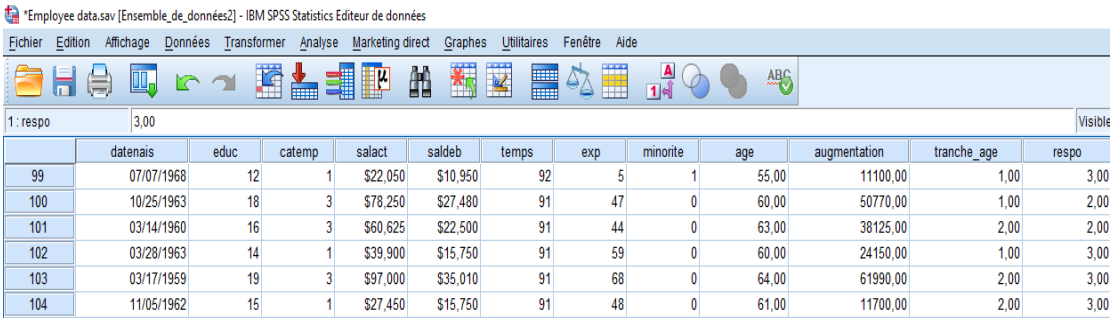
- On clique en bas de l'écran sur Si, et on choisit « Inclure lorsque l'observation remplit la condition : »,



- On introduit la condition « l'individu est cadre (catemp = 3) et possède une expérience inférieure à 60 mois (exp < 60) »,

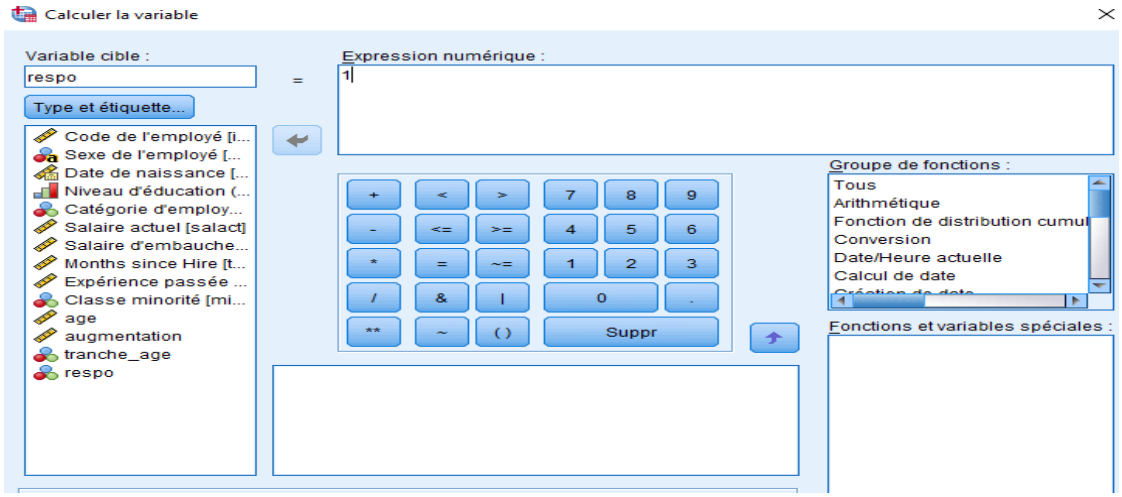


- On clique sur Poursuivre, ensuite sur OK,

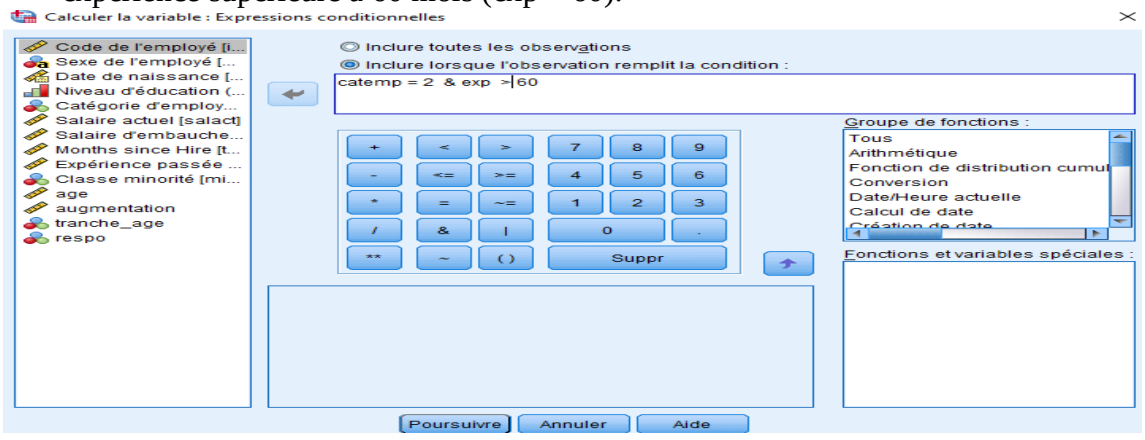


	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	age	augmentation	tranche_age	respo
99	07/07/1968	12	1	$22,050	$10,950	92	5	1	55,00	11100,00	1,00	3,00
100	10/25/1963	18	3	$78,250	$27,480	91	47	0	60,00	50770,00	1,00	2,00
101	03/14/1960	16	3	$60,625	$22,500	91	44	0	63,00	38125,00	2,00	2,00
102	03/28/1963	14	1	$39,900	$15,750	91	59	0	60,00	24150,00	1,00	3,00
103	03/17/1959	19	3	$97,000	$35,010	91	68	0	64,00	61990,00	2,00	3,00
104	11/05/1962	15	1	$27,450	$15,750	91	48	0	61,00	11700,00	2,00	3,00

- On clique à nouveau sur **Transformer > Calculer la variable …**, pour introduire la 3ème valeur pour la variable « respo » ; cette valeur est associée à une condition,



- On clique en bas de l'écran sur Si, et on choisit « Inclure lorsque l'observation remplit la condition : », "respo = 1" si l'individu est cadre (catemp=2) et possède une expérience supérieure à 60 mois (exp > 60).

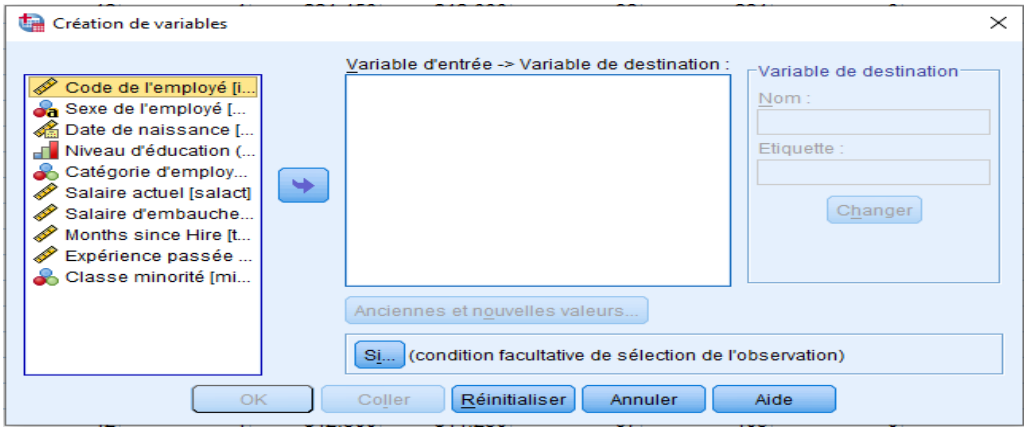


- On clique sur Poursuivre ensuite sur OK, la valeur 1 de respo apparait aussi dans le fichier de données,

110	10/29/1952	15	1	$28,350	$18,000	91	151	1	71,00	10350,00	3,00	3,00
111	11/27/1940	12	2	$30,750	$9,000	91	314	1	83,00	21750,00	3,00	1,00
112	06/21/1948	12	2	$30,750	$15,000	91	240	1	75,00	15750,00	3,00	1,00
113	10/06/1959	16	3	$54,875	$27,480	90	68	0	64,00	27395,00	2,00	3,00
114	08/25/1961	14	1	$37,800	$16,500	90	60	0	62,00	21300,00	2,00	3,00
115	05/12/1961	15	1	$33,450	$14,100	90	85	0	62,00	19350,00	2,00	3,00
116	06/09/1962	15	1	$30.300	$16.500	90	16	0	61.00	13800.00	2.00	3.00

VII.3 Création de variables

Cette fonctionnalité permet de créer de nouvelles variables (généralement catégorielles) à partir de variables du fichier de données, pour cela on utilise le menu **Transformer > Création de variables > …,**



Pour illustrer cette fonctionnalité, on utilise le fichier de données employees data.sav :

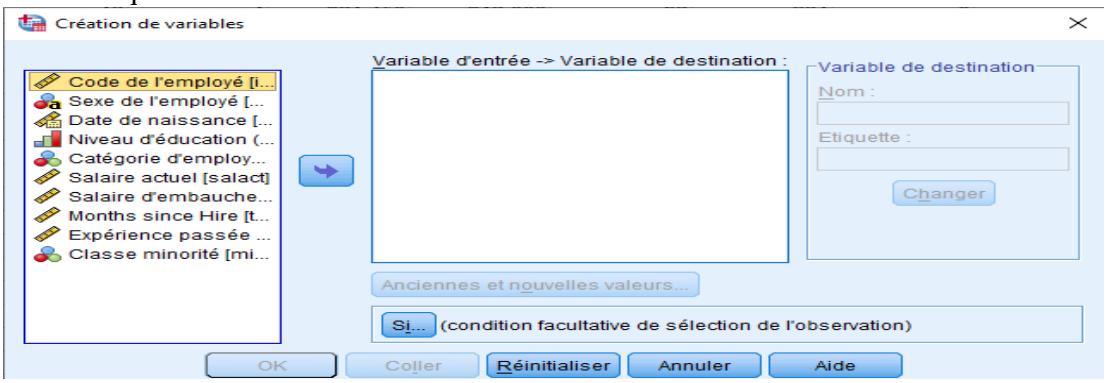
- Créer une nouvelle variable tranche_age en utilisant le menu Transformer → création de variables qui permet d'affecter les valeurs suivantes :

  1 : si 40 <= age <60 ;

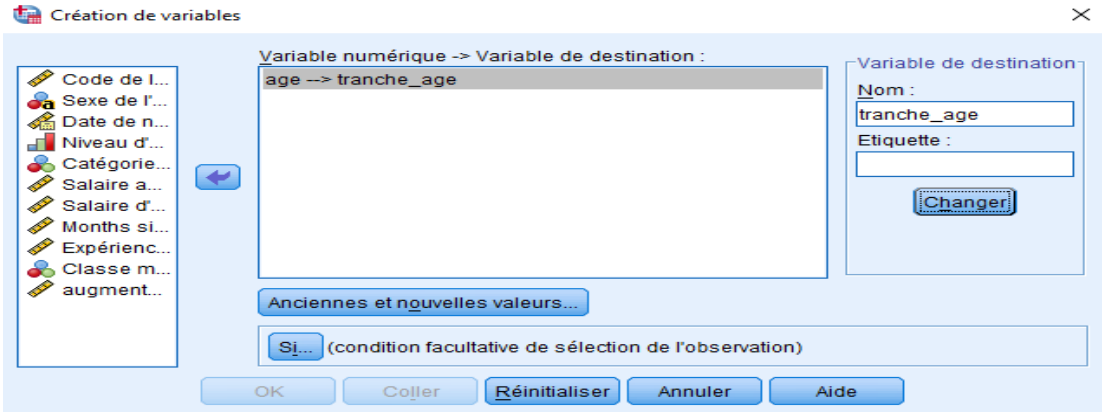
  2 : si 60 <= age <70 ;

  3 : si age >= 70 ;

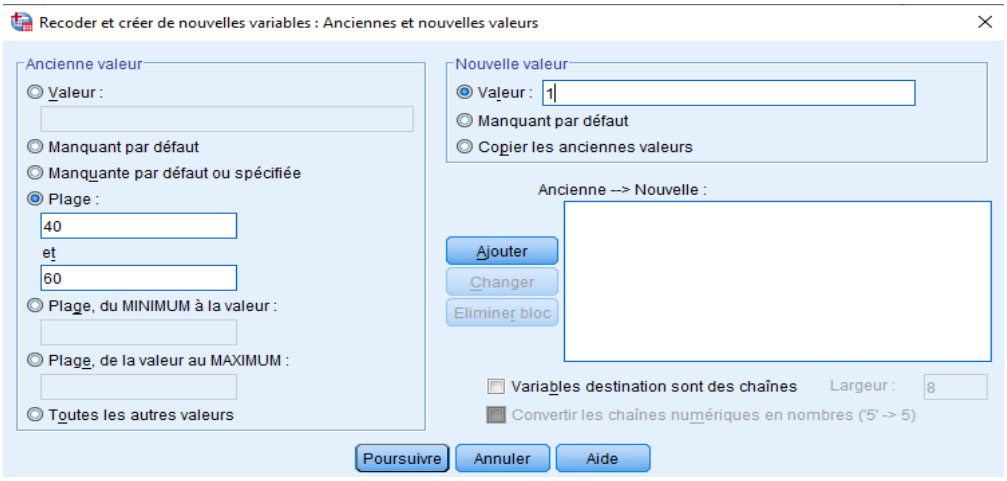
  - Ouvrir le fichier employee data.sav : **Fichier > Ouvrier > …**
  - On clique sur **Transformer > Création de variables …**



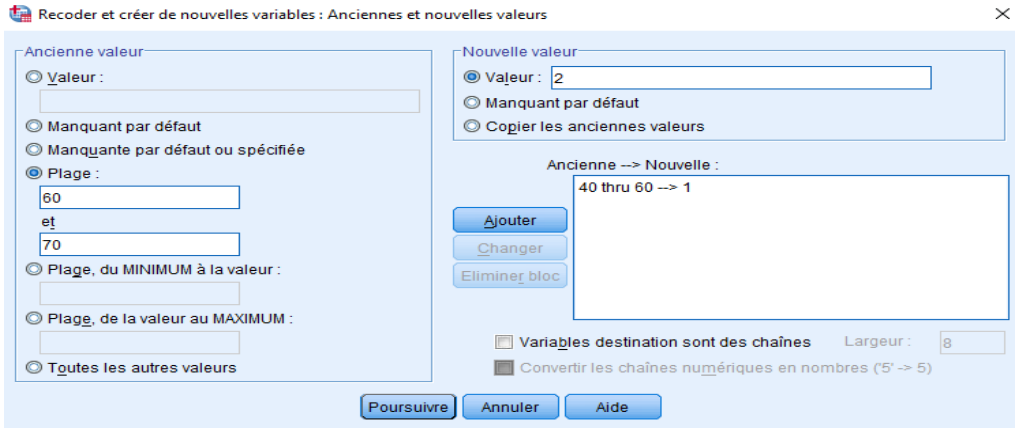
  - On met la variable Age dans la zone Variable d'entrée -> Variable de destination, on met la variable tranche_age à droite dans Variable de destination et on clique sur le bouton Changer,



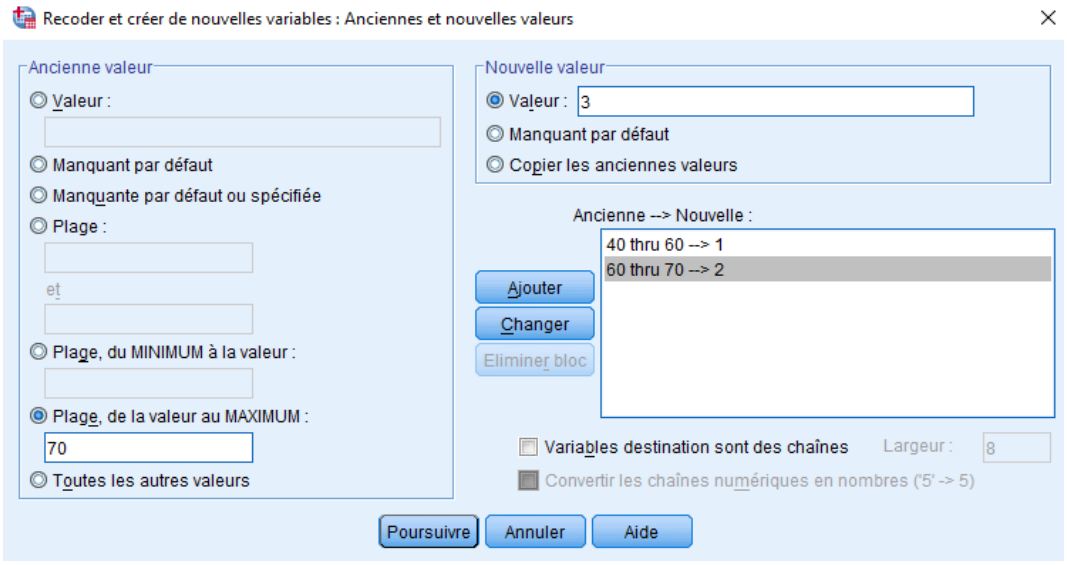
- On clique sur Anciennes et nouvelles valeurs, à droite on met la valeur et à gauche on met l'intervalle de la valeur ensuite on clique sur Ajouter,



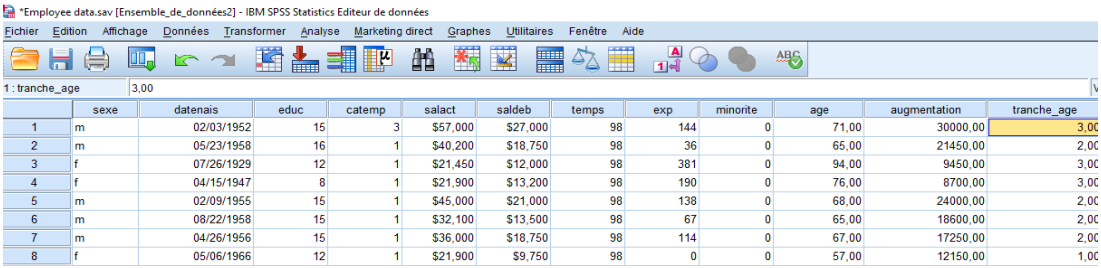
- On introduit la 2ème valeur, et on clique sur ajouter,



- Enfin on introduit la 3ème valeur, on n'a pas un intervalle on choisit à gauche l'option Plage de la valeur au maximum, on clique sur Ajouter et puisque c'est la dernière valeur on clique sur Poursuivre,



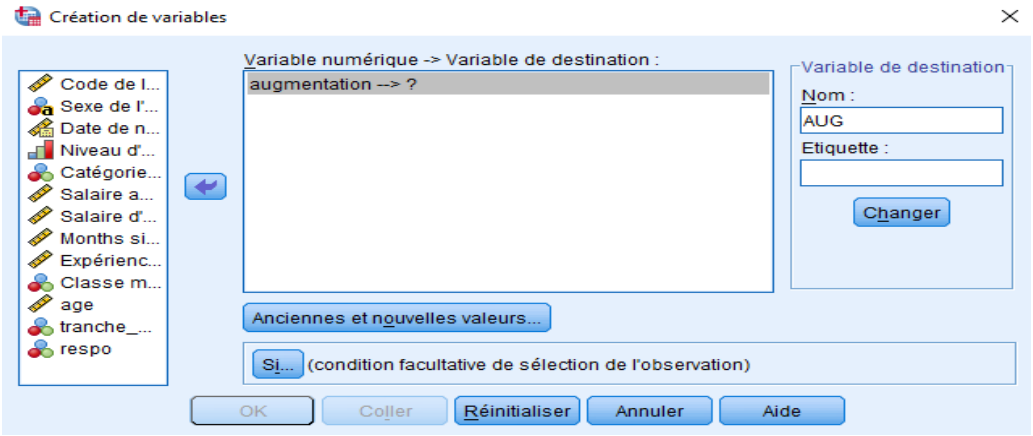
- Une variable tranche_age est ajoutée au fichier de donnée et qui prend trois valeurs.



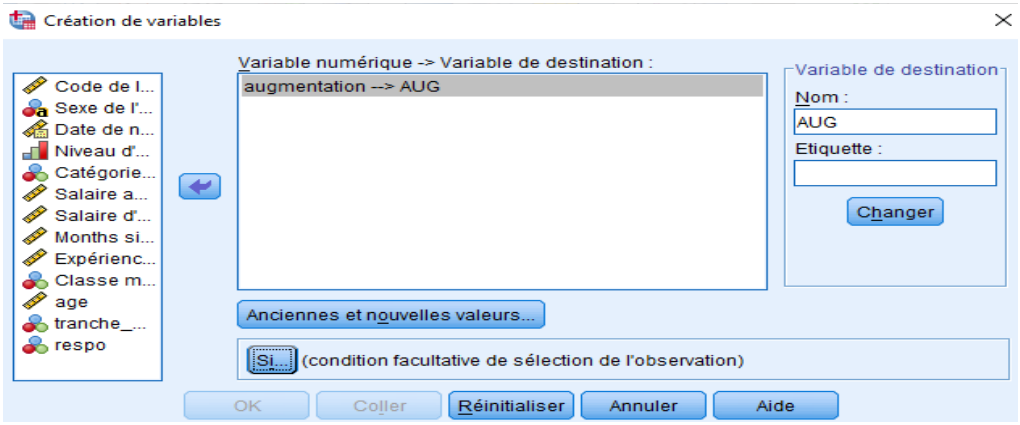
	sexe	datenais	educ	catemp	salact	saldeb	temps	exp	minorite	age	augmentation	tranche_age
1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0	71,00	30000,00	3,00
2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0	65,00	21450,00	2,00
3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0	94,00	9450,00	3,00
4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0	76,00	8700,00	3,00
5	m	02/09/1955	15	1	$45,000	$21,000	98	138	0	68,00	24000,00	2,00
6	m	08/22/1958	15	1	$32,100	$13,500	98	67	0	65,00	18600,00	2,00
7	m	04/26/1956	15	1	$36,000	$18,750	98	114	0	67,00	17250,00	2,00
8	f	05/06/1966	12	1	$21,900	$9,750	98	0	0	57,00	12150,00	1,00

- Créer une nouvelle variable AUG qui regroupe les employés en fonction de la variable *Augmentation* en trois classe :
  - [1 000,10 000[Devient "class1"
  - [10 000,..., 20 000[Devient "class2"
  - [20 000,...,100 000[Devient "class3"

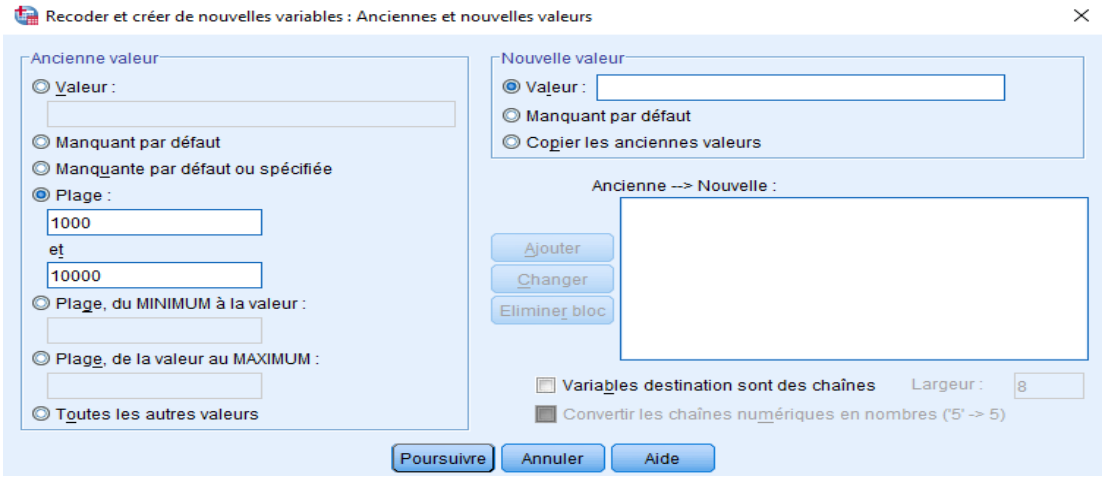
- On clique sur **Transformer > Création de variables …**, on clique sur le bouton Réinitialiser pour effacer la fonction précédente,
- On transfert la variable augmentation vers la zone au milieu, et à droite on écrit le nom de la nouvelle variable AUG,



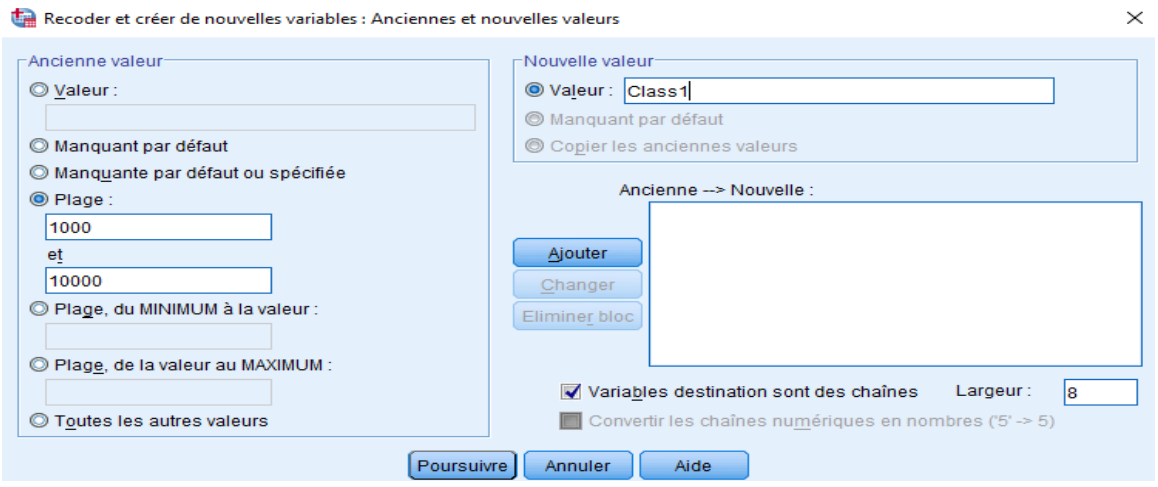
- On clique dur changer, on obtient la fenêtre suivante :



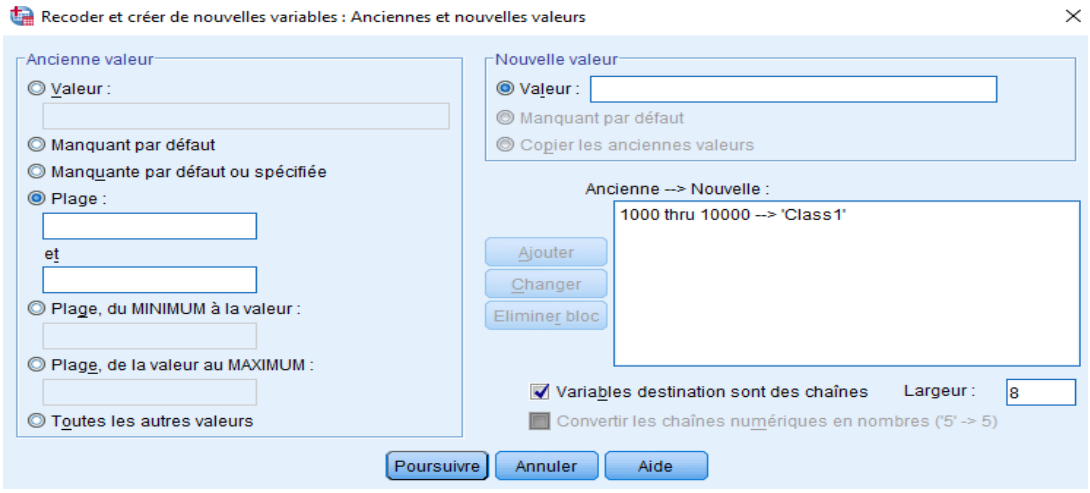
- Cliquer sur Anciennes et nouvelles valeurs, on choisit l'option plage pour introduire le premier intervalle,



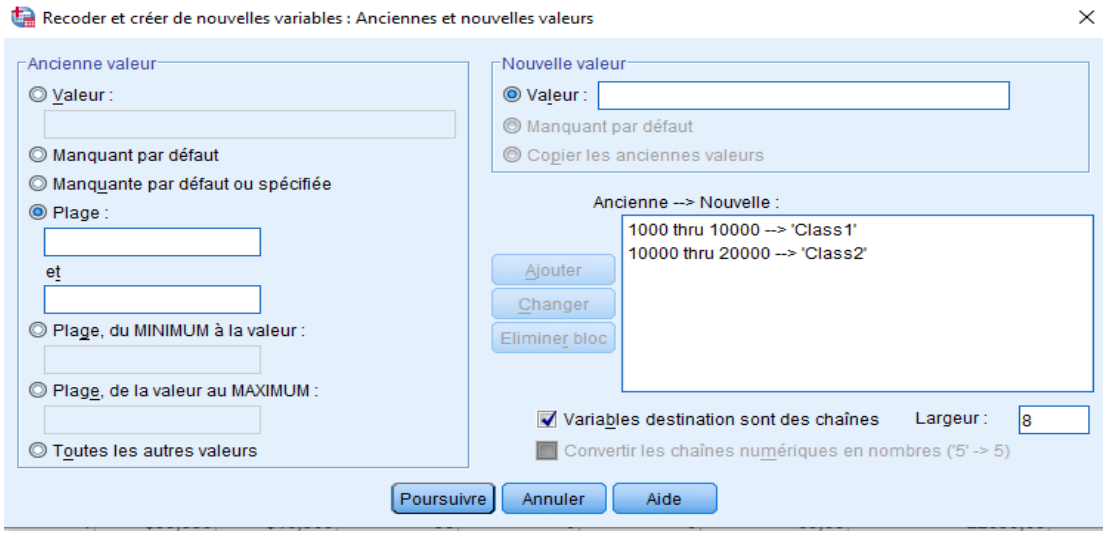
- En bas à droite, cocher « variables de destination sont des chaines » pour pouvoir introduire la valeur « Class1 » dans valeur à droite en haut de la fenêtre,



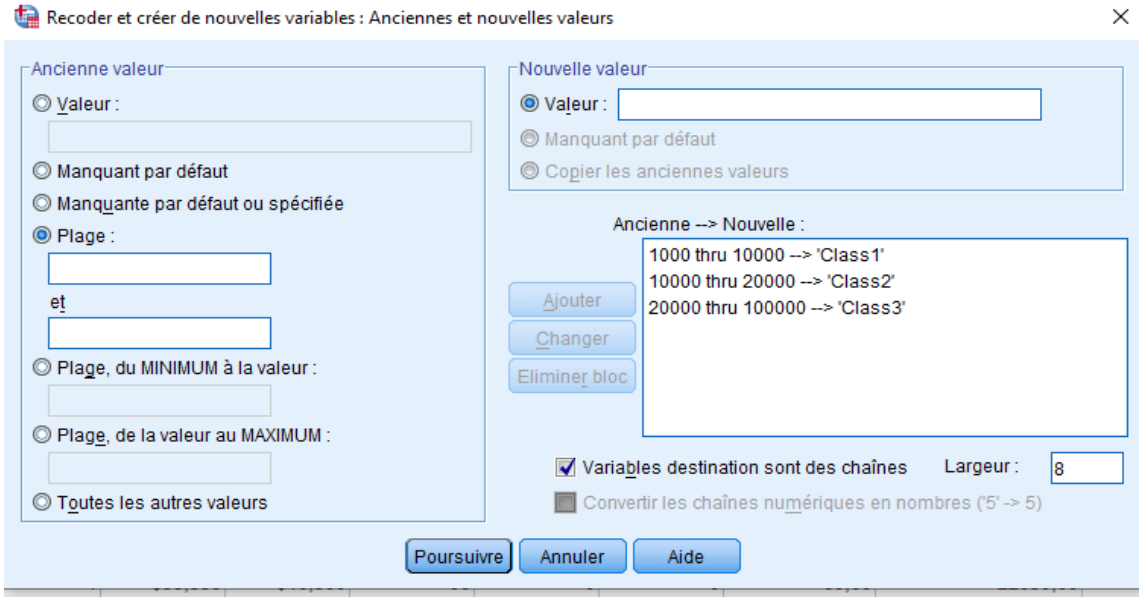
- Cliquer sur Ajouter,



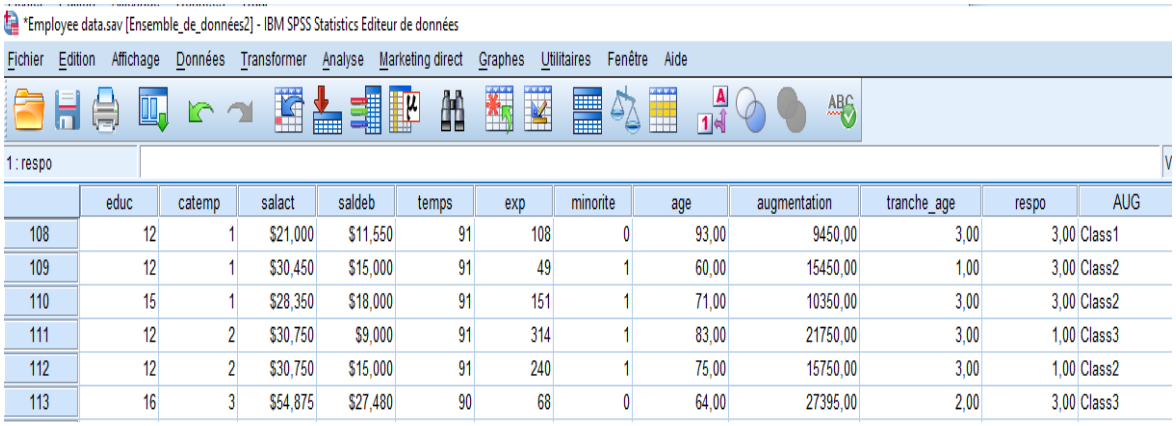
- Introduire le 2ème intervalle, pour la nouvelle valeur Class2 et cliquer sur Ajouter,



- Introduire la 3ème et dernière valeur, cliquer sur Poursuivre ensuite OK,



- Une variable AUG est ajoutée au fichier de données et qui prend les valeurs, Class1, Class2 et Class3,



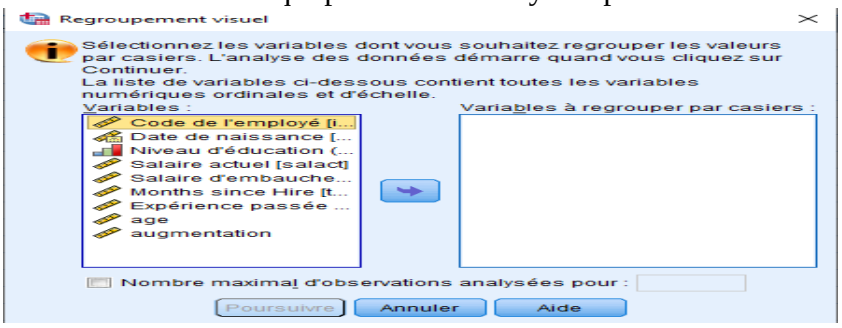
	educ	catemp	salact	saldeb	temps	exp	minorite	age	augmentation	tranche_age	respo	AUG
108	12	1	$21,000	$11,550	91	108	0	93,00	9450,00	3,00	3,00	Class1
109	12	1	$30,450	$15,000	91	49	1	60,00	15450,00	1,00	3,00	Class2
110	15	1	$28,350	$18,000	91	151	1	71,00	10350,00	3,00	3,00	Class2
111	12	2	$30,750	$9,000	91	314	1	83,00	21750,00	3,00	1,00	Class3
112	12	2	$30,750	$15,000	91	240	1	75,00	15750,00	3,00	1,00	Class2
113	16	3	$54,875	$27,480	90	68	0	64,00	27395,00	2,00	3,00	Class3

VII.4 Création d'une variable catégorielle à partir d'une variable d'échelle

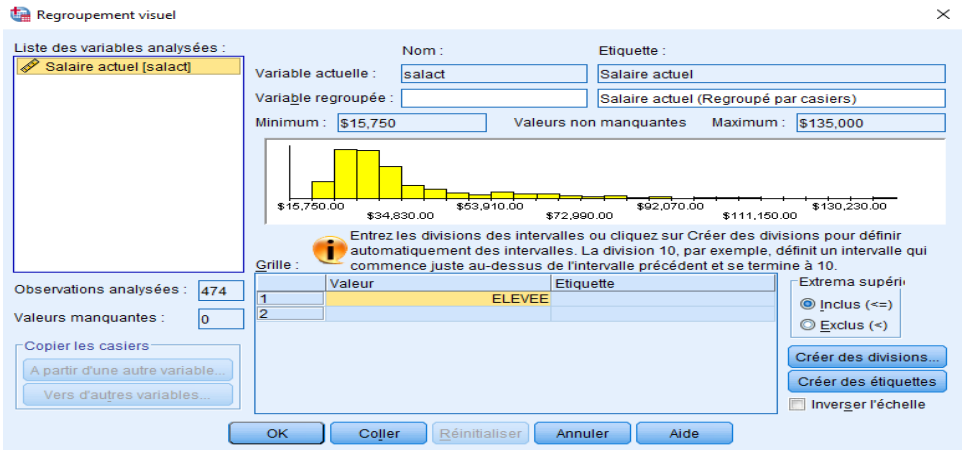
Pour créer la variable catégorielle à partir d'une variable d'échelle, on suit les étapes suivantes:

1. Ouvrir le fichier employee data.sav : **Fichier > Ouvrier > …**
2. A partir des menus de la fenêtre de l'éditeur de données, sélectionnez : **Transformer > Regroupement visuel > ...**,
3. Dans la boîte de dialogue initiale Regroupement visuel, sélectionner les variables d'échelle et/ou ordinales pour lesquelles on souhaite créer des variables regroupées. Le regroupement consiste à prendre plusieurs valeurs contiguës et à les regrouper dans une même catégorie.

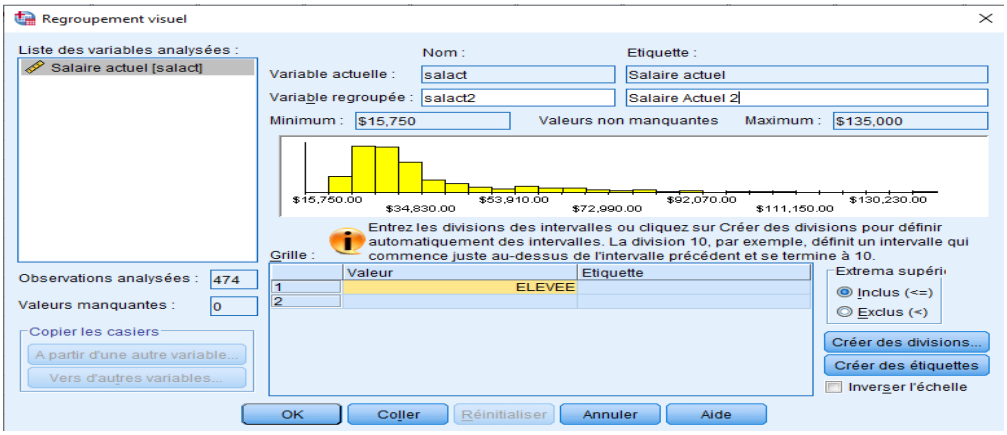
Si le fichier de données contient un nombre d'observations important, cette opération peut prendre un certain temps. Par conséquent, la boîte de dialogue initiale vous permet également de limiter le nombre d'observations à lire (« analyser »). Pour notre cas, le fichier employees data.sav contient 474 observations ce qui permet une analyse rapide.



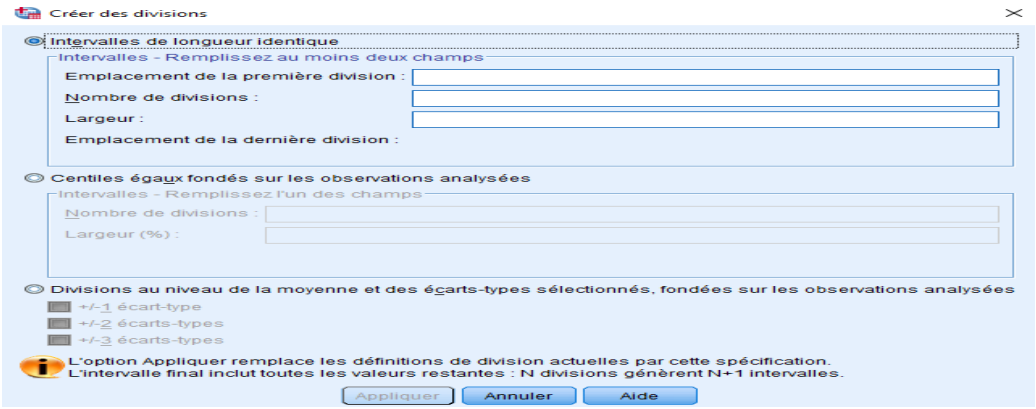
4. Faire glisser Salact (salaire actuel) de la liste Variables vers la liste Variables à regrouper, puis cliquer sur Poursuivre.



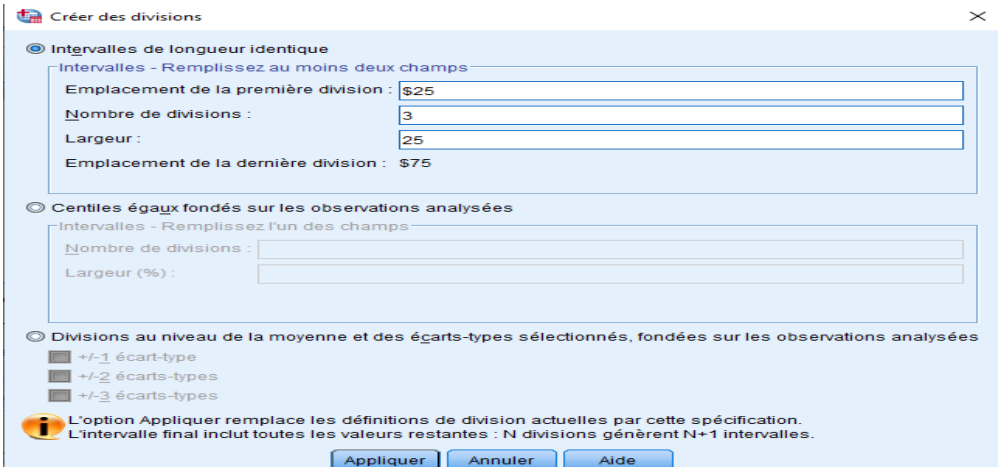
5. Affecter le nom Salact2 à la nouvelle variable regroupée et sélectionnez le libellé de variable Salaire Actuel 2.



6. Cliquer sur Créer des divisions.



7. Sélectionner Intervalles de longueur identique.
8. Saisir 25 pour l'emplacement de la première division, 3 pour le nombre de divisions et 25 pour la largeur.

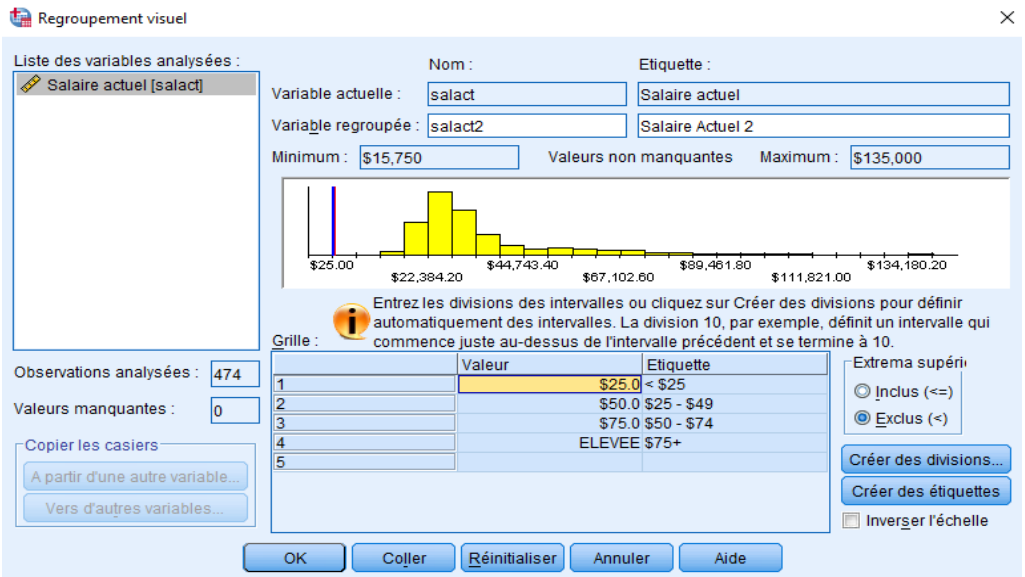


Le nombre de catégories regroupées correspond au nombre de divisions, plus 1. Ainsi, dans cet exemple, la nouvelle variable regroupée comportera quatre catégories, les trois premières couvrant une plage de 25 et la dernière toutes les valeurs supérieures à celle de la dernière division 75.

9. Cliquer sur Appliquer.

Les valeurs qui figurent à présent dans la grille représentent les divisions définies, à savoir les points finaux supérieurs de chaque catégorie. Par ailleurs, les lignes verticales de l'histogramme indiquent l'emplacement des divisions. Par défaut, les valeurs de division sont incluses dans les catégories correspondantes. Par exemple, la première valeur (25) inclurait toutes les valeurs inférieures ou égales à 25. Dans cet exemple toutefois, nous voulons définir les catégories suivantes : inférieur à 25, 25–49, 50–74 et 75 et plus.

10. Dans le groupe Points finaux supérieurs, sélectionner Exclus (<).
11. Cliquer ensuite sur Créer des libellés.



12. Cliquer sur OK pour créer la variable regroupée.

La nouvelle variable est affichée dans l'éditeur de données. Etant donné que la variable est ajoutée à la fin du fichier, elle apparaît dans la colonne la plus à droite dans Vue de données et dans la dernière ligne de la vue de variable.

Chapitre VIII : Création et modification de graphes

On peut créer et modifier des types de graphique divers et variés. Dans cette section, nous allons créer et modifier des graphiques à barres. On peut appliquer les principes à n'importe quel type de graphique.

VIII.1 Création de graphes

La boîte de dialogue Générateur de graphiques est une fenêtre interactive qui permet d'obtenir l'aperçu d'un graphique avant de le générer.

Pour créer un graphique on suit les étapes suivantes :

- On sélectionne dans le menu **Graphes > Générateur de diagrammes > ...**

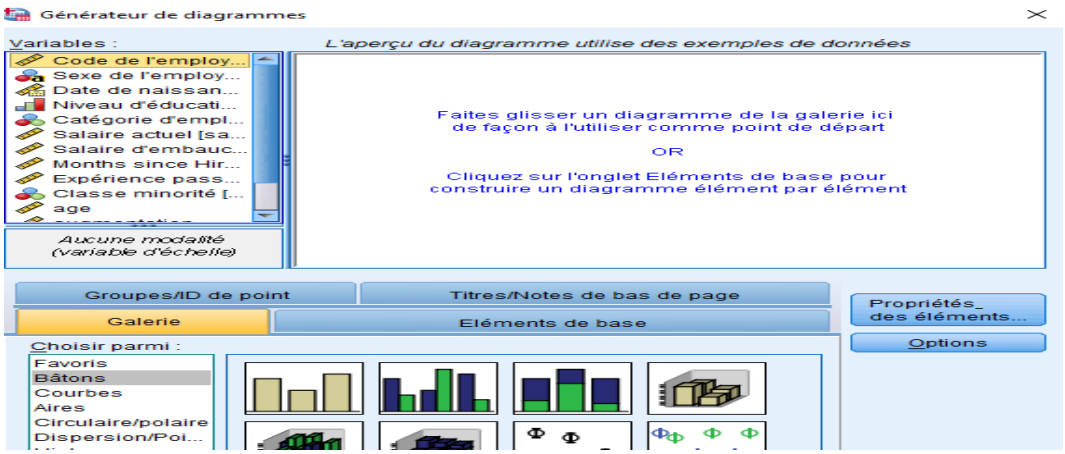
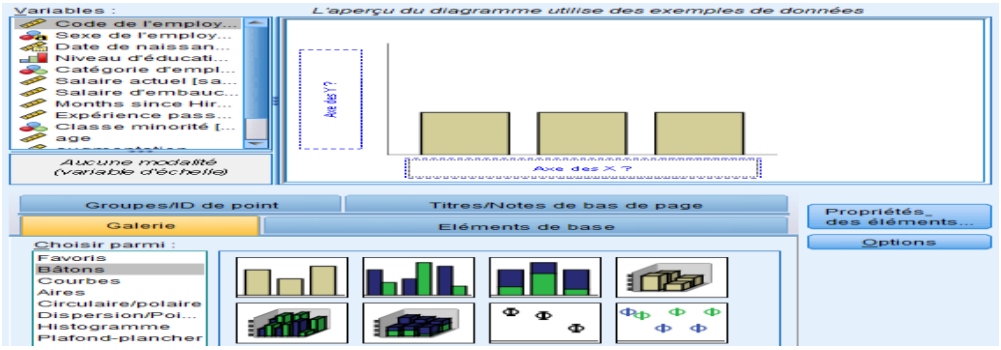
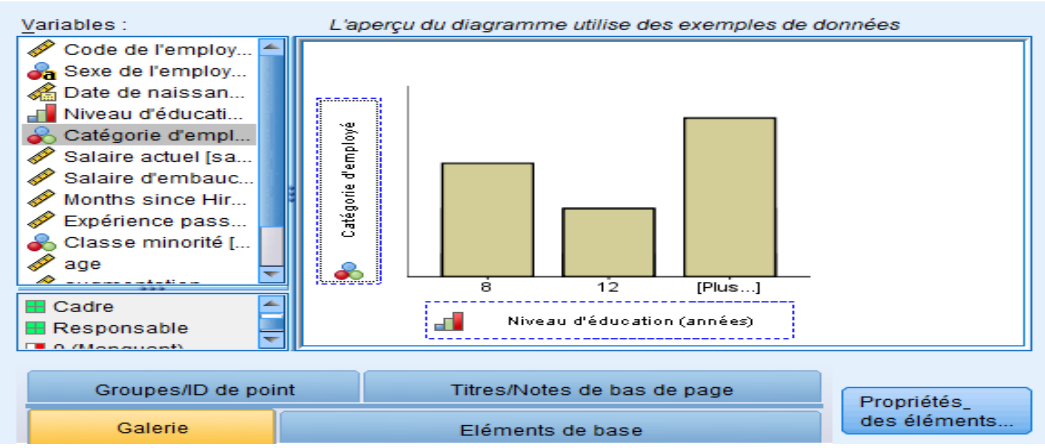


Figure VIII.1. Générateur de diagrammes.

- Cliquer sur l'onglet Galerie, la galerie inclut plusieurs graphiques différents prédéfinis, qui sont organisés par type de graphique. L'onglet Eléments de base fournit également des éléments de base (comme les axes et les éléments graphiques) pour créer des graphiques en partant de zéro, mais il est plus facile d'utiliser la galerie.
- Cliquer sur Bâtons s'il n'est pas sélectionné. Les icônes représentant les graphiques à barres disponibles dans la galerie apparaissent dans la boîte de dialogue. Les images doivent fournir suffisamment d'informations pour identifier le type de graphique spécifique.
- Faire glisser l'icône du graphique à barres simples sur le « canevas », qui est en fait la zone étendue au-dessus de la galerie. Le Générateur de graphiques affiche un aperçu du graphique.



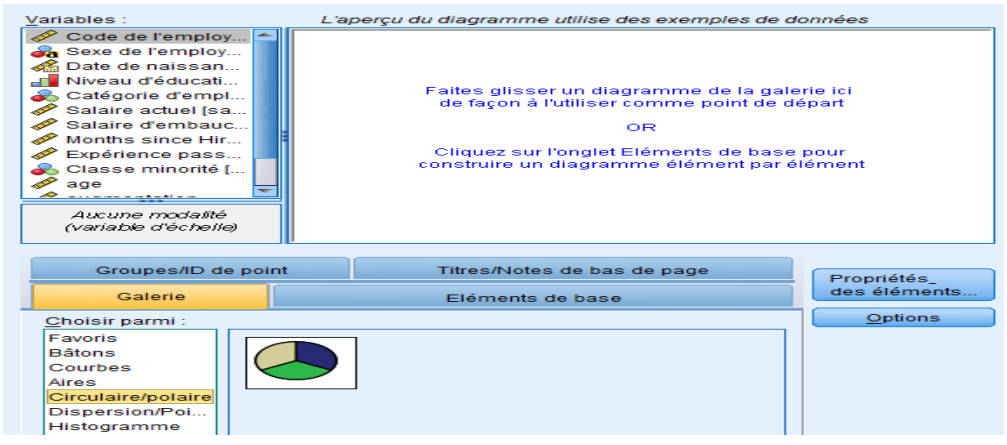
On fait glisser les variables du côté gauche en haut vers le graphe pour définir l'axe des Y et l'axe des X, par exemple on voudrait créer un diagramme en bâtons qui représente le la catégorie de l'employer en fonction du niveau d'études.



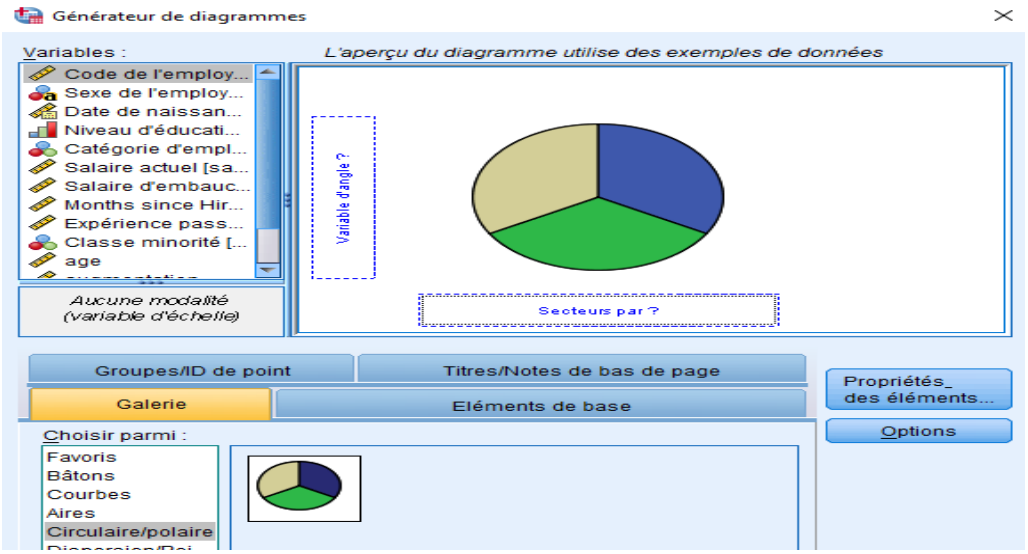
On peut aussi ajouter du texte dans un graphe ; par exemple des titres et des notes de bas de page en cliquant sur l'onglet Titres/Notes de bas de page.

Considérons le fichier de données ***employee data.sav***.

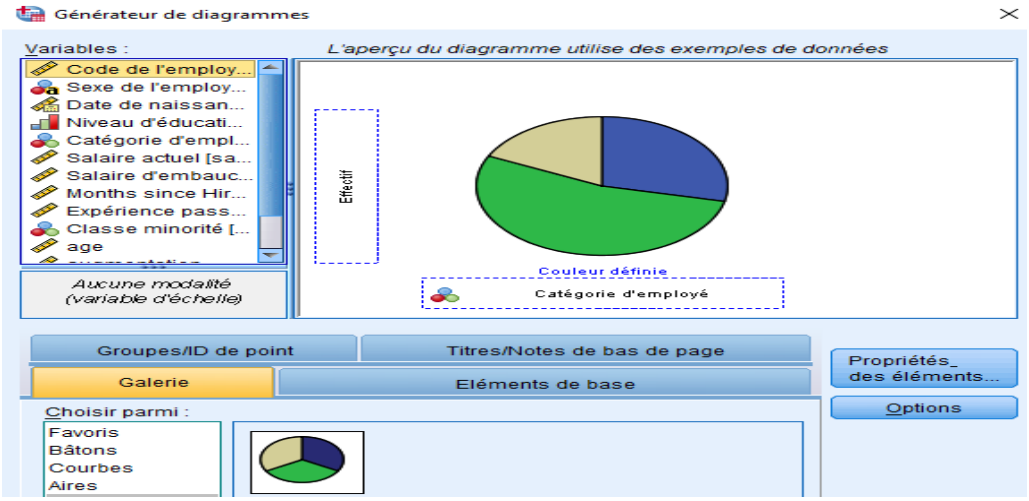
- Quelle est la répartition des employés (générer diagramme secteur) selon la catégorie de l'emploi ?
  - On clique sur le menu **Graphes > Générateur de diagrammes > ...**
  - On clique sur Circulaire/polaire ;



  - On glisse le schéma dans la zone de texte ;



- On glisse la variable Catégorie d'employer dans la zone du graphe, et on clique sur OK ;



- Le graphe est généré dans la fenêtre Résultats ;

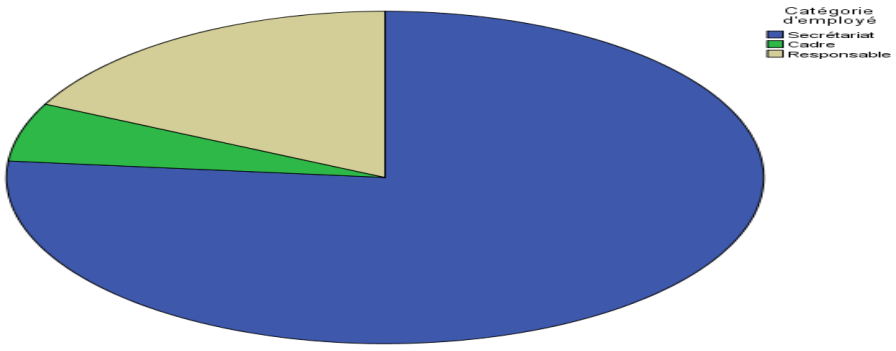


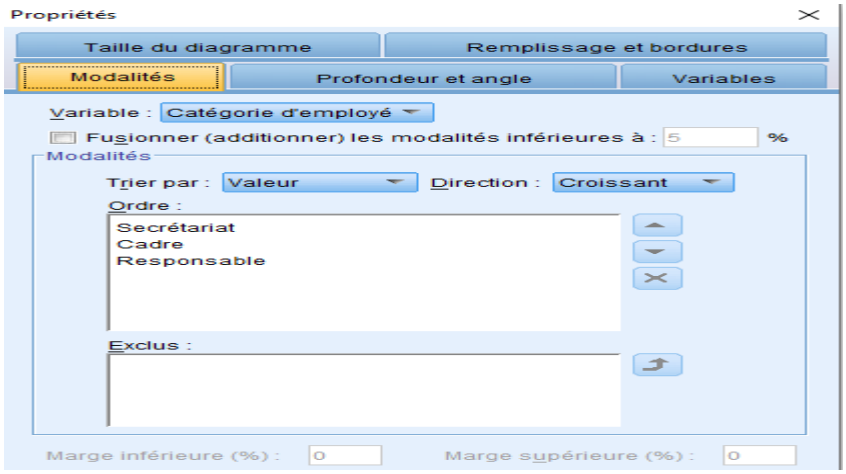
Figure VII.2. Diagramme circulaire pour la variable Catégorie d'emploi.

VIII.2 Modification de graphes

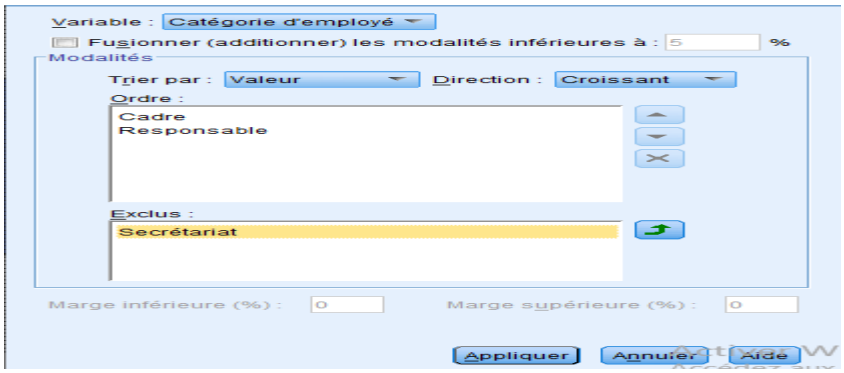
Plusieurs modifications peuvent être apportées à un graphe déjà créé en ajoutant des légendes, des titres, modifier des couleurs, on peut même modifier le type du graphique.

Dans le digramme précédent, on veut exclure la catégorie « Secrétariat », la réinsérer et donner un titre au diagramme avec des pourcentages et des effectifs.

- On double clique sur le diagramme (plus précisément sur les bords),



- On sélectionne la valeur à exclure ici « Secrétariat » et on clique sur le bouton  à droite, ensuite sur Appliquer,



- On obtient le schéma suivant,

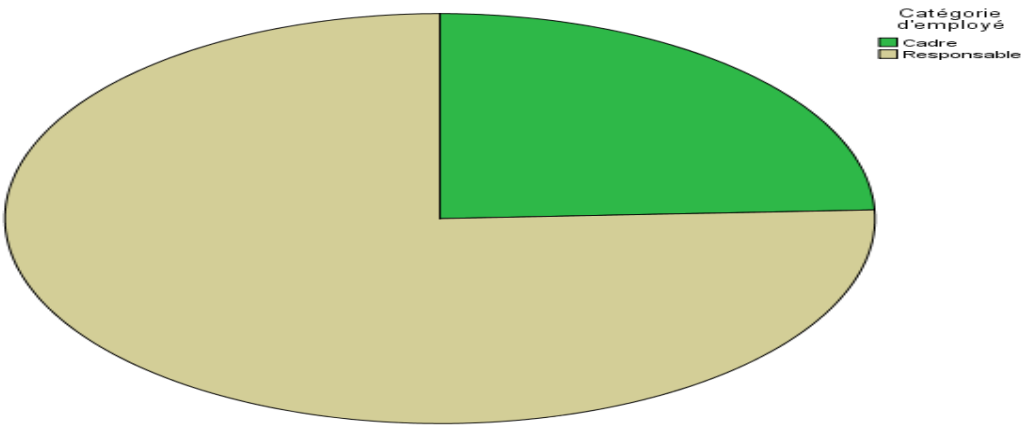
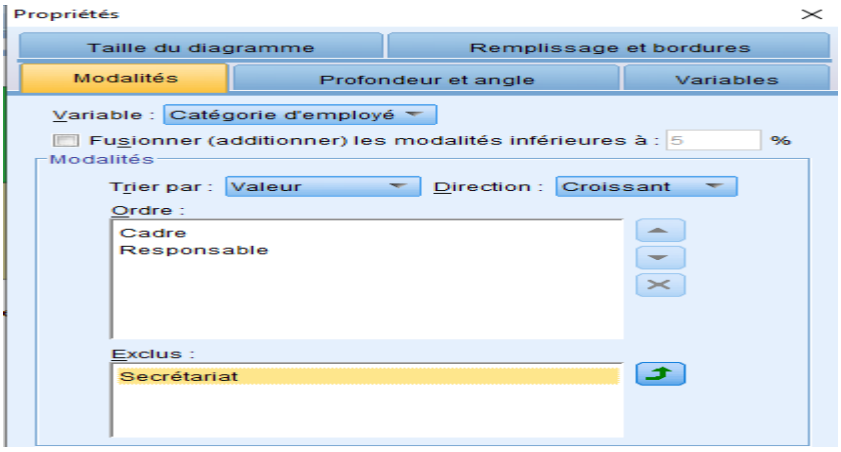


Figure VII.3. Diagramme circulaire pour la variable Catégorie d'emploi sans la modalité Secrétaire.

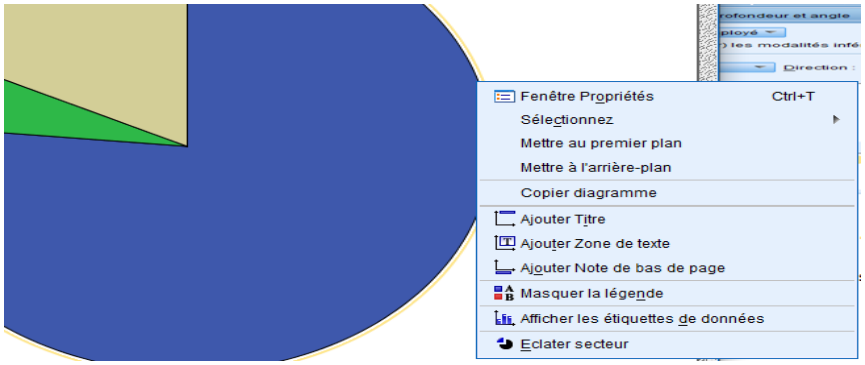
- Pour restaurer le diagramme de départ, on double clique sur le diagramme (plus précisément sur les bords),



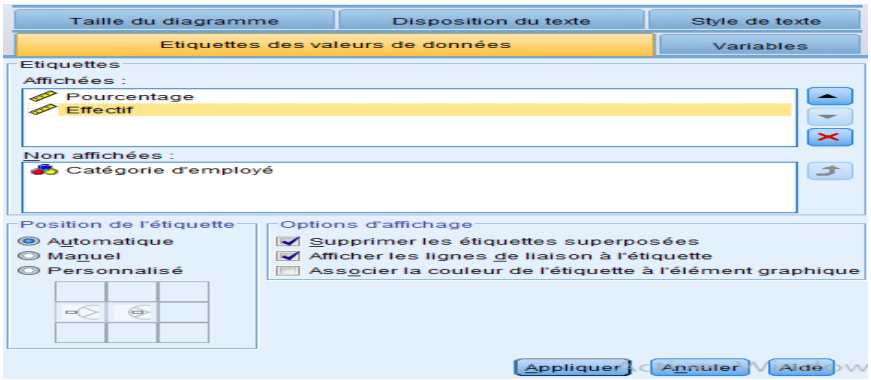
- On sélectionne Secrétariat exclue précédemment, et on clique sur le bouton 

Pour afficher des pourcentages et/ou des effectifs dans les secteurs su diagramme,

- On clique avec le bouton droit sur le diagramme,



- Choisir « Afficher les étiquettes de données »,



- On sélectionne Pourcentage et Effectif,

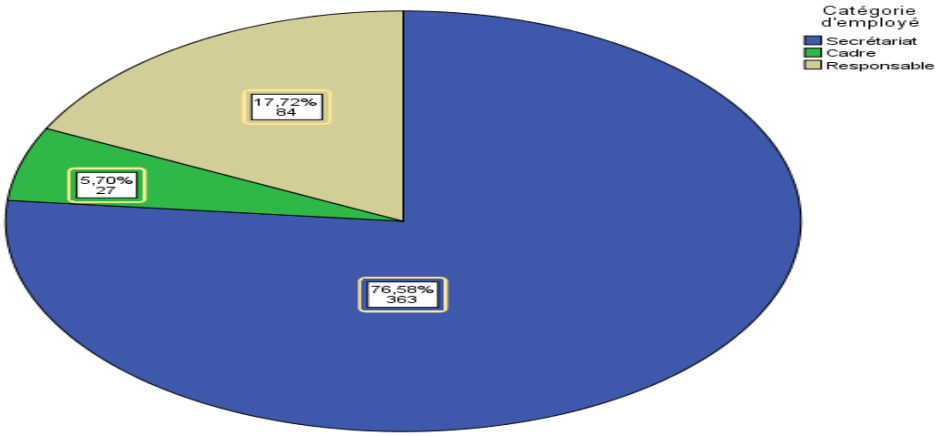


Figure VII.4. Diagramme circulaire pour la variable Catégorie d'emploi avec les effectifs et les pourcentages.

Pour toutes les autres modifications concernant les graphes on procède de la même façon (c'est-à-dire en accédant à la fenêtre propriétés).

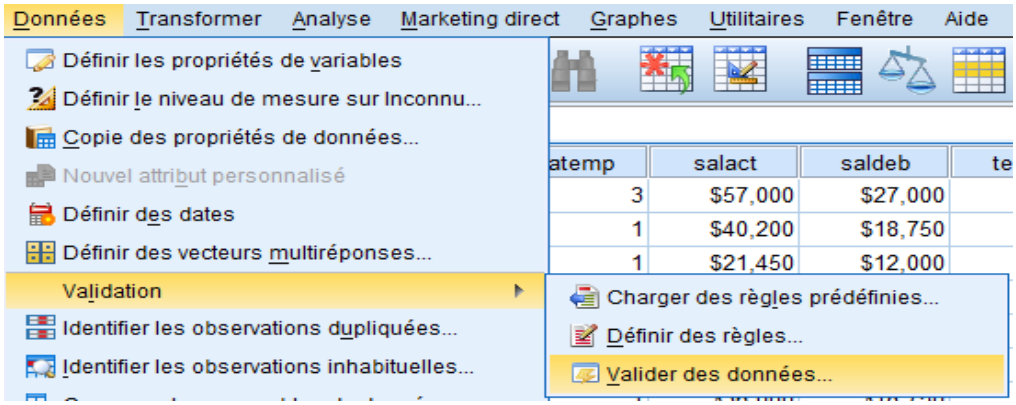
Chapitre IX : Validation de données

Dans cette section, on va présenter différentes manières de détecter des données dupliquées, incohérentes ainsi que des cas d'erreur en utilisant la validation disponible dans SPSS.

IX.1 Opération de validation, détection d'erreurs

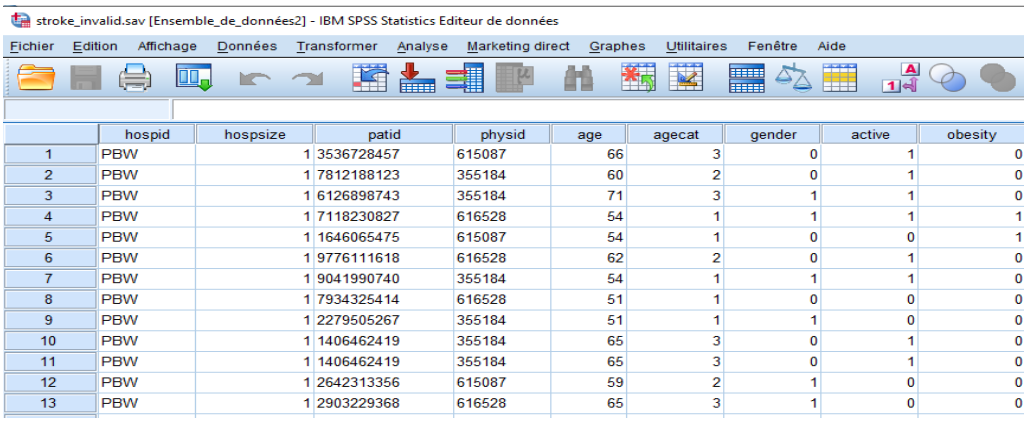
La validation permet de vérifier si le contenu du fichier de données est cohérent (ne contient pas de données aberrantes), elle permet aussi de détecter la présence de données doubles.

- Pour valider les données d'un fichier on utilise le menu **Données > Validation > Valider des données > ...**



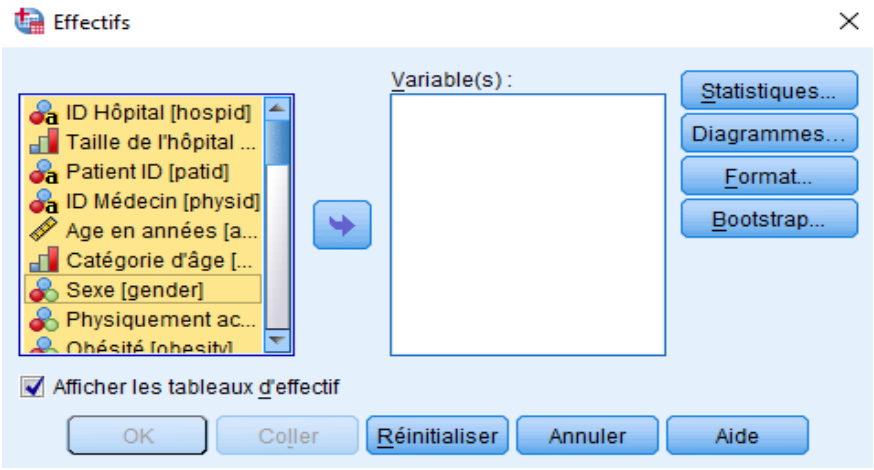
Il existe des fichiers de données dans SPSS qui sont invalides (contiennent des erreurs), l'objectif consiste à les utiliser pour voir l'intérêt et les démarches de validation des données. Pour cela on considère le fichier ***Stroke_invalid.sav*** qui décrit un ensemble de patients victimes d'apoplexie.

- On clique sur le menu **Fichier > Ouvrir > Données >…** le fichier ***Stroke_invalid.sav*** contient 1183 patients (lignes du tableau) et 39 variables.

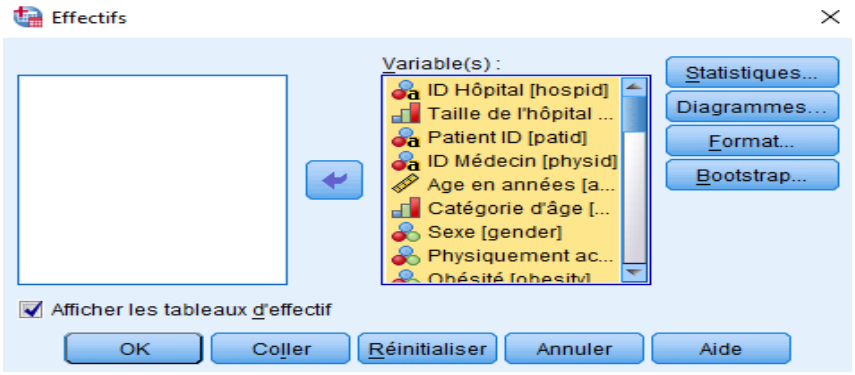


	hospid	hospsize	patid	physid	age	agecat	gender	active	obesity
1	PBW	1	3536728457	615087	66	3	0	1	0
2	PBW	1	7812188123	355184	60	2	0	1	0
3	PBW	1	6126898743	355184	71	3	1	1	0
4	PBW	1	7118230827	616528	54	1	1	1	1
5	PBW	1	1646065475	615087	54	1	0	0	1
6	PBW	1	9776111618	616528	62	2	0	1	0
7	PBW	1	9041990740	355184	54	1	1	1	0
8	PBW	1	7934325414	616528	51	1	0	0	0
9	PBW	1	2279505267	355184	51	1	1	0	0
10	PBW	1	1406462419	355184	65	3	0	1	0
11	PBW	1	1406462419	355184	65	3	0	1	0
12	PBW	1	2642313356	615087	59	2	1	0	0
13	PBW	1	2903229368	616528	65	3	1	0	0

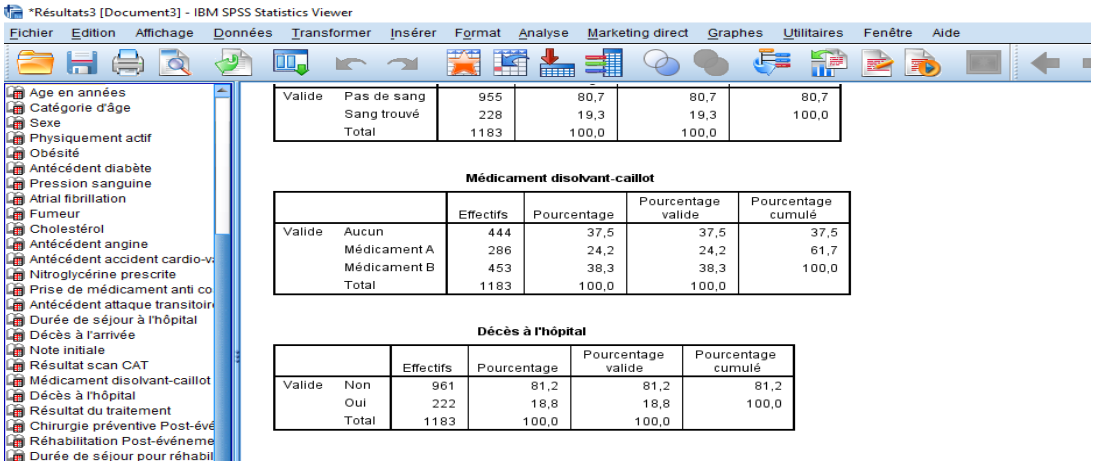
- Afficher les tableaux des fréquences de toutes les variables du fichier Stroke_invali.sav.
  - On clique sur **Analyse > Statistiques descriptives > Effectifs > …**
  - On sélectionne toutes les variables dans le coté gauche,



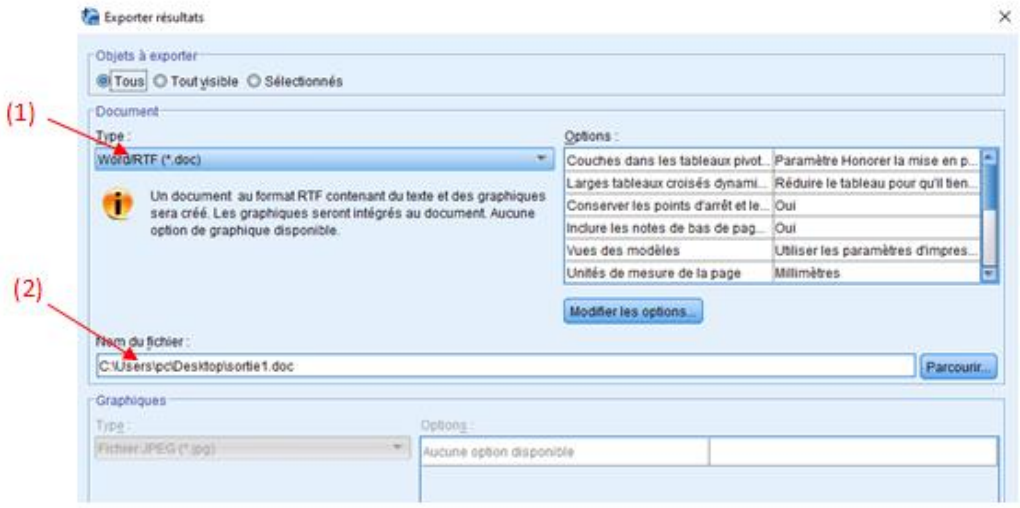
- On clique sur la flèche au milieu pour les déplacer dans la zone Variables à droite, et on clique sur OK,



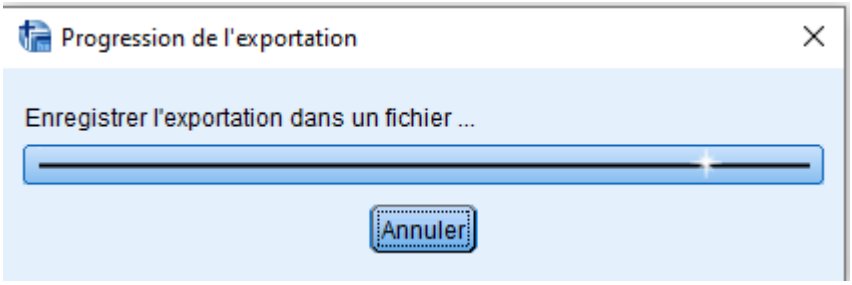
- On obtient dans le fichier résultats les 39 tableaux de fréquences associées à toutes les variables,



- Enregistrer les tableaux générés dans un document Word.
  - On choisit le menu **Fichier > Exporter > …**



- On choisit le type de fichier dans lequel on exporte les résultats dans (1), et le chemin et le nom du fichier dans (2) ensuite on clique sur OK, on obtient une fenêtre qui indique la progression de l'exportation,



- On va chercher le fichier dans l'emplacement indiqué.

- Si on considère les tableaux Catégorie d'âge, Sexe, Durée de séjour à l'hôpital, Décès à l'arrivée. Que peut-on dire ?

Catégorie d'âge		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	45-54	258	21,8	21,8	21,8
	55-64	424	35,8	35,9	57,7
	65-74	343	29,0	29,0	86,7
	75+	157	13,3	13,3	100,0
	Total	1182	99,9	100,0	
Manquante	Système manquant	1	,1		
Total		1183	100,0		

Pour les statistiques, 21.8% des patients possèdent une tranche d'âge entre 45 et 54 ans, 35.8% avec une tranche d'âge entre 55 et 64, 29% entre 65 et 74, et enfin 13.3% qui ont plus de 75 ans, le pourcentage cumulé consiste à additionner le pourcentage courant et le pourcentage juste au dessus. Cette variable compte une seule valeur manquante.

Sexe		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Homme	592	50,0	50,1	50,1
	Femme	590	49,9	49,9	100,0
	Total	1182	99,9	100,0	
Manquante	Système manquant	1	,1		
Total		1183	100,0		

Pour la variable Sexe, 50% des patients sont des hommes et 49.9% des femmes avec une valeur manquante qui désigne un patient sans sexe qui est une anomalie à corriger.

Durée de séjour à l'hôpital		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	1	150	12,7	12,7	12,7
	2	302	25,5	25,6	38,3
	3	339	28,7	28,7	67,0
	4	209	17,7	17,7	84,7
	5	121	10,2	10,2	94,9
	6	60	5,1	5,1	100,0
	Total	1181	99,8	100,0	
Manquante	Système manquant	2	,2		
Total		1183	100,0		

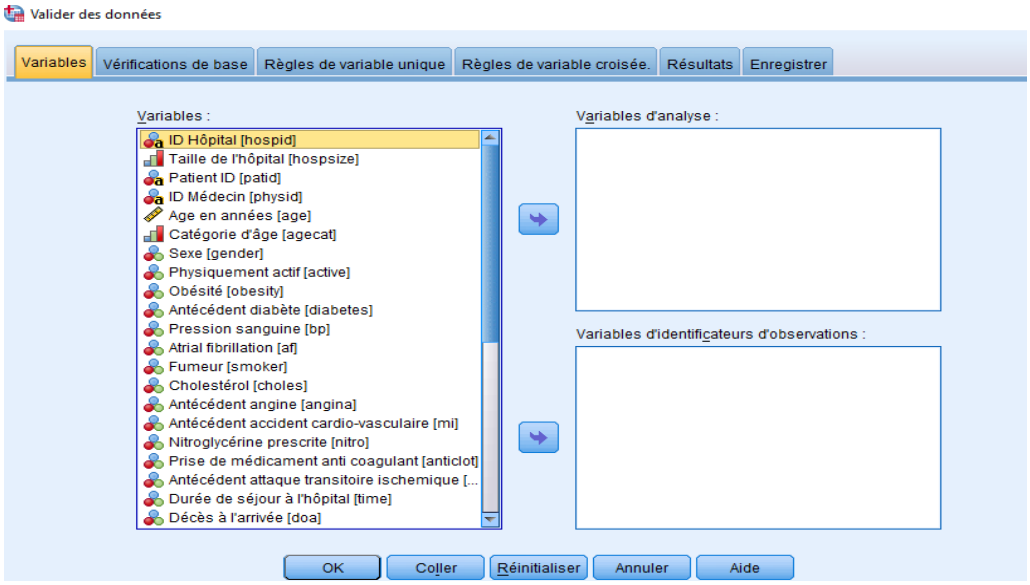
Pour le séjour à l'hôpital, il y a deux patients qui sont dans le fichier de données et quine se sont jamais rendus à l'hôpital ce qui est identifié par les 2 valeurs manquantes ce qui est incohérent car un patient doit se rendre au moins une fois à l'hôpital.

Décès à l'arrivée		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Non	1084	91,6	91,7	91,7
	Oui	98	8,3	8,3	100,0
	Total	1182	99,9	100,0	
Manquante	Système manquant	1	,1		
Total		1183	100,0		

La variable décès à l'arrivée indique si un patient est décédé à son arrivée à l'hôpital ou non donc ce serai illogique de trouver une valeur manquante car le patient soit meurt soit il sort de l'hôpital.

- Expliquer les erreurs

- On peut expliquer les erreurs en cliquant sur **Données > Validation > Valider les données > ....**



- On met les identifiants dans la zone en bas,

- On met les variables à analyser dans la zone en haut, on peut sélectionner toutes les variables restantes, on clique sur OK.

- On obtient les tableaux dans la fenêtre résultat,

Vérifications d'identificateur

Identificateurs incomplets

Observation	Identificateur		
	ID Hôpital	Patient ID	ID Médecin
288	OZN		125304
573		6137798782	790697
774		2322241867	176466

On remarque dans ce tableau que le patient qui se trouve dans la ligne 288 est affecté à un hôpital identifié par son ID (OZN) avec un médecin dont ID est (125304) mais le patient n'a pas d'identifiant ce qui est illogique.

Identificateurs en double

Groupe d'identificateurs en double	Nombre de doublons	Observations ayant des identificateurs en double	Identificateur		
			ID Hôpital	Patient ID	ID Médecin
1	2	10, 11	PBW	1406462419	355184
2	2	14, 15	PBW	2191527525	355184
3	2	21, 22	PBW	7237535360	616528
4	2	28, 29	NHV	4592215163	942982
5	2	30, 31	NHV	7628592330	371884
6	2	64, 65	NHV	0300750006	371884
7	2	83, 84	QWS	4590625286	215041
8	2	86, 87	QWS	6272818258	817329
9	2	96, 97	QWS	1959349605	215041
10	3	100, 101, 102	QWS	5856145337	817329
11	3	104, 105, 106	QWS	1543897849	817329
12	2	122, 123	QWS	9535631975	215041
13	2	144, 145	RLD	0052710039	560175
14	2	151, 152	RLD	5058356558	560175
15	2	156, 157	RLD	7779910241	695521
16	2	164, 165	OZN	2970608839	139142
17	2	168, 169	OZN	0165873576	125304

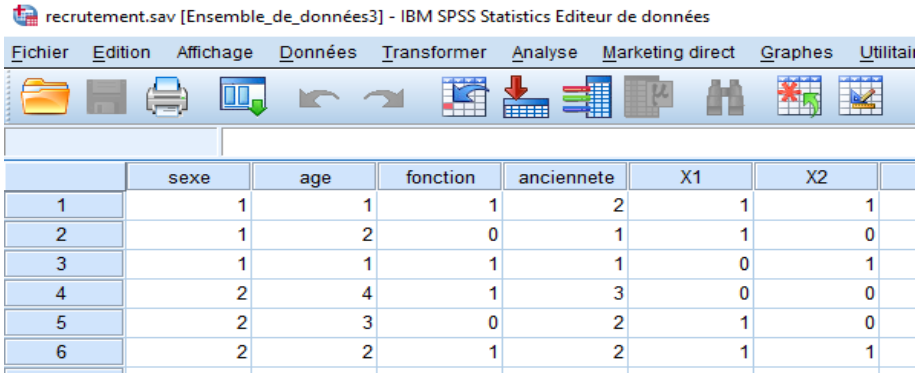
On remarque que le fichier de données contient des duplications par exemple les lignes 10 et 11 sont identiques et les lignes 100, 101, et 102, un patient doit figurer une seule fois dans le fichier de données.

IX.2 Détection de cas doubles

On veut créer le fichier SPSS recrutement.sav avec : Sexe={1 : homme, 2 :femme}, Fonction = {1 : enseignant ; 0 : Autre}, âge={1 :15-20 ;2 :20-30 ;3 :30-40 ;4 :plus de 40}, Ancienneté = {1 : <5ans ; 2 : 5-10 ans ; 3 : >10 ans}, X1 et X2 avec les valeurs {1 : Oui ; 0 : Non}

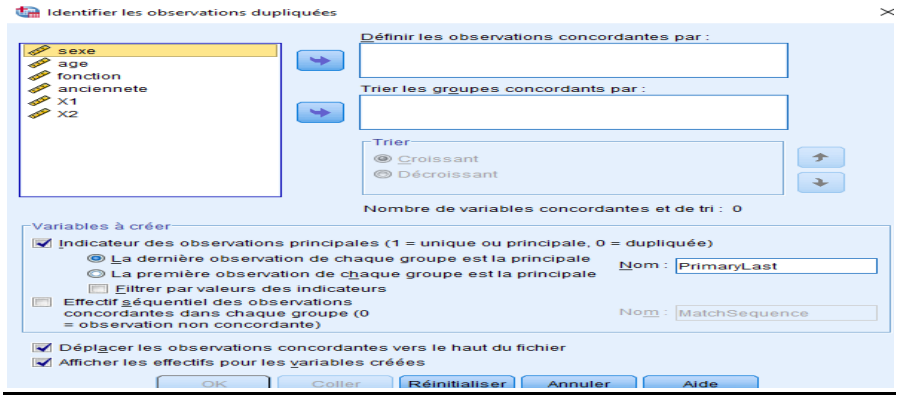
Sexe	Age	Fonction	Ancienneté	X1	X2
1	1	1	2	1	1
1	2	0	1	1	0
1	1	1	1	0	1
2	4	1	3	0	0
2	3	0	2	1	0
2	2	1	2	1	1

- On obtient le fichier qu'on peut visualiser dans la fenêtre Affichage de Données de SPSS,

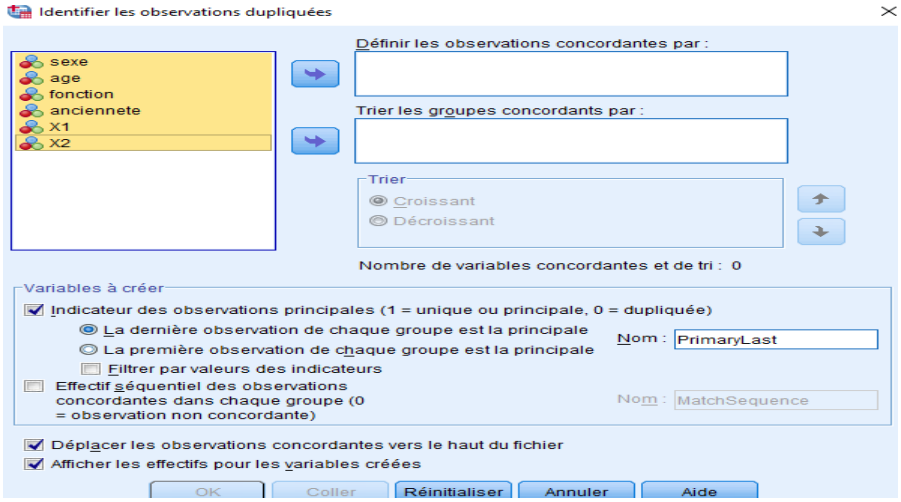


	sexe	age	fonction	ancienneté	X1	X2
1	1	1	1	2	1	1
2	1	2	0	1	1	0
3	1	1	1	1	0	1
4	2	4	1	3	0	0
5	2	3	0	2	1	0
6	2	2	1	2	1	1

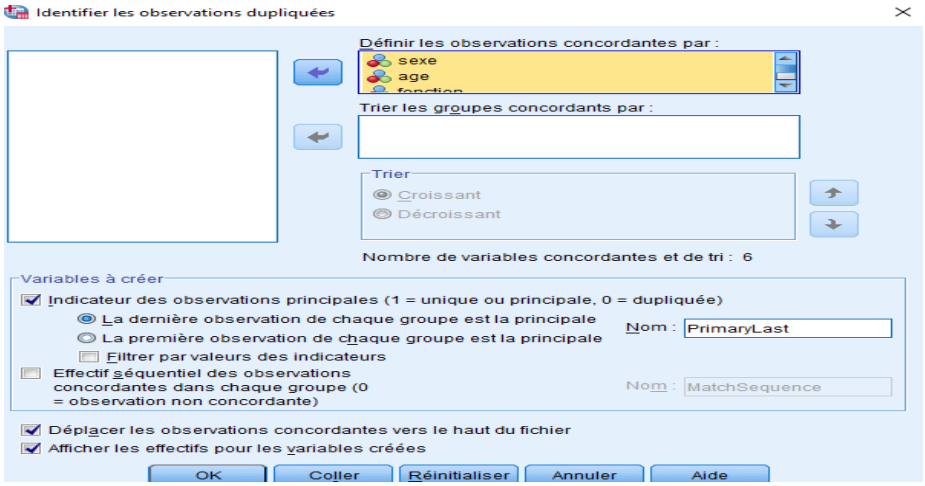
- Comment identifier les données dupliquées ?
  - On clique dur le menu **Données > identifier des observations dupliquées > …**



  - On sélectionne toutes les variables du côté gauche,



- On transfert les variables vers la zone en haut identifiée par « Définir les observations concordantes par : », et on clique dur OK,



- On obtient dans la fenêtre résultats le tableau ci-dessous,

Indicateur de chaque dernière observation concordante comme Principale

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Observation principale	6	100,0	100,0	100,0

Il est indiqué sur le tableau que le fichier de données est composé de 6 lignes qui sont désignées comme observations principales (ce qui indique qu'il n'y a pas de duplications).

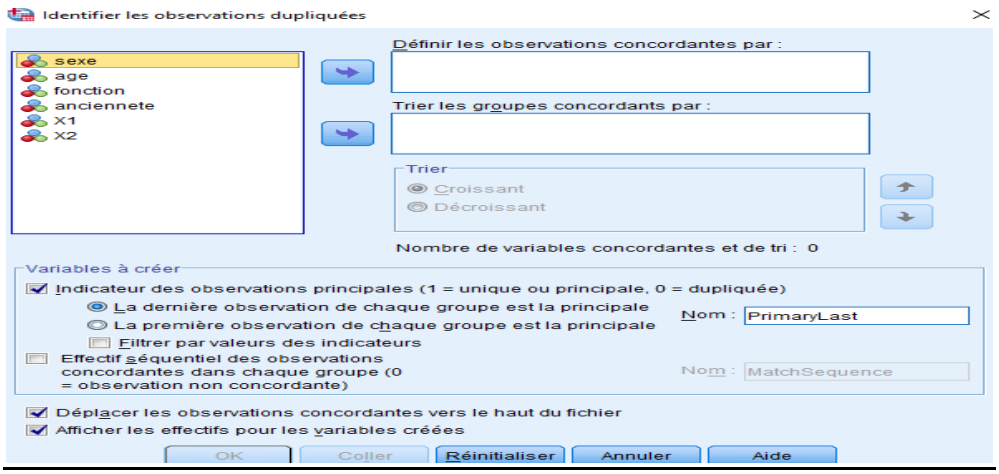
- Dans le fichier de données est ajoutée une variable PrimaryLast qui prend la valeur 1 si la ligne est principale et 0 si c'est une duplication ;

	sexe	age	fonction	anciennete	X1	X2	PrimaryLast
1	1	1	1	1	0	1	1
2	1	1	1	2	1	1	1
3	1	2	0	1	1	0	1
4	2	2	1	2	1	1	1
5	2	3	0	2	1	0	1
6	2	4	1	3	0	0	1

- Dupliquer les observations 2,3,4,5.
  - On restaure le fichier de données initial en supprimant la colonne PrimaryLast,
  - On duplique les lignes indiquées en les copiant et les collants,

	sexe	age	fonction	anciennete	X1	X2
1	1	1	1	1	0	1
2	1	1	1	2	1	1
3	1	2	0	1	1	0
4	2	2	1	2	1	1
5	2	3	0	2	1	0
6	2	4	1	3	0	0
7	1	1	1	2	1	1
8	1	2	0	1	1	0
9	2	2	1	2	1	1
10	2	3	0	2	1	0

- Identifier les observations dupliquées.
  - On cliquer sur **Données > identifier des observations dupliquées > …**



  - On transfert toutes les variables vers la zone en haut et on clique sur OK,

Indicateur de chaque dernière observation concordante comme Principale

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Observation en double	4	40,0	40,0	40,0
	Observation principale	6	60,0	60,0	100,0
	Total	10	100,0	100,0	

Le tableau obtenu indique que nous avons au total 10 observations (lignes) dont 6 sont principales et 4 sont des observations en double, en plus dans le fichier de données les

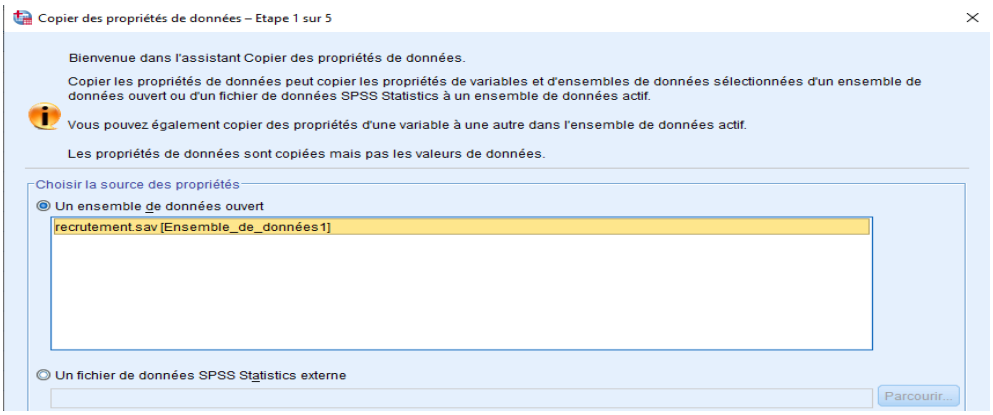
observations sont ordonnées en mettant chaque observation double (identifiée par 0) avant la principale lui correspondant (identifiée par 1).

	sexe	age	fonction	ancienneté	X1	X2	PrimaryLast
1	1	1	1	2	1	1	0
2	1	1	1	2	1	1	1
3	1	2	0	1	1	0	0
4	1	2	0	1	1	0	1
5	2	2	1	2	1	1	0
6	2	2	1	2	1	1	1
7	2	3	0	2	1	0	0
8	2	3	0	2	1	0	1
9	1	1	1	1	0	1	1
10	2	4	1	3	0	0	1

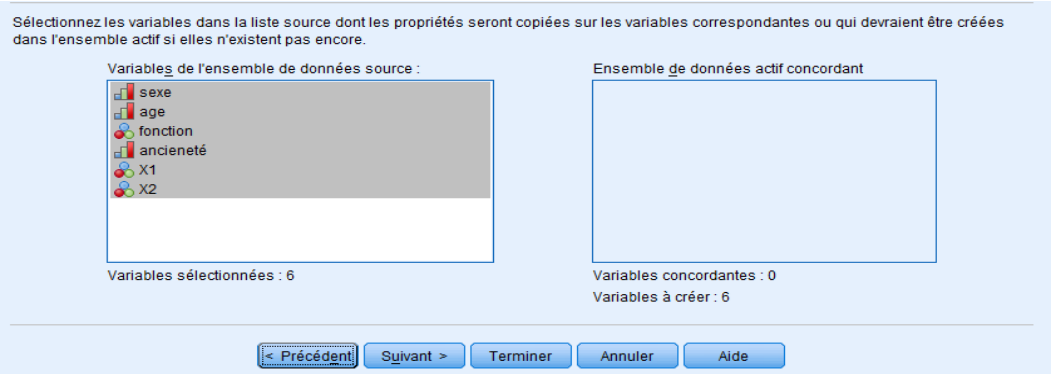
- On veut créer un autre fichier avec les mêmes variables et propriétés du fichier recrutement.sav qu'on va nommer recrutement2.sav en remplaçant les observations 3, 4, et 5 par :

2	1	1	1	0	1
1	4	1	3	0	0
2	4	0	2	1	0

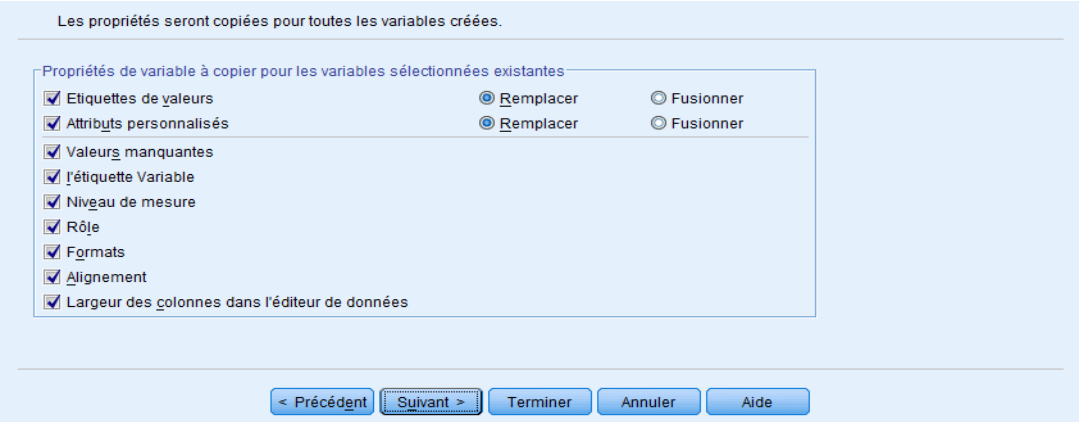
  - On clique sur **Fichier > Nouveau > Données > …**
  - On clique sur **Données > copier les propriétés des données > …** et on sélectionne le fichier de données ouvert « recrutement.sav » à partir duquel on veut copier directement le format (variables, types, valeurs, …),



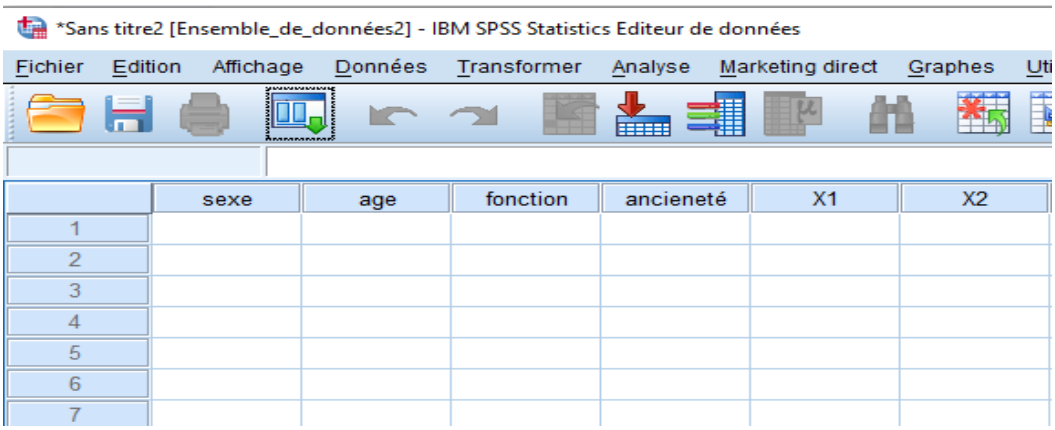
  - On clique sur suivant et on obtient la fenêtre dans laquelle on sélectionne les variables qu'on veut copier,



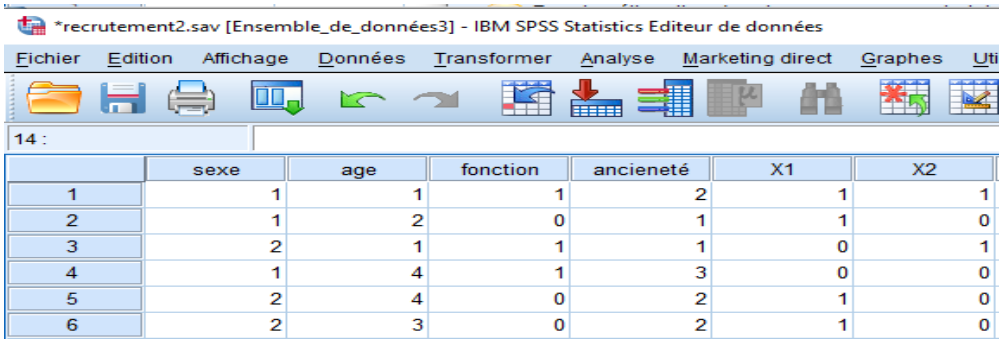
- On clique sur suivant,



- On clique sur Terminer, on obtient un fichier identique à recrutement.sav (sans les données juste la structure), cette opération permet de gagner du temps dans la création.

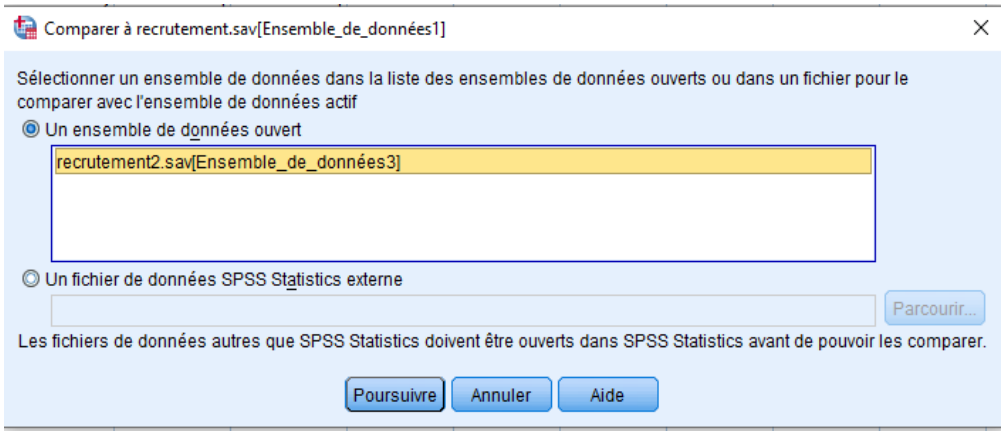


- Comparer les deux fichiers.
  - Le fichier recrutement2.sav se présente de la façon suivante,

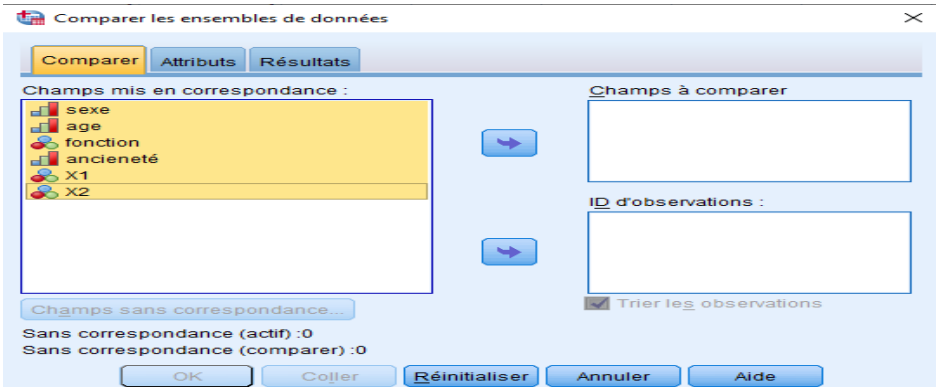


	sexe	age	fonction	ancieneté	X1	X2
1	1	1	1	2	1	1
2	1	2	0	1	1	0
3	2	1	1	1	0	1
4	1	4	1	3	0	0
5	2	4	0	2	1	0
6	2	3	0	2	1	0

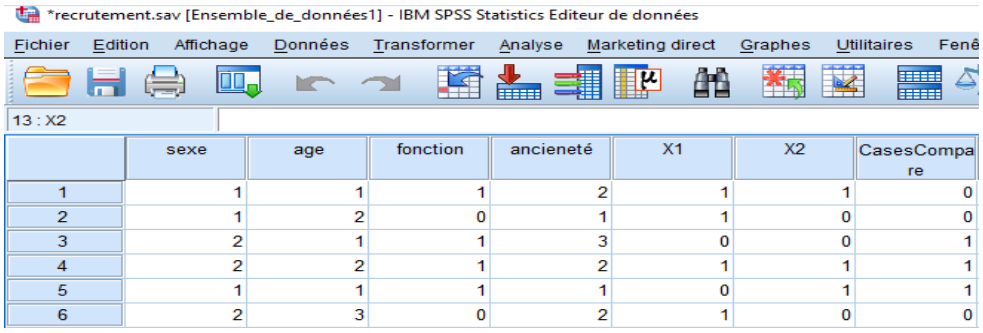
- On clique sur **Données > comparer les ensembles de données > …** on sélectionne le fichier qui sert de base de comparaison et on clique sur Poursuivre,



- On obtient la fenêtre « Comparer les ensembles de données », dans laquelle on sélectionne toutes les variables de la zone gauche et on les transferts dans la zone Champs à comparer,



- Au fichier de données recrutement.sav est ajouté le champ « Casescompare » qui prend la valeur 1 si la ligne est différente et 0 si la ligne est identique comparées aux données de recrutement 2.sav.



	sexe	age	fonction	ancieneté	X1	X2	CasesCompare
1	1	1	1	2	1	1	0
2	1	2	0	1	1	0	0
3	2	1	1	3	0	0	1
4	2	2	1	2	1	1	1
5	1	1	1	1	0	1	1
6	2	3	0	2	1	0	0

Chapitre X : Manipulation de fichiers syntaxe

On peut enregistrer et automatiser de nombreuses tâches courantes grâce au langage de commande. Il fournit également des fonctionnalités qui ne se trouvent ni dans les menus ni dans les boîtes de dialogue. La plupart des commandes sont accessibles depuis les menus et boîtes de dialogue.

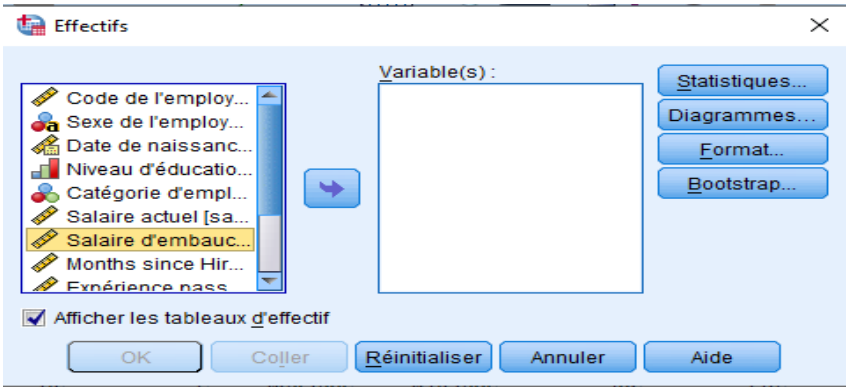
Certaines options et commandes ne sont disponibles qu'en utilisant le langage de commande. Le langage de commande permet également d'enregistrer les différents travaux dans un fichier de syntaxe afin de vous permettre de relancer l'analyse à une date ultérieure.

Un fichier de syntaxe est un fichier texte simple contenant des commandes de syntaxe IBM SPSS Statistics. On peut ouvrir une fenêtre de syntaxe et y entrer des commandes directement. Pour illustration, nous allons utiliser le fichier de données **employees data.sav**.

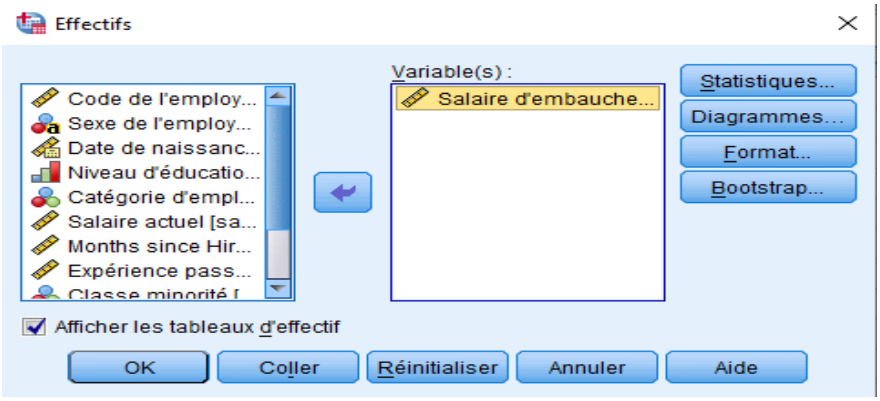
X.1 Utilisation d'une syntaxe (Plaisent M. et al, 2009)

On va tout d'abord exécuter un ensemble d'actions sur le fichier **employees data.sav**.

- On clique sur **Fichier > Ouvrir > Données > …**
- On sélectionne **Analyse > Statistiques descriptives > Effectifs...**

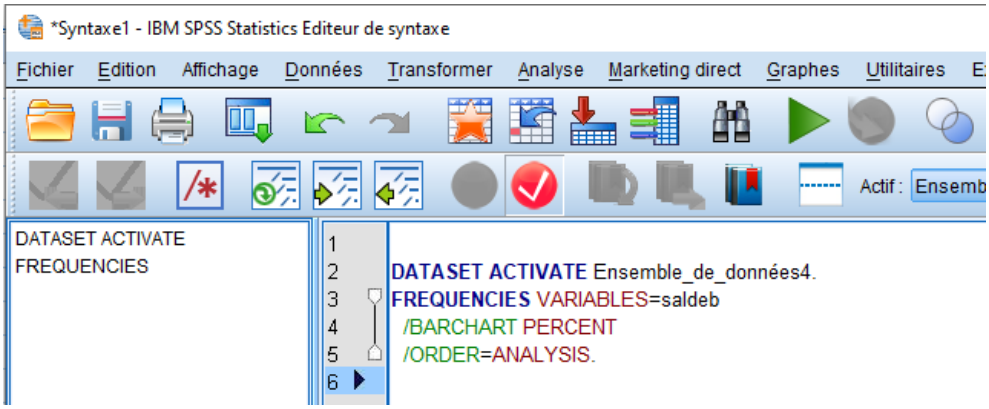


- Sélectionnez la variable Salaire d'embauche et déplacez-la vers la liste Variable(s),



- Cliquer sur diagrammes et sélectionner diagrammes en bâtons, pour valeurs du diagramme choisir Pourcentage, ensuite cliquer sur poursuivre,

- Cliquez ensuite sur Coller pour copier la syntaxe créée grâce aux sélections effectuées dans la boîte de dialogue de l'Editeur de syntaxe.



- Le fichier syntaxe peut être exécuté en choisissant le menu **Exécuter > Sélection > …**

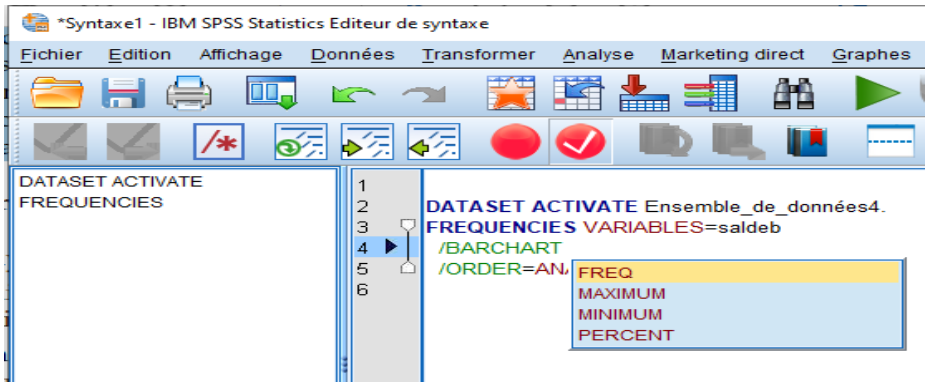
X.2 Modification d'une syntaxe

Dans la fenêtre de syntaxe, on peut toujours modifier la syntaxe. Par exemple, on peut modifier la sous-commande /BARCHART pour afficher des effectifs à la place des pourcentages.

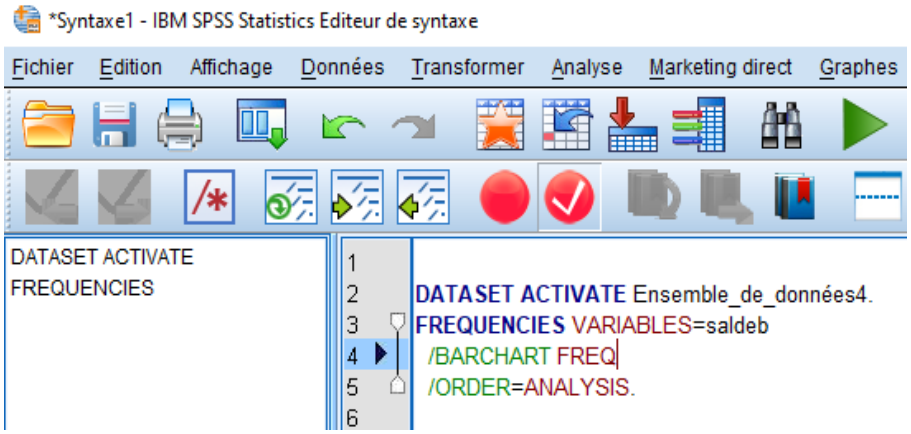
On peut obtenir une liste des mots clés disponibles pour la sous-commande en positionnant le curseur n'importe où après le nom de la sous-commande et en appuyant simultanément sur Ctrl+barre d'espace. Ceci affiche le contrôle de saisie semi-automatique pour la sous-commande.

Pour modifier le pourcentage en fréquences on suit les étapes suivantes :

- Supprimer le mot clé PERCENT de la sous-commande BARCHART,
- Appuyer sur Ctrl+ barre d'espace,



- Cliquer sur l'élément libellé FREQ pour les effectifs. En cliquant sur un élément dans le contrôle de saisie semi-automatique, il sera inséré à l'emplacement actuel du curseur.



X.3 Ouverture et exécution d'un fichier syntaxe

- Pour ouvrir un fichier de syntaxe enregistré, à partir du menu, on sélectionne **Fichier > Ouvrir > Syntaxe...** Une boîte de dialogue standard d'ouverture de fichiers apparaît,
- Sélectionner un fichier de syntaxe. Si aucun fichier de syntaxe n'apparaît, assurez-vous que l'option Syntaxe (*.sps) est sélectionnée en tant que type de fichier à afficher,
- Cliquez sur Ouvrir,
- Utiliser le menu Exécuter de la fenêtre de syntaxe pour exécuter les commandes.

Si les commandes s'appliquent à un fichier de données particulier, on doit d'abord ouvrir ce dernier avant d'exécuter les commandes ou inclure une commande qui ouvre le fichier de données. On peut coller ce type de commande à partir des boîtes de dialogue permettant d'ouvrir les fichiers de données.

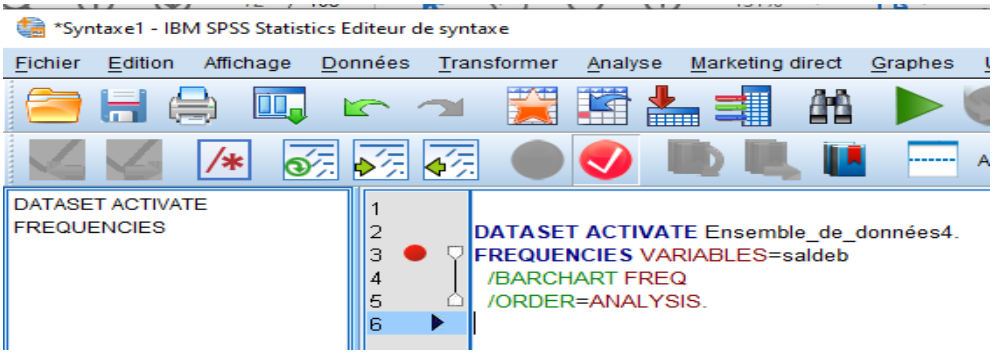
X.4 Utilisation des points de rupture

Les points de rupture permettent d'interrompre l'exécution d'une syntaxe de commande à des points spécifiés à l'intérieur de la syntaxe et de poursuivre l'exécution quand on le veut.

Ceci permet de visualiser les sorties ou les données à un moment intermédiaire du travail de syntaxe, ou d'exécuter la syntaxe de commande affichant les informations sur l'état actuel des données, telles que FREQUENCIES. Les points de rupture ne peuvent être réglés qu'au niveau d'une commande, et non sur les lignes spécifiques au sein d'une commande.

Pour insérer un point de rupture dans une commande on suit les étapes suivantes :

- Cliquer n'importe où dans la zone à gauche du texte associé à la commande,
- Le point de rupture est représenté par un cercle rouge dans la zone à gauche du texte de la commande et sur la même ligne que le nom de la commande, peu importe où on clique.



Lorsqu'on exécute une commande contenant des points de rupture, l'exécution s'interrompt avant chaque commande contenant un point de rupture.

La flèche pointant vers le bas à gauche du texte de commande présente la progression de l'exécution de la syntaxe. Elle couvre la zone s'étendant de la première exécution de commande à la dernière exécution de commande.

Pour reprendre l'exécution après un point de rupture, on sélectionne dans le menu de la fenêtre éditeur de syntaxe **Exécuter > Poursuivre > …**

Chapitre XI : Statistiques récapitulatives pour chaque mesure de variable

Différentes mesures sont adaptées pour les différentes variables, ces mesures peuvent être :

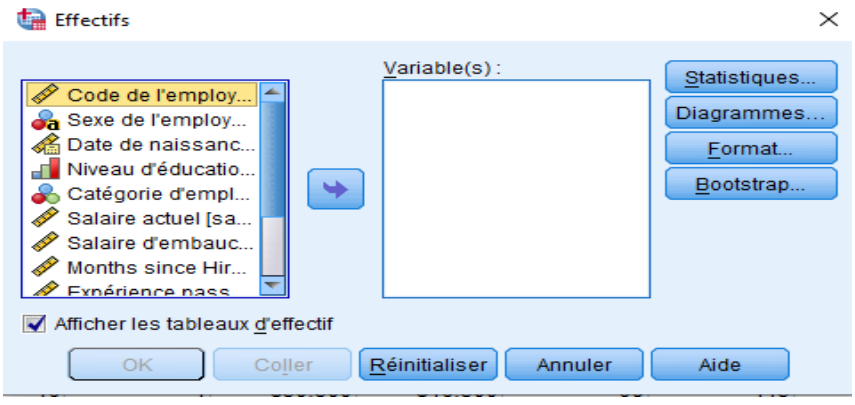
- Qualitatives : Données ayant un nombre limité de valeurs ou de catégories distinctes par exemple, sexe ou situation de famille. Elles sont parfois également qualifiées de variables catégorielles. Ces variables peuvent être des données chaîne (alphanumérique) ou des variables numériques qui utilisent des codes chiffrés pour représenter les catégories par exemple, 0 = Célibataire et 1 = Marié. Il existe deux types essentiels de données catégorielles :
  - Nominal : les valeurs n'ont aucun ordre. Par exemple, une catégorie d'emploi de type ventes n'est pas supérieure ou inférieure à une catégorie d'emploi de type marketing ou étude.
  - Ordinal : les valeurs possèdent un ordre significatif, mais pour lesquelles il n'existe aucune distance mesurable entre les catégories. Par exemple, les valeurs élevées, moyenne et faible, mais il est impossible de calculer la "distance" entre ces valeurs.
- Echelles : Données mesurées sur une échelle d'intervalle ou de rapport, où les valeurs de données indiquent à la fois l'ordre des valeurs et la distance qui les sépare.
  Par exemple, un salaire de 58 160 € est supérieur à un salaire de 42 212 € et la distance entre les deux valeurs est de 15 948 €. Ces données sont aussi appelées données quantitatives ou données continues.

XI.1 Statistiques pour variables catégorielles (

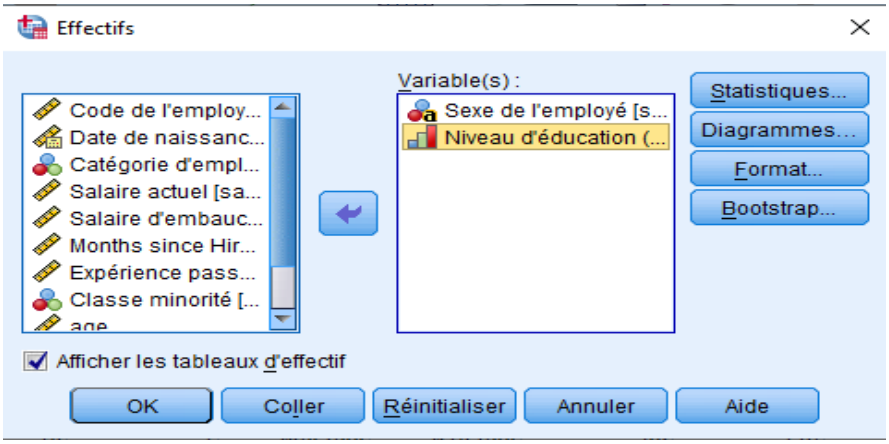
Pour les données catégorielles, la mesure récapitulative la plus courante est le nombre ou le pourcentage d'observations dans chaque catégorie. Pour les données ordinales, la médiane (valeur au-dessus ou au-dessous de laquelle se trouve la moitié des observations) peut également être une mesure récapitulative utile s'il existe un grand nombre de catégories.

La procédure Fréquences produit des tables de fréquences qui affichent le nombre et le pourcentage d'observations pour chaque valeur observée d'une variable.

- On sélectionne **Analyse > Statistiques descriptives > Effectifs...**



- Choisir les variables Sexe de l'employer et Niveau d'éducation et les déplacer vers la liste Variable(s) ; et on clique sur OK,



- On obtient les tableaux de fréquences suivants :

Tableau de fréquences

Sexe de l'employé

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Féminin	216	45,6	45,6	45,6
	Masculin	258	54,4	54,4	100,0
	Total	474	100,0	100,0	

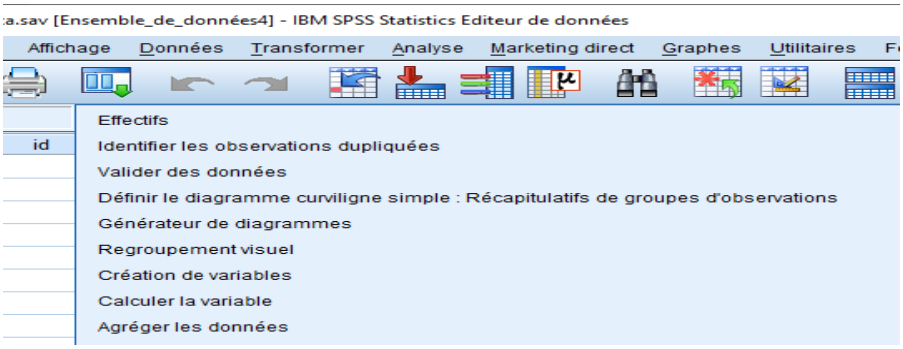
Niveau d'éducation (années)

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	8	53	11,2	11,2	11,2
	12	190	40,1	40,1	51,3
	14	6	1,3	1,3	52,5
	15	116	24,5	24,5	77,0
	16	59	12,4	12,4	89,5
	17	11	2,3	2,3	91,8
	18	9	1,9	1,9	93,7
	19	27	5,7	5,7	99,4
	20	2	,4	,4	99,8
	21	1	,2	,2	100,0
	Total	474	100,0	100,0	

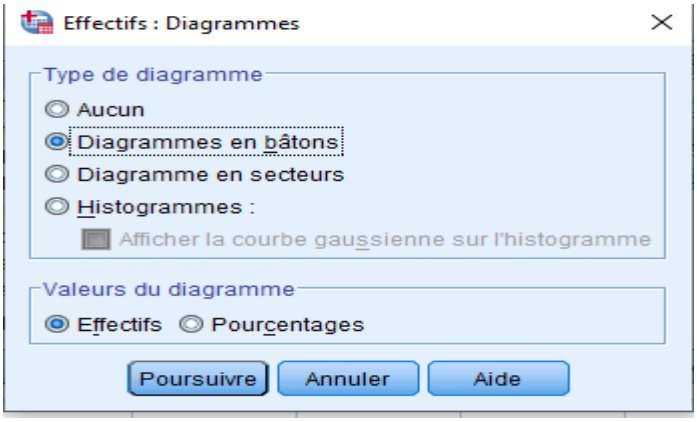
XI.2 Graphiques pour données catégorielles

Il est possible d'afficher graphiquement les informations d'une table de fréquences avec un graphique à barres ou un graphique circulaire.

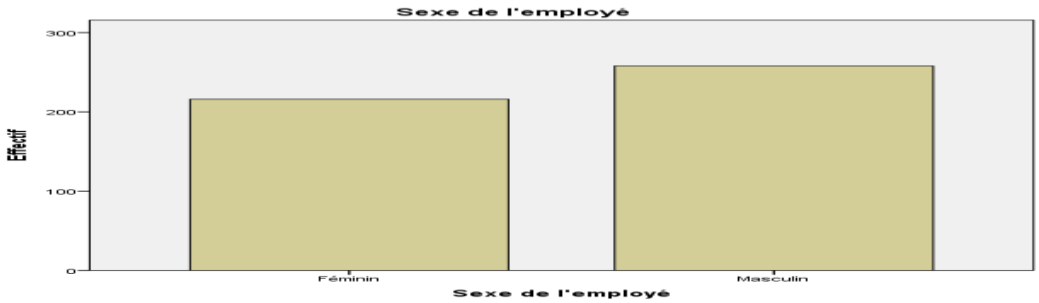
- Ouvrir la boîte de dialogue Fréquences. (Les deux variables doivent toujours être sélectionnées.)
- On peut utiliser le bouton Rappeler boîte de dialogue de la barre d'outils pour revenir rapidement aux dernières procédures utilisées.



- Cliquer sur Diagrammes, et choisir Diagramme en bâtons,



- Cliquer sur Poursuivre, ensuite sur OK, on obtient graphes et tableaux sur la feuille résultats.



XI.3 Statistiques pour variables d'échelle

Plusieurs mesures statistiques sont disponibles pour les variables échelle :

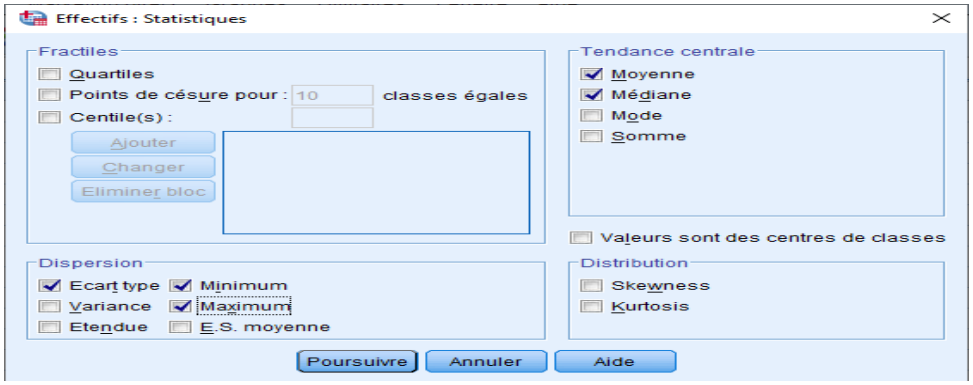
- Mesure de la tendance centrale : ces mesures sont la moyenne (moyenne arithmétique) et la médiane (valeur au dessus ou au dessous de laquelle se trouve la moitié des observations).
- Mesure de dispersion : mesure la quantité de variation ou de dispersion dans les données ; on peut citer par exemple, écart type, maximal, et minimal.

Pour ces variables les tableaux de fréquences ne sont pas d'un grand intérêt car il peut exister autant de valeurs distinctes que d'observations.

Pour afficher les mesures de dispersion et de tendance centrale, on suit les étapes suivantes :

- Ouvrir le fichier employees.sav,
- Choisir le menu **Analyse > Statistiques descriptives > Effectifs...**
- Cliquer sur réinitialiser pour effacer les paramètres précédents,
- Sélectionner une variable échelle dans l'ensemble des variables (par exemple la variable âge) et décocher l'option « Afficher les tableaux d'effectifs »,

- Cliquer sur le bouton « statistiques… » et choisir les mesures qu'on souhaite afficher,



- On obtient le tableau des mesures sur la fenêtre résultats,

Statistiques

age

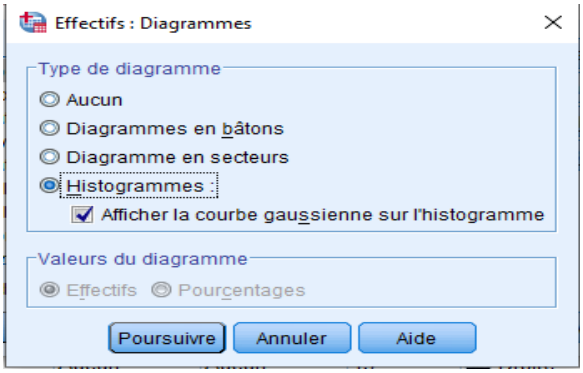
N	Valide	473
	Manquante	1
Moyenne		66,6723
Médiane		61,0000
Ecart-type		11,78409
Minimum		52,00
Maximum		94,00

XI.4 Histogrammes pour variables d'échelle

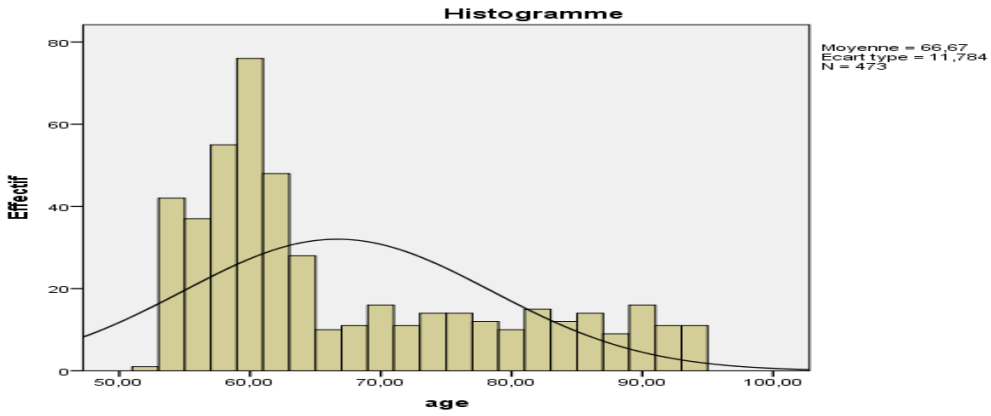
Pour construire un histogramme pour une variable d'échelle, on suit les étapes suivantes :

- Ouvrir le fichier employees.sav,
- Choisir le menu **Analyse > Statistiques descriptives > Effectifs...**
- Sélectionner la variable Echelle « âge » et cliquer sur Diagramme,

- La fenêtre ci-dessous s'affiche, dans laquelle on sélectionne Histogrammes et on coche l'option « Afficher la courbe gaussienne sur l'histogramme ».



- Cliquer sur Poursuivre ensuite sur OK, sur la fenêtre Résultats s'affiche le graphe correspondant à la variable âge.



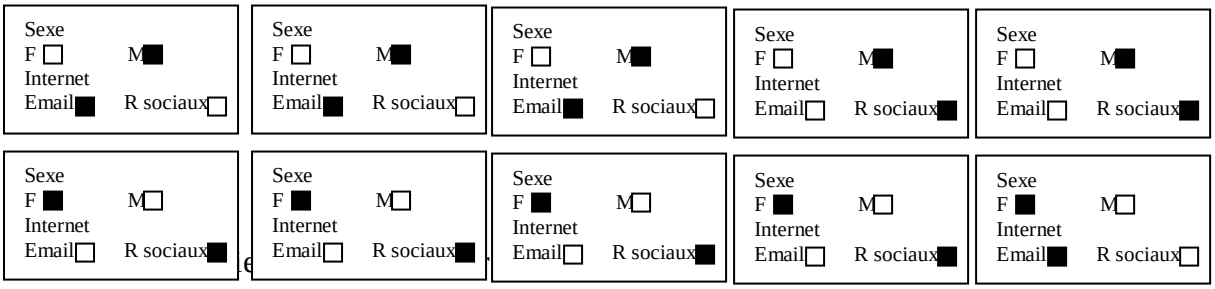
Chapitre XII : Questionnaires et pondération

XII.1 Questionnaire, représentation sous SPSS

Un questionnaire est une série de questions méthodiquement posées afin de définir un cas, une situation, une demande parmi un grand nombre de possibilités. Un questionnaire peut être l'objet d'un formulaire, ou d'un jeu de menus informatiques. On peut le considérer comme modèle d'un parcours administratif.

Les questionnaires sont aussi des outils de recherche pour les sciences humaines et sociales, en particulier la psychologie, la sociologie, le marketing et la géographie.

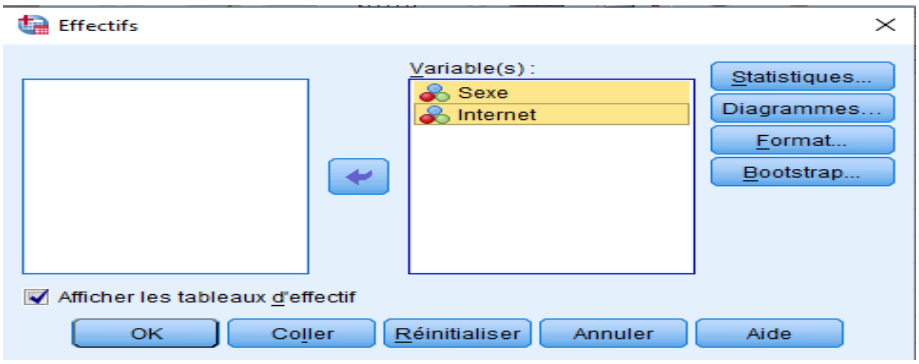
- On considère les 10 questionnaires correspondant à l'utilisation d'internet, composé de deux questions, Sexe du répondant et la manière avec laquelle il utilise internet soit par réseaux sociaux ou par email, on dispose de la configuration de réponse suivante :



- Le fichier SPSS est composé de deux variables Sexe avec les valeurs {Femme=0 et Male=1} et Internet avec les valeurs {Email=0 et R sociaux=1} et 10 observations (une case noire correspond à une réponse).

	Sexe	Internet
1	Male	Email
2	Male	Email
3	Male	Email
4	Male	R Sociaux
5	Male	R Sociaux
6	Femme	R Sociaux
7	Femme	R Sociaux
8	Femme	R Sociaux
9	Femme	R Sociaux
10	Femme	Email

- Générer le tableau des fréquences.
  - On clique sur **Analyse→ statistique descriptive→effectifs…**,



- On obtient les tableaux de fréquences suivants pour chacune des variables,

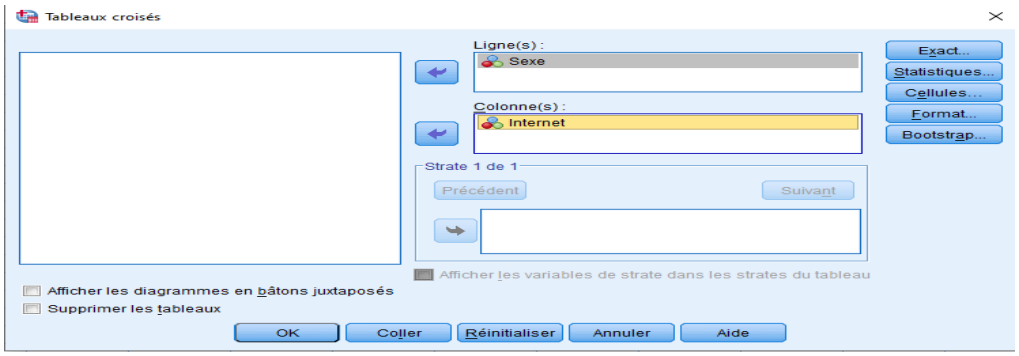
Sexe

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Femme	5	50,0	50,0	50,0
	Male	5	50,0	50,0	100,0
	Total	10	100,0	100,0	

Internet

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Email	4	40,0	40,0	40,0
	R Sociaux	6	60,0	60,0	100,0
	Total	10	100,0	100,0	

- Générer le tableau croisé,
  - On clique sur Analyse→ statistique descriptive→tableaux croisés,
  - On choisit une des variables pour représenter les lignes et la seconde pour les colonnes (peut importe le choix, cela n'influence pas l'analyse),



- On obtient le tableau suivant,

Récapitulatif du traitement des observations

	Observations					
	Valide		Manquante		Total	
	N	Pourcent	N	Pourcent	N	Pourcent
Sexe * Internet	10	100,0%	0	0,0%	10	100,0%

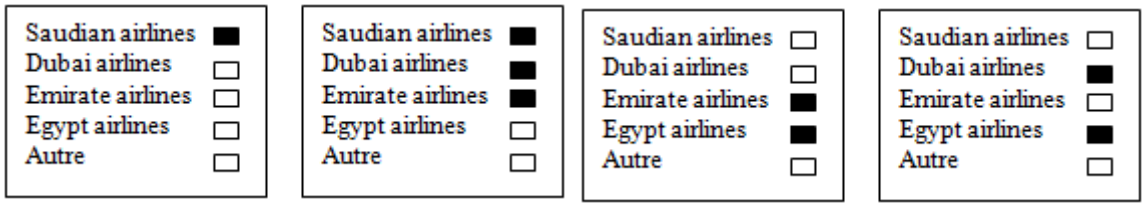
Tableau croisé Sexe * Internet

Effectif

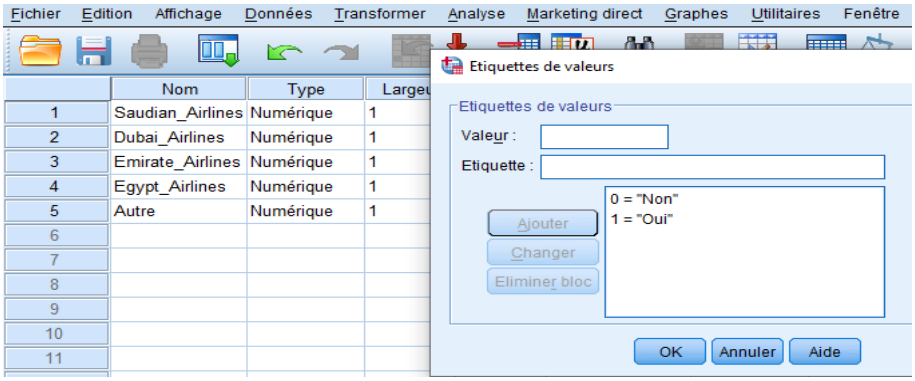
		Internet		Total
		Email	R Sociaux	
Sexe	Femme	1	4	5
	Male	3	2	5
Total		4	6	10

On dispose de la question choix multiples suivante : Quelles sont les lignes aériennes que vous avez utilisées l'année dernière ? Les réponses associées appartiennent à l'ensemble suivant : {Saudian airlines, Dubai airlines, Emirate airlines, Egypt airlines, Autre}. Construire le fichier vols.sav.

- Remplir le fichier vols.sav questionnaire par les 4 cas suivants :



- On a 4 répondants donc 4 observations, et 5 réponses possibles pour une question avec des réponses pouvant être combinées, alors chaque réponse correspond à une variable (au total on aura 5 variables, qui prennent les valeurs 0 dans le cas ou la case n'est pas cochée et 1 dans le cas contraire).



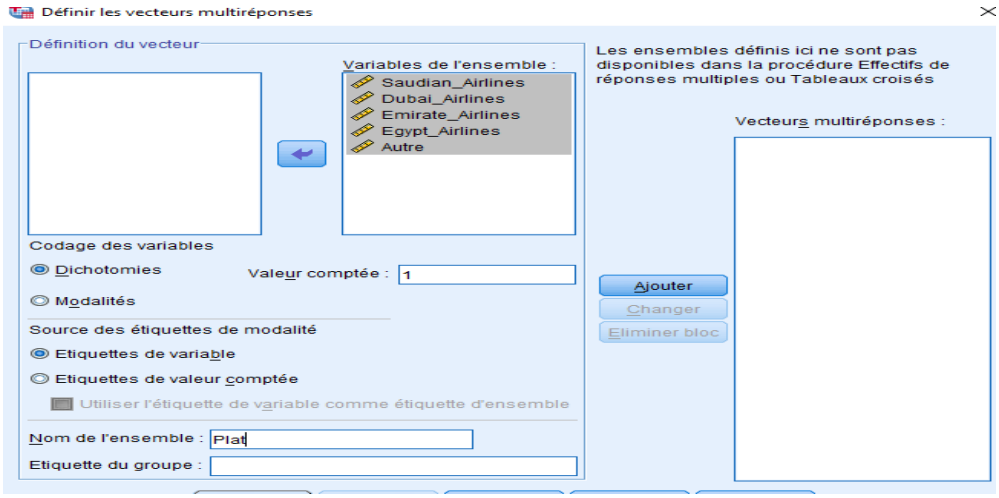
- Au final, le fichier résultat se présente comme suit :

	Saudian_Airlines	Dubai_Airlines	Emirate_Airlines	Egypt_Airlines	Autre
1	1	0	0	0	0
2	1	1	1	0	0
3	0	0	1	1	0
4	0	1	0	1	0
5					

- Définir pour SPSS que le fichier est multi réponses.

Pour définir une question multi réponses, on suit les étapes suivantes :

- On clique sur **Données > Définir vecteur multiréponses > …**,



- On déplace toutes les variables vers la zone « variables de l'ensemble », on attribut à valeur comptée la valeur 1 et on donne un nom à l'ensemble tout en bas ensuite on clique sur Ajouter,

Vecteurs de réponses multiples

Nom	Codé comme	Valeur comptée	Type de données	Variables élémentaires
$Plat	Dichotomies	1	Numérique	Saudian_Airlines Dubai_Airlines Emirate_Airlines Egypt_Airlines Autre

L'intérêt de définir un vecteur multi réponses c'est de considérer les réponses comme appartenant à une même question et ceci pour ne pas fausser l'analyse.

XII.2 Pondération

On utilise à la place du tableau précédent, le tableau ci-dessous qui représente les fréquences, saisir le tableau tel quel dans SPSS :

Sexe	Internet	fréquences
M	email	3
M	R sociaux	2
F	email	1
F	R sociaux	4

	Sexe	Internet	Frequences
1	M	email	3
2	M	R sociau	2
3	F	email	1
4	F	R sociau	4

- Est ce que le tableau est correct ?

Pour vérifier la justesse des données :

- On clique sur **Analyse→ statistique descriptive→tableaux croisés…**

Récapitulatif du traitement des observations

	Observations					
	Valide		Manquante		Total	
	N	Pourcent	N	Pourcent	N	Pourcent
Sexe * Internet	4	100,0%	0	0,0%	4	100,0%

Tableau croisé Sexe * Internet

Effectif

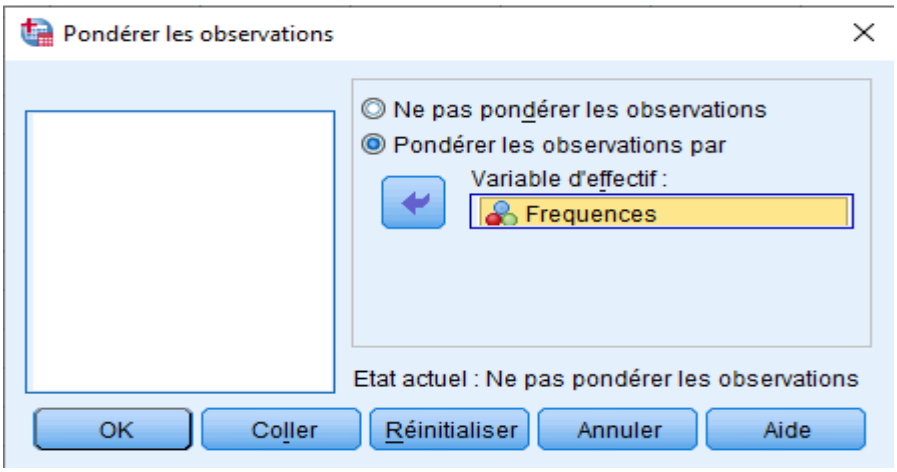
		Internet		
		email	R sociau	Total
Sexe	F	1	1	2
	M	1	1	2
Total		2	2	4

Selon l'analyse avec le tableau croisé, on remarque que le nombre de répondant de sexe féminin est 2 ainsi que le nombre de répondants de sexe masculin. Le nombre total de répondants est de 4. Alors que ce n'est pas le cas avec le tableau de départ qui indique 10 répondants au total répartis en 5 femmes et 5 hommes.

- Comment corriger l'erreur ?

On corrige l'erreur en utilisant la variable fréquence comme critère de pondération.

- On clique sur le bouton  de la barre d'outils ou on choisit le menu **Données > Pondérer les observations > …**, on sélectionne « Pondérer les observations par » et on choisit Frequences comme critère de pondération.



- Pour vérifier la justesse des résultats on clique sur **Analyse→ statistique descriptive→tableaux croisés…**,

Tableau croisé Sexe * Internet

Effectif

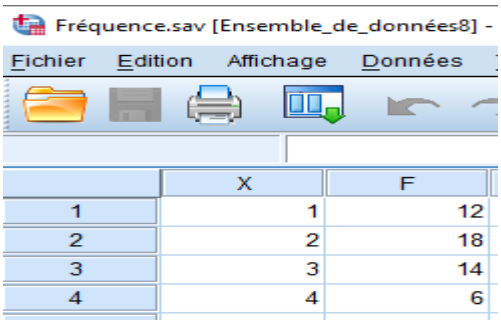
		Internet		
		email	R sociau	Total
Sexe	F	1	4	5
	M	3	2	5
Total		4	6	10

On remarque que les résultats sont plus cohérents avec le tableau de départ.

On considère le tableau suivant qui représente des numéros de questions avec des fréquences d'apparition associées à chaque question :

X	F
1	12
2	18
3	14
4	6

- Créer un fichier frequence.sav,



- Vérifier la justesse du fichier crée,
  - Cliquer sur **Analyse→statistique descriptive→effectifs**…, on choit X pour pouvoir analyser ses valeurs,

Statistiques

X

N	Valide	4
	Manquante	0

X

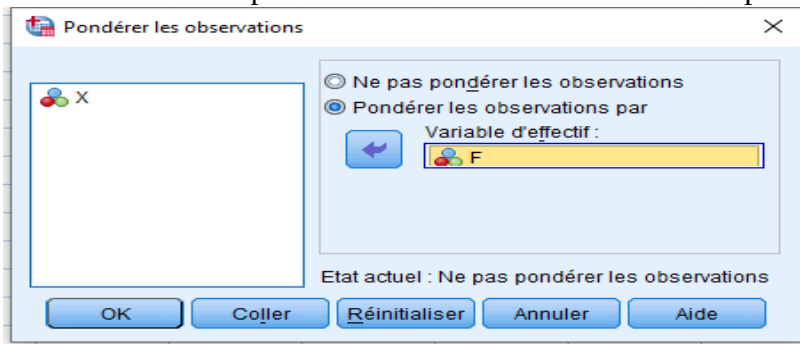
		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	1	1	25,0	25,0	25,0
	2	1	25,0	25,0	50,0
	3	1	25,0	25,0	75,0
	4	1	25,0	25,0	100,0
	Total	4	100,0	100,0	

On remarque que le vrai contenu du tableau de données initial n'est pas reflété sur l'analyse.

- Comment corriger l'erreur ?

On réalise la pondération de X par F,

- On clique sur le menu **Données > Pondérer les observations > …**, on sélectionne « Pondérer les observations par » et on choisit F comme critère de pondération,



Pour vérifier la cohérence de des résultats avec le tableau de départ, on clique sur **Analyse→statistique descriptive→effectifs**…,

Statistiques

X

N	Valide	50
	Manquante	0

X

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	1	12	24,0	24,0	24,0
	2	18	36,0	36,0	60,0
	3	14	28,0	28,0	88,0
	4	6	12,0	12,0	100,0
	Total	50	100,0	100,0	

Références

CARICANO (M.) and POUJOL (F.) [2009], « Analyse de données avec SPSS », ISBN : 978-2-7440-4075-7, Pearson Education France, 2009.

KAMBOU (H.K.) [2021], « Manuel d'initiation au traitement de données sous SPSS », Expertise France, AFRISTAT, 2021.

CUSSON (F.) and CORNEAU (M.) [2010], « Guide d'introduction au logiciel SPSS », Guide élaboré pour les étudiants du cours CRI-1600 G : Initiation aux méthodes quantitatives, Université de Montréal, 2010.

JALBY (V.) [2015], « Introduction à SPSS statistics 22 », Faculté de Droit et de Sciences Économiques, Université de Limoges, 2015.

PLAISENT (M.) and BERNARD (P.) and DAGHFOUS (N.) and FAVREAU (S.) [2009], « Introduction à l'analyse des données de sondage avec SPSS », Presses de l'Université du Québec, 2009.

BACCINI (A.) [2010], « Statistique Descriptive Multidimensionnelle (pour les nuls) », Publications de l'institut de Mathématiques de Toulouse, 2010.

TILLE (Y.) [2023], « Résumé du Cours de Statistique Descriptive avec une Introduction au Calcul de Probabilités », Statistique Descriptive. DEUG. Statistique Descriptive, Suisse, pp.226. ffhal-04132484, 2023.

GILLES (I.) and G.T.GREEN (P.) and RICCIARDI JOOS (P.) and SCHEIDEGGER (R.) and STORARI (C.) and TUESCHER (T.) and WAGNER EGGER (P.) [2008], « Fascicule SPSS », Institut de Mathématiques Appliquées Faculté des S.S.P, Université de Lausanne, 2008.

Annexe

Il existe des fichiers d'exemple installés avec le produit figurent dans le sous-répertoire Echantillons du répertoire d'installation et un dossier distinct au sein du sous-répertoire Echantillons pour chacune des langues suivantes : anglais, français, allemand, italien, japonais, coréen, polonais, russe, chinois simplifié, espagnol et chinois.

Voici la description de quelques fichiers exemple déjà élaborés pour réaliser des tests :

- advert.sav : Ce fichier de données d'hypothèse concerne les mesures qu'entreprend un détaillant pour examiner la relation existant entre l'argent dépensé dans la publicité et les ventes résultantes. Pour ce faire, il collecte les chiffres des ventes passées et les coûts associés à la publicité.
- aflatoxin.sav : Ce fichier de données d'hypothèse concerne le test de l'aflatoxine dans des récoltes de maïs. La concentration de ce poison varie largement d'une récolte à l'autre et au sein de chaque récolte. Un processeur de grain a reçu 16 échantillons issus de 8 récoltes de maïs et a mesuré les niveaux d'aflatoxine en parties par milliard (PPB).
- anorectic.sav : En cherchant à développer une symptomatologie standardisée du comportement anorexique/boulimique, des chercheurs1 ont examiné 55 adolescents souffrant de troubles alimentaires. Chaque patient a été observé quatre fois sur une période de quatre années, soit un total de 220 observations. A chaque observation, les patients ont été notés pour chacun des 16 symptômes. En raison de l'absence de scores de symptôme pour le patient 71/visite 2, le patient 76/visite 2 et le patient 47/visite 3, le nombre d'observations valides est de 217.
- bankloan.sav : Ce fichier de données d'hypothèse concerne les mesures qu'entreprend une banque pour réduire le taux de défaut de paiement. Il contient des informations financières et démographiques sur 850 clients existants et éventuels. Les premières 700 observations concernent des clients auxquels des prêts ont été octroyés. Les 150 dernières observations correspondant aux clients éventuels que la banque doit classer comme bons ou mauvais risques de crédit.
- bankloan_binning.sav : Ce fichier de données d'hypothèse concerne des informations financières et démographiques sur 5 000 clients existants.
- breakfast.sav : Au cours d'une étude classique 3, on a demandé à 21 étudiants en MBA (Master of Business Administration) de l'école de Wharton et à leurs conjoints de classer 15 aliments du petit-déjeuner selon leurs préférences, de 1= « aliment préféré » à 15= « aliment le moins apprécié ». Leurs préférences ont été enregistrées dans six scénarios différents, allant de "Préférence générale" à "En-cas avec boisson uniquement".
- breakfast-overall.sav : Ce fichier de données contient les préférences de petit-déjeuner du premier scénario uniquement, "Préférence générale".

- broadband_1.sav : Ce fichier de données d'hypothèse concerne le nombre d'abonnés, par région, à un service haut débit. Le fichier de données contient le nombre d'abonnés mensuels de 85 régions sur une période de quatre ans.
- broadband_2.sav : Ce fichier de données est identique au fichier broadband_1.sav mais contient les données relatives à trois mois supplémentaires.
- car_insurance_claims.sav : Il s'agit d'un jeu de données présenté et analysé ailleurs 4 qui concerne des actions en indemnisation pour des voitures. Le montant d'action en indemnisation moyen peut être modelé comme présentant une distribution gamma, à l'aide d'une fonction de lien inverse pour associer la moyenne de la variable dépendante à une combinaison linéaire de l'âge de l'assuré, du type de véhicule et de l'âge du véhicule. Le nombre d'actions entreprises peut être utilisé comme pondération de positionnement.
- car_sales.sav : Ce fichier de données contient des estimations de ventes hypothétiques, des barèmes de prix et des spécifications physiques concernant divers modèles et marques de véhicule. Les barèmes de prix et les spécifications physiques proviennent tour à tour de edmunds.com et des sites des constructeurs.
- car_sales_uprepared.sav : Il s'agit d'une version modifiée de car_sales.sav qui n'inclut aucune version transformée des champs.