

Introduction à l'analyse de données à l'aide du logiciel SPSS

SUPPORT ÉCRIT POUR LES TRAVAUX
PRATIQUES DU COURS LPOLS 1221

LPOLS1221 : Introduction à l'analyse de données à l'aide du logiciel SPSS.

<u>AVANT-PROPOS</u>	2
<u>CHAPITRE 1 : INTRODUCTION AU LOGICIEL SPSS</u>	3
 DATA EDITOR	3
 SYNTAX	6
 OUTPUT VIEWER	7
GÉNÉRALITÉS À PROPOS DU LOGICIEL	7
IMPORTATION DE DONNÉES	7
<u>CHAPITRE 2 : DATA MANAGEMENT DANS SPSS</u>	8
DATA	8
TRANSFORM	10
UTILITIES.....	17
TRAITEMENT DES DONNÉES MANQUANTES.....	17
<u>CHAPITRE 3 : ÉLÉMENTS D'ANALYSE STATISTIQUE DESCRIPTIVE</u>	18
BREF RAPPEL DE STATISTIQUES DE BASE.....	18
LE SOUS-MENU DESCRIPTIVE STATISTICS	20
CRÉATION DE GRAPHIQUE	25
<u>CHAPITRE 4 : LES TESTS INFÉRENTIELS BIVARIÉS</u>	28
DISTRIBUTION D'EFFECTIFS : LE TEST DU KHI-CARRÉ	28
COMPARAISON DE MOYENNES : LE TEST-T	32
ANALYSE DE VARIANCE : L'ANOVA ET LES TESTS POST-HOC	37
LE TEST DE CORRÉLATION	41
RÉCAPITULATIF DES TESTS BIVARIÉS	44
<u>CHAPITRE 5 : EXPLORATION DE DONNEES</u>	45
L'ANALYSE EN COMPOSANTES PRINCIPALES.....	45
VALIDATION D'INDICATEURS : L'ANALYSE DE FIDÉLITÉ.....	53
TECHNIQUES DE CLUSTERING	56
CLUSTERING SUR ANALYSE EN COMPOSANTES PRINCIPALES	66
RÉCAPITULATIF DES MÉTHODES D'EXPLORATION DES DONNÉES	69
<u>RESSOURCES SUPPLÉMENTAIRES</u>	70

Avant-propos

Ce syllabus est destiné à documenter les travaux pratiques du cours LPOLS1221 (Analyse quantitative en sciences sociales). Il est structuré en 5 chapitres dont certains comportent une partie de rappel théorique. Le rappel de la théorie est là pour vous aider à vous souvenir des manipulations que vous allez devoir effectuer à l'aide du logiciel SPSS. En aucun cas, ce syllabus ne remplace le cours magistral donné par le professeur titulaire du cours ex cathedra.

Le premier chapitre aura pour but de vous familiariser avec le logiciel statistique SPSS. Le deuxième chapitre, dans la continuité du premier, introduira les bases du data management (comprenez, gestion de données). Ces deux premiers chapitres sont distincts des trois suivants dans le sens où il ne s'agit pas de chapitres analytiques. Les autres chapitres introduiront respectivement, l'analyse statistique descriptive, quelques tests inférentiels bivariés et enfin, quelques techniques d'exploration de données. Chacune de ces analyses ont un but bien précis. Si, pour la démonstration et la compréhension de chacune des différentes méthodes, nous vous proposons des développements mathématiques, ceux-ci ne font pas partie de la matière. En revanche, vous devez être aptes à comprendre ce que les différentes méthodes testent et dans quels cadres elles s'appliquent.

Vous trouverez les séries d'exercices au sein des fiches hebdomadaires. Ceux-ci, sont pour la plupart basés sur la base de données : European Values Study (EVS) de 2012. Cette base de données aura préalablement été téléchargée par nos soins. Sachez cependant, qu'un nombre important de bases de données sont disponibles en ligne et sont également documentées. Vous trouverez plus d'information sur l'EVS à partir du lien suivant : <http://www.gesis.org/en/services/data-analysis/survey-data/european-values-study/>. Si les exercices sont structurés en fonction de la matière vue, cela ne veut pas dire que l'exercice en soi ne sera focalisé que sur cette seule matière. Vous devez être aptes à monopoliser à chaque moment, les éléments de statistique descriptive, inférentielle ou d'exploration vus lors des séances précédentes.

Enfin, tout au long du syllabus, vous remarquerez que certains passages sont écrits en italiques. Il s'agit, la plupart du temps, d'explications supplémentaires, de trucs et astuces de conseils et d'opérations sur lesquelles il faut être vigilant. Vous verrez apparaître de temps en temps les symboles suivants : ✓ ou ✗. Le premier symbole vous indiquera, bien évidemment, les bonnes opérations à faire alors que le deuxième vous indiquera des erreurs couramment commises par les étudiants. Ces symboles sont là pour vous éviter de reproduire ces erreurs.

Afin de bien étudier ce cours et pour le bon déroulement des séances, il vous est demandé de revoir la matière vue pour les séances suivantes. Les assistants sont disposés à répondre à vos questions, que ce soit en classe ou par mail – voire lors d'une permanence. Veillez cependant à relire votre matière, consulter les ressources bibliographiques et les différents liens présents en bout de ce document avant de venir solliciter les assistants. Bien souvent, les réponses sont à votre disposition ou trouvables en quelques secondes via internet. Pour ceux qui le souhaitent, une version étudiante du logiciel SPSS est disponible pour le prix de 10 euros (plus d'informations : <http://www.uclouvain.be/359151.html>). Au sinon, les salles informatiques sont accessibles aux étudiants 24 heures sur 24¹.

Bon travail,

N. Gurnet et I. Rautu

¹ En dehors des heures d'ouverture du bâtiment, l'accès aux salles est possible via votre carte magnétique.

CHAPITRE 1 : Introduction au logiciel SPSS

SPSS est un logiciel de traitement statistique. Dans les salles informatiques de l'UCL, vous trouverez le logiciel, à partir du menu démarrer, dans l'onglet « Mathématiques et Statistiques » et dans le dossier IBM SPSS. Lorsque SPSS s'ouvre, ignorez les messages d'accueil.

SPSS est un logiciel qui comprend **3 types de fenêtres** : le Data Editor (l'éditeur de données), la syntaxe (le code informatique correspondant aux actions exécutées) et la fenêtre des outputs (résultats).

Data Editor

La fenêtre d'accueil du logiciel est composée de deux onglets situés en bas à gauche : d'une part le « data view » (vue des données) et d'autre part le « Variable view » (vue des variables).

34	v30	Numeric	1	C
35	v31	Numeric	1	C
36	v32	Numeric	1	C

Below the table, the tabs 'Data View' and 'Variable View' are shown, with 'Variable View' being the active tab.

FIGURE 1 : AFFICHAGE INFÉRIEUR GAUCHE DE LA FENÊTRE D'ACCUEIL (DATA EDITOR)

« Data view » ou vue des données

Il s'agit de l'affichage brut des données. **Chaque ligne représente une observation et chaque colonne correspond à une variable.** En sciences sociales, les observations correspondent généralement à des individus ayant répondu à un questionnaire et les variables, à la transcription informatique d'une réponse à une question par les différents individus.

Cette fenêtre ressemble assez fortement à l'affichage de logiciels tels qu'Excel ou autres. Vous pouvez donc, introduire directement des données manuellement dans cette fenêtre que ce soit de l'encodage ou à partir d'un « copier-coller » depuis un autre fichier. Notez cependant que la numérotation des informations est déjà disponible – et qu'il est donc inutile de renseigner une variable à cet égard. Dans le même ordre d'idée, les noms des variables ne seront pas renseignés dans cette partie du logiciel (voir le titre suivant : Variable View). Il est donc également inutile de renseigner les noms des variables (comme dans un fichier Excel par exemple) dans cette vue – sans quoi, vous risquez de compromettre le traitement de certaines données.

« Variable view » ou vue des variables

Cette vue permet de paramétriser les variables de votre base de données. Les variables sont dès lors énumérées dans une seule colonne, à l'instar des observations dans la vue des données. Chacune des colonnes suivantes représentent une caractéristique de la variable voulue. En d'autres termes, cet affichage correspond à ce que l'on appelle traditionnellement un *code book* en informatique. Voici une énumération des différentes propriétés des variables disponibles dans SPSS :

Name (nom) : Il s'agit d'un code attribué (généralement une abréviation ou un nom court) à la variable afin d'être rapidement repérée par l'utilisateur. Le nom des variables créées doit suivre certaines règles comme, par exemple, ne pas comprendre d'espace, de caractères spéciaux, ne pas commencer par des chiffres, etc. Le nom des variables est celui qui est affiché dans la vue des données.

Type : Cette caractéristique permet au logiciel de reconnaître et de pouvoir traiter les symboles contenus dans la variable. Il faut donc lui renseigner le format des données collectées. Peu importe ce que peut vouloir signifier la variable dans la réalité, **c'est le format encodé qui importe ici**. Si la variable contient uniquement des nombres, renseignez le type numérique (numeric). Si la variable contient des chaînes de caractères, renseignez le type « string ». D'autres formats sont également disponibles mais seront rarement utilisés dans

le cadre du cours. Par exemple, le format « date », permet à SPSS de traiter des données temporelles, qui répondent à d'autres lois que celles des mathématiques traditionnelles (60 secondes, 60 minutes, 24 heures, x jours par mois, etc.).

Faites attention. Si vous renseignez le mauvais type de variable, il se peut que SPSS fasse disparaître des données. Si votre variable d'origine contient des chaînes de caractères et que vous avez renseigné le type numérique pour la variable en question, SPSS va alors considérer que les données lettrées sont nulles car il s'attend à lire des nombres. Il va donc faire disparaître les lettres au profit d'une case vide (données manquantes).

Width (largeur) : Il s'agit du nombre de caractères maximum autorisé au sein de la variable. Par défaut, SPSS affiche le nombre de caractère de la donnée la plus longue.

Decimals : Nombre de décimales apparentes désiré. Restreindre le nombre de décimales ne supprimera pas la donnée de la base de données mais la restreindra à l'affichage de la base de données et lors de la production d'outputs.

Label (libellé) : Item complet (descriptif) de la variable, aussi appelé « étiquette ». Lors du traitement des données, le **nom** diminué de la variable est plus souvent utilisé pour une raison de rapidité. Cependant, lors de la production de rapport, il est préférable d'avoir des intitulés de variables complets. Lorsque vous produirez des résultats, vous verrez le libellé apparaître et non, le nom.

Si le nom d'une variable est « Gndr » et le libellé, « Sexe de l'individu », seul le deuxième apparaîtra lors de la production de résultats et le premier sera celui affiché au niveau de la vue des données.

Utiliser le nom des variables rend l'utilisation du logiciel plus facile et plus rapide. Généralement, le logiciel affiche par défaut la liste des variables par leur libellé dans ses différents menus, ce qui n'est pas très pratique – le libellé des variables étant généralement trop long que pour être affiché entièrement. N'hésitez pas à effectuer un clic droit sur une liste de variables dans un menu divers et cliquer sur « Display Variables names » comme le représente la Figure 2. Vous pouvez dès lors demander à SPSS d'afficher le nom des variables (Display Variable Names), à la place du libellé ou trier les variables par ordre alphabétique (Sort Alphabetically) afin de rendre la manipulation des variables plus aisée. Vous pouvez également demander des informations sur la variable (Variable information...). Cette opération va afficher les valeurs (values : voir ci-dessous) renseignées pour la variable (ex : 1 = 'homme' ; 2 = 'femme').

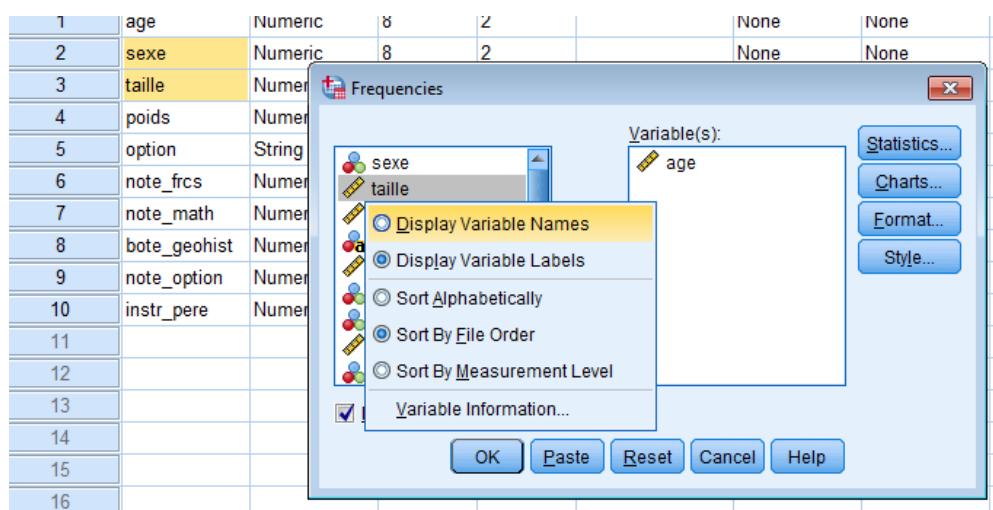


FIGURE 2 : MENU DE FACILITATION DE LA GESTION DES VARIABLES (CLIC DROIT)

Values (valeurs) : Répertoire des valeurs/modalités possibles pour la variable et attribution d'un libellé aux différentes valeurs/modalités. Cette option a la même fonction que le libellé mais pour les données et est donc souvent utilisé pour des variables nominales ou ordinaires. Grâce à cet outil, il est possible par exemple de signaler que la valeur 1 enregistrée dans la base de données de la variable « Gndr », désigne la modalité « Femme » tandis que la valeur 2 désigne la modalité « Homme ».

Si vous avez oublié quel label est associé à quelle valeur, vous serez renseignés dans cet onglet de la vue des variables ou alors en effectuant un clic droit sur une liste de variable dans un menu divers et en sélectionnant « Variable information ».

Missing (Données Manquantes) : Permet de signaler à SPSS que certaines valeurs de variables sont à considérer comme absentes lors de l'analyse. Il arrive souvent dans la construction des questionnaires de prévoir des modalités « je ne sais pas » ou « la personne n'a répondu à la question ». Ces réponses sont dès lors codées avec une valeur – généralement élevée (ex : 98, 998, etc.) ou négative (ex : -1, -98, etc.). Si ces données ne sont pas communiquées comme manquantes, elles peuvent dès lors biaiser bon nombre d'analyse.

Imaginons une question de consensus à une affirmation allant de 0 (pas du tout d'accord) à 10 (tout à fait d'accord), les différentes valeurs manquantes étant codées 97, 98 et 99. Si elles ne sont pas renseignées au logiciel comme manquantes, celles-ci vont considérablement gonfler la moyenne et peut éventuellement gonfler la moyenne au-delà de 10 si le nombre de données manquantes est considérable.

Measure (mesure) : Définition de la nature de la variable. Trois possibilités existent :

- (1) Variable **nominale**, c'est-à-dire une variable catégorielle nominale ;
- (2) Variable **ordinale**, c'est-à-dire variable catégorielle ordinaire ;
- (3) Variable **d'échelle** (scale), c'est-à-dire une variable quantitative. NB : SPSS ne fait pas la distinction entre variables quantitatives discrètes ou continues.

*Faites attention à bien paramétrier vos variables et à comprendre la différence fondamentale entre le type et la mesure d'une variable. Normalement, SPSS le fait automatiquement pour vous mais une vérification est souhaitable. Vérifiez tout d'abord le **Type de variable** renseigné. SPSS a besoin en effet de savoir ce qu'il va devoir lire dans la colonne pour pouvoir par la suite faire des opérations sur les variables. Il ne s'agit pas ici de renseigner des variables nominales, ordinaires ou quantitatives, mais de renseigner à SPSS s'il va devoir lire des nombres (numeric) ou des chaînes de caractères (string) dans la variable concernée. Si vous avez des formats spéciaux (date, monnaie, etc.), vous pouvez les renseigner dans cette colonne. Enfin, la **mesure** est un moyen pour vous de mieux repérer les variables nominales, ordinaires et quantitatives mais également d'annoncer à SPSS les tests statistiques possibles à partir de telle ou telle variable. Bien évidemment, dès qu'une variable est renseignée comme étant de type « string », vous ne pourrez jamais renseigner une mesure d'échelle vu que la calculabilité n'est pas possible sur des caractères.*

Niveaux de mesure des variables dans SPSS (excepté les variables dichotomiques)

Variable nominale	Variable ordinaire	Variable d'échelle (Scale)
Ex: Couleur préférée	Ex: Niveau d'instruction	Ex: Âge
Pas de calculabilité	Calculabilité possible	Calculabilité

FIGURE 3 : RAPPEL DES NIVEAUX DE MESURE EN STATISTIQUE.

Pour rappel, les variables ordinaires sont des variables avec des informations de type qualitatif mais hiérarchisable. Il s'agit dès lors de variables ambidextres qui peuvent à la fois revêtir le rôle de variable nominale et le rôle de variable quantitative en certaines occasions. Une bonne façon de distinguer la variable ordinaire d'une variable quantitative est de tester l'additivité des valeurs qui la composent. Par exemple, il est possible d'additionner deux âges ($18+20 = 38$) alors que c'est statistiquement insensé d'additionner un niveau d'instruction primaire avec un niveau d'instruction secondaire. Ceci n'empêche cependant pas de calculer une moyenne me renseignant sur le niveau d'éducation général d'une population par exemple.

Il s'agit d'une fenêtre de commandes SPSS. On y encode directement les commandes (exige une connaissance du langage de programmation SPSS), soit indirectement (par la fonction « *Coller/Paste* » disponible dans de nombreux menus). **Attention : Prenez le réflexe d'enregistrer la syntaxe pour toute opération effectuée plutôt que les résultats obtenus.** Cela permet d'automatiser la reproduction de vos procédures et donc, d'effectuer un gain de temps parfois non négligeable. En dehors de cela, tenir à jour une syntaxe permet d'avoir une trace écrite de ce qui a été effectué et permet donc aux personnes extérieures au projet de comprendre ce qui a été fait.

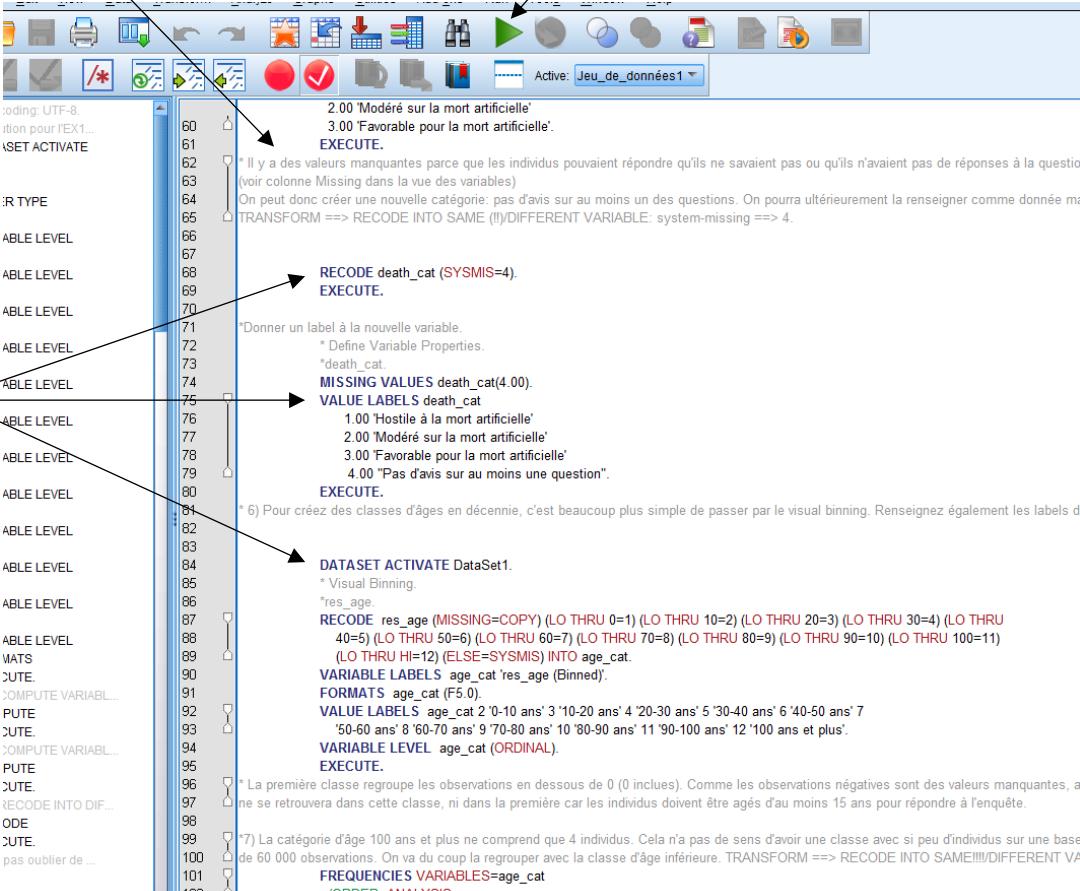
Pratiquement toutes les procédures de SPSS vous permettent d'obtenir un code de syntaxe via le bouton « Coller/Paste ». Vous cliquez sur ce bouton à la place d'exécuter votre procédure. Le code apparaît alors dans la fenêtre de syntaxe. Il vous suffit ensuite de sélectionner le code apparu et de l'exécuter (via le bouton Play).

Des commentaires ou des annotations peuvent être ajoutés votre syntaxe, ne fut-ce que pour décrire la commande suivante. Il suffit pour cela de commencer votre ligne par un astérisque et de la terminer par un point. Le texte sera automatiquement grisé et ne pourra pas être exécuté comme un code statistique.

Commentaire commençant par

* et finissant par un point.

Bouton d'exécution du code



```

* Il y a des valeurs manquantes parce que les individus pouvaient répondre qu'ils ne savaient pas ou qu'ils n'avaient pas de réponses à la question (voir colonne Missing dans la vue des variables)
On peut donc créer une nouvelle catégorie: pas d'avis sur au moins un des questions. On pourra ultérieurement la renseigner comme donnée manquante.
TRANSFORM => RECODE INTO SAME (!) / DIFFERENT VARIABLE: system-missing => 4.

RECODE death_cat (SYSMIS=4).
EXECUTE.

* Donner un label à la nouvelle variable.
* Define Variable Properties.
*death_cat.
MISSING VALUES death_cat(4.00).
VALUE LABELS death_cat
 1.00 'Hostile à la mort artificielle'
 2.00 'Modéré sur la mort artificielle'
 3.00 'Favorable pour la mort artificielle'
 4.00 "Pas d'avis sur au moins une question".
EXECUTE.

* 6) Pour créer des classes d'âges en décennies, c'est beaucoup plus simple de passer par le visual binning. Renseignez également les labels de
DATASET ACTIVATE DataSet1.
* Visual Binning.
'res_age.
RECODE res_age (MISSING=COPY) (LO THRU 0=1) (LO THRU 10=2) (LO THRU 20=3) (LO THRU 30=4) (LO THRU
 40=5) (LO THRU 50=6) (LO THRU 60=7) (LO THRU 70=8) (LO THRU 80=9) (LO THRU 90=10) (LO THRU 100=11)
 (LO THRU Hi=12) (ELSE=SYSMIS) INTO age_cat.
VARIABLE LABELS age_cat 'res_age (Binned)'.
FORMATS age_cat (F5.0).
VALUE LABELS age_cat 2 '0-10 ans' 3 '10-20 ans' 4 '20-30 ans' 5 '30-40 ans' 6 '40-50 ans' 7
 '50-60 ans' 8 '60-70 ans' 9 '70-80 ans' 10 '80-90 ans' 11 '90-100 ans' 12 '100 ans et plus'.
VARIABLE LEVEL age_cat (ORDINAL).
EXECUTE.

* La première classe regroupe les observations en dessous de 0 (0 inclus). Comme les observations négatives sont des valeurs manquantes, a
ne se retrouvera dans cette classe, ni dans la première car les individus doivent être âgés d'au moins 15 ans pour répondre à l'enquête.

*7) La catégorie d'âge 100 ans et plus ne comprend que 4 individus. Cela n'a pas de sens d'avoir une classe avec si peu d'individus sur une base
de 60 000 observations. On va du coup la regrouper avec la classe d'âge inférieure. TRANSFORM => RECODE INTO SAME!!!! / DIFFERENT VAI
FREQUENCIES VARIABLES=age_cat
/ORDER=ANALYSIS.

```

FIGURE 4 : EXEMPLE DE SYNTAXE

Un code de syntaxe commence toujours par un mot-clé annonçant la procédure amorcée (ex : RECODE, FREQUENCIES) et termine par EXECUTE.

Output viewer

Cette fenêtre présente les résultats des analyses exécutées par SPSS. Lorsque vous exécutez des procédures ne produisant pas d'outputs, la fenêtre d'outputs affichera tout de même le code informatique de la procédure qu'elle vient d'effectuer.

Généralités à propos du logiciel

Les trois différentes fenêtres qui viennent d'être présentées peuvent être sauvegardées sous forme de fichiers différents : une base de données, un fichier d'output et une syntaxe. Hormis si vous êtes sûrs de vous, nous vous conseillons de plutôt prendre le réflexe de sauvegarder des syntaxes et non des bases de données. Le principe est simple : si vous faites une erreur lors du traitement de vos données et que vous sauvegardez la base de données, les données originales seront écrasées et il vous sera difficile de revenir en arrière alors que le code de la syntaxe est modifiable et ré-exécutable rapidement.

Le logiciel regroupe les différentes procédures dans différents menus ayant des caractéristiques communes. Lorsque vous recherchez une procédure particulière, demandez-vous toujours ce que vous recherchez précisément et les titres de menus vous aideront allègrement.

- File (Fichier) : ouvrir, fermer et sauvegarder les fichiers de données, imprimer et quitter SPSS
- Edit (Édition) : copier et corriger des données, modifier les options
- View (Affichage) : barres d'outils, affichage des étiquettes
- Data (Données) : définir les variables, manipuler les fichiers
- Transform (Transformer) : modifier et créer de nouvelles variables
- Analyze (Analyse) : exécuter les procédures statistiques
- Graphs (Graphes) : créer des graphiques
- Utilities (Utilitaires) : informations sur les variables, les préférences et les commandes
- Add-ons : pour les modules complémentaires, si certains sont disponibles
- Window (Fenêtre) : examiner les fenêtres actives, changer de fenêtre
- Help (Aide) : aide SPSS

Importation de données

L'importation des données est possible via le menu FILE, le sous-menu OPEN et DATA. Vous noterez également qu'il est possible d'ouvrir un fichier de syntaxe ou une fenêtre d'outputs. Par défaut, le logiciel vous propose de rechercher des bases de données enregistrées sur votre ordinateur au format SPSS (.sav). Si vous recherchez un fichier en un autre format (Excel par exemple), il vous suffit de modifier le type de fichier recherché comme illustré à la Figure 5.

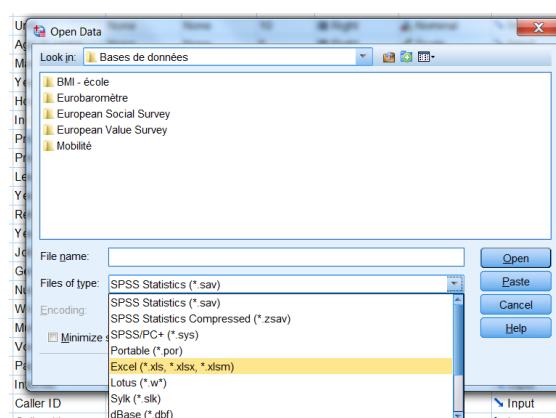


FIGURE 5 : FENÊTRE D'IMPORTATION DE DONNÉES

CHAPITRE 2 : Data management dans SPSS

Le data management ou traitement de données est la partie de l'analyse quantitative qui s'occupe de la modification et à la préparation des données. La manipulation de la base de données peut consister en plusieurs types d'opérations comme modifier les variables, créer des nouvelles variables, la gestion des données (tri, suppression, réorganisation), etc. Les principaux menus de SPSS menant à des procédures de data management sont le menu DATA (données) et le menu TRANSFORM (transformer). Accessoirement, nous verrons que certaines procédures du menu UTILITIES (utilitaires) permettent également d'améliorer le confort de l'utilisateur dans la manipulation de SPSS. Les différentes procédures de data management abordées dans les TP vont être énumérées et détaillées par les différents menus du logiciel. En complément, une section de ce chapitre sera dédiée au traitement des données manquantes.

Data

Define variables properties (Définir les propriétés des variables) : Nous avons vu au chapitre 1 comment paramétrier des variables depuis la vue des variables pour une première prise en main du logiciel. Ce type de paramétrage ne permet cependant pas de générer (via *PASTE*) du code de syntaxe et donc de pouvoir reproduire le paramétrage des variables. C'est pourquoi il existe une procédure spécialement conçue pour le paramétrage des variables. Nous verrons d'autant plus que cette procédure possède quelques avantages considérables qui favoriseront son utilisation au paramétrage dans la vue des variables.

Lorsque vous cliquez sur la procédure, un premier menu apparaît où le logiciel vous demande de lui signaler les variables à modifier. Vous avez donc sur la gauche une liste comprenant toutes les variables de la base de données et vous signalez les variables à modifier en les faisant glisser sur la liste de droite.

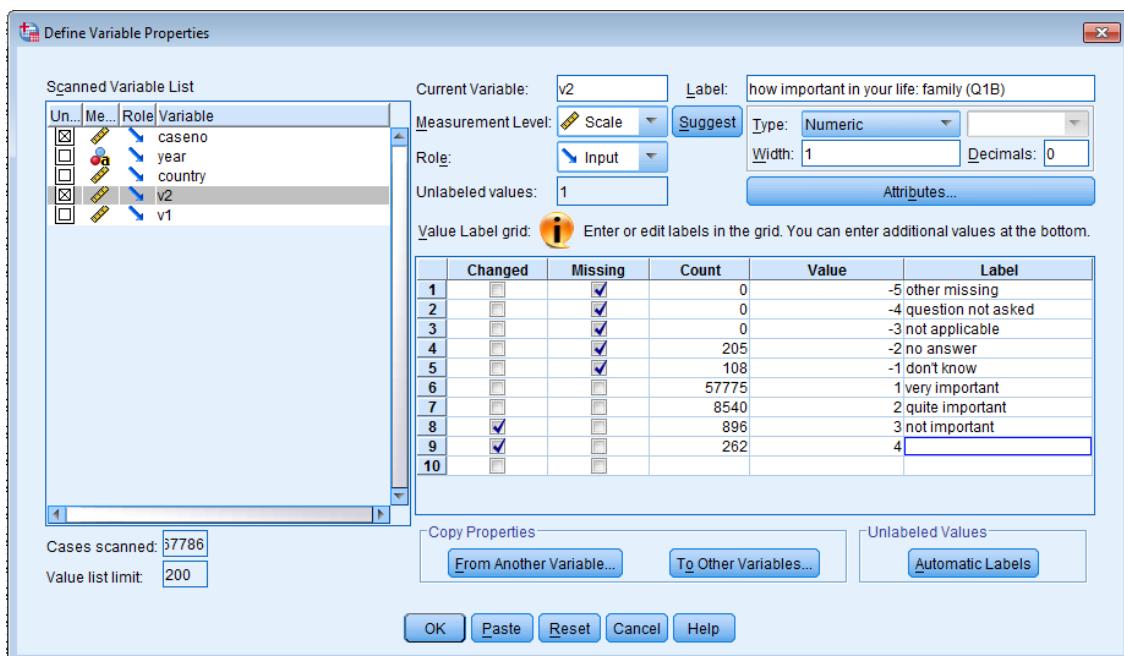


FIGURE 6 : MENU DE PARAMÉTRAGE DES VARIABLES

Vous arrivez ensuite au menu représenté dans la Figure 6 dans lequel vous retrouverez à gauche la liste des variables à modifier, vous les modifierez une à une. En haut à droite du menu, vous retrouverez les différents paramètres des variables déjà vu au chapitre 1 : label de la variable, type de variable, largeur, décimales, niveau de mesure, etc.

Notez le bouton SUGGEST à côté du niveau de mesure. SPSS va analyser le nombre de modalités présentes dans la variable et va vous suggérer un niveau de mesure en fournissant une explication. Cet outil peut vous aider si vous avez éprouvé certaines difficultés à définir les niveaux de mesure mais vous devriez être apte à le faire de façon autonome lors de l'examen.

En dessous, SPSS vous affiche la liste des valeurs présentes au sein de la variable. Vous pouvez cochez les valeurs à attribuer aux valeurs manquantes et définir un label correspondant aux différentes valeurs.

La liste des valeurs présentes dans la variable peut être un outil parmi d'autres de repérer qu'il n'y a pas par exemple de chaînes de caractères dans votre variable alors que vous avez renseigné le type de la variable comme étant numérique.

Une fois votre paramétrage de variable terminé, vous pouvez soit exécuter la commande via OK ou générer le code de syntaxe via PASTE. Pour rappel, pour exécuter la syntaxe, vous devez sélectionner le code voulu et l'exécuter.

Select Cases (Sélectionner des observations) : Vous pouvez proposer à SPSS de ne conserver que certaines observations pour le bien de vos analyses². Le logiciel sélectionne par défaut toutes les observations (ALL CASES). Vous pouvez dans ce menu conserver des données qui remplissent une condition logique (IF CONDITION IS SATISFIED). Vous pouvez alors spécifier la condition à respecter dans le menu IF.

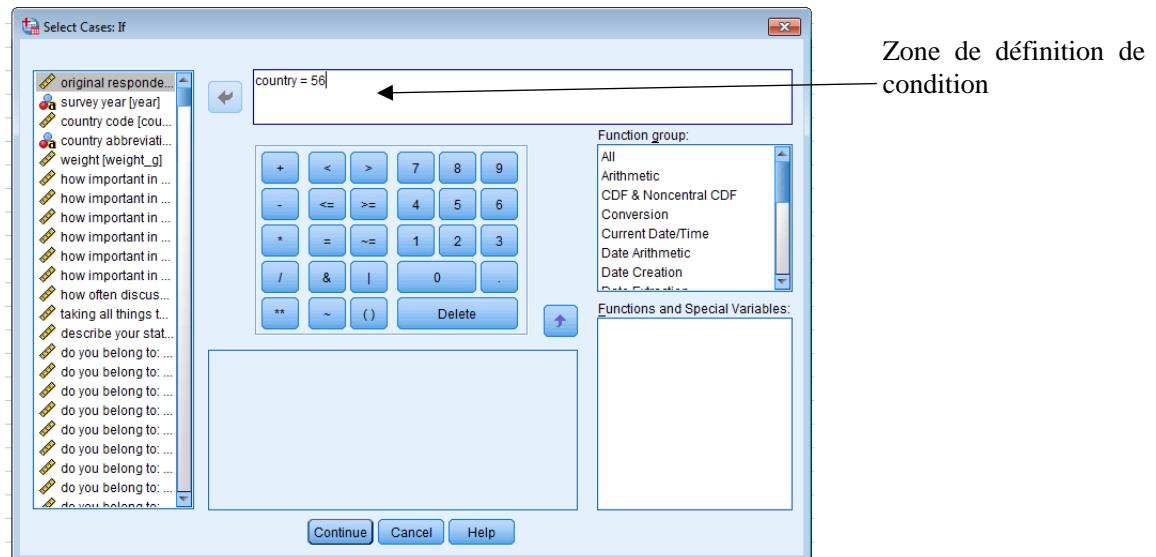


FIGURE 7 : POSER UNE CONDITION DE SÉLECTION D'OBSERVATIONS

Définir une condition revient à dire qu'une ou plusieurs valeurs sont présentes dans une ou plusieurs variables spécifiques. Dans le cadre d'une condition simple, on écrit équationnellement la condition de la façon suivante : Variable - Opérateur arithmétique – valeur. Les opérateurs arithmétiques les plus utilisés sont les suivants : <, >, =, <=, >=, <>.

Si par exemple, on ne veut garder que les ressortissants de Belgique³ (valeur 56 depuis la variable Country), on posera la condition comme ceci : Country = 56

Si on veut poser plusieurs conditions simples, on peut également utiliser des opérateurs logiques : AND, OR.

Si par exemple, on ne veut garder que les ressortissants de Belgique ou de France (valeur 250 depuis la variable Country) et âgés de moins de trente ans, on posera la condition comme ceci : (Country = 56 OR Country=250) AND age < 30

L'utilisation des parenthèses est ici nécessaire : le OR va opérer avant le AND.

! Attention. Si vous demandez à SPSS de retrouver des valeurs de caractères (par exemple BE pour Belgique), entourez ces caractères d'apostrophes lors de la pose de condition : 'BE'.

² SPSS ne va pas supprimer les observations mais simplement les retirer des futures procédures d'analyse.

³ Il faut au préalable déjà connaître la valeur attribuée au pays, qu'elle soit de type numérique ou chaîne de caractères.

Lorsque vous avez fini de travailler sur les données filtrées, il vous suffit de retourner dans la procédure et de re-sélectionner ALL CASES ou de modifier votre condition si vous voulez travailler sur un autre sous-échantillon.

Utilisez la sélection des observations est important afin de ne pas biaiser vos résultats. Si vous disposez d'une base de données Européenne et que votre analyse ne porte que sur la Belgique, il s'agit d'une erreur fondamentale d'utiliser l'entièreté de la base de données. Il faudra restreindre votre base de données aux seules données issues de Belgique et non de l'Europe.

D'autres possibilités de filtrage sont proposées dans cette procédure que nous ne verrons pas dans le cadre de ce cours : constitution d'un échantillon aléatoire de votre base de données, à partir du numéro d'observation ou encore via une variable de filtre (rejetant les données manquantes ou égales à 0). Il vous est également possible de supprimer les observations non conservées mais bien entendu, nous ne vous conseillons pas cela.

Split File (Scinder un fichier) : Si vous le désirez, le logiciel vous donne également l'opportunité de sortir des résultats (voir chapitres ultérieurs) en fonction des modalités d'une variable spécifique. Cette option est également très pratique pour gagner du temps !

Si vous désirez obtenir le résultat d'un test précis pour chacun des 42 pays que constitue votre base de données européenne, le fait de scinder le fichier à partir de la variable de pays fera que vous n'aurez qu'à lancer votre test qu'une seule fois et non 42 fois. Attention toutefois à faire cette opération uniquement si cela a du sens.

Par défaut, le logiciel ne sépare pas les résultats (Analyze all cases, do not create groups). Vous pouvez ordonner à SPSS de regrouper les résultats en fonction des modalités présentes dans la variable (Organize output by groups) ou de comparer directement chaque résultat de chaque modalité en même temps (Compare groups).

Transform

Le menu TRANSFORM va regrouper les procédures qui vont permettre la création/modification de variables (et donc des données) ainsi que la gestion des valeurs manquantes.

Compute variable (calculer la variable) : Cette procédure permet de créer une variable à partir d'opérations mathématiques sur une ou plusieurs variables (voire sur aucune variable, mais cela est plus rare).

Vous devez ainsi renseigner au logiciel le nom de la nouvelle variable qui sera créé. En dessous de cet espace, un onglet vous permet de définir les caractéristiques de la variable⁴. Comme pour l'expression des conditions de sélection d'observation vue ci-dessus, vous devez renseigner l'équation souhaitée. Vous pouvez dès lors additionner, soustraire, diviser, multiplier des variables à votre guise. Notez que l'exponentielle est représentée par « ** » dans SPSS et que l'opérateur logique AND est remplacé par « & » et l'opérateur logique OR par « | ».

Moins utilisés dans ce cours, vous remarquerez sur la droite des menus de fonctions. A l'instar de logiciel comme Excel, vous pouvez renseigner une fonction précise dont l'utilisation est décrite dans l'encadré en dessous de la calculette. Ainsi, les deux équations renseignées dans la Figure 8 feront la même opération⁵.

⁴ Ou alors vous procédez au paramétrage dans la vue des variables ou dans la fonction *Define Variables properties*.

⁵ Notez que le « OU » a été rajouté pour l'exemple. Vous ne pouvez pas rajouter ce genre de texte dans la procédure.

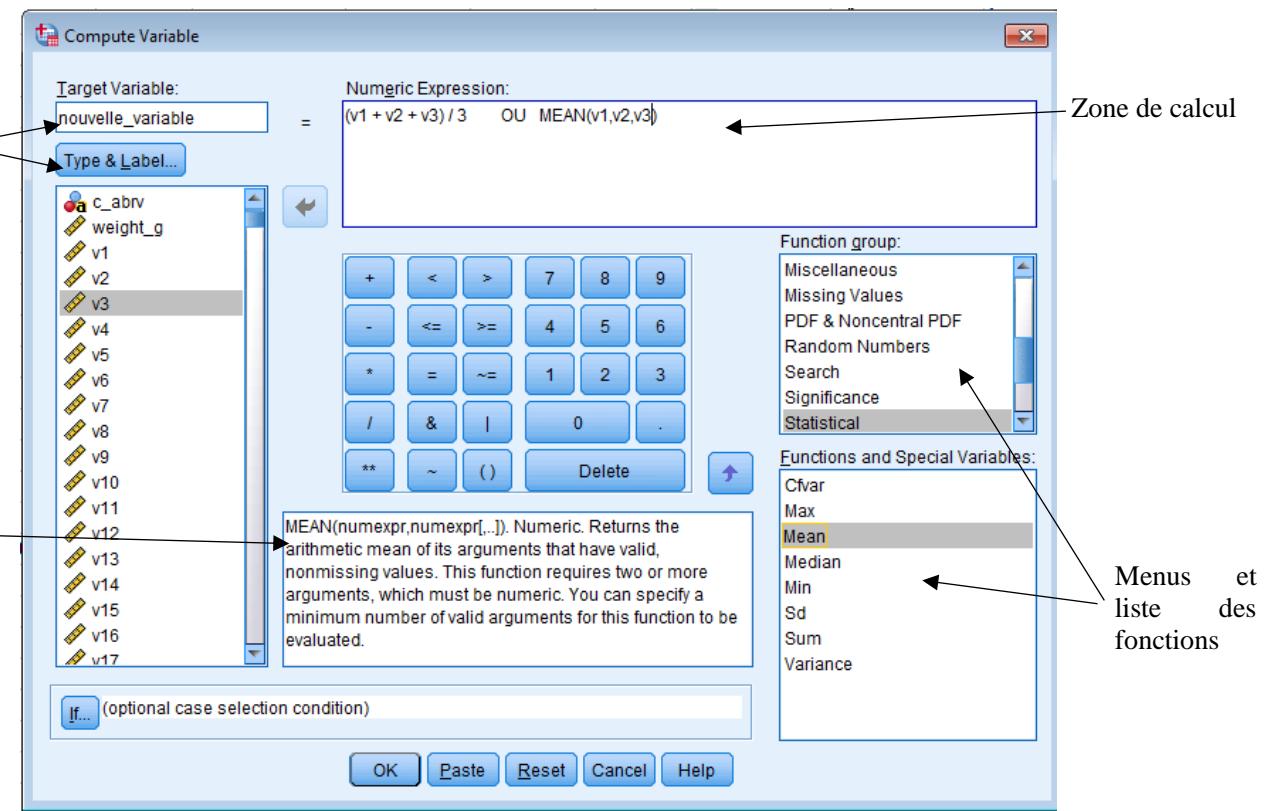


FIGURE 8 : FENÊTRE DE LA PROCÉDURE COMPUTE VARIABLE

Fonction idéale pour créer des variables « proxy » : si plusieurs variables mesurent un même phénomène. On peut dès lors obtenir un score total en sommant les scores obtenus sur chaque variable. Faites attention au type de variable utilisé puisqu'il s'agit d'une opération qui va utiliser la calculabilité de la variable.

- ✓ $BMI = \text{poids} / (\text{taille}) ^{**2}$;
- ✓ $\text{Proxy} = \text{var1} + \text{var2} + \text{var3} + \text{var4}$;
- ✓ $\text{Proxy} = \text{sum}(\text{var1}, \text{var2}, \text{var3}, \text{var4})$;
- ✗ $\text{Couleur_preferee} / 2 + 4$;

Recode into different variables (Création de variables) : Cette procédure permet de créer une nouvelle variable en substituant une valeur ou une série de valeur par une valeur bien précise.

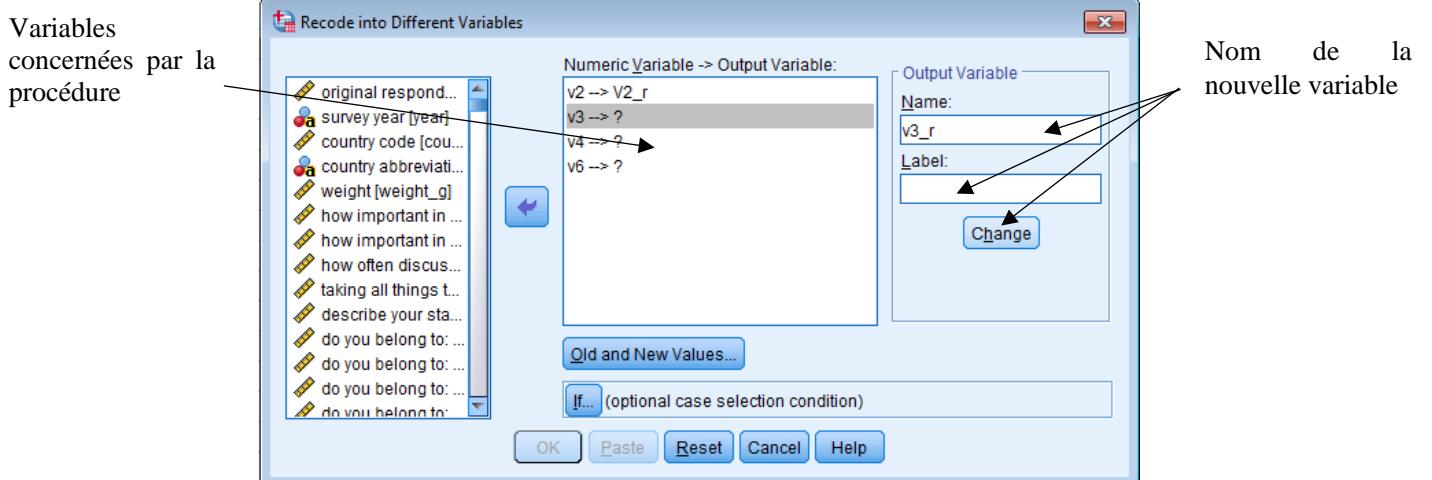


FIGURE 9 : MENU D'ACCUEIL DE LA FONCTION RECODE INTO DIFFERENT VARIABLES

Dans le premier menu présenté par le logiciel, vous devez définir les variables à partir desquelles SPSS va apporter les modifications souhaitées. Plusieurs variables peuvent être modifiées en même temps, créant ainsi autant de variables.

Faire cependant attention à ce que toutes les variables modifiées en même temps vont subir les mêmes opérations. Si chaque variable nécessite un traitement de données spécifique, refaire l'opération pour chaque variable.

Vous devez, comme à chaque fois que l'on crée une nouvelle variable, spécifier le nom des nouvelles variables. N'oubliez pas de cliquer sur *Change*. Tant que cela n'est pas fait, vous verrez au sein de l'encadré central, le nom de votre variable suivi d'une flèche avec un point d'interrogation (voir Figure 9). Une fois que vous aurez fini de renommer toutes les nouvelles variables, les boutons d'exécution et de génération du code de syntaxe s'activeront.

Afin d'annoncer les valeurs à remplacer, vous devez cliquer dans l'onglet « Old and New Values ». Ce nouveau menu est séparé en deux : à gauche, un encadré dans lequel il faut spécifier à SPSS les valeurs à rechercher dans la variable d'origine et à droite, la nouvelle valeur que vont avoir les observations concernées dans la nouvelle variable (Figure 10).

Les différentes options du menu des anciennes valeurs (old value) sont les suivantes :

- ❖ Value : permet de substituer une valeur unique de la variable d'origine.
- ❖ System-missing : permet de substituer les données manquantes de la variable d'origine.
- ❖ System- or user- missing : Idem mais en tenant compte des valeurs renseignées comme missing par l'utilisateur également.
- ❖ Range : permet de définir une étendue de valeurs depuis la variable d'origine.
- ❖ Range, Lowest through value : permet de définir une étendue de valeur depuis la valeur la plus basse comprise dans la variable d'origine jusqu'à la valeur indiquée comprise.
- ❖ Range, value through highest : Idem mais partant de la valeur indiquée inclue jusqu'à la valeur la plus élevée de la variable.
- ❖ All other values : Permet de substituer toutes les valeurs non renseignées.

! Si vous recodez (en utilisant les options range ou all other values) sans utiliser l'option System- or user- missing, le logiciel va copier la valeur des possibles données manquantes (-1 ;98 ;998 ; etc.) dans la nouvelle variable sans pour autant les renseigner en données manquantes.

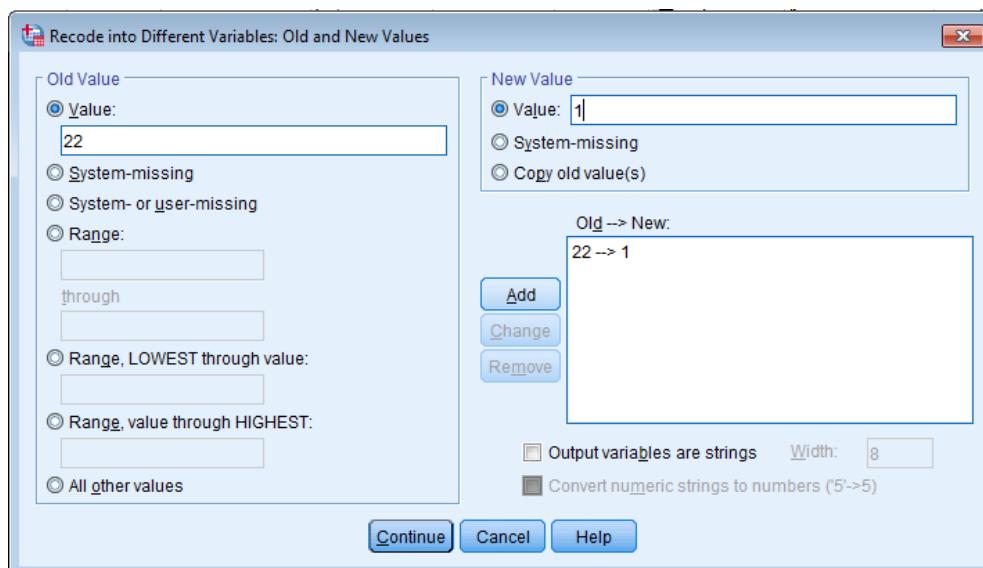


FIGURE 10 : MENU DE SUBSTITUTION DES VALEURS

Les options du menu de la nouvelle variable (New Value) sont moins nombreuses :

- ❖ Value : valeur de substitution.
- ❖ System-missing : renseigne les valeurs en données manquantes.
- ❖ Copy old value(s) : Copie tout simplement la valeur d'origine dans la nouvelle variable.

Cette procédure est donc idéale pour catégoriser des variables quantitatives ou pour recoder des valeurs.

- ✓ Convertir une variable quantitative (la taille en cm) en variable de 3 catégories de taille :
 - ✓ Range, Lowest through 150 = 1
 - ✓ Range 151 through 180 = 2
 - ✓ Range, value through Highest 181 = 3
 - ✓ System – or User missing = System-missing
- ✓ Remplacer une valeur (3 doit devenir 2) dans une variable :
 - ✓ Value 3 = Value 2
 - ✓ System – or User missing= System-missing
 - ✓ All other values= Copy old value(s)
- ✗ Convertir une variable quantitative (taille en cm) en variable quantitative (taille en mètres)
 - ✗ Value 101 = Value 1,01
 - ✗ Value 102 = Value 1,02
 - ✗ Etc...

Recode into same variables (Recoder la variable) : Alors que la fonction *Recode into Different variables* (création de variables) permet de créer une nouvelle variable afin de conserver la variable d'origine, cette fonction écrase les données d'origine. Il s'agit donc d'être particulièrement vigilant lors de l'utilisation de cette procédure !

Il s'agit ensuite d'indiquer à SPSS quelles valeurs de la variable d'origine (anciennes valeurs) doivent être codées dans la nouvelle variable (nouvelles valeurs) comme vu précédemment dans la procédure *Recode into different variables*.

Cette fonction, bien que dangereuse pour les distraits ou les non-avertis, trouve son utilité dans le recodage de valeur au sein d'une variable. Si une erreur d'encodage est advenue par exemple, au lieu de créer une nouvelle variable, je vais remplacer la valeur erronée par la bonne valeur.

- ✓ Recoder dans la variable (Taille en mètre), la valeur 2,01 par 201.

Count values within cases (Compter les occurrences) : Cette procédure permet de créer des variables de comptage (dichotomiques / proxy) en recherchant une ou plusieurs modalités parmi une ou plusieurs variables.

Le principe est assez simple : vous spécifier une ou plusieurs valeurs et le logiciel va scanner les variables renseignées. S'il trouve la valeur, il va ajouter 1 à la nouvelle variable. Si le logiciel ne détecte aucune des valeurs renseignées parmi les variables scannées, l'observation concernée aura la valeur 0 au sein de la nouvelle variable. Si vous renseignez plusieurs variables à scanner, la procédure va additionner les 0 et les 1 au sein de la nouvelle variable.

La façon de renseigner les valeurs à détecter par SPSS est la même que pour la procédure *Recode Into Different variables* :

- Value : détecter une valeur spécifique isolée lors du scan de variable(s)
- System-missing : détecter les valeurs manquantes lors du scan de variable(s)
- System-or-user missing : détecter les valeurs manquantes ou renseignées comme manquantes
- Range : détecter une étendue de valeurs
- Range, Lowest through : détecter une étendue de la valeur minimale jusqu'à la valeur comprise
- Range, value through highest : détecter étendue de la valeur à la valeur maximale.

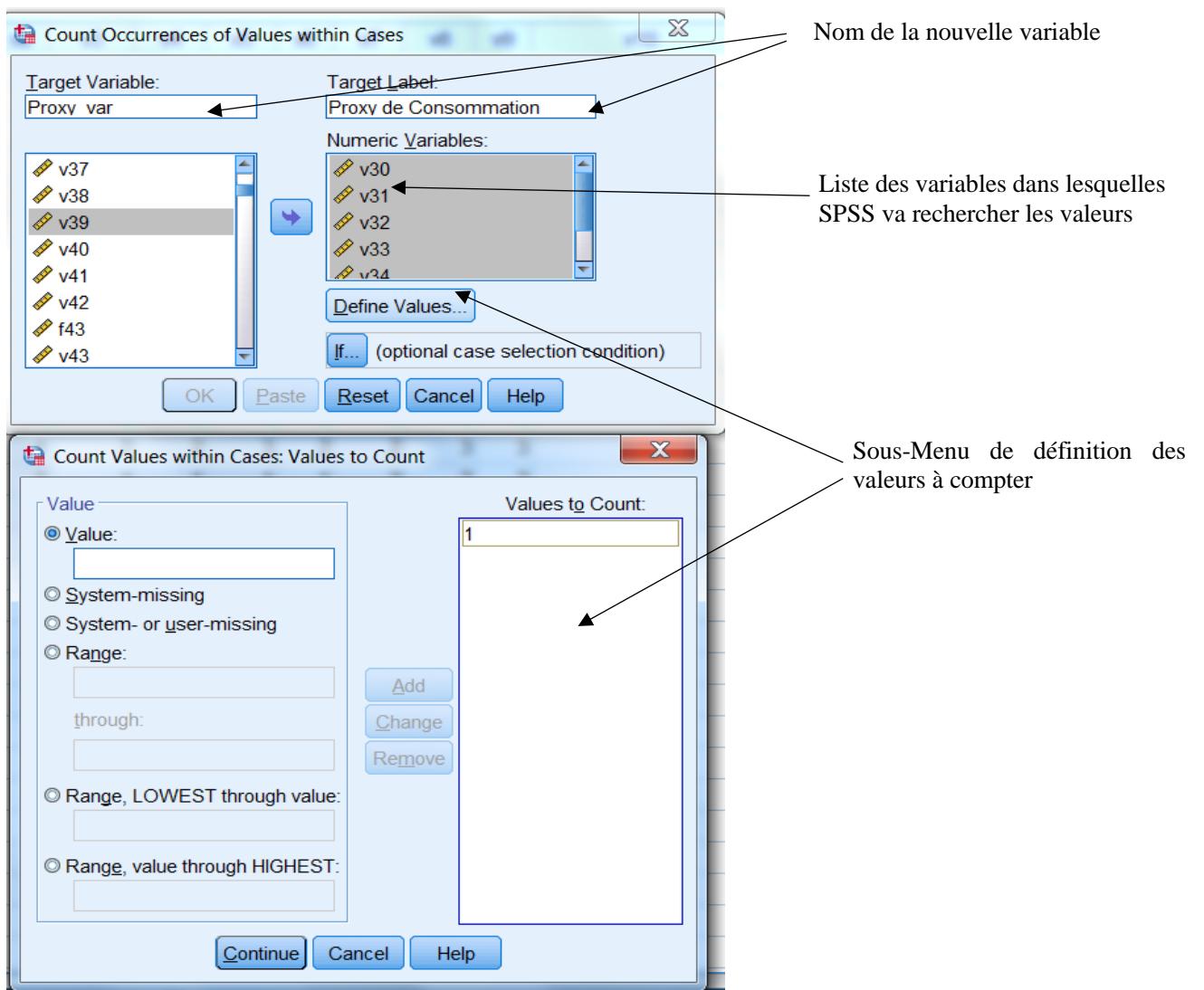


FIGURE 11 : MENU DE LA PROCÉDURE DU COMPTAGE DES OCCURRENCES

Cette procédure, bien que faisable via la procédure RECODE INTO DIFFERENT VARIABLE ou COMPUTE VARIABLE, permet cependant un gain de temps non négligeable lors de traitement de données spécifique.

- ✓ J'ai par exemple une variable renseignant le pays de l'interviewé (Country) de type de chaîne de caractères. Si je veux créer une variable de filtre pour ne conserver que les ressortissants de France, Allemagne et Belgique. Je demande à SPSS de détecter les valeurs 'FR'; 'BE'; 'AL'. Tous les ressortissants de ces pays auront donc la valeur 1 et les autres 0.
- ✓ J'ai 10 variables demandant à chaque répondant s'il possède un objet électroménager ou non (0=non ; 1=où). Je veux créer une variable de proxy m'indiquant combien de produits possèdent chaque répondant. Je demande alors de scanner à SPSS les 10 variables et je renseigne 1 comme valeur à détecter.

Visual Binning (regroupement en classes visuelles) : Cette procédure permet de créer des classes (divisions) au sein de variables quantitatives de manière rapide et efficace tout en tenant compte des données présentes au sein de chaque variable.

Vous renseignez les variables que vous voulez découper en classes et poursuivez pour arriver au menu d'accueil présenté à la Figure 12.

Nom de nouvelle variable

Liste des variables à scanner

Distribution des données de la variable

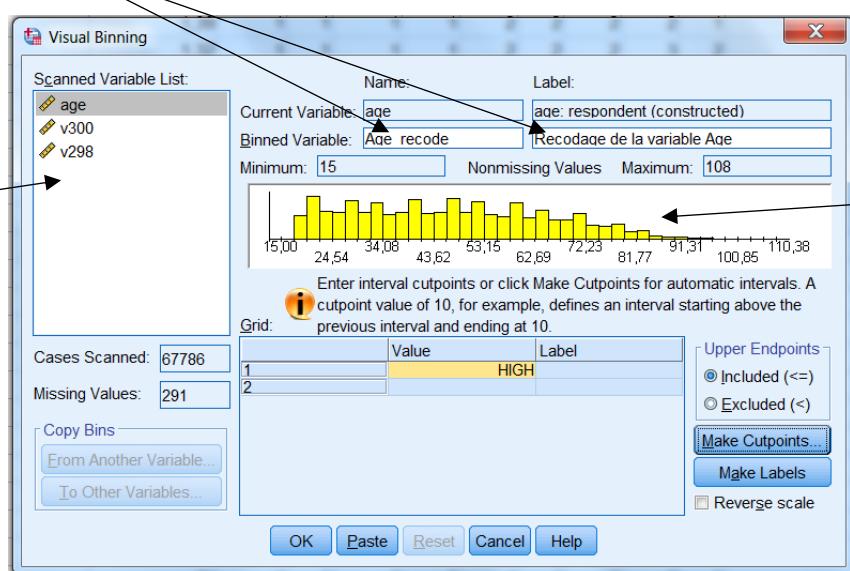


FIGURE 12 : MENU D'ACCUEIL DU VISUAL BINNING

Vous pouvez de base voir l'histogramme – soit la distribution des données de la variable au centre de la fenêtre. Ensuite, dans l'onglet *Make cutpoints* (créer des divisions), vous pouvez créer vos classes. SPSS vous propose trois sortes de divisions (voir Figure 13) qui ont chacune leurs avantages et inconvénients, à vous de faire le tri de ce dont vous avez besoin. N'oubliez pas que si vous désirez des classes vraiment spécifiques, il est peut-être préférable de passer par la fonction *Recode Into Different variables*. Cette procédure va plutôt servir à créer des variables de classes de façon automatique.

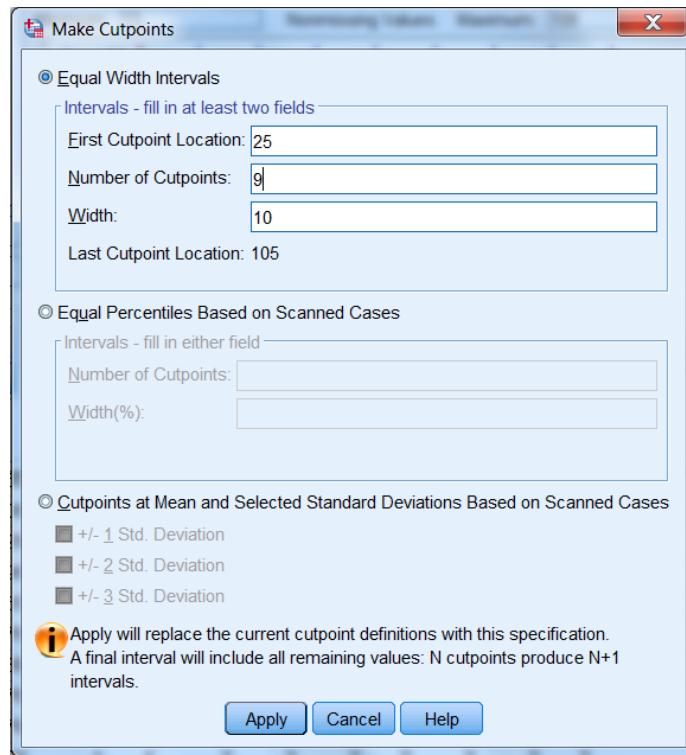


FIGURE 13 : SOUS-MENU DE PARAMÉTRAGE DES DÉCOUPES DE CLASSE

Equal Width Intervals (Intervalles de longueur identiques) : Avec ce mode de découpe, SPSS va créer des classes qui auront un même nombre de valeur au sein de chaque classe. Il s'agit donc de classes à intervalle fixe. Vous devez tout d'abord mentionner un point de découpe initial, et ensuite choisir entre spécifier la longueur d'intervalle ou le nombre de découpe à effectuer (nombre de classe désiré -1).

- ✓ *Cette option est particulièrement utile pour créer par exemple des classes de tranches salariales (par tranche de 100\$) ou des classes par tranches d'âge (10 ans).*

Equal Percentiles Based on Scanned Cases (Centiles égaux fondés sur les observations analysées) : Avec ce mode de découpe, SPSS va créer des classes qui auront un même nombre d'observation au sein de chacune d'elles. Il est à noter que, vu que l'on parle bel et bien de percentiles, SPSS va opérer la découpe à partir d'une valeur qui va se rapprocher le plus proche du percentile demandé.

Si je veux créer 2 classes reprenant chacune 50% de la population, le logiciel va opérer la découpe au niveau de la médiane. Cependant cette valeur peut être présente chez plusieurs observations, rendant compliqué la découpe en deux groupes de même taille. Il se peut donc que la classe 1 comprenne 49.7% des observations et la classe 2, 50.3%. Cela étant, cette valeur sera celle qui va diviser les classes de telle façon à ce qu'elles soient le plus proche de 50%.

- ✓ *Idéal pour créer des classes en fonction de quartiles, déciles, etc.*

Cutpoints at Mean and Selected Standard Deviations Based on Scanned Cases (Classes en fonction des écarts-types à la moyenne) : Comme la procédure l'indique, le logiciel va ici créer des classes en fonction de l'écart des observations par rapport à la moyenne. En fonction du nombre d'écart-types (x) sélectionné, vous aurez un nombre de classes égal à $2x + 2$.

Si je sélectionne l'option 1 & 2 écart-types, j'aurai donc 6 classes : les observations à plus de 2 EC en dessous et au-dessus de la moyenne, les observations entre 1 et 2 EC en dessous et au-dessus de la moyenne et enfin, les observations à maximum 1 EC en dessous et au-dessus de la moyenne.

- ✓ *Cette option est donc utile pour créer des classes en fonction de la distribution des données, permettant ainsi de créer des classes d'individus proches de la moyenne et d'autres classes plus éloignées. Cette option est également utile pour repérer les observations extrêmes.*

Découpages en bleu,
classe sélectionnée
en rouge

Numéro des
classes

Borne supérieur
de la classe

Renseignez un
label à la classe
ici

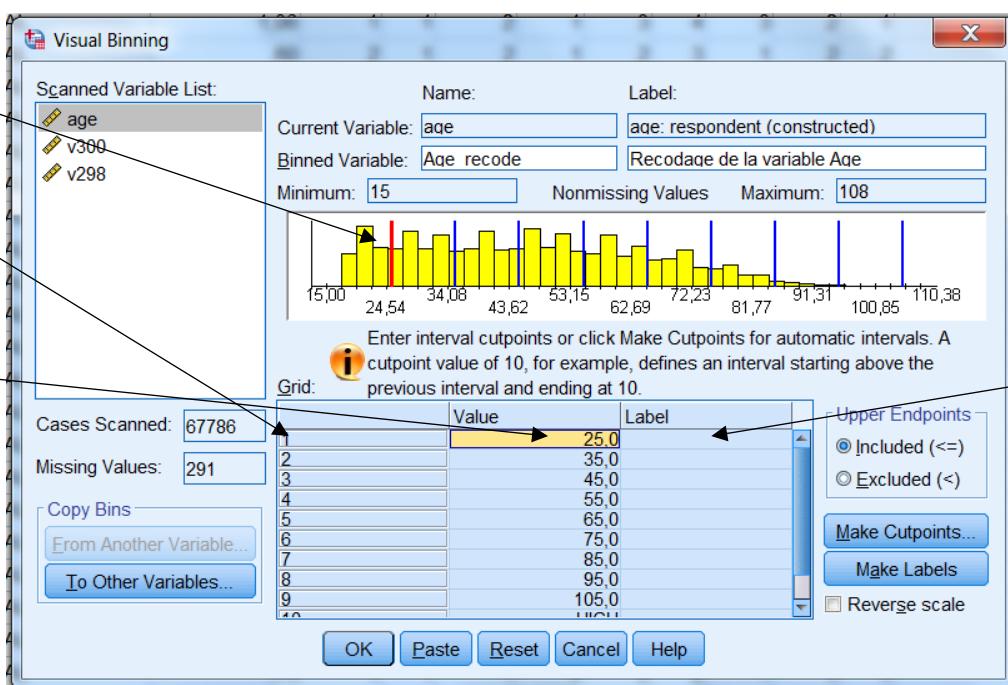


FIGURE 14 : VISUAL BINNING APRÈS DÉCOUPE

Une fois le découpage de classes effectué (peu importe la méthode), SPSS vous affichera une représentation visuelle du découpage au niveau de l'histogramme (voir Figure 14). En dessous de celui-ci, vous verrez apparaître les valeurs de découpe de chacune des découpes. La valeur indiquée étant toujours la borne

supérieure de la classe⁶. Vous pouvez dès lors dans la colonne *Label* ajouter un libellé aux différentes classes nouvellement créées.

Utilities

Define variable sets (Définir des jeux de variables)⁷: Afin de réduire le nombre de variables visibles et ainsi améliorer votre gestion de la base de données⁸, vous pouvez créer un jeu de variables. Ce groupe de variable a besoin d'un nom. Ce nom importe peu, il n'est utile que pour votre utilisation personnelle. Une fois que le jeu est créé, vous devez lui attribuer les variables que vous désirez conserver pour vos analyses.

Si votre étude comporte plusieurs volets (ex : opinion politique ; pratiques de consommation ; opinion sur les minorités ; etc.), vous pouvez créer plusieurs jeux de variables regroupant les variables en relation directe avec vos volets. Vous pourrez ainsi travailler sur chaque volet sans être « envahi » par les variables non en relation avec ce volet précis.

Use variable sets (Utiliser des jeux de variables) : Afin d'utiliser le jeu, vous devez utiliser le jeu de variable fraîchement créé. Par défaut SPSS sélectionne toutes les variables (*ALL VARIABLES*). N'oubliez pas de désélectionner cette option, sans quoi SPSS affichera toutes les variables assez logiquement. L'option « *NEW VARIABLES* » concerne les variables que vous allez créer. Si vous désélectionnez cette option, SPSS ne montrera donc pas les nouvelles variables. Créer ou utiliser un jeu de variables ne supprimera jamais de données dans la base de données. Pour afficher à nouveau l'ensemble des variables, cocher *ALL VARIABLES* à nouveau dans la même option.

Notez que si vous venez d'encoder manuellement les données dans SPSS, toutes les variables sont considérées comme nouvelles variables puisque créées durant la session.

Traitement des données manquantes

L'option **Replace missing values** (Remplacer des valeurs manquantes) dans l'onglet **transformer** permet de remplacer des valeurs manquantes présentes dans une variable automatiquement. Plusieurs méthodes sont proposées : moyenne, médiane, tendance linéaire.

Cependant, avant de remplacer des valeurs manquantes, il est toujours bon d'essayer de comprendre pourquoi il y a des valeurs manquantes, surtout lorsqu'elles sont en nombre important (la question était-elle compréhensible ? Toutes les modalités de réponses étaient-elles présentes ? La question était-elle correctement formulée ? Les possibilités de réponses étaient-elles adaptées ? Ai-je bien ciblé ma population ? etc.). Il est également important de cerner que la modification des données manquantes, quand bien même via l'utilisation d'algorithmes complexes peut amener à une distorsion de la réalité. Il n'est donc pas conseillé d'imputer des valeurs aux données manquantes lorsque celles-ci sont présentes en grand nombre.

- ✓ *Une variable possède 2 valeurs manquantes parmi les 3000 observations. Je remplace la valeur manquante par la médiane.*
- ✗ *Remplacer les valeurs manquantes par la moyenne alors que ma variable comporte 2532 valeurs manquantes parmi 3000 observations.*

Une fois que cette première étape a été établie, on peut prendre une décision quant aux données manquantes présentes dans la base de données. Les principaux traitements étant soit d'exclure les données manquantes, soit de les modifier. Il n'existe cependant pas de guideline bien précis pour le traitement des données manquantes, chaque situation étant spécifique. Notez que dans la plupart des procédures de sorties de résultats, les données manquantes seront exclues des analyses et que les données manquantes peuvent compliquer le traitement de certaines données.

⁶ La première classe étant toujours par défaut de la valeur la plus basse jusqu'à la valeur renseignée et les autres classes commençant à la valeur de la classe antérieure jusqu'à la valeur renseignée.

⁷ Notez qu'il n'existe pas de génération de code de syntaxe pour la constitution des groupes de variables malheureusement. Par contre, ceux-ci seront sauvegardés si vous enregistrez la base de données.

⁸ Sachant qu'on travaille rarement sur toutes les variables d'une base de données à la fois.

CHAPITRE 3 : Eléments d'analyse statistique descriptive

L'analyse descriptive est une étape incontournable de n'importe quelle étude statistique. Elle permet de dresser un état des lieux de la base de données. Cependant, aussi simple que peut paraître à premier abord l'analyse descriptive, celle-ci n'est toujours aisée car elle peut s'avérer longue.

Nous allons donc vous présenter les principales méthodes d'analyse descriptive fonction par fonction. Comme vous vous en douterez, les principales fonctions d'analyse descriptive se retrouvent dans le menu **Analyze** (Analyse) et dans le sous-menu **Descriptive Statistics** (Statistiques descriptives). Enfin, les graphiques peuvent également servir d'analyse descriptive visuelle, bien que sélectionnables dans les différentes procédures d'analyse, ceux-ci sont également disponibles dans le menu **Graphs** (graphiques).

Lors de n'importe quelle procédure d'analyse, il faut toujours rester vigilant au type de variable à analyser (nominale, ordinaire, quantitative, dichotomique, etc.). Nous n'utiliserons en effet pas les mêmes indicateurs pour chacun des types de variable. De ce fait, nous allons commencer par dresser un bref rappel des différentes statistiques utilisées dans le cadre de l'analyse descriptive.

Bref rappel de statistiques de base

Analyse descriptive des variables à caractère qualitatif :

Notions de statistique nécessaires à l'analyse des variables à caractère qualitatif :

- ✓ Effectif : Nombre d'occurrences.
- ✓ Pourcentage : Rapport entre l'effectif observé et un total d'effectifs défini.
- ✓ Mode : Modalité la plus fréquente.

Ce qui va nous intéresser dans l'analyse des variables qualitatives va être tout simplement la répartition des réponses au sein de chacune des modalités de réponse. On va donc essentiellement se centrer autour des notions d'effectif et de pourcentage.

Bien faire attention à lier l'effectif aux pourcentages. Si, en effet, les pourcentages facilitent la lecture et la compréhension des informations contenues au sein des variables analysées, un pourcentage de 50% n'aura pas la même importance interprétative sachant que l'effectif est de 20 ou de 1000 individus.

Analyse descriptive des variables à caractère quantitatif :

Les notions de statistique nécessaires à l'analyse des variables à caractère quantitatif sont réparties entre les indices de tendance centrale, les indices de dispersion et optionnellement, les indices de normalité.

Les indices de tendance centrale :

- ✓ Moyenne : $\bar{X} = \frac{\sum x_i}{n}$
- ✓ Médiane : Valeur qui sépare l'échantillon en deux groupes de tailles égales.
- ✓ Mode : modalité la plus fréquente. Un indice plutôt utilisé dans l'analyse des variables à caractère qualitatif.

Les indices de dispersion :

- ✓ Variance : $s^2 = \frac{\sum(x_i - \bar{x})}{n-1}$
- ✓ Écart-type : $s = \sqrt{s^2} = \sqrt{\frac{\sum(x_i - \bar{x})}{n-1}}$
- ✓ Étendue : écart entre valeur minimale et valeur maximale
- ✓ Quartiles / Déciles / Percentiles : Valeurs qui séparent l'échantillon en 4/10/x groupes de tailles égales. On nomme généralement les quartiles comme suit : Q1 ou Q25 ; Q2 ou Q50 ou médiane ; Q3 ou Q75. Les percentiles sont notés P.

- ✓ Ecart interquartile (mesure de dispersion/d'asymétrie) : $EI = Q_3 - Q_1$: étendue entre le troisième et le premier quartile.

L'analyse des variables à caractère quantitatif va donc se centrer sur des indices. Ce qui va nous intéresser va être de définir la tendance centrale des variables concernées et de regarder comment les données varient autour de celle-ci. Au plus la variance et l'écart-type seront petits, au plus les données auront tendance à être homogènes, au plus ces indices vont être grands, au plus les données auront un caractère hétérogène.

Ainsi, imaginons que nous possédons l'ensemble des résultats des étudiants à un test de mathématiques, par exemple. Si la moyenne est à 52%, le fait d'avoir un écart-type de 1.7 ou de 15.2, va renseigner des profils très différents d'étudiants. Dans le premier cas, les étudiants auront tendance à avoir des résultats généraux très proches de la moyenne avec donc, une concentration d'étudiants très moyen ; alors que dans le second cas, les écarts à la moyenne seront beaucoup plus importants, impliquant des profils d'étudiants qui ont très bien réussi le test et d'autres étudiants qui auront échoué au test de manière flagrante.

Les indicateurs de normalité des données :

Les lois de probabilités statistiques permettent de décrire de manière théorique un phénomène aléatoire. L'une de ces lois de probabilité les plus utilisée est la loi de distribution normale (ou loi de Gauss). La loi normale stipule que la moyenne soit égale à la médiane et au mode et que la dispersion autour de cette tendance centrale se fait de telle sorte qu'à 1 écart-type au-dessus et en dessous de celle-ci, l'on retrouve 68% des observations, à 2 écart-types, l'on retrouve 95% des observations et à 3 écart-types, l'on retrouve 99% des observations.

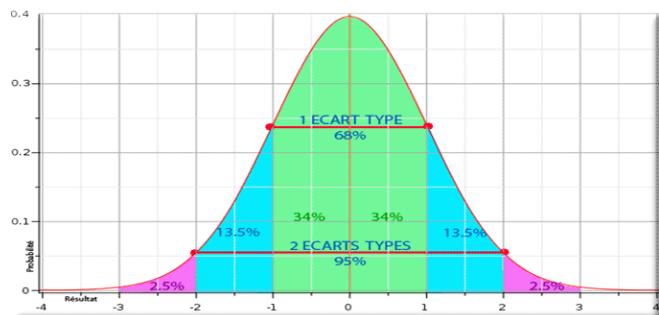


FIGURE 15 : COURBE DE DISTRIBUTION GAUSSIENNE

Il existe plusieurs indicateurs de normalité. Tout d'abord, il y a les indices de symétrie (Skewness) et d'aplatissement (Kurtosis). Comme présenté à la figure ci-contre, les données doivent suivre une courbe symétrique à gauche et à droite de la tendance centrale et avoir un certain nivelingement. Au plus ces indices vont se rapprocher de 0, au plus les données vont se rapprocher de la loi normale. Généralement, si le rapport entre la valeur de ces indices et leur erreur-type varie entre -2 et 2, on considérera qu'il n'y a pas de problèmes de normalité⁹.

Les indicateurs de normalité, en dehors de stipuler si les données vont suivre une courbe de Gauss (voir Figure 15), vont également nous être utiles pour savoir ultérieurement si des tests statistiques, requérant la distribution normale des données, seront applicables ou non.

Statistiques complémentaires

Dans la quasi-totalité des procédures de résultats (comprenez, les procédures reprises dans le menu *Analyze*), un tableau récapitulatif des données traitées pour l'analyse sera présent (comme par exemple, illustré par la Figure 16). Ainsi, les données valides représentent les données non-manquantes et utilisées dans l'analyse. Cette statistique peut être, pour certains tests, la somme des données non manquantes sur plusieurs variables¹⁰. A l'inverse, les données manquantes (Missing) comprennent le nombre d'observations dont l'information est tout simplement manquante (signalée par un . dans la vue des données) ainsi que les données renseignées comme manquantes dans les caractéristiques des variables¹¹.

⁹ Site de SPSS à l'UDES.

¹⁰ Par exemple, et nous le verrons ultérieurement, l'analyse en composantes principales exclura de l'analyse toute observation dont au moins une donnée est manquante sur l'une des variables renseignées dans l'analyse.

¹¹ Voir premier Chapitre.

Statistics		
sex respondent (Q86)		
N	Valid	67774
	Missing	12

FIGURE 16 : TABLEAU RÉCAPITULATIF DES DONNÉES UTILISÉES POUR L’ANALYSE

Ainsi dans ce tableau, issu d'une analyse de fréquences, 67774 observations ont été retenues pour l'analyse, alors que 12 données ont été mises de côté.

Il est toujours important de faire attention à ce tableau car un grand nombre d'observations manquantes non pris en compte lors de l'analyse peut mener à des erreurs d'interprétation, des extrapolations non-fondées, etc.

Le sous-menu Descriptive Statistics

Un certain nombre de procédures d'analyse descriptive existent dans SPSS, applicable généralement à chaque type de variable, pouvant ainsi semer le doute quant à la sélection de la procédure à appliquer pour décrire telle ou telle variable. Si chaque procédure possède ses avantages et ses inconvénients, nous conseillons à l'étudiant débutant d'utiliser la procédure de fréquences pour les variables à caractère qualitatif et les procédures « means » ou « explore » pour l'analyse des variables quantitatives. Souvenez-vous que les variables ordinaires peuvent, en certains cas, revêtir les deux statuts.

Frequencies (Fréquences) : Une analyse de fréquence est généralement utilisée pour des variables nominales ou ordinaires. Une analyse de fréquence va tout simplement fournir un tableau révélant les effectifs de chacune des valeurs présentes au sein de la variable. Renseignez les variables à explorer dans la colonne de droite. Si vous renseignez plusieurs variables, SPSS effectuera plusieurs fois l'analyse pour chaque variable, on reste donc bien dans de l'analyse univariée.

sex respondent (Q86)					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	male	30161	44,5	44,5	44,5
	female	37613	55,5	55,5	100,0
	Total	67774	100,0	100,0	
Missing	other missing	11	,0		
	no answer	1	,0		
	Total	12	,0		
Total		67786	100,0		

FIGURE 17 : TABLEAU DE SORTIE DE LA PROCÉDURE DE FRÉQUENCES

Le tableau décrit la répartition des données entre les différentes valeurs de la variable¹², le pourcentage et pourcentage valide (c'est-à-dire, sans tenir compte des données manquantes) que chacun de ces effectifs représentent, et enfin le pourcentage cumulé. Vous noterez que le tableau est divisé en deux : la description des données valides dans la partie supérieure et la description des données manquantes dans la partie inférieure.

Si cette procédure est essentiellement destinée aux variables de type qualitatif, SPSS est en mesure également de vous fournir différentes statistiques pour l'étude des variables à caractère quantitatif¹³ dans le menu d'option **statistics** : des indices de tendance centrale, de dispersion et de normalité des données.

Vous apercevrez que, par défaut, au sein de la fenêtre d'accueil de la procédure, l'option **Display Frequency tables** (Afficher la table de fréquences) est sélectionnée. Cette option est à utiliser de manière générale avec des variables nominales ou ordinaires. Si cependant, vous voulez connaître le contenu d'une variable sans avoir à regarder l'entièreté des données de la variable dans la vue des données, la table de fréquences est une

¹² Grâce aux labels associés aux valeurs (Chapitre 1), les valeurs 1 et 2 ont été remplacées par « male » et « female » dans le tableau de sortie.

¹³ Nous verrons cependant qu'il existe des procédures bien plus adaptées aux variables quantitatives.

bonne option. Lorsque vous utilisez cette procédure pour décrire des variables à caractère quantitatif, veillez à décocher cette option.

On ne s'intéressera en effet plus de savoir combien de personnes ont répondu avoir 34 ou 36 ans pour une variable d'âge par exemple mais bel et bien d'obtenir des indices de dispersion et de tendance centrale. De plus, le tableau risquera d'être très compliqué à lire, surtout avec des variables très précises au niveau quantitatif où beaucoup de valeurs risques de se retrouver isolées et le tableau kilométrique.

Il est également possible de demander à SPSS de produire des graphiques via le menu d'option **Charts**. Veillez à demander un diagramme en barre pour les variables qualitatives et un histogramme pour les variables quantitatives si vous désirez une production de graphe¹⁴.

Explore (Explorer) : Cette fonction est très intéressante pour la description des variables quantitatives. Elle vous fournit d'emblée les indicateurs suivants : moyenne, moyenne tronquée à 95% (calcul de la moyenne en retirant 5% des valeurs les plus basses et les élevées), médiane, écart-type, valeurs minimales et maximales, l'étendue, l'étendue interquartile, les deux indices de normalité des données : la symétrie (skewness) et l'aplatissement (kurtosis).

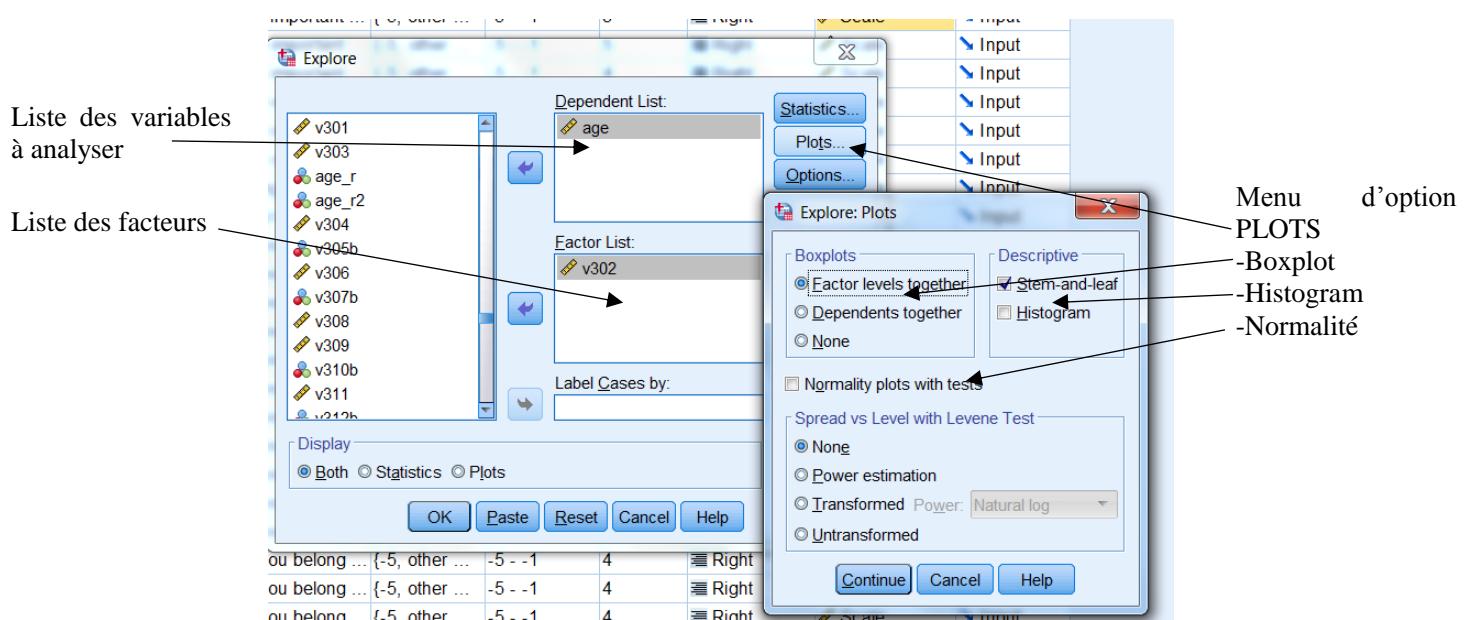


FIGURE 18 : FONCTION EXPLORE

La variable dépendante en statistiques est la variable d'analyse. Vous devez donc renseigner dans l'onglet *Dependent List* les variables à analyser. De base, la fonction explore, propose une analyse univariée des variables dépendantes. Cependant, il est possible de demander des analyses bivariées via l'onglet *Factor List*. Un facteur en statistiques, est un synonyme pour décrire une variable qualitative. Si un facteur est renseigné, SPSS va fournir des statistiques de la variable dépendante pour chacune des modalités du facteur.

Cela peut être utile pour analyser par exemple, les dispersions salariales en fonction du sexe, du niveau d'éducation ou autre.

Via l'option **Plots** (Diagrammes), vous pouvez également demander une série de graphiques (voir Figure 18) :

1. Le *Boxplot* (ou boîte à moustaches) : montre la répartition de la variable dépendante en fonction des quartiles. Les données symbolisées avec un rond, représentent des données fortement éloignées du reste des données (entre 1.5 et 3 écarts-interquartile du premier ou troisième quartile). Les données symbolisées par une étoile représentent les données extrêmes (situées au-delà de 3 écarts-interquartile du premier ou du

¹⁴ Pour plus d'informations sur les graphiques, aller directement à la section de ce chapitre destinée aux graphiques.

troisième quartile. Si vous avez renseigné un facteur, le graphique effectuera plusieurs boîtes à moustaches en fonction des modalités du facteur.

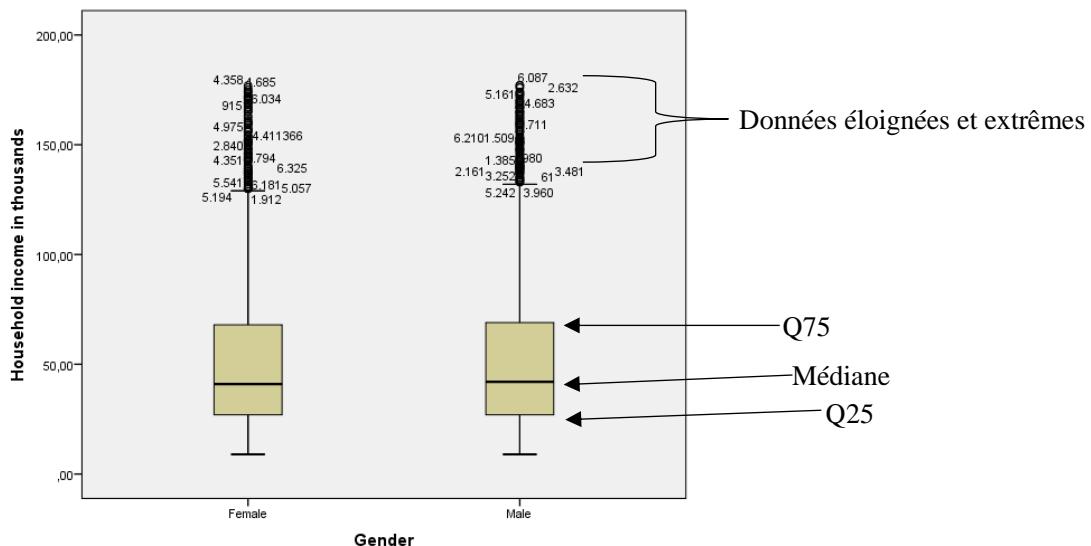


FIGURE 19 : BOXPLOT DU SALAIRE ANNUEL EN FONCTION DU GENRE

Les deux boîtes à moustaches se ressemblent, on peut donc déjà conclure que d'un point de vue descriptif, il n'y a pas de différences flagrantes de salaire entre les sexes au sein de l'échantillon a priori. Ensuite, on peut analyser la distribution en tant que telle : on peut remarquer que 50% des individus se concentre entre les valeurs les plus basses (~10 000 euros par an et la médiane ~40 000 euros par an). La distribution des 50% de données restantes se partage, elle, une étendue de salaire bien plus importante, allant de 40 000 euros à près de 175 000 euros par an, le dernier quartile (les données au-delà de Q75) indiquant une étendue encore plus importante. Ce graphe montre donc des différences flagrantes de salaires au sein de l'échantillon.

2. Un histogramme et un diagramme à feuilles (*Steam-and-Leaf*), sélectionné par défaut par SPSS. Vous pouvez décocher ce deuxième graphique, nous ne le verrons pas dans le cadre de ce cours.

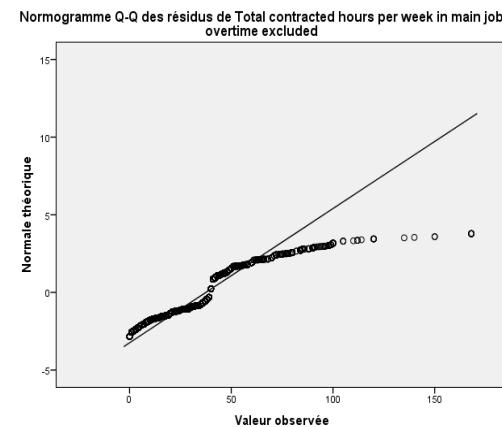
3. Des graphes de normalités, appelés *QQ plot* (ou tracé quartile-quartile) ainsi qu'un test de normalité associé via l'option *Normality plot with tests* (graphe de répartition gaussien avec test).

Le test de Kolmogorov-Smirnov mesure la distribution des observations et teste la normalité de la distribution de votre variable. Celui-ci teste l'hypothèse H_1 : à savoir que les données sont distribuées de façon non-normale. H_0 représente l'antithèse de H_1 : les données sont distribuées normalement. En statistique, un test est toujours lié à une p-valeur renseignant sur l'acceptabilité ou le rejet de H_1 . Une p-valeur ou significativité du test est une valeur oscillant entre 0 et 1 et indiquant le pourcentage de chance de se tromper d'affirmer que H_1 est vrai. En statistique, il existe plusieurs seuils d'acceptabilité : 0.001, 0.01 et le plus courant : 0.05. Cela signifie que si la p-valeur renseignée dépasse 0.05, on rejette H_1 et on accepte dès lors H_0 (autrement dit : une probabilité supérieure à 5% de se tromper est jugée trop risquée pour généraliser un propos à une population donnée). Ainsi, dans le cadre du test de Kolmogorov-Smirnov, on considérera les données comme normalement distribuées si la **significativité** (Sig.) associée au test dépasse 0.05.

Le tracé quartile-quartile dessine, via une ligne droite, la façon dont devrait être dispersées si elles suivaient une distribution normale. Si les données suivent parfaitement la diagonale du graphe, les données sont normales. Dans le cas contraire, il y a fort à parier que ce n'est pas le cas. Un deuxième tracé quartile-quartile vous est également fourni, vous montrant la distance des données en écart-type (Scores Z, aussi appelées données standardisées) à la droite de distribution normale.

Les outils de type qualitatif (estimations des tracés quartiles-quartiles, de l'aplatissement et de la symétrie) sont à utiliser en combinaison avec les outils plus déterminants comme le test de Kolmogorov-Smirnov.

Mauvaise distribution sur QQplot



Bonne distribution sur QQplot

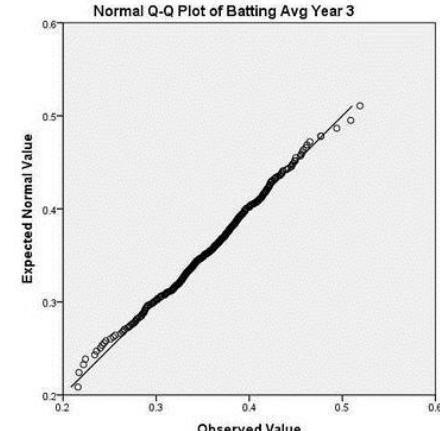


FIGURE 20 : TRACÉS QUARTILE-QUARTILE

Descriptives : Fonction assez basique permettant des statistiques destinées à l'analyse des variables quantitatives. Cette fonction fournit beaucoup moins d'information que la fonction *explore* et n'est donc idéale à utiliser. Cependant, plus rapide d'exécution, si vous n'avez pas besoin de peu d'informations, cette fonction conviendra parfaitement

Crosstable (tableau croisé) : Cette option permet de sortir une analyse bivariée de variables qualitatives. SPSS vous propose deux encadrés dans lesquels entrer les variables que l'on veut étudier. Vous pouvez les renseigner en ligne (row) ou en colonne (column). Le tableau étant le résultat de l'intersection des modalités des variables placées en colonne et en ligne. Dans l'onglet *cells*, vous pouvez demander à SPSS d'afficher les pourcentages en colonne et en ligne.

sex respondent (Q86) * education level (Q110) (recoded) Crosstabulation

	education level (Q110) (recoded)			Total
	Lower	Middle	Upper	
sex (Q86)male Count	8037	14751	7084	29872
% within sex respondent (Q86)	26,9%	49,4%	23,7%	100,0%
% within education level (Q110)	40,3%	47,0%	44,7%	44,5%
femaleCount	11930	16607	8756	37293
% within sex respondent (Q86)	32,0%	44,5%	23,5%	100,0%
% within education level (Q110)	59,7%	53,0%	55,3%	55,5%
Total Count	19967	31358	15840	67165
% within sex respondent (Q86)	29,7%	46,7%	23,6%	100,0%
% within education level (Q110)	100,0%	100,0%	100,0%	100,0%

FIGURE 21 : TABLEAU CROISÉ DU GENRE EN FONCTION DU NIVEAU D'EDUCATION

Les pourcentages en ligne et en colonne représentent des rapports d'effectifs au sein de la cellule étudiée en fonction d'un total particulier (total de la ligne ou de la colonne). Dans l'exemple ci-dessus, la première statistique montre la **distribution en effectif** au sein de chacune des cellules du tableau. Il y a donc 8037 hommes ont un degré d'éducation bas, 14751 un degré d'éducation moyen et 7084 ont un degré d'éducation élevé, pour un total 29872 hommes, etc. Le pourcentage en ligne est indiqué **en rouge** dans le tableau. Il s'agit d'une traduction des effectifs en pourcentage en fonction du total de la modalité en ligne prise en compte. Soit, dans l'exemple ci-dessus, 26,9% est le rapport entre 8037 (le nombre d'hommes ayant un degré d'éducation faible) et le nombre total d'hommes : 29872. Le pourcentage en colonne est représenté **en vert** dans le tableau ci-dessus. Considérant la même cellule, il s'agit du rapport entre les 8037 hommes ayant un degré d'éducation faible et l'ensemble des personnes ayant un degré d'éducation faible (19967). Dit autrement : parmi les individus ayant un degré d'éducation faible, il y a 40,3% d'hommes.

Si jamais vous n'arrivez pas à vous souvenir quel pourcentage est celui en ligne ou en colonne, notez que le pourcentage est dénommé de la façon suivante : '% within' suivi d'un nom de variable. Il vous suffit dès lors de repérer dans le tableau si vous avez renseigné votre variable en ligne ou en colonne.

Notez également que pour dénombrer le pourcentage total d'une modalité, on regardera le pourcentage total inverse (ligne si la variable est située en colonne, ou en colonne si la variable est renseignée en ligne). Si on reprend la Figure 21, le pourcentage total d'homme est bien défini par le pourcentage en colonne car il s'agit du rapport entre le total de la modalité (29872) et le total du tableau (67165) représentant ainsi 44.5% d'hommes.

QQ PLOT & PP PLOT : SPSS vous permet également de réaliser des tracés quartile-quartile ou percentile-percentile afin de mesurer la normalité de vos données en dehors de la fonction *Explore*.

La procédure means : Afin de produire des statistiques descriptives bivariées sur des variables quantitative et qualitative de manière rapide, vous pouvez également utiliser la procédure « Means ». La procédure produira des statistiques en fonction des variables croisées. Celle-ci ne se trouve pas dans le sous-menu « Descriptive statistics » mais au sein du sous-menu « Compare Means ». Vous renseignerez ainsi une (ou plusieurs) variable quantitative d'étude au sein de l'onglet « Dependent list » et des variables catégorielles au sein de l'onglet « Independent list ». Vous pourrez ensuite sélectionner les statistiques désirées au sein dans les *Options*. Par défaut, SPSS affichera la moyenne, les effectifs et l'écart-type. Vous obtiendrez un résultat identique à la Figure 22.

Report			
Age in years			
Job satisfaction	Mean	N	Std. Deviation
Highly dissatisfied	28,23	506	13,676
Somewhat dissatisfied	35,75	504	15,667
Neutral	42,78	462	16,966
Somewhat satisfied	47,68	453	15,962
Highly satisfied	51,76	283	14,484
Total	40,00	2208	17,469

FIGURE 22 : OUTPUT DE LA PROCÉDURE MEANS : L'ÂGE EN FONCTION DU DEGRÉ DE SATISFACTION AU TRAVAIL

Création de graphique

SPSS vous donne également la possibilité de créer vos propres graphiques via le menu *Graphs*. Si plusieurs procédures sont disponibles dans ce menu, nous allons seulement nous arrêter au sous-menu *Legacy Dialogs* (Boîte de dialogues). Dans ce sous-menu, une série de graphiques différents sont disponibles. Il est important de noter que pour chaque type de graphique correspond un type et un nombre de variable déterminé. Nous vous conseillons généralement d'utiliser les graphiques de la manière suivante (illustrés à la Figure 24):

- ✓ Représenter visuellement une variable qualitative : un diagramme en barres simples (Bar/simple)
- ✓ Représenter visuellement une variable quantitative : un histogramme
- ✓ Représenter deux variables qualitatives : un diagramme en barres juxtaposées (Bar/Clustered) ou un tracé de ligne (Line/Multiple).
- ✓ Représenter une variable qualitative et quantitative : une boîte à moustache (Boxplot/Simple)
- ✓ Représenter deux variables quantitatives : un nuage de points (Scatter/dot / Simple)

Au sein de chacune des procédures de graphiques, il existe une fonctionnalité qui permet de créer des matrices de graphiques qui est représentée par l'encadré *Panel By*. En fonction, de variables à modalités renseignées dans les onglets lignes ou colonnes, SPSS va sortir un graphique spécifique aux observations rencontrant les modalités de réponse des variables renseignées en ligne et en colonne (à l'instar des tableaux croisés).

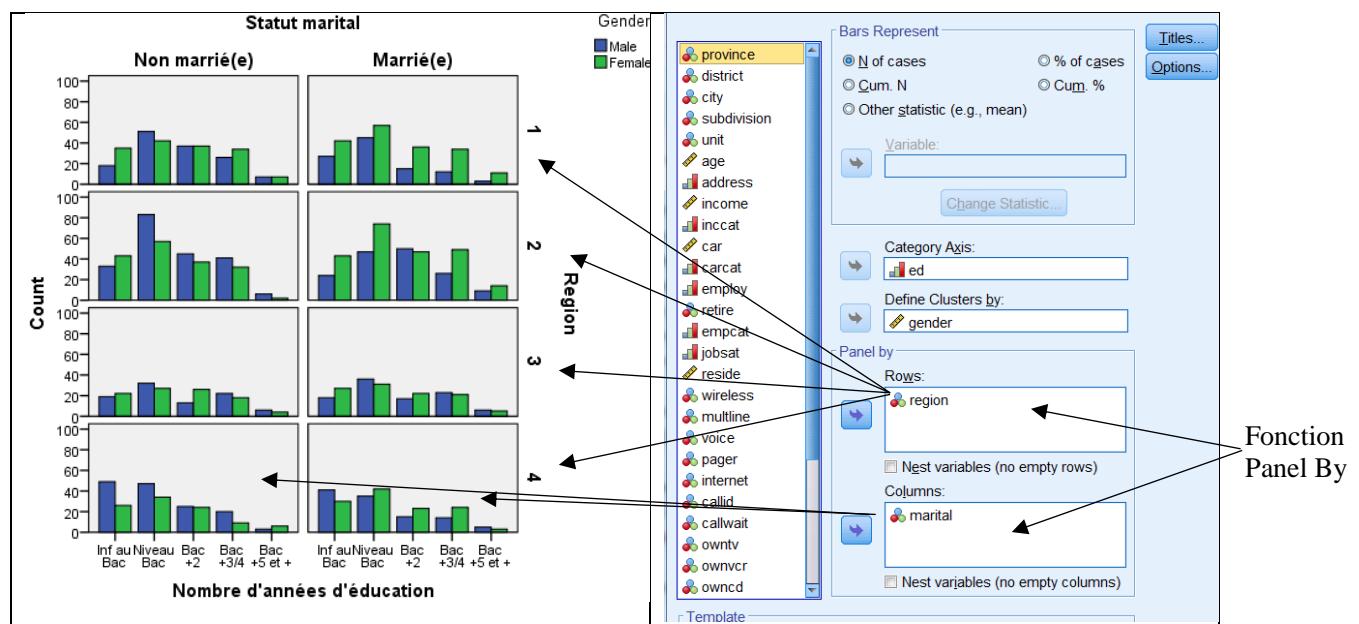
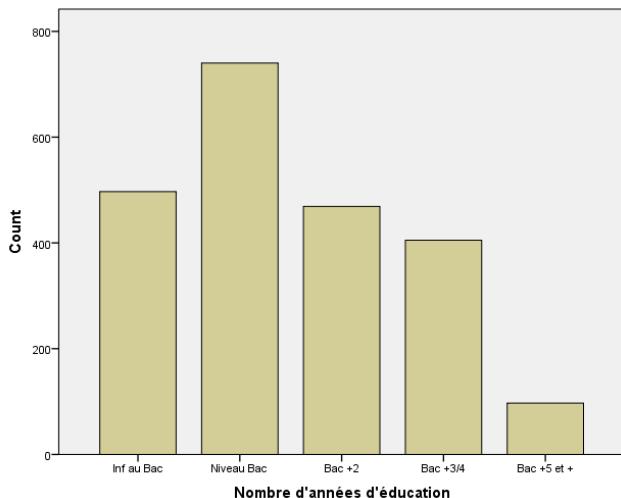


FIGURE 23 : CRÉATION DE MATRICE DE GRAPHIQUE

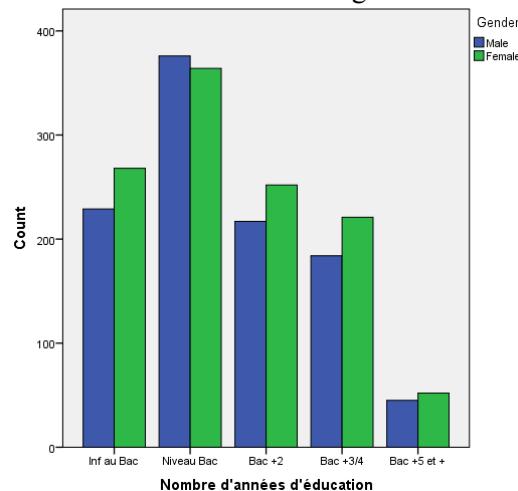
Au sein de la procédure Scatter/dot, il vous est également possible d'attribuer des couleurs aux observations en fonction d'une variable qualitative via l'onglet *Set Markers by* (Définir les marques par).

Enfin, il vous est possible en double-cliquant sur n'importe quel graphique de personnaliser vos graphiques à votre guise : changer le titre des axes, la couleur de fond, la couleur des barres, la police d'écriture, etc.

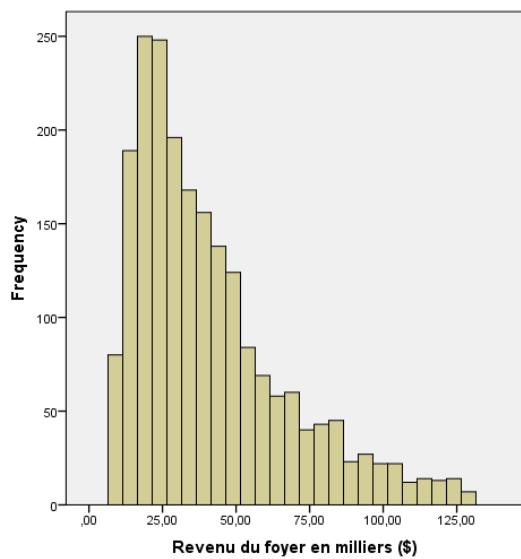
1. Diagramme en barres simples : niveau d'éducation



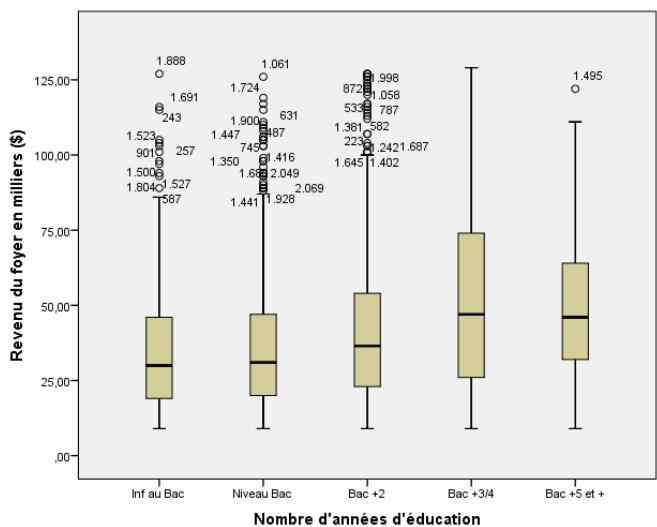
2. Diagramme en barres juxtaposées : niveau d'éducation en fonction du genre



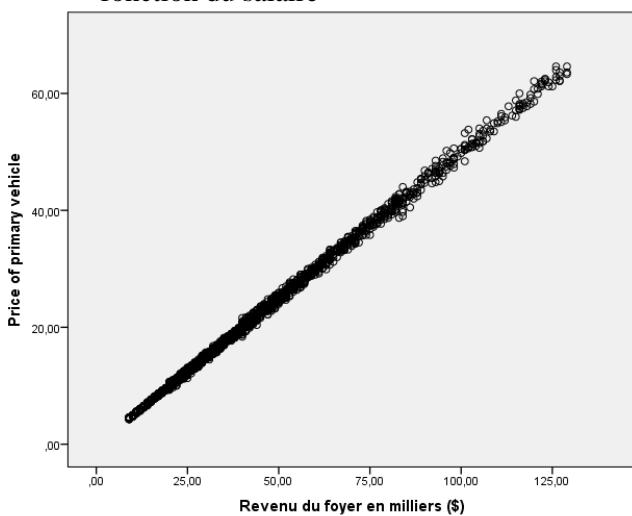
3. Histogramme : salaire annuel perçu



4. Boxplot : revenu annuel en fonction du niveau d'éducation



5. Scatter/Dot : Prix du véhicule principal en fonction du salaire



6. Tracé en lignes : Niveau d'éducation en fonction de la tranche salariale

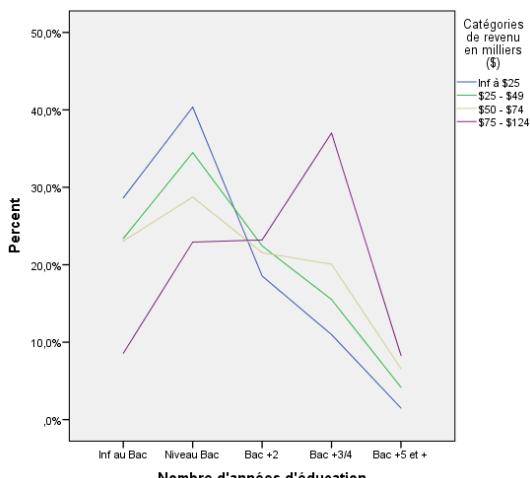


FIGURE 24 : DIFFÉRENTS TYPES DE GRAPHIQUE

Faites attention, si vous ne respectez les suggestions que nous vous faisons ci-dessus et que vous faites une mauvaise utilisation des graphiques en fonction des types de variables, vous risquez d'obtenir des graphiques incompréhensibles comme ceux-ci :

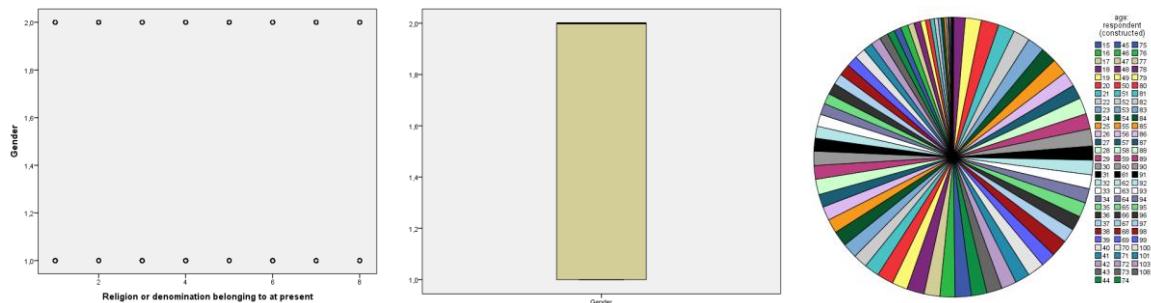


FIGURE 25 : MAUVAISE UTILISATION DES GRAPHIQUE EN FONCTION DES TYPES DE VARIABLES

D'un point de vue pratique, nous vous recommandons également d'éviter les graphiques dont l'information n'est pas directement perceptible pour le lecteur tels que des graphiques en tartes, en barres pilées et surtout, les graphiques en trois dimensions. Un graphique est avant tout une aide visuelle et doit donc être facile à lire et à comprendre.

Edition des graphiques

Vous constaterez qu'il y a une panoplie d'options disponibles dans le menu de paramétrage des graphiques sur lesquels nous n'allons pas nous arrêter dans le cadre de ce cours. Notez toutefois, et cela nous sera utile ultérieurement, qu'il est possible d'ajouter des axes horizontaux et verticaux au graphique via les options montrées à la Figure 26. Ceux-ci peuvent en effet faciliter la lecture des nuages de points par exemple.

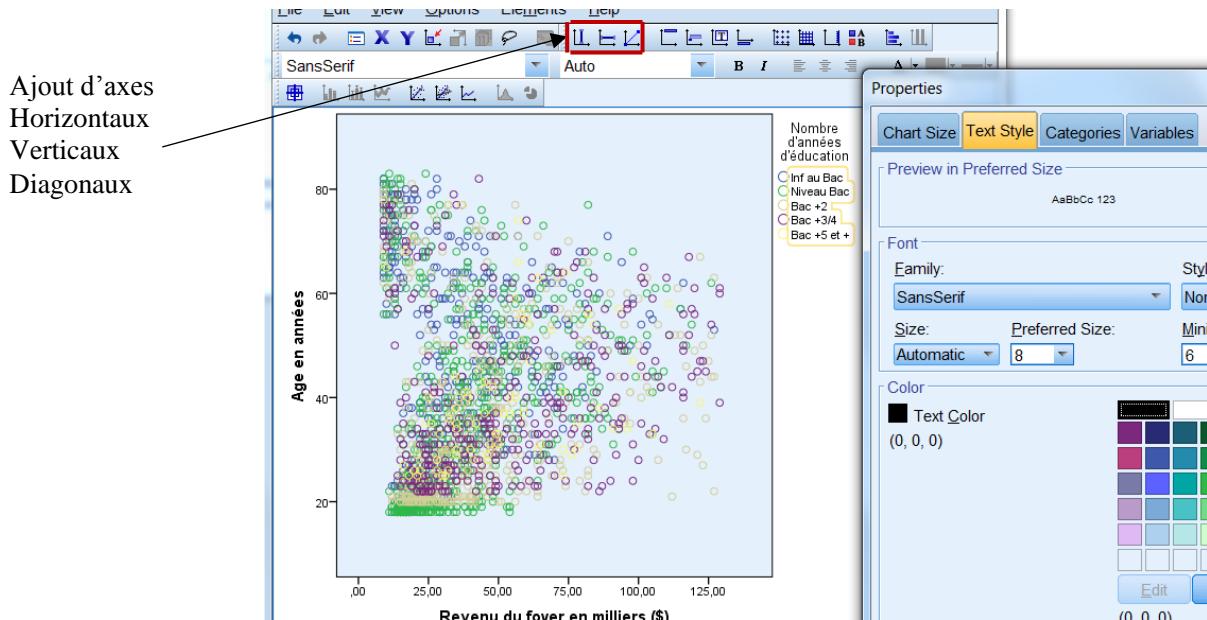


FIGURE 26 : PARAMÉTRAGE D'UN NUAGE DE POINTS AVEC COLORATION DES POINTS (MARKERS BY).

CHAPITRE 4 : Les tests inférentiels bivariés

Un test inférentiel permet de généraliser un phénomène observé au sein de l'échantillon à la population. Un test inférentiel (dans le sens entendu de modèle statistique) est toujours multivarié et va tenter de cerner les effets de variables sur d'autres variables. Dans le cadre de ce cours, nous verrons essentiellement des tests inférentiels bivariés. En d'autres termes, tester si une relation observée entre deux variables à partir d'un échantillon est due au hasard ou non. Les hypothèses associées sont spécifiques à chaque test, dépendant ainsi du paramètre testé. Cependant, on peut résumer les hypothèses de la façon suivante : H_0 : les paramètres testés sont égaux ; H_1 : les paramètres testés sont inégaux.

Pour rappel, chaque test est associé à une p-valeur (ou significativité du test). Cette statistique varie entre 0 et 1 et renseigne sur l'acceptabilité de H_1 . Il s'agit donc d'un pourcentage d'erreur. En statistique, il existe plusieurs seuils d'acceptabilité : 0.001, 0.01 et le plus courant : 0.05. Cela signifie que si la p-valeur renseignée dépasse 0.05, on rejette H_1 et on accepte dès lors H_0 .

Dit autrement : imaginons une p-valeur de 0.02. Cela signifie que j'ai 2% de chance de me tromper de dire que H_1 est vrai. A un seuil de 0.05, j'accepte H_1 alors qu'à un seuil 0.01, je rejette H_1 . Tout dépend de la marge d'erreur que je suis prêt à accepter dans mon étude. La fixation des seuils d'acceptabilité varie bien souvent en fonction de la taille d'échantillon. Si je veux minimiser ma marge d'erreur, je dois augmenter mon échantillon et donc augmenter les frais d'une enquête.

Distribution d'effectifs : Le test du Khi-carré

Rappel Théorique

Le test du Khi-carré est un test statistique inférentiel bivarié tentant de mesurer l'association entre des variables de type qualitatif, associé dans SPSS à la procédure des tableaux croisés (*Crosstables*). Ce test a pour hypothèses :

H_0 : absence de relation entre les deux variables (distribution égale des effectifs entre les cellules)
 H_1 : relation significative entre les deux variables (distribution inégale des effectifs entre les cellules)

Ce test est fondé sur la comparaison statistique entre les effectifs observés – E.O. (effectifs présents dans l'échantillon) et les effectifs théorique – E.T. (Distribution théorique des effectifs si les variables n'avaient pas d'effet l'une sur l'autre). Le calcul du Khi-deux s'effectue de la manière suivante :

$$\chi^2 = \sum \frac{(E.O. - E.T.)^2}{E.T.}$$

Imaginons un tableau croisé, combinant les informations collectées sur le genre et le niveau de salaire des individus dont les effectifs seraient distribués de la façon suivante :

Effectifs observés		Niveau de salaire		Total
		Elevé	Faible	
Genre	Femmes	60	60	120
	% colonne	50%	75%	60%
	Hommes	60	20	80
	% colonne	50%	25%	40%
Total		120	80	200

L'échantillon comprend 40% d'hommes et 60% de femmes. Si on n'observait pas de différences salariales en fonction du genre, on devrait logiquement conserver les proportions 60/40 à l'intérieur des deux modalités de niveau de salaire. Les effectifs théoriques devraient donc se distribuer de la façon suivante :

Effectifs théoriques		Niveau de salaire		Total
		Elevé	Faible	
Genre	Femmes	72	48	120
	% colonne	60%	60%	60%
	Hommes	48	32	80
	% colonne	40%	40%	40%
	Total	120	80	200

La valeur du χ^2 est obtenue à partir des écarts entre E.T. et E.O. au sein de chacune des cellules :

$$\begin{aligned}\chi^2 &= \sum \frac{(EO - ET)^2}{ET} = \frac{(60 - 72)^2}{72} + \frac{(60 - 48)^2}{48} + \frac{(60 - 48)^2}{48} + \frac{(20 - 32)^2}{32} \\ &= \frac{144}{72} + \frac{144}{48} + \frac{144}{48} + \frac{144}{32} = 2 + 3 + 3 + 4.5 = 12.5\end{aligned}$$

Au plus les écarts entre E.T. et E.O. vont tendre vers 0, au plus on va avoir tendance à accepter H_0 et à prôner qu'il n'existe pas de relations entre ces deux variables. A l'inverse, Au plus les écarts entre E.T et E.O. vont être élevés, au plus H_1 risque d'être acceptée.

Une p-valeur est associée au test afin de pouvoir évaluer la grandeur de la statistique du Khi-carré. Comme déjà expliqué précédemment, une p-valeur est une statistique variant entre 0 et 1, indiquant le pourcentage de se tromper d'accepter H_1 . Cette p-valeur ou significativité du test est retrouvable dans une table statistique du Khi-deux. Il existe différents seuils d'acceptabilité en statistiques, nous retiendrons, dans le cadre de ce cours, un seuil d'erreur maximal de 5%. Si la significativité du test est supérieure à ce seuil, nous rejeterons alors H_1 et nous conserverons H_1 dans le cas contraire.

La significativité du X^2 est directement calculée par SPSS. Si nous devions retrouver cette p-valeur manuellement, nous devrions la retrouver dans une table statistique. Dans cette table, deux informations sont importantes à regarder :

- En colonne : le degré de liberté (dl) est calculé comme (le nombre de modalités en colonne – 1), multiplié par (le nombre de modalités en ligne – 1). Dans ce cas :

$$dl = (2 - 1) * (2 - 1) = 1$$
- En ligne : le seuil (α) de p-valeur. Si la valeur du X^2 pour le degré de liberté associé est supérieure à la valeur correspondante dans la table, cela signifie que l'on peut accepter H_1 .

dl	0,05	0,02	0,01	0,001
1	3,841	5,412	6,635	10,827
2	5,991	7,824	9,210	13,815
3	7,815	9,837	11,345	16,266
4	9,488	11,668	13,277	18,467
5	11,070	13,388	15,086	20,515

Dans notre cas, afin d'avoir une erreur maximale de 5%, il faudrait que la valeur du Khi-carré soit supérieure à 3,841 ; avec un seuil d'erreur maximale de 1%, il faudrait que la valeur du Khi-carré soit supérieure à 6,635, etc. Nous avons donc dans cette situation précise très peu de chance (<0,001) d'émettre une erreur d'affirmer qu'il existe des disparités salariales en fonction du genre.

Cela signifie donc que même avec une probabilité de 0.1% de se tromper, on peut rejeter l'hypothèse nulle. Autrement dit, on peut affirmer avec une probabilité d'erreur inférieure à 0.1% de se tromper qu'il y a une association entre niveau de salaire et genre.

Afin de pouvoir appliquer un test du Khi-carré, trois conditions doivent être respectées :

- 1) Le tableau peut contenir un maximum de 20% de cellules (cellules de totaux exclues), ayant un effectif théorique inférieur à 5.
- 2) Aucune cellule du tableau ne peut contenir d'effectif théorique inférieur à 1.
- 3) Indépendance entre les variables

Si jamais ces conditions ne sont pas remplies, cela signifie que certaines modalités de vos variables sont représentées par très peu d'individus. Deux solutions s'offrent alors à nous : renoncer au test ou fusionner des modalités (cf. Chapitre 2) afin d'effacer le problème d'effectif. Si vous optez pour la deuxième option, faites attention de faire des regroupements qui ont du sens.

Si le calcul du chi-carré et la significativité associée peuvent prétendre affirmer ou infirmer le fait que deux variables soient en relation, cela ne nous renseigne pas sur la force de cette relation. En effet, le calcul du Khi-carré étant basé sur les écarts entre E.T. et E.O. mis au carré, lorsque l'on a une base de données importante, il sera aisément d'obtenir des valeurs de Chi-carré significatives à des seuils d'erreur très bas alors que dans l'absolu, les écarts ne seront pas grands.

Afin de mesurer cette force de relation, nous utiliserons une statistique supplémentaire appelée le V de Cramer. Il existe cependant d'autres mesures d'associations que celle-ci applicables à des situations particulières (données pairées par exemple).

Le V de Cramer se calcule de la manière suivante : $V = \sqrt{\frac{\chi^2}{N * \min(r-1, c-1)}}$ où $\min(r-1, c-1)$ représente la valeur minimale entre le nombre de colonne ou de ligne diminuée de 1. Lorsque vous demandez la statistique du V de Cramer à SPSS, celui-ci fournit également la statistique Phi (Φ). Le coefficient Φ mesure l'association entre deux variables mais pour des tableaux de dimension de 2x2. Le Φ est calculé de la façon suivante: $\Phi = \sqrt{\frac{\chi^2}{N}}$.

Techniquement, vous l'aurez compris, le Phi et le V de Cramer ont la même formule à l'exception que l'énoncé $\min(r-1, c-1)$ est remplacé par 1 dans le calcul du Phi (Vu que $r-1=1$ et $c-1=1$). C'est pourquoi, la plupart du temps, les résultats sont identiques entre les deux tests. On ne vous demande pas d'être des experts des indicateurs d'association. Retenez de toujours regarder le V de Cramer, vu que le résultat sera identique au Phi dans le cas d'un tableau de dimension 2x2.

La statistique du V de Cramer (et du Φ) variera entre 0 et 1. Au plus la statistique sera proche de 1, au plus la force de la relation sera forte (1 étant une relation parfaite). À l'inverse, au plus la statistique sera proche de 0, au plus la relation sera considérée comme insignifiante. Afin de vous aider à décoder cette statistique, voici un baromètre sur lequel les chercheurs en sciences sociales s'accordent :

- Autour de 0.1 : Faible relation
- Autour de 0.3 : Relation modérée
- Autour de 0.5 : forte relation

*Dans notre exemple, le V de Cramer, tout comme le Phi, est égal à : $\sqrt{\frac{12,5}{200*1}} = 0.25$, l'effet du genre sur le salaire est donc modéré. Cela signifie que si, en effet, il y a une tendance à ce que les femmes gagnent moins que les hommes, le fait d'appartenir à un des deux genres n'est pas une condition suffisante que pour expliquer la tranche salariale.*

Manipulations dans SPSS

Pour réaliser un test du Chi-carré, effectuez un tableau croisé (Cf. Chapitre 3) et sélectionnez dans le sous-menu *Statistics*, les options *Chi-square* et *Phi and Cramer's V*. Si vous le désirez, vous pouvez également demander à SPSS d'afficher les effectifs théoriques. Pour cela, vous devez vous rendre dans le sous-menu *Cells*, et vous devez sélectionner *Expected Counts*. N'oubliez pas de demander les pourcentages pour en ligne (et en colonne éventuellement) pour l'analyse du tableau.

Exemple : Sélectionnez les observations pour la Belgique. Croisez ensuite les variables V5 (*How important is in your life : politics*) et V336_r (*education level recoded*). Une fois l'analyse exécutée dans SPSS, on obtient les résultats suivants :

Tableau croisé: education level * how important in your life: politics

		how important in your life: politics				Total
		very important	quite important	not important	not at all important	
education level	Lower Effectif	25	87	177	195	484
	% education level	5,2%	18,0%	36,6%	40,3%	100,0%
	% important: politics	25,8%	22,4%	29,9%	45,6%	32,2%
Middle Effectif		30	128	214	164	536
	% education level	5,6%	23,9%	39,9%	30,6%	100,0%
	% important: politics	30,9%	32,9%	36,2%	38,3%	35,6%
Upper Effectif		42	174	200	69	485
	% education level	8,7%	35,9%	41,2%	14,2%	100,0%
	% important: politics	43,3%	44,7%	33,8%	16,1%	32,2%
Total	Effectif	97	389	591	428	1505
	% education level	6,4%	25,8%	39,3%	28,4%	100,0%
	% important: politics	100,0%	100,0%	100,0%	100,0%	100,0%

L'analyse des pourcentages en ligne montre que l'intérêt pour la politique est plus marqué selon le niveau d'éducation des répondants. L'analyse des pourcentages en colonne montre quant à elle que les classes les plus éduquées sont celles qui représentent le plus l'intérêt à la politique. Nous émettons alors l'hypothèse qu'il y a une relation entre l'intérêt pour la politique et le niveau d'instruction des individus.

Passons maintenant à l'analyse du tableau du Khi-deux. Vous remarquerez qu'en dessous de ce tableau, il y a une annotation. Celle-ci nous renseigne si les conditions ont été remplies. Nous avons donc 0% de cellules avec un effectif théorique inférieur à 5 et aucun effectif théorique inférieur à 1. Les conditions sont remplies, nous pouvons analyser le test.

Tests du khi-deux

	Valeur	ddl	Sig. asymptotique (bilatérale)
khi-deux de Pearson	97,458 ^a	6	,000
Rapport de vraisemblance	101,634	6	,000
Association linéaire par linéaire	79,572	1	,000
N d'observations valides	1505		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 31,19.

Le test du χ^2 nous permet de vérifier notre hypothèse, via l'analyse de la significativité (p-valeur) associée au test. Celle-ci étant de 0,000, elle nous indique que nous avons moins de 0,001% de chance de nous tromper en affirmant qu'il y a bel et bien une relation significative entre le niveau d'instruction et l'intérêt porté à la politique.

Le deuxième tableau indique les statistiques pour le Phi et le V de Cramer. Il est important ici de regarder la valeur de la statistique et non la p-valeur associée, cette dernière étant celle calculée pour le test du Khi-deux.

Mesures symétriques

	Valeur	Signification approx.
Nominal par Nominal	Phi ,254	,000
	V de Cramer ,180	,000
N d'observations valides	1505	

Le V de Cramer nous indique que cette relation est de faible ampleur. Le niveau d'éducation n'est donc pas le déterminant fondamental de l'intérêt porté à la politique bien que ce n'est pas un hasard si les personnes plus éduquées s'intéressent en général plus à la politique en général.

Comparaison de moyennes : le test-t

Rappel Théorique

Le test-t de Student est un test statistique inférentiel qui permet de mesurer si **deux** moyennes issues d'échantillons (ou de sous-groupes) différents proviennent d'une même population. Les types de variables indiqués pour ce test sont donc une variable quantitative et une variable qualitative (dichotomique). Il existe plusieurs types de Test-t : le test pour échantillon unique et pour échantillons indépendants¹⁵. Le test-t à échantillon unique va comparer la moyenne de l'échantillon à une moyenne d'une population connue mais dont l'écart-type est inconnu. Le test-t à échantillons indépendants va comparer deux moyennes au sein d'un même échantillon. Ces différents test-t ont cependant les mêmes hypothèses :

H_0 : les moyennes ne sont pas significativement différentes.

H_1 : les moyennes sont significativement différentes.

Le calcul du test-t se base sur la statistique t qui est la résultante du quotient du différentiel de moyenne par l'erreur-type de la moyenne. Ce calcul est cependant sensiblement différent selon que l'on utilise un test-t à échantillon unique ou à échantillons indépendants.

L'erreur-type à la moyenne se calcule de la façon suivante dans le cadre d'un test-t à échantillons indépendants : $ETM = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$.

On voudrait connaître les différences de salaire entre les hommes et les femmes. Les moyennes, écart-types et effectifs de la variable salaire est distribuée de la façon suivante en fonction du genre :

	hommes	femmes
Moyenne	2928,00	1958,04
ET	1288,63	861,07
Effectifs	20	23

L'erreur-type à la moyenne est donc de : $ETM = \sqrt{\frac{1288^2}{20} + \frac{861,07^2}{23}} = 339,5$

Une fois l'erreur-type à la moyenne calculée, il est possible de calculer la statistique t sur base du calcul suivant : $t = \frac{(\bar{X}_1 - \bar{X}_2)}{ETM}$

Avec nos données, cela donne : $t = \frac{(2928-1958,04)}{339,5} = 2,86$

Ensuite, comme pour le calcul du Khi-carré, après consultation d'une table statistique de loi t¹⁶, une p-valeur est associée en fonction du degré de liberté permettant de confirmer/infirmer H_0 .

Le nombre de degré de liberté est égal à n-1. Soit 43-1= 42.

¹⁵ Notez également la présence dans le menu d'SPSS du test-t pour échantillons appariés que nous n'allons pas voir dans le cadre de ce cours.

¹⁶ La loi t est une loi de distribution statistique ressemblant à la loi normale mais prenant en compte l'imprécision de l'écart-type dans un cadre inférentiel.

TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

df	Tail probability <i>p</i>											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496

*Nous n'avons pas la valeur précise pour un degré de liberté de 42 mais nous pouvons voir cependant que notre statistique *t*, pour 40 ou 50 degrés de libertés est suffisamment grande que pour confirmer H_1 avec un seuil d'erreur inférieur à 0.005. Soit, moins de 0.5% de chance de se tromper d'affirmer que les hommes sont mieux payés que les femmes. Pour rappel, la valeur du *t* doit être supérieure à celle indiquée dans la table pour connaître la *p*-valeur.*

Enfin, si le test-t permet de confirmer s'il y a bel et bien un effet du fait d'appartenir à telle ou telle catégorie d'une variable qualitative dichotomique sur une variable quantitative, il ne permet pas de connaître la force de cette relation. Pour cela, il nous faut calculer l'Eta-carré dont la formule est la suivante :

$$\eta^2 = \frac{t^2}{t^2 + (N_1 + N_2 - 2)}$$

Si la valeur de η^2 est proche de 0.01, l'effet est de petite taille ; si la valeur est proche de 0.06, l'effet est de taille moyenne ; si la valeur est proche ou supérieure à 0.14, l'effet est de grande taille.

L'Eta-carré est donc $\eta^2 = \frac{2.86^2}{2.86^2 + (20+23-2)} = 0.16$. On peut donc considérer qu'en plus d'avoir un très faible pourcentage de chance de se tromper d'affirmer H_1 , que cette relation est très forte et que le fait d'être un homme ou une femme va avoir un impact très fort sur le salaire espéré.

Le calcul de l'Eta-carré n'est bien entendu utile que si le résultat du test-t est significatif. Un test-t non significatif donnera systématiquement un η^2 très petit.

Plusieurs conditions sont nécessaires à l'application du test-t :

- La variable dépendante doit être de type quantitatif
- La variable dépendante doit être normalement distribuée si les effectifs sont faibles ($n < 50$)¹⁷
- Les variances doivent être égales (homoscédastiques)¹⁸
- Indépendance entre les groupes¹⁹

L'homoscédasticité²⁰ des variances peut se vérifier via le test d'homogénéité des variances de Lévene. Comme pour les autres tests, H_0 va assumer des variances égales et donc homoscédastiques alors que H_1 va assumer des variances hétéroscédistiques. Ce test est intégré à la procédure SPSS. Vous verrez cependant que cette condition n'est pas indispensable mais qu'il est important de savoir de savoir si les variances sont égales ou non. Si la signification de ce test est supérieure au seuil, les variances seront considérées comme homogènes et hétérogènes dans le cas inverse.

¹⁷ Cf. Chapitre 3

¹⁸ Dans le cadre d'un test-t à échantillons indépendants

¹⁹ Idem.

²⁰ Terme statistique pour parler d'égalité des variances. Variances égales, homogènes et homoscédastiques sont des synonymes. L'équivalent antinomique étant des variances inégales, hétérogènes ou hétéroscédistiques.

Manipulation avec SPSS

Vous trouverez les différents test-t dans le menu *Analyze*, puis dans le sous menu *Compare Means*. Nous allons voir chacune des deux procédures de test-t vues dans le cadre du cours.

1) One-Sample T-test (Test-t à échantillon unique)

On sait par le gouvernement belge que le revenu brut moyen des ménages en Belgique était de 2819 euros²¹ par mois en 2012. Nous voudrions vérifier cela avec notre base de données de l'EVS datant également de 2012. La variable V353MM de la base de données est une variable ordinaire du revenu mensuel des ménages dont les valeurs sont les suivantes²² :

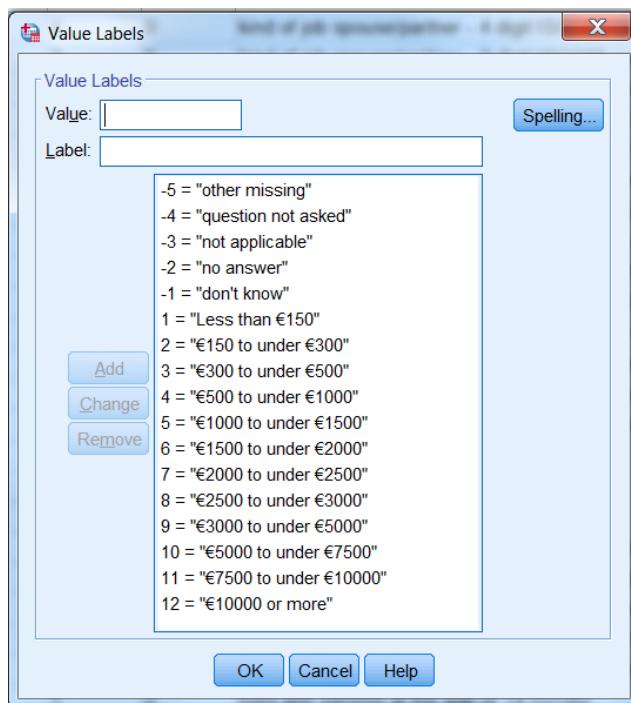


FIGURE 27 : DÉTAILS DES MODALITÉS DE LA VARIABLE V353MM

La modalité 8 va de 2500 euros à 3000 euros. Il nous est désormais possible de ramener notre moyenne de 2819 euros sur la même échelle de la base de données : $3000-2500=500$ et $319/500=0.638$. Notre moyenne est donc de 8,638.

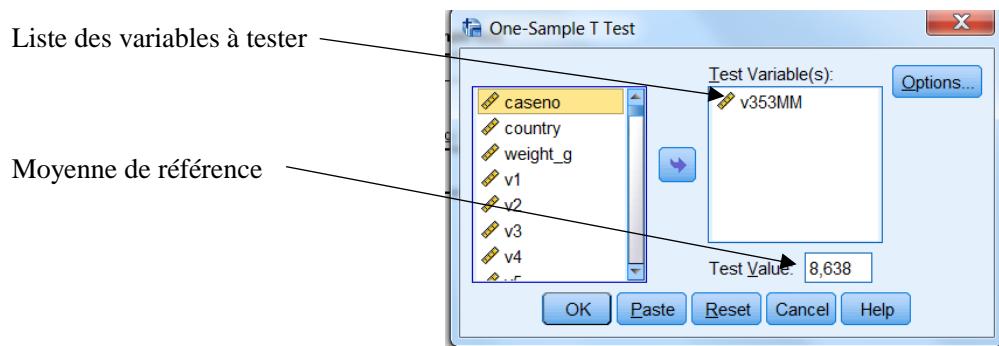


FIGURE 28 : DÉTAILS DU MENU DU TEST-T À ÉCHANTILLON UNIQUE

Voici ensuite les différentes sorties d'SPSS :

²¹ <http://www.express.be/business/fr/economy/43-des-menages-belges-se-retrouvent-parfois-a-court-dargent-avant-davoir-reu-leur-salaire/199059.htm>

²² Pour rappel, vous pouvez connaître les détails des modalités dans la vue des variables. Cf. Chapitre 1.

	N	Mean	Std. Deviation	Std. Error Mean
monthly household income (Q125)	1354	7,05	1,876	,051

FIGURE 29 : STATISTIQUES DESCRIPTIVES DE LA VARIABLE TESTÉES

Ce premier tableau descriptif nous permet déjà de savoir que notre test-t sera sans doute très significatif vu que la moyenne de la variable de revenu est de 7.05, soit après transformation, de 2025 euros par mois.

		One-Sample Test					
		Test Value = 8.638					
		t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
monthly household income (Q125)		31,067	1353	,000	-1,584	Lower -1,68	
						Upper -1,48	

FIGURE 30 : RÉSULTATS DU TEST-T À ÉCHANTILLON UNIQUE

Enfin, le deuxième tableau est le résultat du test-t en soi. Votre attention va donc se porter essentiellement dans un premier temps sur la significativité du test (Sig.). Celle-ci est de 0.000. On peut donc accepter H_1 : la moyenne de l'échantillon est significativement plus faible que la moyenne de référence.

2) Independent Samples T-test (Test-t à échantillons indépendants)

Nous allons ici tenter de comparer les moyennes de revenu ménager mensuel (V353MM) en fonction du niveau d'éducation (V336_r).

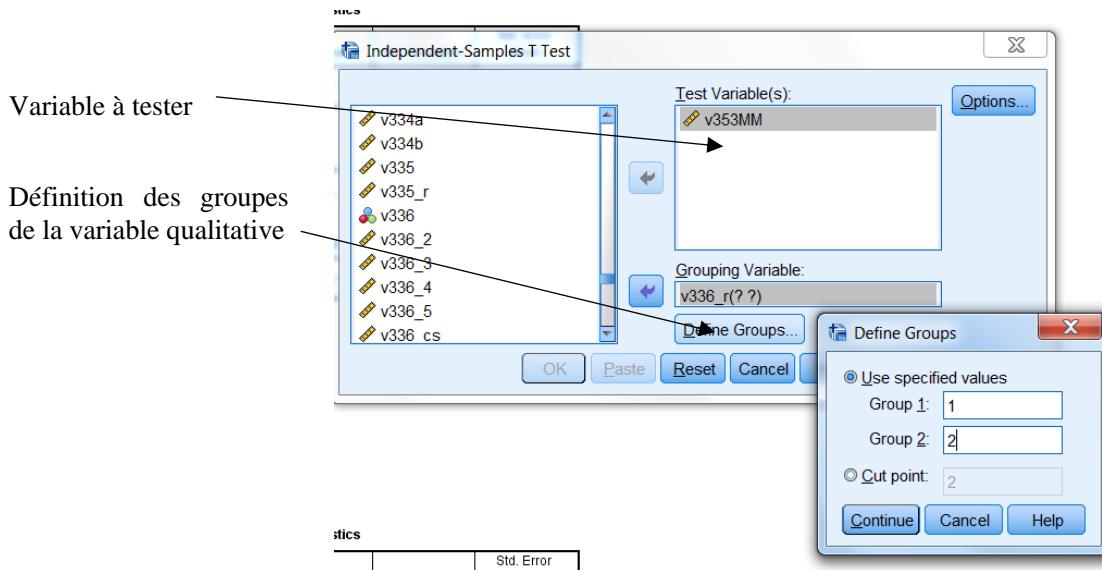


FIGURE 31 : MENU DE LA PROCÉDURE POUR TEST-T À ÉCHANTILLONS INDÉPENDANTS

La variable qualitative doit donc être glissée dans l'onglet *Grouping Variable*. Après cela, vous devez définir les groupes. Le test-t permet de comparer deux moyennes différentes. La variable qualitative ne doit donc pas spécifiquement être dichotomique mais dès lors, il faut mentionner les groupes à comparer. Deux options sont possibles : définir chacune des modalités dans les onglets Group1 et Group2 ou définir un point de césure (*Cut Point*). Le point de césure va constituer deux groupes bien distincts à partir d'une valeur renseignée. Le premier groupe comprendra toutes les observations dont la valeur de la variable est strictement inférieure à la valeur renseignée et le deuxième groupe reprendra les observations dont les valeurs sont égales ou supérieures à la valeur renseignée. Cela permet ainsi de dichotomiser n'importe quelle variable quantitative très rapidement sans créer de nouvelles variables.

Group Statistics

	education level (Q110) (recode)	N	Mean	Std. Deviation	Std. Error Mean
monthly household income (Q125)	Lower	445	6,00	1,617	,077
	Middle	471	7,07	1,732	,080

Le premier tableau est identique à celui vu précédemment à l'exception qu'il reprend les statistiques pour chacune des modalités testées. On voit donc que la moyenne salariale des personnes ayant un niveau d'éducation faible est inférieure à celle des individus ayant une éducation moyenne.

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality of Means						
	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
								Lower	Upper
monthly household income (Q125)	Equal variances assumed	6,352	,012	-9,62	914	,000	-1,068	,111	-1,285 -,850
	Equal variances not assumed			-9,65	913,866	,000	-1,068	,111	-1,285 -,851

FIGURE 32 : TABLEAU DE TEST-T POUR ÉCHANTILLONS INDÉPENDANTS

Le deuxième tableau reprend les résultats du test en soi et est composé de deux tests différents à lire dans l'ordre suivant. La partie rouge du tableau représente le test de Lévène qui va nous permettre de mesurer si la condition d'homoscédasticité est respectée. Si le test valide H₁, il faudra alors lire la seconde ligne du tableau (*Equal variances not assumed*) et dans le cas contraire où on accepte H₀, on lira la première ligne du tableau (*Equal variances assumed*) pour la suite de l'analyse. La significativité du test (Sig.) est de 0.012 et inférieure à 0.05. Nous devons donc rejeter H₀ : les variances ne sont pas homogènes, on lira donc la seconde ligne du tableau. Ensuite, nous pouvons lire le tableau du test-t (partie en vert) en fonction du résultat du test de Lévène. Ici dans notre cas, la significativité (Sig 2-tailed) est de 0.000, on peut donc dire qu'il y a bel et bien une différence significative de salaire en fonction des niveaux d'éducation faibles et moyens.

Vous remarquerez que les résultats entre la première et deuxième ligne du tableau sont souvent identiques. Dans notre cas-ci, les deux p-valeurs sont de 0.000. Il est cependant important de tenir compte de ce détail car le calcul du test-t est sensiblement différent selon que la condition d'homoscédasticité soit remplie ou non. Cela pourra amener alors à des résultats différents lorsque la différence n'est pas flagrante. Par exemple une p-valeur de 0.06 pour la condition d'homoscédasticité remplir et de 0.035 pour la condition d'homoscédasticité non remplie. La décision statistique ne sera, dans ce cas-là, pas la même.

A cela, il faut encore calculer l'éta-carré qui n'est malheureusement pas calculé par SPSS. Cependant cela, n'est pas très compliqué vu que la statistique t et les différents effectifs, vous sont donnés par les sorties SPSS. Le premier tableau nous indique que les effectifs sont de 445 et 471 et le deuxième tableau nous dit que la statistique t est de -9.65.

$$\eta^2 = \frac{t^2}{t^2 + (N_1 + N_2 - 2)} = \frac{-9.65^2}{-9.65^2 + (445 + 471 - 2)} = 0.09$$

L'effet est donc de taille moyenne. Etre de niveau d'éducation moyenne a un certain impact sur le revenu perçu par le ménage mais n'est pas non plus totalisant.

Analyse de variance : l'anova et les tests post-hoc

Rappel théorique

L'ANOVA est un test complexe dont il existe une multitude de déclinaisons. Nous allons uniquement nous intéresser à L'ANOVA à 1 facteur (dite « simple ») à k groupes (One-way ANOVA en anglais). L'ANOVA à 1 facteur est un test inférentiel bivarié qui permet de comparer les moyennes de plusieurs groupes. Le principe de l'ANOVA repose sur le fait que les moyennes vont varier en fonction de la variance et de l'effectif présent au sein de chaque classe. A l'inverse du test-t, nous ne sommes ici pas limités à deux groupes uniquement. Les hypothèses du test sont donc sensiblement identiques à celles du test-t :

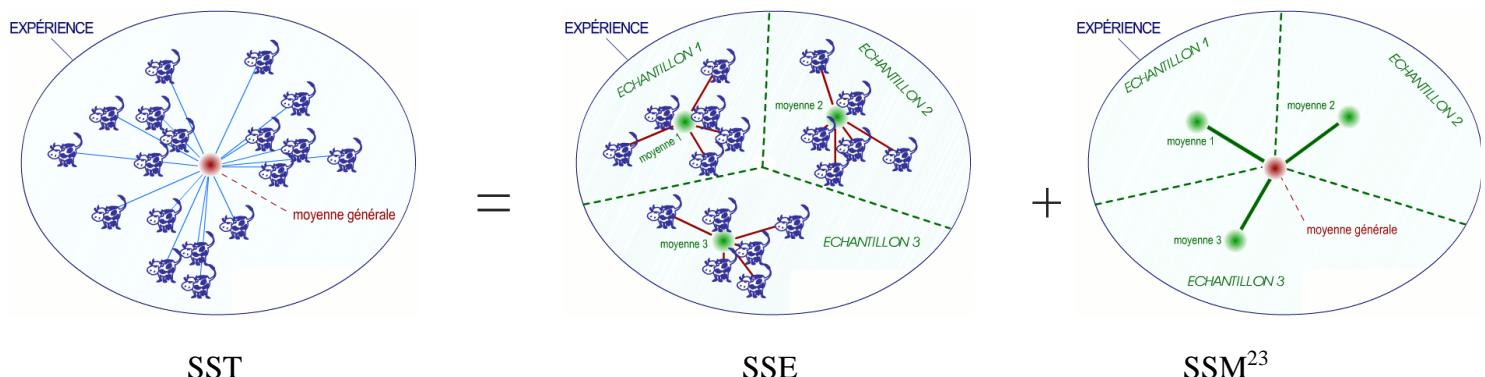
$$H_0 : \text{Les moyennes sont égales}$$

$$H_1 : \text{Les moyennes sont inégales}$$

Le calcul de la statistique F, centrale dans l'analyse de variance se base essentiellement sur le concept de décomposition de variance. La somme des carrés des écarts (SS) est le numérateur de l'équation de la variance alors que le dénominateur représente le nombre de degré de liberté.

$$SS = \sum_{ijk} (x_{ijk} - \bar{X})^2 \quad \text{d'où} \quad S_{n-1}^2 = \frac{SS}{n-1}$$

Nous pouvons donc conserver la somme des carrés des écarts comme indicateur de variance. Cette somme des carrés aux écarts est partitionnée en trois composantes de la variabilité : La variabilité totale (SST – *Sum of Square Total*) qui est la somme de la variabilité intra-classe (SSE – *Sum of square Errors*) [soit, la somme des carrés des écarts à la moyenne de la modalité], et de la variabilité interclasse (SSM – *Sum of square Model*) [soit, la somme des carrés des écarts des moyennes des différentes modalités à la moyenne générale pondérée par le nombre d'observation de chaque classe].



$$\text{Où } SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_k)^2 \text{ et } SSM = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2$$

La statistique F, issue du test de Fisher va se calculer à partir du rapport des variabilités interclasses sur les variabilités intra-classe, chacune divisées par le nombre de degré de liberté qui leur est propre. Ces degrés de libertés sont de $k-1$ pour le numérateur (k = nombre de classes) et de $n-k$ pour le dénominateur. Soit, on va faire le rapport de la variance entre les groupes par la variance à l'intérieur des groupes.

$$F = \frac{SSM/k - 1}{SSE/(n - k)}$$

Imaginons ce petit tableau de données reprenant les points à un examen de mathématiques sur 100 de 3 classes (a, b, c) de secondaire.

²³ Images trouvées sur le site : <http://webapps.fundp.ac.be/biostats/biostat/modules/module140/page7.html>

a	95	92	83	94	72	99	75	21
b	8	65	75	74	79	4		
c	10	21	18	45	24	87	25	100

Les moyennes des trois classes sont respectivement de 78,8%, 51% et 38%. La somme des carrés des écarts intra-classe va donc se calculer en comparant chaque score au sein de la classe avec la moyenne de la classe mise au carré. Ainsi pour la classe b, $SSE = (8-51)^2 + (65-51)^2 + (75-51)^2 + (74-51)^2 + (79-51)^2 + (4-51)^2 = 6143$. On fait de même pour a et c et on somme les trois résultats. La somme totale des carrés des erreurs est donc de $4494 + 6143 + 8788 = 19425$.

SSM est calculé par le même processus mais à partir des moyennes des groupes à la moyenne générale en multipliant par le nombre d'observation présent au sein de la classe. La moyenne générale de réussite au test est de 55%. $SSM = 8 * (78,8 - 55)^2 + 6 * (51 - 55)^2 + 10 * (38 - 55)^2 = 7320$.

$$F = \frac{SSM/k - 1}{SSE/(n - k)} = \frac{7320/2}{19425/21} = 3.95$$

Une fois la statistique F calculée, on peut se référer à une table statistique de la loi F pour repérer le degré de significativité de la relation observée. A l'inverse des tables statistiques présentées pour les tests de Chi² et Test-T, la table de la loi F sera sensiblement différente vu que l'on a deux types de degrés de libertés. Il y a donc plusieurs tables pour chaque niveau de significativité. La table suivante montre les statistique F minimales à obtenir pour satisfaire à un seuil de 0.05 de significativité. Il vous suffit alors de rechercher le nombre de degrés de liberté adéquat pour le numérateur et le dénominateur.

F - Distribution ($\alpha = 0.05$ in the Right Tail)

df ₂	df ₁	Numerator Degrees of Freedom						
		1	2	3	4	5	6	7
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	2
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	
4	7.7086	9.9443	6.5914	6.3882	6.2561	6.1631	6.0942	
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	

Dans notre cas, avec des degrés de libertés de 2 au numérateur et de 21 au dénominateur, la statistique F doit être supérieure à 3.4668 pour satisfaire un seuil de significativité de 5%. Notre statistique étant de 3.95, on peut accepter H_1 et prétendre que les moyennes obtenues dans les différentes classes sont significativement différentes.

Enfin, comme pour les autres tests, il nous est possible de calculer un indicateur de l'effet de ces différences observées. Cet indicateur s'appelle le R, résultant de la racine du R² qui mentionne la part de la variance expliquée par le modèle ANOVA. Celui-ci n'est pas calculé par SPSS, il faudra donc le calculer à la main.

$$R^2 = \frac{\text{Somme des carrés Inter-groupe}}{\text{Somme des carrés Totale}} = \frac{\sum N_g (\bar{X}_g - \bar{X}_t)^2}{\sum (X_i - \bar{X}_t)^2} \quad \text{et } R = \sqrt{R^2}$$

Si le R est aux alentours de 0.1, les différences observées seront considérées comme étant de faible ampleur, autour de 0.3, d'ampleur moyenne, autour et au-delà de 0.5, de forte ampleur. Il est évident que si le résultat de l'ANOVA n'est pas significatif, il est inutile de calculer le R.

Dans notre exemple, le R² est donc égal à $\frac{7320}{7320+19425} = 0.27$, soit le modèle explique 27% de la variance. Ce qui nous donne un effet R de 0.52. Soit, les différences de moyennes observées sont grandes.

Le test ANOVA nécessite certaines conditions à son application :

- 1) Variable dépendante continue et normalement distribuée si l'effectif est restreint (n < 50).
- 2) Variable explicative catégorielle.
- 3) Condition d'indépendance : indépendance entre les moyennes et entre les deux groupes
- 4) Condition d'homoscédasticité : variances des groupes égales (test de Lévene). L'ANOVA est beaucoup plus sensible à la condition d'homoscédasticité que ne l'est le Test-t.

Cette dernière condition est souvent difficile à obtenir dans le cadre de données sociales. Si jamais le test de Lévene n'est pas satisfaisant, il est possible d'utiliser le test de Brown-Forsythe. Celui-ci calcule des écarts à la médiane et non à la moyenne et va venir remplacer la lecture du test ANOVA.

Cela fait, L'ANOVA ne nous a donné qu'une seule information : à savoir si les moyennes sont égales ou si au moins une moyenne est différente des autres. Cela ne permet d'aller suffisamment loin dans l'analyse. Il est donc vivement conseillé d'effectuer un test post-hoc. Les tests post-hoc sont des tests qui permettent d'étayer le résultat de l'ANOVA en comparant les moyennes deux à deux. De plus, ces tests post-hoc sont analysables même si la condition d'homoscédasticité (et donc le résultat de l'anova) n'est pas rencontrée. Il existe toute une série de test post-hoc. Le cadre de ce cours n'inclue pas la compréhension totale des tests post-hoc. Nous allons donc nous limiter à deux tests post-hoc selon que la condition d'homoscédasticité soit respectée (Test de Tukey) ou non (Test de Dunnett). Ces tests, bien que ressemblant, ne sont PAS des test-t²⁴. Les hypothèses de ces tests sont les suivantes :

- H₀ : les deux moyennes comparées sont identiques
H₁ : les deux moyennes comparées sont différentes.

Manipulations dans SPSS

Vous trouverez le test ANOVA dans le menu *Analyze, Compare Means (comparer les moyennes), One-Way ANOVA* (Anova à 1 facteur). A l'exercice précédent sur le test-t à échantillons indépendants, nous avions comparé le revenu mensuel des ménages belges en fonction de deux modalités de niveau d'éducation. Nous allons maintenant tenter d'effectuer une analyse comparative de toutes les modalités d'éducation. Rentrez donc dans le menu de l'ANOVA, la variable **V353MM** (revenu mensuel du ménage) comme variable dépendante (*Dependent variable*) et la variable **V336_r** (niveau d'éducation) comme facteur (*Factor*).

Ouvrez ensuite le sous-menu *Post-Hoc*. Vous verrez, alors une série de tests post-hoc organisé en fonction de présence d'homoscédasticité ou d'hétéroscédisticité des variances. Sélectionnez les tests de *Tukey* et le

²⁴ Pour les plus curieux, le test de Tukey se base sur une comparaison entre la différence des moyennes divisées par l'erreur-type à la moyenne $q = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s_1 + s_2}{n_1 + n_2}}}$ et la différence de moyenne studentisée, telle que : $q_t = \frac{\bar{Y}_{max} - \bar{Y}_{min}}{s \sqrt{2/n}}$ si

$q > q_t$ alors les moyennes seront considérées comme différentes. Autrement, on prend la valeur de q_t et on va rechercher la p-valeur associée dans une table de Tukey (similaire à la table t mais en prenant compte les groupes).

T3 de Dunnett. Si vous le désirez, vous pouvez ajuster le seuil de significativité tout en dessous des choix de test.

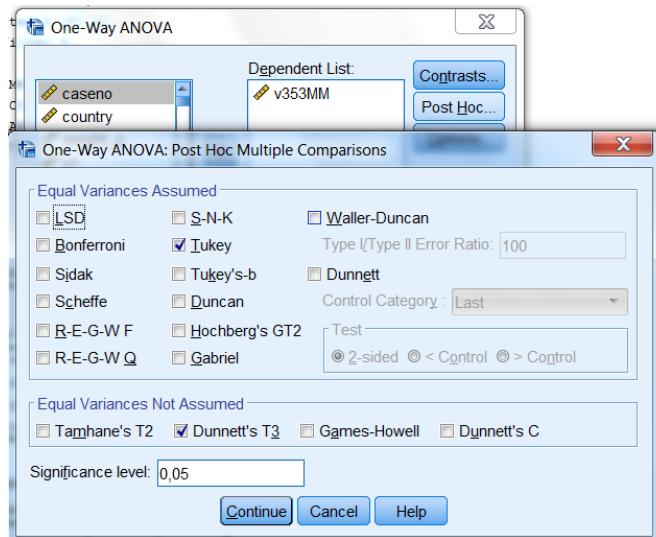


FIGURE 33 : SOUS-MENU DE SÉLECTION DES TEST POST-HOC

Allez ensuite dans le sous-menu *Options* et sélectionnez les options suivantes :

- *Descriptive* : Fournit des statistiques descriptives sur les différents groupes (moyenne, écart-type, etc.)
- *Homogeneity of variance test* : Fournit le test de Lévene d'homoscédasticité
- *Brown-Forsythe* : Test remplaçant le résultat de l'ANOVA en cas d'hétérosécédasticité.

Exécutez la commande et vous obtiendrez les résultats de sortie suivants :

Descriptives									
	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	
					Lower Bound	Upper Bound			
Lower	445	6,00	1,617	,077	5,85	6,15	1	12	
Middle	471	7,07	1,732	,080	6,91	7,23	1	12	
Upper	438	8,11	1,668	,080	7,95	8,26	1	12	
Total	1354	7,05	1,876	,051	6,95	7,15	1	12	

Ce premier tableau affiche donc les statistiques descriptives des revenus moyens en fonction des niveaux d'éducation. Les moyennes sont respectivement de 6, 7,07 et 8,11. Il nous est permis de transformer ces valeurs en euros, grâce aux informations de la variable, préalablement affichés à la Figure 27. Ce qui nous fait donc, 1500€ par mois pour les niveaux d'éducation faible, 2035€ par mois pour les niveaux d'éducation moyen et enfin, 2555 € pour les individus dotés d'un niveau d'éducation élevé. Cette transformation faite, ces différences qui pouvaient nous paraître minimes à première vue, semblent bel et bien conséquentes.

Test of Homogeneity of Variances				
monthly household income (Q125)				
Levene Statistic	df1	df2	Sig.	
3,289	2	1351	,038	

L'on regarde ensuite, les résultats du test de Lévene. Celui-ci est inférieur à 0,05, nous ne pouvons donc pas lire le résultat de l'ANOVA présenté ci-dessous car les variances sont hétérosécédastiques :

ANOVA						
monthly household income (Q125)						
	Sum of Squares	df	Mean Square	F	Sig.	
Between Groups	976,210	2	488,105	174,137	,000	
Within Groups	3786,855	1351	2,803			
Total	4763,064	1353				

Si le résultat de Lévene avait été concluant, nous aurions alors pu lire la significativité du test anova qui est dans ce cas, de 0.000. Comme nos variances sont hétéroscédastiques, nous devons donc aller regarder le résultat du test de Brown-Forsythe.

Robust Tests of Equality of Means monthly household income (Q125)				
	Statistic ^a	df1	df2	Sig.
Brown-Forsythe	174,509	2	1349,395	,000

a. Asymptotically F distributed.

La p –valeur du test de Brown-Forsythe est de 0.000, ce qui signifie que nos moyennes de nos trois groupes sont significativement différentes les unes des autres. Nous pouvons maintenant calculer l'effet de ces différences, en retournant au résultat de l'ANOVA. $R^2 = \frac{976}{4763} = 0.2$ et $R = \sqrt{0.2} = 0.45$. Soit, le modèle explique 20% de la variance et les différences observées entre les moyennes des groupes sont de forte ampleur. L'analyse du test ANOVA en tant que tel, s'arrête ici. Nous savons donc qu'au moins une moyenne d'un des groupes est significativement différente des autres et de façon très prononcée. Pour détailler ce résultat, il nous faut aller voir le résultat du test post-hoc.

Multiple Comparisons								
				Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
		(I) education level (Q110) (recoded)	(J) education level (Q110) (recoded)				Lower Bound	Upper Bound
Tukey HSD	Lower	Middle		-1,068*	,111	,000	-1,33	-,81
		Upper		-2,103*	,113	,000	-2,37	-1,84
	Middle	Lower		1,068*	,111	,000	,81	1,33
		Upper		-1,035*	,111	,000	-1,30	-,77
	Upper	Lower		2,103*	,113	,000	1,84	2,37
		Middle		1,035*	,111	,000	,77	1,30
Dunnett T3	Lower	Middle		-1,068*	,111	,000	-1,33	-,80
		Upper		-2,103*	,111	,000	-2,37	-1,84
	Middle	Lower		1,068*	,111	,000	,80	1,33
		Upper		-1,035*	,113	,000	-1,30	-,77
	Upper	Lower		2,103*	,111	,000	1,84	2,37
		Middle		1,035*	,113	,000	,77	1,30

*. The mean difference is significant at the 0.05 level.

Ce tableau est coupé en deux : les 3 lignes supérieures représentent les résultats du test de Tukey et les 3 lignes inférieures, les résultats du test de Dunnett. Comme nous sommes dans un cas de variances inégales, nous devons donc regarder les résultats du test de Dunnett. La lecture de ce tableau est simple, chaque ligne compare une modalité (celle de gauche) avec les autres modalités (à droite). Ainsi, la première ligne du test de Dunnett, compare les individus de niveau d'éducation faible avec ceux de niveau d'éducation moyen et élevé ensuite. Pour chacune de ces comparaisons sont ensuite calculées les différences de moyennes, erreurs standards et une significativité associée (Sig). Toutes les significativités du test de Dunnett sont de 0.000. Nous pouvons donc conclure que toutes les moyennes comparées sont significativement différentes les unes des autres.

Imaginons qu'une de ces p-valeurs était non-significative. Cela signifierait que les deux moyennes ne sont pas considérées comme différentes.

Le test de corrélation

Rappel Théorique

Le test de corrélation est un test inférentiel bivarié qui permet de mesurer la dépendance entre deux variables quantitatives (si la variable X augmente d'une unité, la variable Y augmentera ou diminuera de x unité). Le coefficient de corrélation se calcule sur la covariance entre les deux variables visées, divisé par le produit des deux écarts-types. La covariance est une statistique permettant de connaître les écarts conjoints d'une observation vis-à-vis de deux variables ou plus. Son calcul est donc sensiblement différent à celui de la variance :

$$COV(x,y) = \frac{1}{n-1} \sum_{i=1}^N (x_i - \bar{X})(y_i - \bar{Y})$$

Le calcul de la corrélation est donc le suivant :

$$r = \frac{COV(x,y)}{S_x S_y}$$

Ce coefficient est donc compris entre -1 (association négative) et 1 (association positive). Un score proche de 0 renverra à une absence de relation. Au plus ce coefficient sera proche de 1 ou de -1, au plus il décrira la force de la relation linéaire entre les variables quantitatives. En sciences sociales, l'on a tendance à considérer les coefficients de corrélation selon la règle suivante :

Autour de (-) 0.1 : corrélation faible

Autour de (-) 0.3 : corrélation moyenne

Autour de (-) 0.5, au-dessus de 0.5 ou en-dessous de -0.5 : corrélation forte

Il n'en est bien sûr pas forcément de même dans d'autres domaines scientifiques.

Une significativité est associée au coefficient r, vous permettant d'établir la présence de relation entre les deux variables ou non. Le test de corrélation a pour hypothèse :

H_0 : Absence de dépendance entre les deux variables.

H_1 : Il y a dépendance entre les deux variables.

Le calcul de la p-valeur se fait comme suit: $t = r \sqrt{\frac{n-2}{1-r^2}}$. On ira ensuite chercher, au nombre de degré de liberté associé, la p-valeur dans une table t.

Imaginons deux variables quantitatives suivantes mesurant les résultats de 5 étudiants à deux tests différents (a et b) sur 20 :

a	6	7	10	12	13
b	8	6	10	14	16

La moyenne des tests a et b sont donc respectivement de 9.6 et 10.8 et leurs écarts-types de 3 et de 4.1. Nous pouvons donc calculer la covariance et ensuite la corrélation.

$$COV(x,y) = \frac{1}{5-1} \sum (6 - 9.6)(8 - 10.8) + (7 - 9.6)(6 - 10.8) + \dots = \frac{1}{4} 47.6 = 11.9$$

$$r = \frac{11.9}{3 * 4.1} = 0.96$$

Le résultat de la corrélation est donc ici très fort (presque parfait). Cela signifie que plus un étudiant fera un bon score au test a, plus il fera également un bon score au test b (ou inversement).

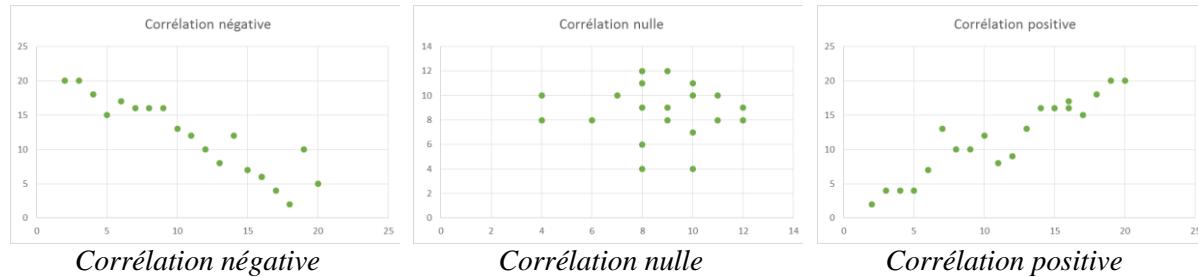
$$\text{Calcul de la p-valeur : } t = 0.96 \sqrt{\frac{5-2}{1-0.96^2}} = 5.9$$

TABLE B: t -DISTRIBUTION CRITICAL VALUES

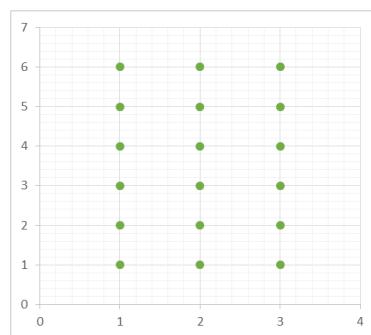
df	Tail probability p											
	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959

Le nombre de degré de liberté = $n-2 = 3$. Si la statistique t est supérieur à 5.841 pour un tel nombre de degré de liberté, on peut s'assurer une p-valeur d'au plus 0.005. Nous pouvons donc affirmer sans aucun doute²⁵ dans notre démonstration ci-dessus que les résultats aux tests a et b sont très fortement corrélés.

La représentation graphique d'une corrélation via un nuage de points est assez parlante d'un point de vue interprétatif et analytique. De manière générale, si vous repérez une droite au sein du nuage de points, c'est très probablement qu'une corrélation est présente ; alors qu'un nuage de point sans distinction linéaire amènera à l'hypothèse d'absence de corrélation :



Prenez garde cependant que, si en effet, le calcul de la corrélation est possible sur des données ordinaires ou discrètes, la représentation graphique de celles-ci a de grandes chances de ne pas être interprétable (et d'autant plus que le nombre de valeurs est faible) comme le démontre le graphique ci-dessous où le calcul d'une corrélation est pourtant possible.



Manipulation dans SPSS

Allez dans le menu *Analyze*, puis *Correlate*, et enfin, *Bivariate*. Nous allons mesurer la corrélation entre le revenu mensuel du répondant (**V353MM**) avec le diplôme le plus élevé obtenu par le père (ou la mère) (**V355_4**). Le menu est assez implicite, il vous suffit de glisser les variables à analyser dans l'encadré. Si vous le désirez, il est possible d'obtenir quelques statistiques descriptives via le sous-menu *Options*.

Correlations			
	monthly household income (Q125)	highest educational level attained father/mother (8 categories) (Q127)	
monthly household income (Q125)	Pearson Correlation		,306**
	Sig. (2-tailed)		,000
	N	1354	1241
highest educational level attained father/mother (8 categories) (Q127)	Pearson Correlation	,306**	1
	Sig. (2-tailed)	,000	
	N	1241	1371

²⁵. Correlation is significant at the 0.01 level (2-tailed). La procédure de corrélation va ainsi produire un tableau croisé des variables rentrées dans l'analyse. Pour chacun des croisements des variables testées, SPSS vous fournit la statistique de corrélation (R), la significativité associée et l'effectif (N). Il va de soi que lorsque vous lisez le croisement d'une variable avec elle-même, la corrélation sera de 1 et aucune p-valeur ne sera associée. Ainsi dans l'exemple ci-dessus, le

²⁵ Exception faite du faible effectif pour la facilité de la démonstration.

diplôme obtenu par le père est corrélé positivement et de manière modérée ($r=0.306$, p -valeur= 0.000) avec le revenu des répondants.

Lorsque vous rentrez plusieurs variables dans l'analyse, SPSS sortira donc une matrice de plus grande ampleur et les corrélations seront toujours effectuées deux à deux. Il ne faudra pas nécessairement regarder tous les croisements de variables : les résultats au-dessus de la diagonale étant le reflet des résultats en dessous de celle-ci.

Récapitulatif des tests bivariés

	Khi-carré	T-Test	Anova	Corrélation
<i>Variables</i>	2 qualitatives	1 qualitative (2 modalités) 1 quantitative	1 qualitative 1 quantitative	2 quantitatives
<i>Conditions</i>	Max 20% ET<5 0% ET<1 Indépendance	Normalité (N<50) Indépendance Homoscédasticité	Normalité (N<50) Indépendance Homoscédasticité	Indépendance
<i>Test (H_0)</i>	ET = EO	$\bar{X}_1 = \bar{X}_2$	$\bar{X}_1 = \bar{X}_2 = etc = \bar{X}_n$	Pas de relation linéaire
<i>Test d'association</i>	V de Kramer Calculé par SPSS	Eta-carré $\eta^2 = \frac{t^2}{t^2 + (N_1 + N_2 - 2)}$	R $R = \sqrt{\frac{SSM}{SST}}$	R Calculé par SPSS
<i>Interprétation du test d'association</i>	~ 0.1 : faible ~ 0.3 : moyen ~ 0.5 : élevé	~ 0.01 : faible ~ 0.06 : moyen ~ 0.14 : élevé	~ 0.1 : faible ~ 0.3 : moyen ~ 0.5 : élevé	~ 0.1 : faible ~ 0.3 : moyen ~ 0.5 : élevé
<i>Représentation Graphique²⁶</i>	Barres juxtaposées	Box-plot	Box-plot	Nuage de points

FIGURE 34 : TABLEAU RÉCAPITULATIFS DES SPÉCIFICITÉS DES TESTS BIVARIÉS

²⁶ Cf. Chapitre 3 : éléments de statistique descriptive.

CHAPITRE 5 : EXPLORATION DE DONNEES

Nous avons donc pour l'instant vu deux types d'analyse statistique : l'analyse descriptive simple et les tests inférentiels (bivariés). Nous allons maintenant voir des techniques d'analyse multivariée (et donc plus complexes) qui vont nous permettre de détecter des structures (visibles ou cachées) au sein des données. Ces techniques vont – dans le cadre de ce cours- être utilisées à titre essentiellement descriptif mais vont monopoliser vos connaissances sur la statistique descriptive simple et potentiellement les tests inférentiels bivariés.

L'analyse en composantes principales

Rappel théorique

Lorsque le nombre de variables dans une base de données est important, les techniques de statistiques descriptives simples peuvent s'avérer longues et pénibles, voire inefficaces pour décrire un phénomène : centaines de graphiques différents, de tableaux et d'interprétations variées, etc. C'est pourquoi, il peut être utile de créer des synthèses de l'information en recherchant s'il existe des corrélations négatives ou positives entre variables, des groupes de variables corrélées entre elles, des individus qui se ressemblent, des typologies de variables, des typologies d'individus, etc. L'analyse en composantes principales (ACP) est une technique multivariée de réduction de l'information qui peut se définir de la manière suivante : « Produire un résumé d'information au sens de l'ACP, c'est établir une similarité entre les individus, chercher des groupes d'individus homogènes, mettre en évidence une typologie d'individus. Quant aux variables c'est mettre en évidence des bilans de liaisons entre elles, moyennant des variables synthétiques et mettre en évidence une typologie de variables. L'ACP cherche d'une façon générale à établir des liaisons entre ces deux typologies » (Kouani et al.²⁷). L'analyse en composantes principales fait partie de la grande famille des analyses factorielles.

L'analyse en composantes principales est donc une technique multivariée qui va tenter de représenter les individus sur un plan à moindre dimensions (et donc avec un degré d'interprétation plus grand) tout en essayant de minimiser la perte d'information liée à la synthèse. Les axes factoriels seront donc des variables fictives qui vont résumer l'information partagée par plusieurs variables corrélées entre elles et qui sont décorrélés les uns par rapport aux autres (comprenez que les axes factoriels sont orthogonaux entre eux et donc indépendants). Ces facteurs sont des variables de type quantitative dont les valeurs attribuées aux individus sont centrées-réduites (moyenne de 0 et écart-type de 1). Lors du calcul mathématique, on va construire les axes factoriels de manière à ce qu'ils reprennent la plus grande variance possible. En d'autres termes, qu'ils colportent le plus grand pourcentage d'information sur les données utilisées. En statistique, l'on appelle valeur-propre, ce pourcentage de variance, d'information que reprennent les composantes principales. Le nombre de valeur propre est égal au nombre de variables (une variable ayant un potentiel initial de 100% d'information, a une valeur propre de 1) et c'est grâce à cet indicateur que l'on mesure la condensation de l'information. Le pourcentage de variance attribué à chaque axe va être distribué de façon décroissante de telle sorte à ce que le 1^{er} facteur aura un potentiel d'explication supérieur au 2^{ème} facteur, lui-même ayant un potentiel d'explication supérieur au troisième facteur, etc.

Une composante principale ayant 3.4 valeur propres, expliquera donc le même montant d'information que 3.4 variables. Il y a donc bel et bien eu une condensation de l'information.

Afin de véritablement réalisé une synthèse de l'information, viendra la question de la sélection des axes factoriels car l'ACP va, de base, distribuer l'information sur un nombre d'axes factoriels égal au nombre de variables entrées dans l'analyse. Il existe globalement deux méthodes pour cela. La première est le critère de Kaiser où l'on conserve les axes factoriels avec une valeur-propre supérieur à 1 (on conserve donc les axes factoriels sur lesquels nous avons réalisé un gain d'information). La deuxième est le coude de Catell : si l'on représente sur un plan les axes factoriels en abscisse et les valeurs propres en ordonnées, il est possible d'obtenir un graphique en ligne. Le coude de Catell stipule que l'on conserve le nombre d'axes factoriels avant que l'on constate une perte d'information (« un coude ») significative. Cette méthode, certes plus subjective, est également une méthode plus conservatrice dans le sens où lorsque le nombre de variables est

²⁷ Article complet disponible en ligne : <http://www.radisma.info/docannexe.php?id=522>

très élevé, le nombre d'axes factoriels avec une valeur propre supérieur à 1 risque d'être élevé également. Cette méthode permettra de conserver moins d'axes factoriels. Dans tous les cas, il paraît évident que conserver des axes factoriels avec des valeurs-propres inférieures à 1 va à contre-sens du but de l'ACP.

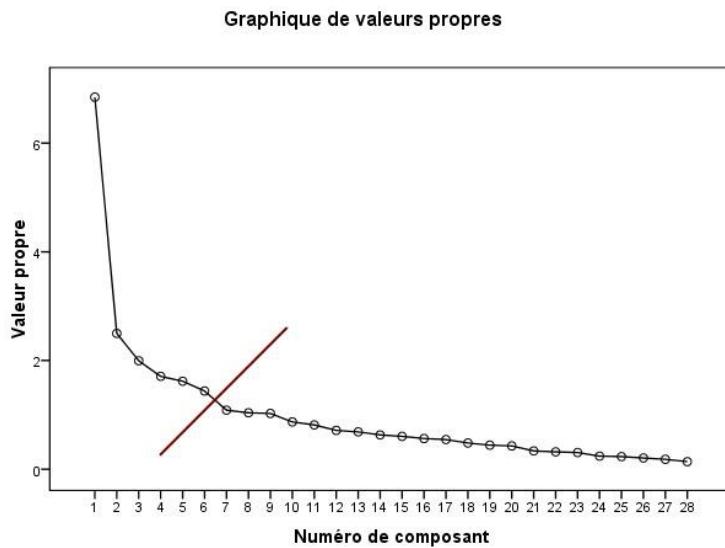


FIGURE 35 : REPRÉSENTATION DE LA MÉTHODE DU COUDE DE CATTELL²⁸

Ensuite, afin de faciliter l'interprétation de ces métavariables nouvellement créées, il est vivement conseillé d'effectuer une rotation. La rotation (orthogonale dans le cadre de ce cours), va viser à maximiser les corrélations des variables fortement corrélées et à minimiser les corrélations des variables faiblement corrélées aux axes factoriels. Dans un second temps, cette rotation va également redistribuer la variance entre les axes factoriels. Les axes factoriels peuvent ensuite être interprétés en fonction des variables positivement et négativement fortement corrélées. L'exercice sera donc de retrouver une cohérence interne des axes factoriels (quel phénomène commun expliquent ces variables fortement corrélées ?) et une cohérence externe (comment différencier les axes factoriels les uns des autres ?).

Vous l'aurez donc bien compris que l'analyse en composantes principales n'est pas un test²⁹ mais bel et bien d'un instrument multivarié à dessein essentiellement descriptif qui permet d'atteindre différents objectifs :

- a) Comprendre la structure (visible ou cachée) des variables rentrées dans l'analyse.
- b) Réduire le nombre de variables nécessaires à l'analyse.
- c) Construire ou affiner des indicateurs³⁰.

La construction d'indicateur part d'un postulat hypothético-déductif. Lors de l'élaboration d'un questionnaire, il est possible de théoriser un indicateur qui pourrait être mesuré par différentes dimensions de telle sorte que cet indicateur serait la résultante de la somme de ces différentes dimensions, elles-mêmes mesurées par différentes questions et donc variables. Dans ce cas-là, il est utile de passer par l'analyse en composantes principales afin de vérifier que le modèle théorique est bel et bien démontré par l'ACP (comprenez que les variables mesurant les dimensions se retrouvent corrélées sur un même axe factoriel). (Cf. Chapitre 5 : L'alpha de Cronbach).

Les conditions d'applications d'une analyse en composantes principales sont les suivantes :

- ❖ Les variables rentrées dans l'analyse doivent être quantitatives³¹.
- ❖ Il est préférable que le nombre de variable soit élevé.
- ❖ Il est également préférable d'avoir un effectif minimal de 10 par variable introduite dans l'analyse.

²⁸ <http://spss.espaceweb.usherbrooke.ca/pages/interdependance/analyse-en-composantes-principales.php>

²⁹ Il n'y a donc pas d'hypothèses à tester ni de variables dépendantes ou indépendantes

³⁰ Essentiellement utilisé dans d'autres domaines d'études comme la psychologie.

³¹ Dérogation pour les variables dichotomiques et ordinaires.

- ❖ Il est obligatoire d'avoir des corrélations minimales entre les variables renseignées dans l'analyse.

Sur ce dernier point, les tests de Kaiser-Meyer-Olkin (KMO) et de sphéricité de Bartlett permettent de vérifier les corrélations minimales à l'acceptation des résultats de l'ACP. Le test KMO donne un aperçu global des corrélations entre variables et se situe entre 0 et 1. Il est admis de considérer les seuils suivants pour l'interprétation du test :

0,80 et plus	Excellent
0,70 et plus	Bien
0,60 et plus	Médiocre
0,50 et plus	Misérable
Moins de 0,50	Inacceptable

Le test de sphéricité de Bartlett quant à lui, renvoie une p-valeur. Si le test n'est pas significatif, on considère alors que les variables sont indépendantes les unes des autres et qu'une analyse en composante principale n'a pas lieu d'être. Si le résultat du test de Kaiser-Meyer-Olkin ou de sphéricité de Bartlett ne sont pas concluants, il peut être conseillé alors de rajouter des variables dans l'analyse.

Manipulations SPSS

Nous allons tenter d'interpréter la tolérance des belges vis-à-vis des spécificités de leurs voisins. Comme nous avons un nombre de variables important (15), nous allons, grâce à l'ACP, essayer de voir s'il n'y a pas des redondances entre variables et une structure cohérente de celle-ci. Sélectionnez donc les ressortissants de Belgique.

Pour exécuter une analyse en composantes principales simple dans SPSS, on exécute les fonctions suivantes : « Analyse » (Analyze) → « Réduction des dimensions » (Dimension reduction) → « Analyse factorielle » (Factor).

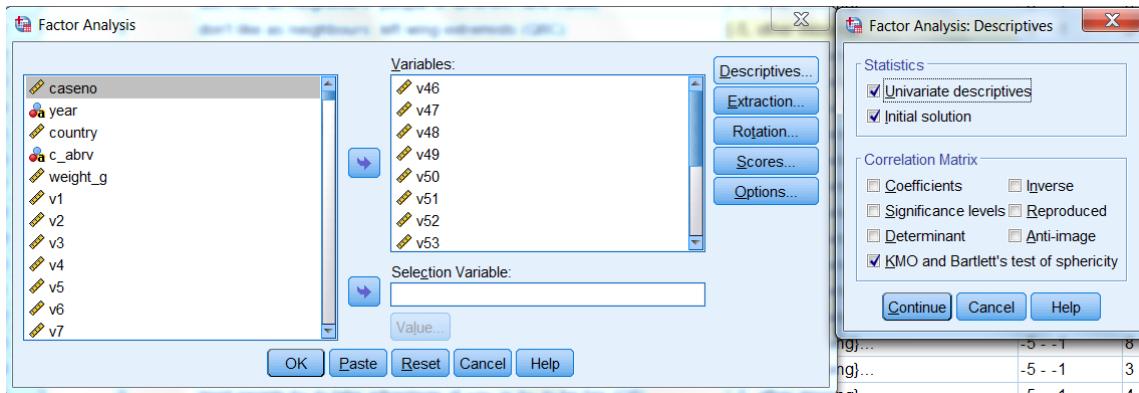


FIGURE 36 : MENU DE L'ANALYSE EN COMPOSANTES PRINCIPALES

Dans l'encadré « Variable(s) », insérer la liste des variables numériques, dichotomiques ou ordinaires sur lesquelles vous voulez effectuer l'ACP. Dans l'exemple qui nous occupe, rentrez les variables V46 à V60.

N'oubliez pas de bien vérifier le sens des variables au sinon vous risquez de faire des erreurs d'interprétation. Dans l'exemple ci-dessus, l'on a rentré une série de variable où l'on demandait aux individus, leur tolérance vis-à-vis de certaines caractéristiques de leurs voisins. Par défaut, les étudiants pensent qu'avoir un faible score induit une intolérance des citoyens vis-à-vis de ces pratiques et inversement. Pourtant, il est probable que ce soit l'inverse : un faible score indiquera une grande tolérance et un grand score une intolérance totale. Dans notre exemple, un score élevé indiquera une grande intolérance aux voisins possédant telle ou telle caractéristique (Cf. Figure 37)

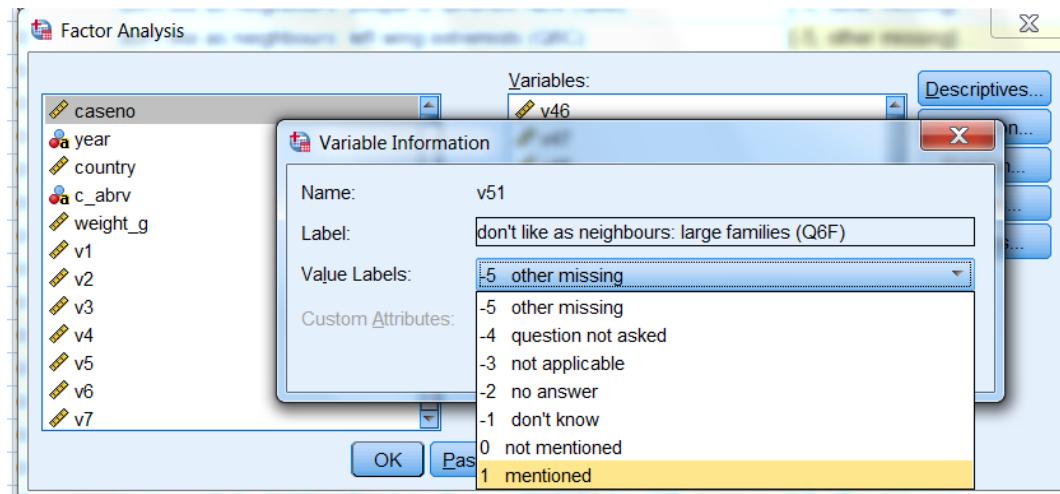


FIGURE 37 : VÉRIFIER LE SENS DES VARIABLES ORDINALES

Dans les onglets suivants, sélectionnez :

Descriptives :

- ❖ Univariate descriptives : vous donne un résumé descriptif des variables insérées dans l'analyse
- ❖ Initial solution : donne un résumé de l'analyse en composantes principales avant rotation
- ❖ KMO and Bartlett's test of sphericity : permet de vérifier les conditions de corrélations minimales
- Optionnel : Coefficients (affiche la matrice de corrélation) et signification levels (affiche la matrice des p-valeurs des corrélations)

Extraction :

- ❖ méthode : « principal components »³²
- ❖ Unrotated factor solution : Affiche les corrélations des variables aux axes avant rotation
- ❖ Scree plot : Permet d'afficher le graphique en ligne tenant compte de la valeur propre répartie sur chaque axe (cf. Figure 35)
- ❖ Dans l'encadré extract, SPSS choisi par défaut le critère de Kaiser comme seuil d'extraction (Les axes factoriels avec une valeur propre supérieure à 1). Vous pouvez changer cette valeur dans l'encadré ou choisir un nombre de facteur fixe à extraire dans la deuxième option.

Par défaut, utilisez le critère de Kaiser comme méthode. Si vous vous rendez compte après coup que le nombre de facteurs n'était pas idéal, utilisez alors la méthode de Catell et limitez le nombre d'axes factoriels manuellement.

Rotation :

- ❖ Varimax : rotation orthogonale (les axes factoriels restent indépendants les uns aux autres)
- ❖ Rotated Solution : Permet d'obtenir la table des corrélations de variables aux axes après rotations.

Scores (Facteurs) :

- ❖ Save as variable (Enregistrer dans des variables) : choisir « regression » (régresseurs). Permet de sauvegarder les composantes principales en variables dans la base de données afin d'être réutilisées ultérieurement.

Lancez l'analyse.

³² Normalement SPSS sélectionne cette méthode par défaut. Assurez-vous que ce soit bien le cas.

Interprétation

Le premier tableau reprend donc un résumé descriptif des 15 variables rentrées dans l'analyse en affichant leur moyenne et écart-type et le nombre de données valides.

	Descriptive Statistics		
	Mean	Std. Deviation	Analysis N
don't like as neighbours: people with criminal record (Q6A)	,27	,444	1501
don't like as neighbours: people of different race (Q6B)	,05	,226	1501
don't like as neighbours: left wing extremists (Q6C)	,25	,436	1501
don't like as neighbours: heavy drinkers (Q6D)	,38	,485	1501
don't like as neighbours: right wing extremists (Q6E)	,38	,485	1501
don't like as neighbours: large families (Q6F)	,02	,155	1501
don't like as neighbours: emotionally unstable people (Q6G)	,14	,346	1501
don't like as neighbours: muslims (Q6H)	,15	,352	1501
don't like as neighbours: immigrants/foreign workers (Q6I)	,06	,241	1501
don't like as neighbours: people with AIDS (Q6J)	,05	,217	1501
don't like as neighbours: drug addicts (Q6K)	,54	,498	1501
don't like as neighbours: homosexuals (Q6L)	,07	,251	1501
don't like as neighbours: jews (Q6M)	,04	,193	1501
don't like as neighbours: gypsies (Q6N)	,26	,440	1501
don't like as neighbours: christians (Q6O)	,00	,068	1501

FIGURE 38 : STATISTIQUES UNIVARIÉES (ACP)

Rappelez-vous que la variable est dichotomique 0(caractéristique pas mentionnée comme dérangeante), 1 (caractéristique mentionnée comme dérangeante). Les belges n'aiment pas avoir comme voisin en moyenne, des drogués, des buveurs et les extrémistes de droite. A l'inverse, les voisins qui gênent le moins les belges sont les chrétiens, les juifs et les personnes de race différente.

Le deuxième tableau reprend les tests de Kaiser-Meyer-Olkin et le test de sphéricité de Bartlett, nous permettant de vérifier les conditions de corrélations minimales à la réalisation d'une analyse en composantes principales.

KMO and Bartlett's Test			
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,757	
Bartlett's Test of Sphericity	Approx. Chi-Square	3120,773	
df		105	
Sig.		,000	

FIGURE 39 : KMO ET TEST DE SPHÉRICITÉ DE BARTLETT (ACP)

L'indice KMO est de 0.757 – ce qui est un bon score et la significativité du test de sphéricité de Bartlett est nettement inférieure à 0.05, ce qui indique que les corrélations minimales sont remplies. De manière générale, privilégiez l'interprétation de l'indice KMO qui est plus précis que le test de Bartlett.

Ensuite nous avons le tableau des communalités ou autrement dit, de la qualité de représentation des variables dans la solution factorielle. Si une variable est trop peu représentée, il peut être intéressant de la retirer de l'analyse. Les communalités sont expliquées en % de valeur propre extraite.

Communalities		Initial	Extraction
don't like as neighbours: people with criminal record (Q6A)	1,000	,311	
don't like as neighbours: people of different race (Q6B)	1,000	,564	
don't like as neighbours: left wing extremists (Q6C)	1,000	,754	
don't like as neighbours: heavy drinkers (Q6D)	1,000	,514	
don't like as neighbours: right wing extremists (Q6E)	1,000	,779	
don't like as neighbours: large families (Q6F)	1,000	,426	
don't like as neighbours: emotionally unstable people (Q6G)	1,000	,413	
don't like as neighbours: muslims (Q6H)	1,000	,597	
don't like as neighbours: immigrants/foreign workers (Q6I)	1,000	,543	
don't like as neighbours: people with AIDS (Q6J)	1,000	,637	
don't like as neighbours: drug addicts (Q6K)	1,000	,556	
don't like as neighbours: homosexuals (Q6L)	1,000	,634	
don't like as neighbours: jews (Q6M)	1,000	,493	
don't like as neighbours: gypsies (Q6N)	1,000	,376	
don't like as neighbours: christians (Q6O)	1,000	,637	

Extraction Method: Principal Component Analysis.

FIGURE 40 : COMMUNALITÉS –EXTRACTION DES VARIABLES (ACP)

Dans notre exemple, les variables les moins extraites dans l'analyse factorielle sont les voisins avec un casier criminel (0.311) et les gitans (0.376).

La Figure 41 montre le résumé de l'extraction des axes factoriels. Comme expliqué précédemment, l'analyse en composantes principales va créer autant de composantes qu'il y a de variables insérées dans l'analyse et va redistribuer l'information de manière inégales sur ceux-ci, il est dès lors important de regarder les valeurs propres attribuées à chaque axe ainsi que le % de variance associé.

Component	Total Variance Explained								
	Initial Eigenvalues			Extraction Sums of Squared			Rotation Sums of Squared		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3,121	20,804	20,804	3,121	20,804	20,804	2,129	14,192	14,192
2	1,552	10,346	31,149	1,552	10,346	31,149	1,76	11,735	25,926
3	1,412	9,416	40,565	1,412	9,416	40,565	1,625	10,834	36,76
4	1,139	7,591	48,156	1,139	7,591	48,156	1,551	10,34	47,1
5	1,011	6,737	54,893	1,011	6,737	54,893	1,169	7,793	54,893
6	0,9	5,997	60,89						
7	0,86	5,732	66,622						
8	0,803	5,354	71,976						
9	0,779	5,195	77,171						
10	0,668	4,45	81,621						
11	0,638	4,251	85,872						
12	0,592	3,947	89,82						
13	0,579	3,859	93,678						
14	0,518	3,453	97,131						
15	0,43	2,869	100						

Extraction Method: Principal Component Analysis.

FIGURE 41 : TABLEAU D'EXTRACTION DES AXES FACTORIELS

L'axe factoriel 1 a donc une valeur propre de 3.1, ce qui signifie que cette seule variable explique à elle toute seule 3.1 variables ou 20.8% de la variance totale des variables insérées dans l'analyse. Il y a donc bel et bien condensation de l'information. Avec le critère de Kaiser, nous retiendrons 5 axes factoriels, tous les autres ayant une valeur propre inférieure à 1 (ils ont donc un potentiel d'explication inférieur à 1 variable). Avec 5 axes factoriels, nous pouvons expliquer 54.8% de la variance. Pour rappel, avec 15 variables, il nous aurait fallu un peu moins du double de variable pour expliquer autant de pourcentage de variance. Dans la suite du tableau, vous pouvez voir dans les dernières colonnes comment l'information a été redistribuée après rotation. L'axe 1 n'a plus que 2.1 de valeurs propres et explique 14.1% de la variance.

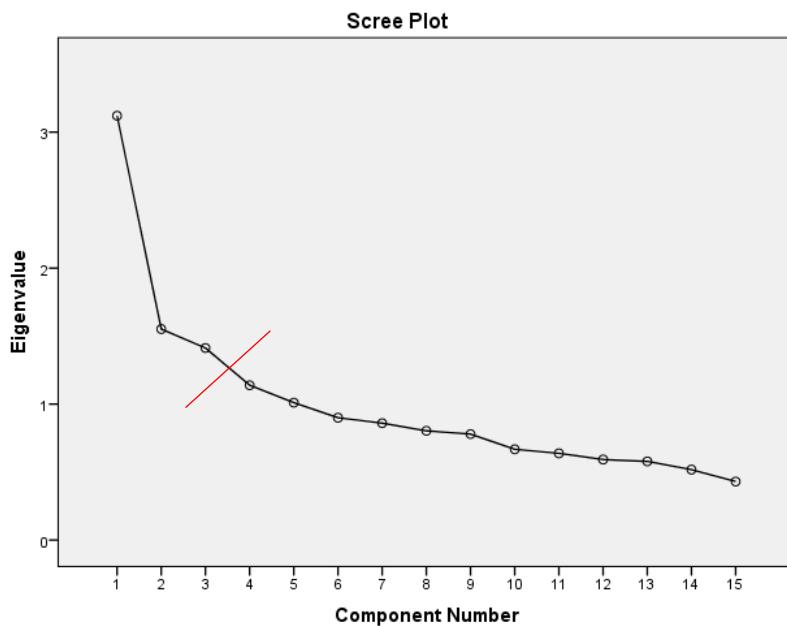


FIGURE 42 : COUDE DE CATELL

Avec la technique du coude de Catell, on garderait dans ce cas-ci seulement 3 facteurs. On constate donc bel et bien que la méthode est plus restrictive que le critère de Kaiser. Dans tous les cas, nous allons d'abord continuer l'analyse en suivant le critère de Kaiser. Si les résultats n'étaient pas concluants, nous relancerions l'analyse en conservant 3 facteurs.

Nous pouvons ensuite analyser les axes factoriels en tant que tel en analysant les fortes et moyennes corrélations³³ des variables aux axes. Au plus la variable sera corrélée à l'axe, au plus elle sera explicative de ce dernier. La matrice des composantes (Component Matrix) est affichée dans SPSS. Vous allez cependant voir que ce résultat n'est pas très intéressant car il est très difficile de trouver des spécificités aux axes factoriels. En effet, permettre de distinguer des axes factoriels sera plus aisé si des variables sont exclusivement corrélés à ceux-ci.

	Component Matrix ^a				
	Component				
	1	2	3	4	5
don't like as neighbours: people with criminal record (Q6A)	.422	.246	.195	-.169	.078
don't like as neighbours: people of different race (Q6B)	.632	-.241	-.227	-.235	-.011
don't like as neighbours: left wing extremists (Q6C)	.299	.671	-.456	.079	-.013
don't like as neighbours: heavy drinkers (Q6D)	.351	.305	.536	.072	-.075
don't like as neighbours: right wing extremists (Q6E)	.179	.670	-.517	.171	.017
don't like as neighbours: large families (Q6F)	.382	-.183	.072	.286	.399
don't like as neighbours: emotionally unstable people (Q6G)	.262	.253	.360	.056	.385
don't like as neighbours: muslims (Q6H)	.609	-.152	-.180	-.413	-.012
don't like as neighbours: immigrants/foreign workers (Q6I)	.570	-.237	-.180	-.353	.071
don't like as neighbours: people with AIDS (Q6J)	.561	-.220	.004	.374	-.367
don't like as neighbours: drug addicts (Q6K)	.389	.314	.545	-.007	-.093
don't like as neighbours: homosexuals (Q6L)	.500	-.161	.028	.475	-.363
don't like as neighbours: jews (Q6M)	.647	-.155	-.140	.173	-.015
don't like as neighbours: gypsies (Q6N)	.484	.132	.152	-.292	.125
don't like as neighbours: christians (Q6O)	.139	-.222	-.142	.391	.629

Extraction Method: Principal Component Analysis.
a. 5 components extracted.

FIGURE 43 : CORRÉLATIONS DES VARIABLES AUX COMPOSANTES AVANT ROTATION

Dans ce cas-ci, on voit en effet que presque toutes les variables sont moyennement ou fortement corrélées à l'axe factoriel 1 ou encore que certaines variables ne sont pas corrélées spécifiquement à un axe (ex : emotionaly unstable people). Il est dès lors difficile de spécifier une structure particulière.

Nous allons donc plutôt nous arrêter auprès de la matrice des corrélations des variables aux composantes après rotation (Rotated Component Matrix). Vous verrez aussitôt que l'analyse y est plus aisée. Il existe dans le logiciel Excel, une fonction bien pratique qui permet de surligner directement les corrélations moyennes ou fortes. Il vous suffit pour cela de copier-coller le tableau dans Excel, de le sélectionner et de choisir une mise en forme conditionnelle. Il vous est alors possible de spécifier de colorer les cases sélectionnées si la valeur au sein de celle-ci dépasse 0.3 ou est inférieure à -0.3. Vous pouvez ensuite rajouter des règles afin d'avoir des jeux de couleurs différents en fonction de la valeur de la corrélation.

	1	2	3	4
how much confidence in: church (Q63A)	0.228	-0.075	0.036	-0.105
how much confidence in: armed forces (Q63B)	-0.048	0.313	-0.006	0.167
how much confidence in: trade union system (Q63C)	0	-0.175	0.5	0.27
how much confidence in: the press (Q63D)	0.262	0.179	-0.044	0.074
how much confidence in: trade unions (Q63E)	0.087	0.036	0.074	0.761
how much confidence in: the police (Q63F)	0.301	0.105	0.274	0.317
how much confidence in: parliament (Q63G)	0.732	0.271	0.08	0.189
how much confidence in: civil service (Q63H)	0.524	0.162	0.291	0.098
how much confidence in: social security system (Q63I)	0.235	0.167	0.725	0.078
how much confidence in: european union (Q63J)	0.395	0.688	0.197	0.081
how much confidence in: the army (Q63K)	0.182	0.059	0.081	0.051
how much confidence in: united nations organisation (Q63L)	0.189	0.044	0.158	0.351
how much confidence in: health care system (Q63M)	0.097	0.189	0.795	0
how much confidence in: justice system (Q63N)	0.438	0.341	0.301	0.174
how much confidence in: major companies (Q63O)	0.293	0.127	0.114	0.088
how much confidence in: environmental organizations (Q63P)	0.221	0.093	0.16	0.481
how much confidence in: political parties (Q63Q)	0.795	0.071	0.008	0.232
how much confidence in: government (Q63R)	0.824	0.168	0.081	0.159

Extraction Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

³³ Pour rappel, une corrélation moyenne est observée à partir de (-) 0.3

	Rotated Component Matrix ^a				
	Component				
	1	2	3	4	5
don't like as neighbours: people with criminal record (Q6A)	,276	,467	,003	,128	,015
don't like as neighbours: people of different race (Q6B)	,706	,001	,238	,036	,085
don't like as neighbours: left wing extremists (Q6C)	,105	,107	,040	,854	-,031
don't like as neighbours: heavy drinkers (Q6D)	-,042	,690	,185	-,007	-,044
don't like as neighbours: right wing extremists (Q6E)	-,013	,006	,021	,882	,013
don't like as neighbours: large families (Q6F)	,146	,150	,194	-,053	,585
don't like as neighbours: emotionally unstable people (Q6G)	,002	,539	-,110	,047	,328
don't like as neighbours: muslims (Q6H)	,761	,081	,095	,042	-,027
don't like as neighbours: immigrants/foreign workers (Q6I)	,726	,027	,088	-,019	,080
don't like as neighbours: people with AIDS (Q6J)	,192	,079	,769	-,007	,044
don't like as neighbours: drug addicts (Q6K)	,025	,722	,163	-,012	-,090
don't like as neighbours: homosexuals (Q6L)	,066	,092	,785	,028	,068
don't like as neighbours: jews (Q6M)	,422	,087	,479	,120	,253
don't like as neighbours: gypsies (Q6N)	,440	,419	-,039	,060	,037
don't like as neighbours: christians (Q6O)	,003	-,121	,020	,019	,788

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization.

a. Rotation converged in 6 iterations.

FIGURE 44 : MISE EN FORME CONDITIONNELLE DANS EXCEL ET RÉSULTAT DU TABLEAU [VERT LÉGER (0.3<R<0.5), VERT (0.5<R<0.7), VERT FONCÉ (R>0.7)]

On voit désormais clairement que certaines variables sont exclusivement corrélées à un axe factoriel, ce qui va faciliter l'analyse. Il va falloir maintenant reprendre chacune de ces fortes corrélations et essayer de repérer ce qu'il y a en commun entre elles. Faisons l'exercice axe par axe en affichant les variables par ordre décroissant de corrélations.

- ❖ Axe 1 : N'aime pas les voisins musulmans, travailleurs étrangers, personne de race différentes, gitans et juifs.
➔ Interprétation de l'axe : Propension à discriminer des personnes sur base raciale ou culturelle.
- ❖ Axe 2 : N'aime pas les voisins qui se droguent, boivent beaucoup, émotionnellement instable, avec un casier judiciaire ou gitans.
➔ Interprétation de l'axe : Propension à discriminer des personnes sur base de comportements troubles³⁴.
- ❖ Axe 3 : N'aime pas les voisins homosexuels, atteints du SIDA et les juifs.
➔ Interprétation de l'axe : Indicateur d'homophobie³⁵.
- ❖ Axe 4 : N'aime pas les voisins qui votent extrême droite et extrême gauche.
➔ Interprétation de l'axe : Propension à discriminer sur base d'opinions politiques extrêmes.
- ❖ Axe 5 : N'aime pas les voisins chrétiens, les larges familles et les personnes émotionnellement instables.
➔ Interprétation de l'axe : Axe difficile à interpréter.

Rappelez-vous, si l'interprétation des axes est compliquée, c'est sans doute que le nombre d'axes retenus dans l'analyse n'était pas adéquat. On pourrait dans ce cas-ci essayé de relancer l'analyse en ne gardant que 3 axes comme le proposait la technique du coude de Catell. Cependant, rappelez-vous également que les derniers axes sont également ceux qui expliquent le moins de variances.

Vous avez désormais 5 nouvelles variables qui apparaissent dans votre jeu de données :

460	FAC1_1	Numeric	11	5	REGR factor score 1 for analysis 1	None
461	FAC2_1	Numeric	11	5	REGR factor score 2 for analysis 1	None
462	FAC3_1	Numeric	11	5	REGR factor score 3 for analysis 1	None
463	FAC4_1	Numeric	11	5	REGR factor score 4 for analysis 1	None
464	FAC5_1	Numeric	11	5	REGR factor score 5 for analysis 1	None
465						

³⁴ Bien que la corrélation avec les gitans puisse apparaître comme non relevant avec les autres variables, ces personnes souffrent souvent de stéréotypes supplémentaires que la différence culturelle, notamment celui de trouble à l'ordre public.

³⁵ Il est par contre plus difficile d'interpréter ici l'intégration des juifs dans cet axe. Un spécialiste sur le sujet pourrait sans doute répondre, l'exercice est ici fait rapidement.

Vous pouvez désormais modifier les caractéristiques de ces variables. De manière générale, le nom de la variable est donné de la façon suivante : FAC pour facteur, le premier chiffre pour le numéro de l'axe et le deuxième chiffre pour le numéro de l'analyse factorielle effectuée depuis l'ouverture de session.

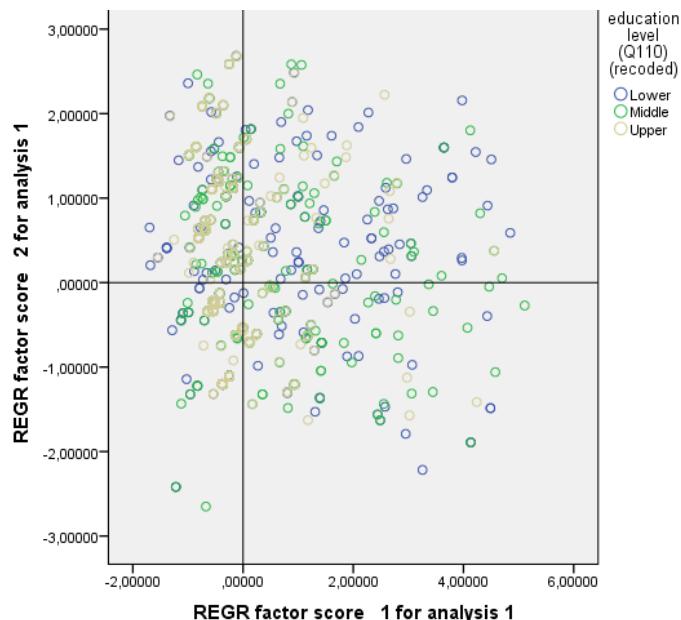
Ainsi, le premier axe factoriel de la deuxième analyse aurait pour nom : FAC1_2

Remarques

1.Toujours vérifier que les échelles des variables aillent dans le même sens par soucis de simplicité dans l'analyse. Si ce n'est pas le cas, procéder à une modification des variables : « Compute variable » → Newvar = – (Oldvar)³⁶.

2.Ne pas hésiter à demander des aides graphiques également. Pour cela : Graphs → Legacy Dialogs → Scatter/dot → Simple :

- ✓ Axe des Y : composante 2 (FAC2_1)
- ✓ Axe des X : composante 1 (FAC1_1)
- ✓ Option : Définir variable par : les modalités d'une variable particulière (V336_r)



Rappelez-vous, les axes factoriels sont des variables centrées-réduites. 0 représente donc la moyenne des individus. Ce qui va nous intéresser dans cette représentation graphique, c'est d'essayer de visualiser les individus qui sont distants de ce centre névralgique. Les personnes distantes de la moyenne X (= discrimination raciale ou culturelle) sont plus nombreuses à être d'éducation moyenne ou faible. Alors que les personnes distantes de la moyenne de l'axe 2 sont plus difficiles à différencier.

Validation d'indicateurs : l'analyse de fidélité

Rappel théorique

L'analyse en composantes principales, lorsqu'elle est utilisée dans un cadre déductif, va permettre de vérifier des indicateurs préconstruits. Le questionnaire sera dès lors organisé de telle sorte à ce que plusieurs questions soient amenées à mesurer un même indicateur. On va donc utiliser l'ACP afin de vérifier que les variables (items) mesurant une même dimension théoriquement, soient bel et bien corrélées exclusivement à un axe factoriel (cf. Figure 45). Si les corrélations de l'ACP confirment ce point, on peut alors passer à l'analyse de fidélité. Cette analyse, centrée autour de la statistique de l'alpha de Cronbach, va vérifier que la cohérence

³⁶ Cf. Chapitre 2 : Data management.

des items entre eux soit satisfaisante. L'analyse de fidélité se distingue de l'analyse en composantes principales dans le sens où les axes factoriels – même si les variables concernées sont bel et bien fortement corrélées à l'axe, vont être influencé par les petites corrélations des autres variables. Si dans une ACP, on ne regarde que les fortes corrélations à l'axe pour le définir, il n'empêche que toutes les variables sont corrélées à l'axe et modifient donc, ne fut-ce que très légèrement, le score Z de la composante. L'analyse de fidélité permet donc en réalité de confirmer que les items supposés mesurent un même indicateur et que, par conséquent, ils peuvent être fusionnés ensemble³⁷ sans tenir compte des infimes corrélations avec les autres items. L'analyse de fidélité et de construction d'indicateur est souvent lourde en terme de ressources et nécessite généralement de multiples enquêtes. Si elle est plus régulièrement utilisée dans le cadre des sciences psychologiques, elle s'applique également à d'autres domaines, notamment celui des sciences sociales.

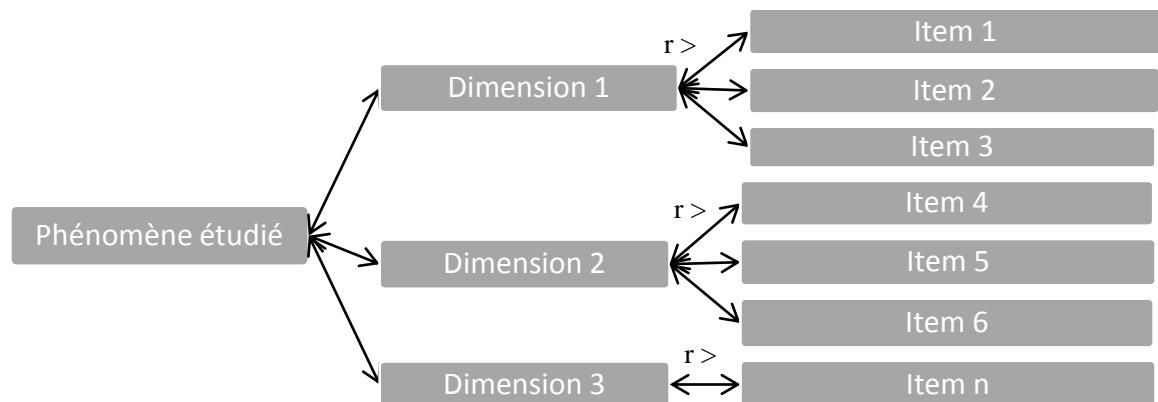


FIGURE 45 : SCHÉMA D'INDICATEURS (DIMENSIONS) PRÉCONSTRUITES

L'alpha de Cronbach est donc une statistique qui va mesurer la cohérence interne d'une dimension composée de plusieurs items. Assez logiquement, l'alpha de Cronbach va donc se baser sur la notion centrale de corrélation moyenne dont le calcul est le suivant :

$$\alpha = \frac{N \bar{r}}{1 + (N - 1)\bar{r}}$$

Si les corrélations entre les items de la même dimension sont fortes, nous postulons que les items ont une cohérence interne forte et sont donc, *de facto*, fidèles à la dimension testée. Les chercheurs s'accordent pour interpréter l'alpha de Cronbach selon la définition de Nannuly (1978). **Les résultats inférieurs à 0.7 seront évités.**

Cronbach's alpha Internal consistency ¹	
$\alpha \geq 0.9$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

¹ Nannuly, 1978.

Lors d'une analyse de fidélité, il est extrêmement important à ce que les variables aillent dans le même sens ! On ne peut dans ce cas-ci pas mélanger des variables négativement corrélées.

Une façon très simple d'inverser le sens d'une variable est d'utiliser la fonction compute variable avec la formule suivante : nouvelle_variable= -(variable).

³⁷ Comprenez par-là, les sommer pour créer une nouvelle variable (cf. Chapitre 2)

Manipulations dans SPSS

Nous allons repartir de l'exercice précédent effectué lors de l'analyse principale. Nous avions donc vu que des variables étaient principalement corrélées à certains axes factoriels. En voici, un rapide résumé pour les deux premiers axes (Cf. L'analyse en composantes principales, p.45) :

Axe 1 : Voisins de race différentes, musulmans et travailleurs étranger, juifs ou gitans (V47, V53, V54, V58, V59) : Indicateur de discrimination du voisinage en fonction de la race ou de l'appartenance culturelle.

Axe 2 : Voisins avec un casier criminel, qui boivent, émotionnellement instable, gitans et qui se droguent (V46, V49, V52, V56, V59) : Indicateur de discrimination du voisinage en fonction de son comportement.

Nous allons donc tester si ces variables mesurent une même dimension : Rendez-vous dans le menu *Analyze*, *Scale* (échelle) et enfin *Reliability analysis* (analyse de fidélité). Vous arrivez ensuite dans le menu illustré par la Figure 46). Insérez dans un premier temps les variables concernant le premier axe (V47, V53, V54, V58, V59). Par défaut, SPSS choisit la méthode « Alpha » (de Cronbach). Vérifiez qu'il n'y ait pas de changement. Dans le menu *statistics*, vous pouvez cocher les options suivantes : *Item*, *Scale* (échelle) et *Scale without Item* (Echelle sans l'élément). Vous pouvez également à voir les *corrélations* ainsi qu'une série de statistiques descriptives dans l'encadré *Summaries* (récapitulatif) comme la moyenne ou l'écart-type. Lancez ensuite l'analyse.

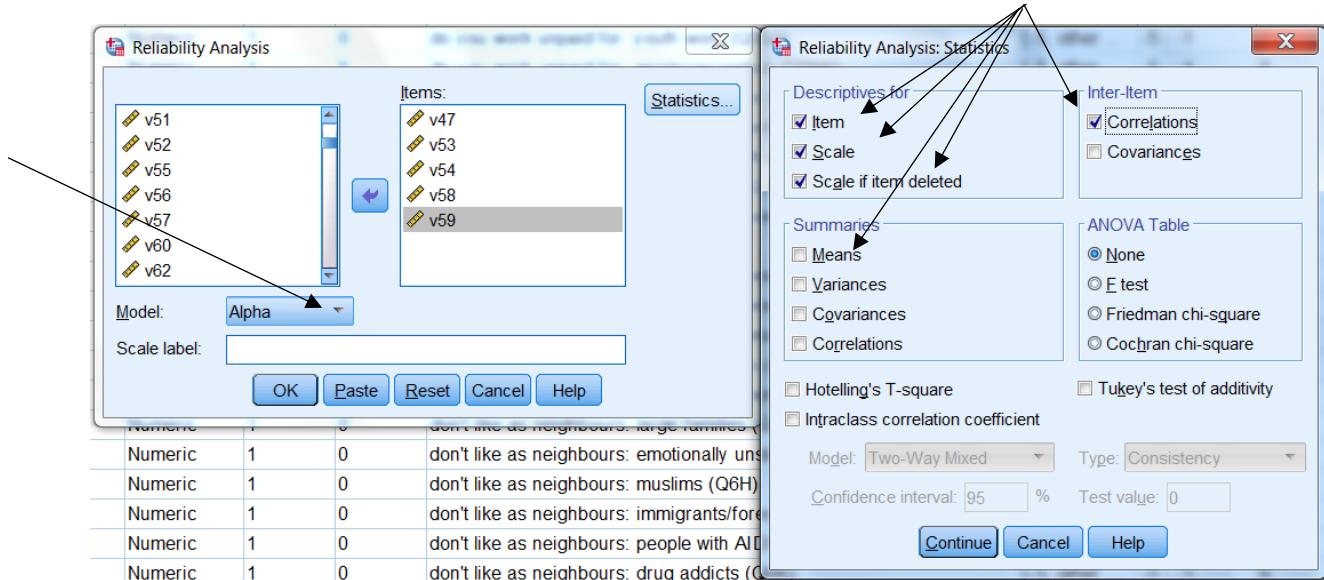


FIGURE 46 : MENU DE L'ANALYSE DE FIDÉLITÉ

Le premier tableau de sortie vous indique le nombre de données prises en compte dans l'analyse. A ce stade-ci, vous devez être accoutumé à voir apparaître ce tableau. Le deuxième tableau de sortie vous indique la statistique de l'alpha de Cronbach.

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.638	.689	5

Dans ce cas-ci, notre indice est assez mauvais : 0.638. SPSS nous indique également que si nous standardisons nos variables, l'alpha de Cronbach augmente à 0.689, ce qui reste un résultat faible.

Les deux tableaux suivants vous présentent des statistiques descriptives et une matrice de corrélation. Au plus les corrélations seront faibles entre vos variables, au plus faible sera votre alpha de Cronbach. Le tableau suivant (*Item-total Statistics*) est sans doute la sortie la plus intéressante après le tableau exposé ci-dessus. Il mesure les variations sur l'indicateur si l'on supprime la variable exprimée en ligne.

Nous pouvons ainsi voir dans la Figure 47 que si l'on retire la variable (je n'aime pas avoir pour voisin des gitans), l'alpha de Cronbach augmente, signifiant ainsi une meilleure cohérence de l'indicateur. Cela est d'autant plus logique que la variable citée est l'une des moins corrélées à

l'ensemble et que cette variable n'est pas aussi fortement corrélée que les trois premières à l'axe factoriel.

Item-Total Statistics					
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
don't like as neighbours: people of different race (Q6B)	,51	,721	,461	,287	,569
don't like as neighbours: muslims (Q6H)	,42	,559	,504	,274	,522
don't like as neighbours: immigrants/foreign workers (Q6I)	,50	,711	,441	,251	,572
don't like as neighbours: jews (Q6M)	,52	,772	,412	,190	,596
don't like as neighbours: gypsies (Q6N)	,30	,547	,319	,113	,674

FIGURE 47 : STATISTIQUES DE L'INDICATEUR SI ON SUPPRIME UN ITEM

Enfin, vous obtenez un tableau résumé de l'échelle si vous sommez les éléments pour en créer un indicateur.

Scale Statistics			
Mean	Variance	Std. Deviation	N of Items
,56	,949	,974	5

Le coefficient de Cronbach est ici trop faible que pour créer la variable. Cependant, s'il avait été correct, elle aurait eu les statistiques descriptives suivantes. Si l'on exécute le même exercice avec les variables du deuxième axe, l'on constate que notre indicateur de Cronbach est encore insatisfaisant. En d'autres termes, ces variables à elles-seules n'ont pas une cohérence interne suffisante que pour expliquer un même indicateur.

Techniques de clustering

L'analyse en cluster a pour but, à contrario de l'analyse en composantes principales, de détecter des structures au sein des données (et non des variables). En anglais, cluster signifie « groupe ». Cette technique va donc nous permettre d'identifier des groupes d'individus partageant une série de paramètres. Les groupes seront constitués à partir d'un algorithme itératif qui va calculer les distances entre individus.

Rappel théorique

La distance/proximité entre les observations est au centre du processus de classification. Il existe plusieurs façons de calculer une distance, la distance euclidienne en est une parmi d'autre. Le calcul euclidien entre deux observations est le suivant :

$$d = \sqrt{\sum_{i=1}^n (a_{var_i} - b_{var_i})^2} \quad \text{avec } \begin{cases} a \{var_1; var_2; var_3; \dots; var_n\} \\ b \{var_1; var_2; var_3; \dots; var_n\} \end{cases}$$

Bien souvent, on utilise plutôt le carré de la distance euclidienne. On va ensuite calculer une matrice de toutes les distances possibles entre observations. Une fois la matrice de distance calculée, il s'agit de savoir comment regrouper les informations entre elles. La **méthode de Ward** propose de regrouper les classes *en maximisant la variance interclasse et en minimisant la variance intra-classe*. Dit autrement, la méthode va tenter de créer des groupes les plus homogènes possibles (variance intraclasse minimale) et les plus distincts les uns des autres (variance interclasse maximale). Pour cela, on calcule la distance euclidienne au carré des barycentres de classes (G) pondéré par le produit des poids (W) de classe divisé par la somme de ces poids. Au départ, toutes les classes contiennent donc une et une seule observation (singleton) et ont donc le même poids et un barycentre égal à l'observation elle-même. La distance de Ward la plus faible entre deux classes constituera la nouvelle agglomération. Lorsque deux classes sont rassemblées, la matrice de distances est alors recalculée en fonction des barycentres de classe et du poids actualisés de chaque classe.

Distance de Ward pour deux classes A et B, de poids W et de barycentre G.

$$\Delta(A, B) = \frac{W_A W_B}{W_A + W_B} d^2(G_A, G_B)$$

Il s'agit donc d'un algorithme récursif (dès que deux observations sont regroupées, il faut recalculer les distances de tous les points pour effectuer le prochain regroupement) et hiérarchique (une fois qu'une opération est réalisée, on ne revient pas en arrière). Il y aura en tout et pour tout $N-1$ calcul pour arriver à l'agglomération totale des observations. Cette ultime agrégation revient donc à ne pas créer de groupes. La représentation graphique du *clustering* hiérarchique s'appelle un **dendrogramme**.

Une fois ces calculs effectués, l'analyste va devoir prendre une décision quant à la délimitation du nombre de classes d'individu final. Pour cela, on va décider de stopper les agrégations de clusters lorsque le coefficient de Ward augmente de manière significative. Une augmentation significative signifie que deux groupes relativement différents vont être fusionnés. Le choix de la délimitation des clusters est donc fait d'une balance entre une perte d'information (lorsque je fusionne deux observations, je perds leurs singularités respectives) et un nombre de classe relativement restreint (le but étant de créer des groupes d'individus significatifs). Une autre façon de faire est de repérer ces « bonds » du coefficient de Ward de manière visuelle, à l'aide du dendrogramme.

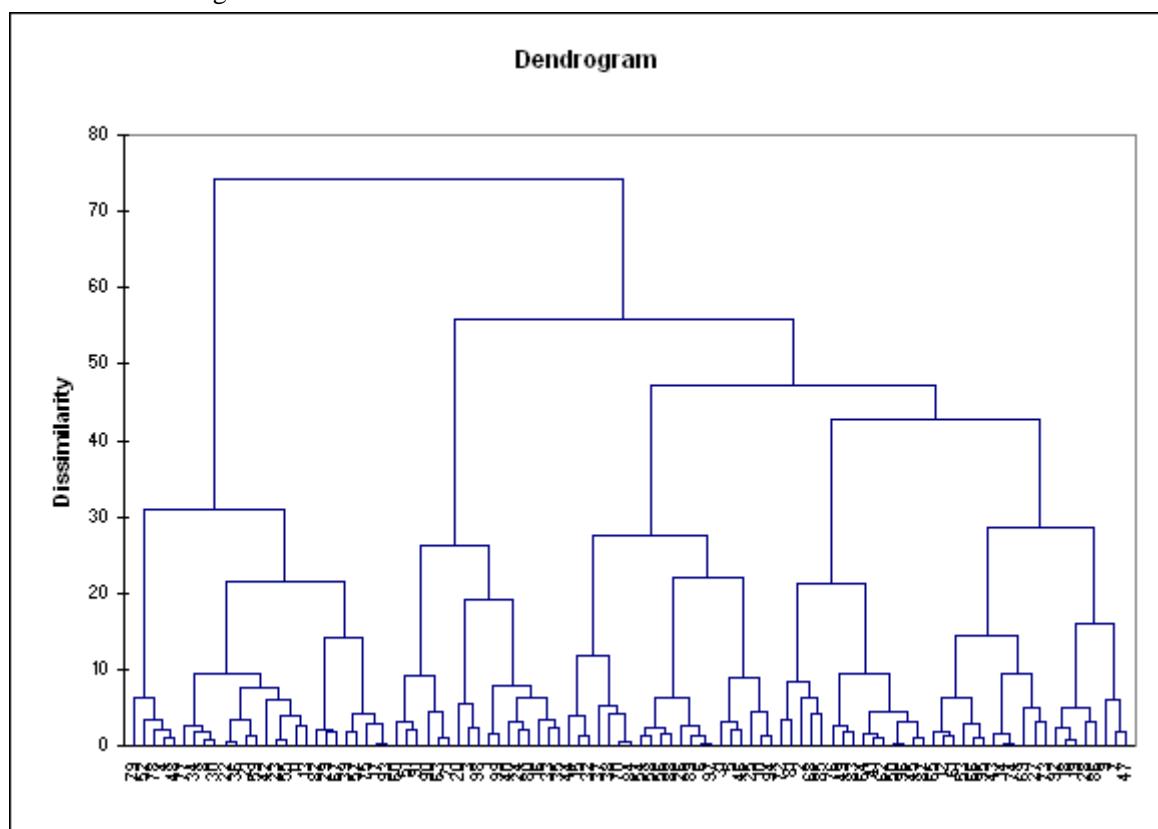


FIGURE 48 : EXEMPLE DE DENDROGRAMME

Lire un dendrogramme : toutes les observations sont représentées en abscisse et l'indice de distance est affiché en ordonnée (dans le cas de la figure ci-dessus, il s'agit d'un indice de dissimilarité et non une distance de Ward). Chaque trait horizontal représente une fusion alors que chacun trait vertical représente un cluster. Dans le cadre de la figure ci-dessus, il semblerait que conserver 5 groupes soit une bonne solution.

Une bonne façon de repérer les bonds significatifs dans la table des agrégations est d'opérer le calcul suivant pour chaque itération (i) : $\text{Ward}(i) / \text{Ward}(i-1)$.

Excel est un excellent outil pour faire ce genre de calcul sur la table d'agrégation (voir plus loin).

Une fois le nombre de cluster défini, il va falloir alors repérer de quoi sont constitués ces nouvelles classes en effectuant une série d'analyses descriptives à partir des variables insérées dans l'analyse mais également d'autres variables pouvant expliquer les caractéristiques du groupe. Une analyse en cluster est généralement plus performante si elle est précédée d'une analyse en composantes principales et donc composée d'axes factoriels.

Manipulations dans SPSS

Pour exécuter une classification hiérarchique dans SPSS, on exécute les fonctions suivantes :

« Analyse » (*Analyze*) → « Classifier » (*Classify*) → « Cluster hiérarchique » (*Hierarchical cluster*)

Dans l'encadré « Variable(s) » (*Variables*), insérer la liste des variables numériques, dichotomiques ou ordinaires sur lesquelles vous voulez effectuer la classification. Dans les onglets suivants, sélectionnez :

Statistiques (Statistics) : Sélectionnez le planning des agglomérations (*Agglomeration Schedule*) qui vous permet de voir les étapes de l'algorithme. C'est grâce à cette sortie que nous pourrons calculer les « sauts » significatifs de coefficient de Ward. Il vous est également possible de sélectionner la matrice des distances (*Proximity Matrix*). Cependant, cette sortie n'est pas analytiquement intéressante et va prendre un temps considérable à être produite lorsque le nombre d'observation est élevé.

Tracé (Plots) : sélectionner l'option « Dendrogramme ».

Méthode (Method) : Sélectionner la Méthode de Ward (*Ward's method*) et dans l'option distance, choisissez carré de la distance euclidienne (*Squared Euclidian Distance*). Standardiser les données pourrait être intéressant mais il est dès lors plus pratique d'effectuer une ACP au préalable. Dans ce cas-là, on n'opère aucune transformation des données.

Enregistrer (Save) : Permet de créer une nouvelle variable avec les affectations de classes. Il vous est possible de sauvegarder plusieurs solutions. Comme pour l'analyse en composantes principales, SPSS va leur donner un nom générique : CLU3_1. « CLU » signifie que la variable a été réalisée à la suite d'une analyse en clusters, le premier chiffre désigne le nombre de classes retenues dans la variable (ici, une variable à 3 classes) et le deuxième chiffre représente le numéro de l'analyse en clusters effectué depuis l'ouverture de session. Ces variables seront des variables nominales.

Pour cet exercice, sélectionnez les ressortissants de Belgique et insérer dans l'analyse les variables V136, V138, V139, V145 et V146. Ces variables traitent de l'importance que les individus accordent dans certaines valeurs au sein du mariage. Lancez l'analyse.

Interprétation des résultats

1) Table des agrégations.

Cette table indique les itérations effectuées par la technique de clustering hiérarchique de Ward. On peut donc voir qu'à la première étape de la figure ci-dessous, le logiciel a fusionné les observations 7488 et 7558 en une nouvelle classe. Le « nom » de ce nouveau groupe portera le numéro de l'observation insérée dans la colonne *Cluster 1*. Le coefficient de Ward est de 0. Cela signifie que les observations fusionnées ont répondu de la même manière sur les 5 variables, vu que la distance est nulle, autrement dit, il n'y a aucune variation entre ces deux observations. Très logiquement, les premières fusions seront identiques, vu que le cluster hiérarchique va privilégier les fusions où le coefficient est minimal. Une fois que le coefficient de distance augmente, cela signifie que nous fusionnons des clusters contenant des différences. Les trois colonnes suivantes vous indiquent quand le cluster (1 ou 2) a été utilisé préalablement (au début, étant donné que toutes les observations sont des clusters, vous verrez apparaître le chiffre 0, signifiant que l'observation n'a pas encore été utilisée) et lorsque ce cluster nouvellement créé sera réutilisé. Pour la première ligne, nous pouvons voir que ce nouveau cluster « 7488 » sera réutilisé à l'itération 71.

Etape	Cluster combiné		Coefficients	Etape de première apparition du cluster		Etape suivante
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	7488	7558	,000	0	0	71
2	7417	7557	,000	0	0	138
3	7326	7556	,000	0	0	226
4	7520	7555	,000	0	0	39
5	7475	7554	,000	0	0	83
6	7551	7553	,000	0	0	8
7	7544	7552	,000	0	0	15
8	6098	7551	,000	0	6	24

Suivent les fusions 9 à 1484...

1485	6050	6054	1122,038	1478	1470	1487
1486	6051	6052	1211,809	1480	1482	1488
1487	6050	6057	1329,553	1485	1483	1491
1488	6051	6075	1447,491	1486	1467	1492
1489	6053	6062	1571,927	1484	1477	1490
1490	6053	6070	1756,911	1489	1474	1491
1491	6050	6053	2228,934	1487	1490	1492
1492	6050	6051	2825,086	1491	1488	0

FIGURE 49 : TABLE DES AGRÉGATIONS

Comme l'analyse en clusters présuppose une perte d'information du fait du regroupement des individus en classes, c'est plutôt la fin du tableau qui va nous intéresser. Ou, autrement dit, vers la fin des agrégations qu'il va prêter une attention plus grande. Nous allons donc être vigilent à ne pas fusionner deux classes trop différentes l'une de l'autre en comparant les incrémentations du coefficient de Ward à chaque étape de fusion. Pour cela deux manières de faire : ou vous regarder directement dans la table des agrégations en comparant « manuellement » les différentes augmentations du coefficient d'étape en étape, ou vous utilisez Excel. Si vous optez pour la deuxième option, copiez le tableau des agrégations et collez-le dans une feuille Excel.

	A	B	C	D	E	F	G	H
1485	1482	6052	6067	909,563	1471	1461	1486	
1486	1483	6057	6066	975,418	1463	1479	1487	
1487	1484	6053	6064	1041,776	1481	1475	1489	
1488	1485	6050	6054	1122,038	1478	1470	1487	
1489	1486	6051	6052	1211,809	1480	1482	1488	
1490	1487	6050	6057	1329,553	1485	1483	1491	
1491	1488	6051	6075	1447,491	1486	1467	1492	
1492	1489	6053	6062	1571,927	1484	1477	1490	
1493	1490	6053	6070	1756,911	1489	1474	1491	
1494	1491	6050	6053	2228,934	1487	1490	1492	
1495	1492	6050	6051	2825,086	1491	1488	0	=D1495/D1494
1496								

FIGURE 50 : INSERTION D'UNE FORMULE DANS EXCEL

Une fois cela fait, rendez-vous à la dernière ligne du tableau et sélectionnez la case de la ligne correspondante d'une colonne vide. Vous allez devoir créer une chaîne de fonctions. Pour cela, commencez par placer un =

dans la case (Excel repèrera ainsi qu'il s'agit d'une fonction). Sélectionnez ensuite la case du coefficient de la ligne (case en bleue dans la Figure 50), divisez (/) et sélectionnez la case du coefficient de la ligne du dessus. Comme indique dans la figure, vous devez obtenir la figure suivante :

$$=D1495/D1494$$

Vous obtenez ainsi un ratio de l'incrément du coefficient de Ward. Il est assez clair qu'au plus on avance dans les itérations, au plus on mélange des groupes différents et que donc le coefficient de Ward grandit. Il ne s'agit donc pas de repérer le plus grand coefficient (fusion finale) ou la plus grande différence d'apport du coefficient mais bel et bien de repérer un incrément du coefficient *relatif*.

Une fois la formule effectué dans la case de la dernière ligne, il vous suffit de la sélectionner et de la faire glisser sur les autres cases de la colonne. Excel va ainsi calculer le même calcul pour chaque itération à partir de la ligne i et en fonction de la ligne i-1 comme présenté dans la Figure 51, dans la colonne orangée.

1481	1478	6050	6110	696,077	1380	1464	1485	1,07398573
1482	1479	6066	6068	746,702	1350	1360	1483	1,07272902
1483	1480	6051	6063	798,698	1466	1379	1486	1,0696342
1484	1481	6053	6143	852,453	1469	1472	1484	1,06730329
1485	1482	6052	6067	909,563	1471	1461	1486	1,0669949
1486	1483	6057	6066	975,418	1463	1479	1487	1,0724029
1487	1484	6053	6064	1041,776	1481	1475	1489	1,06803032
1488	1485	6050	6054	1122,038	1478	1470	1487	1,07704343
1489	1486	6051	6052	1211,809	1480	1482	1488	1,08000709
1490	1487	6050	6057	1329,553	1485	1483	1491	1,09716383
1491	1488	6051	6075	1447,491	1486	1467	1492	1,088705
1492	1489	6053	6062	1571,927	1484	1477	1490	1,08596668
1493	1490	6053	6070	1756,911	1489	1474	1491	1,11767977
1494	1491	6050	6053	2228,934	1487	1490	1492	1,26866643
1495	1492	6050	6051	2825,086	1491	1488	0	1,26746059

FIGURE 51 : ANALYSE DES RATIOS D'INCRÉMENTS DU COEFFICIENT DE WARD

Nous remarquons ainsi que l'avant-dernière étape est l'étape où l'incrément relatif du coefficient de Ward est le plus élevé³⁸. Nous allons donc essayer de ne pas subir cette fusion mélangeant des groupes potentiellement différents. Nous retiendrons donc 3 classes à l'issue de cette analyse.

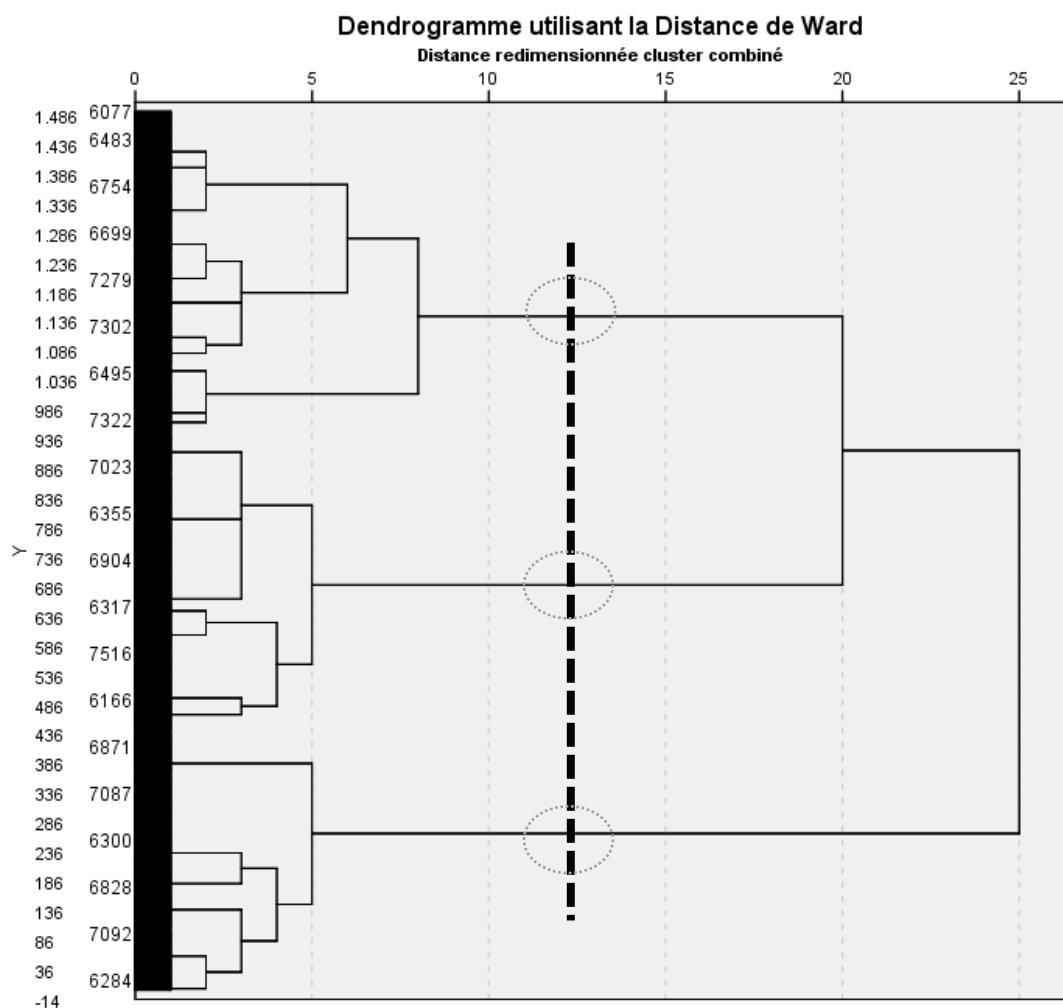
2) Dendrogramme

Cette illustration n'est pas très pratique dans SPSS, il faut généralement la redimensionner. Mais une fois cette étape exécutée, le dendrogramme présente une grande utilité, puisqu'il illustre graphiquement la table des agrégations ou autrement dit, la perte d'informations observée à chaque regroupement de deux classes. La hauteur du dendrogramme est dès lors proportionnelle à la perte d'inertie interclasse de Ward ou à la distance entre deux objets regroupés. Le coefficient de Ward dans le dendrogramme est redimensionné.

Si on ne sait pas exactement combien de classes on veut créer, l'analyse du dendrogramme est un premier bon moyen pour en déterminer le nombre optimal (c'est-à-dire un compromis entre un objectif de synthèse et un objectif de perte raisonnable d'informations).

Dans l'exemple ci-dessous, la césure à 3 classes est évidente. Cette analyse, aboutissant à la création de trois groupes, implique une perte d'information de 1756,911. Cela signifie donc qu'un peu moins de 40% de l'information initiale est conservée.

³⁸ Dans ce cas-ci, le ratio est encore relativement faible. Dans certains cas, vous obtiendrez des ratios de 2 ou 3, rendant le nombre de classes à définir très évident.



Une fois que le nombre optimal de classes est défini, il reste à demander à SPSS de créer ces classes et d'en attribuer une à chaque unité d'analyse. Pour cela, retourner dans le menu de la technique de cluster et sélectionnez l'option « SAVE » et « SINGLE SOLUTION » à 3 classes.

Commence ensuite le travail de description des classes (cf. Chapitre 3). Dépendant du type de variables utilisées pour le regroupement, ou de vos affinités avec certaines méthodes, vous vous dirigerez vers des tableaux croisés ou des moyennes.

1) Un tableau de fréquences de la nouvelle variable « cluster »

Ward Method

		Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Classe 1	506	33,5	33,9	33,9
	Classe 2	440	29,2	29,5	63,4
	Classe 3	547	36,2	36,6	100,0
	Total	1493	98,9	100,0	
Manquant	Système	16	1,1		
	Total	1509	100,0		

L'analyse classificatoire a été réalisée seulement pour les individus ayant participé à l'enquête en Belgique. La répartition des individus au sein des trois classes montre une répartition à peu près équitable : sur les 1493 individus pour lesquels on dispose de réponses valides pour l'ensemble des variables utilisées, 33.9% appartiennent à la première classe, 29.5% composent la deuxième classe et 36.6% se situant quant à eux dans la troisième classe.

2) Moyennes et Ecarts-types calculés pour chaque classe et chaque variable utilisée dans la classification

Statistiques descriptives

Ward Method		N	Moyenne	Ecart type
Classe 1	important in marriage: faithfulness (Q42A)	506	1,22	,465
	important in marriage: same social background (Q42C)	506	2,63	,526
	important in marriage: shared religious beliefs (Q42D)	506	2,94	,240
	important in marriage: children (Q42J)	506	1,35	,478
	important in marriage: discuss problems (Q42K)	506	1,29	,482
	N valide (liste)	506		
Classe 2	important in marriage: faithfulness (Q42A)	440	1,00	,000
	important in marriage: same social background (Q42C)	440	1,85	,683
	important in marriage: shared religious beliefs (Q42D)	440	1,70	,577
	important in marriage: children (Q42J)	440	1,01	,095
	important in marriage: discuss problems (Q42K)	440	1,14	,361
	N valide (liste)	440		
Classe 3	important in marriage: faithfulness (Q42A)	547	1,22	,463
	important in marriage: same social background (Q42C)	547	2,20	,633
	important in marriage: shared religious beliefs (Q42D)	547	2,24	,702
	important in marriage: children (Q42J)	547	2,40	,520
	important in marriage: discuss problems (Q42K)	547	1,18	,386
	N valide (liste)	547		

Pour rappel : vérifiez toujours le sens des échelles. Dans ce cas-ci : 1= très important, 2= moyennement important et 3= pas important. Juste avec des moyennes, nous voyons déjà comment se caractérisent grossièrement les classes : la fidélité est importante aux yeux de tous et l'est unanimement au sein de la classe 2. La classe 2 se démarque en ayant les moyennes relatives les plus basses. La classe 3 veut moins d'enfants alors que la classe 1 ne trouve pas d'une importance capitale le partage de croyances religieuses ou la provenance d'un même milieu social.

3) Tableaux croisés avec chaque variable ayant servi à la classification (+ une ou plusieurs variables de type sociodémographique)

Ward Method * important in marriage: faithfulness (Q42A)

Tableau croisé

Ward Method	Classe 1	Effectif	important in marriage: faithfulness (Q42A)			Total
			very important	rather important	not very important	
			% dans Ward Method	% dans important in marriage: faithfulness	% dans important in marriage: faithfulness	
	Classe 2	Effectif	440	0	0	440
		% dans Ward Method	100,0%	0,0%	0,0%	100,0%
		% dans important in marriage: faithfulness	34,2%	0,0%	0,0%	29,5%
	Classe 3	Effectif	440	95	12	547
		% dans Ward Method	80,4%	17,4%	2,2%	100,0%
		% dans important in marriage: faithfulness	34,2%	51,4%	52,2%	36,6%
	Total	Effectif	1285	185	23	1493
		% dans Ward Method	86,1%	12,4%	1,5%	100,0%
		% dans important in marriage: faithfulness	100,0%	100,0%	100,0%	100,0%

Ward Method * important in marriage: same social background (Q42C)

Tableau croisé

Ward Method	Classe 1	Effectif	important in marriage: same social background (Q42C)			Total
			very important	rather important	not very important	
			% dans Ward Method	% dans important in marriage: same social background (Q42C)	% dans important in marriage: same social background (Q42C)	
	Classe 2	Effectif	139	226	75	440
		% dans Ward Method	31,6%	51,4%	17,0%	100,0%
		% dans important in marriage: same social background (Q42C)	64,7%	32,5%	12,9%	29,5%

	Classe 3 Effectif	65	305	177	547
	% dans Ward Method	11,9%	55,8%	32,4%	100,0%
	% dans important in marriage: same social background (Q42C)	30,2%	43,9%	30,4%	36,6%
Total	Effectif	215	695	583	1493
	% dans Ward Method	14,4%	46,6%	39,0%	100,0%
	% dans important in marriage: same social background (Q42C)	100,0%	100,0%	100,0%	100,0%

Ward Method * important in marriage: shared religious beliefs (Q42D)

Tableau croisé

Ward Method	Classe 1	Effectif	important in marriage: shared religious beliefs (Q42D)			Total
			very important	rather important	not very important	
Ward Method	Classe 1 Effectif	0	31	475	506	
	% dans Ward Method	0,0%	6,1%	93,9%	100,0%	
	% dans important in marriage: shared religious beliefs (Q42D)	0,0%	5,8%	66,2%	33,9%	
Ward Method	Classe 2 Effectif	159	254	27	440	
	% dans Ward Method	36,1%	57,7%	6,1%	100,0%	
	% dans important in marriage: shared religious beliefs (Q42D)	65,2%	47,7%	3,8%	29,5%	
Ward Method	Classe 3 Effectif	85	247	215	547	
	% dans Ward Method	15,5%	45,2%	39,3%	100,0%	
	% dans important in marriage: shared religious beliefs (Q42D)	34,8%	46,4%	30,0%	36,6%	
Total	Effectif	244	532	717	1493	
	% dans Ward Method	16,3%	35,6%	48,0%	100,0%	
	% dans important in marriage: shared religious beliefs (Q42D)	100,0%	100,0%	100,0%	100,0%	

Ward Method * important in marriage: children (Q42J)

Tableau croisé

Ward Method	Classe 1	Effectif	important in marriage: children (Q42J)			Total
			very important	rather important	not very important	
Ward Method	Classe 1 Effectif	328	178	0	506	
	% dans Ward Method	64,8%	35,2%	0,0%	100,0%	
	% dans important in marriage: children	42,5%	36,2%	0,0%	33,9%	
Ward Method	Classe 2 Effectif	436	4	0	440	
	% dans Ward Method	99,1%	0,9%	0,0%	100,0%	
	% dans important in marriage: children	56,5%	0,8%	0,0%	29,5%	
Ward Method	Classe 3 Effectif	8	310	229	547	
	% dans Ward Method	1,5%	56,7%	41,9%	100,0%	
	% dans important in marriage: children	1,0%	63,0%	100,0%	36,6%	
Total	Effectif	772	492	229	1493	
	% dans Ward Method	51,7%	33,0%	15,3%	100,0%	
	% dans important in marriage: children	100,0%	100,0%	100,0%	100,0%	

Ward Method * important in marriage: discuss problems (Q42K)

Tableau croisé

Ward Method	Classe 1	Effectif	important in marriage: discuss problems (Q42K)			Total
			very important	rather important	not very important	
Ward Method	Classe 1 Effectif	363	137	6	506	
	% dans Ward Method	71,7%	27,1%	1,2%	100,0%	
	% dans important in marriage: discuss problems (Q42K)	30,4%	47,4%	66,7%	33,9%	
Ward Method	Classe 2 Effectif	380	58	2	440	
	% dans Ward Method	86,4%	13,2%	0,5%	100,0%	
	% dans important in marriage: discuss problems (Q42K)	31,8%	20,1%	22,2%	29,5%	
Ward Method	Classe 3 Effectif	452	94	1	547	
	% dans Ward Method	82,6%	17,2%	0,2%	100,0%	
	% dans important in marriage: discuss problems (Q42K)	37,8%	32,5%	11,1%	36,6%	
Total	Effectif	1195	289	9	1493	
	% dans Ward Method	80,0%	19,4%	0,6%	100,0%	
	% dans important in marriage: discuss problems (Q42K)	100,0%	100,0%	100,0%	100,0%	

Ces 5 tableaux nous donne grosso modo, les mêmes informations que celles obtenues via la procédure de moyennes. Les 2 tableaux suivants, en revanche, viennent étayer nos résultats.

Ward Method * are you a religious person (Q28)

Tableau croisé

Ward Method	Classe 1		are you a religious person (Q28)			Total
			religious person	not religious person	convinced atheist	
Ward Method	Classe 1	Effectif	227	205	67	499
		% dans Ward Method	45,5%	41,1%	13,4%	100,0%
		% dans are you a religious person	26,5%	44,7%	40,4%	33,7%
	Classe 2	Effectif	339	78	23	440
		% dans Ward Method	77,0%	17,7%	5,2%	100,0%
		% dans are you a religious person	39,6%	17,0%	13,9%	29,7%
	Classe 3	Effectif	290	176	76	542
		% dans Ward Method	53,5%	32,5%	14,0%	100,0%
		% dans are you a religious person	33,9%	38,3%	45,8%	36,6%
Total		Effectif	856	459	166	1481
		% dans Ward Method	57,8%	31,0%	11,2%	100,0%
		% dans are you a religious person	100,0%	100,0%	100,0%	100,0%

Ward Method * sex respondent (Q86)

Tableau croisé

Ward Method	Classe 1		sex respondent (Q86)		Total
			male	female	
Ward Method	Classe 1	Effectif	273	233	506
		% dans Ward Method	54,0%	46,0%	100,0%
		% dans sex respondent (Q86)	37,9%	30,2%	33,9%
	Classe 2	Effectif	184	256	440
		% dans Ward Method	41,8%	58,2%	100,0%
		% dans sex respondent (Q86)	25,5%	33,2%	29,5%
	Classe 3	Effectif	264	283	547
		% dans Ward Method	48,3%	51,7%	100,0%
		% dans sex respondent (Q86)	36,6%	36,7%	36,6%
Total		Effectif	721	772	1493
		% dans Ward Method	48,3%	51,7%	100,0%
		% dans sex respondent (Q86)	100,0%	100,0%	100,0%

Nous pouvons donc constater que la classe 2, unanime sur la question de l'importance de la fidélité, donc la place faite aux enfants est importante, où il est important de partager les mêmes croyances et de provenir d'un même milieu socio-économique, est une classe essentiellement religieuse et composée d'une majorité de femmes. En revanche, les classes 1 et 3, sont moins religieuses, la première est une classe plus masculine alors que la deuxième est plus féminine.

IL VOUS EST POSSIBLE DE RÉALISER UN TABLEAU DE SYNTHÈSE À PARTIR DE L'ENSEMBLE DE CES INFORMATIONS CROISÉES COMME MONTRÉ À LA

Ou avec des symboles résumés :

		Classe 1 N = 506	Classe 2 N = 440	Classe 3 N = 547
Importance de la fidélité		+	++	+
Importance d'avoir un background social identique		-	+	+/-
Importance de partager des croyances religieuses communes		--	+	-
Importance des enfants		+	++	-
Importance de discuter de problèmes		+	++	++
Appartenance religieuse du répondant	Personne religieuse	45.5	77	53.5
	Personne non religieuse/Athéée	55.5	23	46.5
Sexe du répondant	Homme	54.0	41.8	48.3
	Femmes	46.0	58.2	51.7

Figure 52. L'analyse de ce tableau de synthèse permet enfin de décrire les trois classes d'individus qui ont été constituées sur base d'indicateurs évaluant l'importance accordée, dans le cadre du mariage, à la fidélité,

le partage d'une même origine sociale, le partage de croyances religieuses, aux enfants ou encore à la discussion de problèmes.



Pour la description des classes, nous vous recommandons d'attirer l'attention sur trois points particuliers :

- L'importance relative de la classe
- Les particularités spécifiques de la classe
- Ce qui distingue cette classe des autres.

		Classe 1 N = 506	Classe 2 N = 440	Classe 3 N = 547
Importance de la fidélité	Très important	80.0	100.0	80.4
	Moyennement important	17.8	0.0	17.4
	Pas important	2.2	0.0	2.2
Importance d'avoir un background social identique	Très important	2.2	31.6	11.9
	Moyennement important	32.4	51.4	55.8
	Pas important	65.4	17.0	32.4
Importance de partager des croyances religieuses communes	Très important	0.0	36.1	15.5
	Moyennement important	6.1	57.7	45.2
	Pas important	93.9	6.1	39.3
Importance des enfants	Très important	64.8	99.1	1.5
	Moyennement important	35.2	0.9	56.7
	Pas important	0.0	0.0	41.9
Importance de discuter de problèmes ³⁹	Très important	71.7	86.4	82.6
	Moyennement important	27.1	13.2	17.2
	Pas important	1.2	0.5	0.2
Appartenance religieuse du répondant	Personne religieuse	45.5	77.0	53.5
	Personne non religieuse	41.1	17.7	32.5
	Athée convaincu	13.4	5.2	14.0
Sexe du répondant	Homme	54.0	41.8	48.3
	Femmes	46.0	58.2	51.7

Ou avec des symboles résumés :

		Classe 1 N = 506	Classe 2 N = 440	Classe 3 N = 547
Importance de la fidélité		+	++	+
Importance d'avoir un background social identique		-	+	+/-
Importance de partager des croyances religieuses communes		--	+	-
Importance des enfants		+	++	-
Importance de discuter de problèmes		+	++	++
Appartenance religieuse du répondant	Personne religieuse	45.5	77	53.5
	Personne non religieuse/Athéée	55.5	23	46.5
Sexe du répondant	Homme	54.0	41.8	48.3
	Femmes	46.0	58.2	51.7

FIGURE 52 : TABLEAU DE SYNTHÈSE DE DÉFINITION DES CLASSES

³⁹ Vous remarquerez que, concernant la lecture du tableau croisant la variable en classes avec l'importance accordée au fait que l'on discute de problèmes, les conditions d'interprétation du Chi-deux ne sont pas toutes remplies. S'il s'agissait d'une variable non intégrée dans la classification, cela signifierait qu'en l'état, un recodage est tout à fait nécessaire avant de vouloir interpréter les (dis)similitudes entre classes.

Clustering sur Analyse en composantes principales

Les mêmes manipulations peuvent être reproduites, à la différence près que plutôt que d'inclure les variables d'origine pour la classification, on inclut les composantes qui ont été créées au préalable, à partir d'une analyse factorielle. L'avantage est que, dans ce cas, il s'agit de variables directement standardisées.

Résultats de l'analyse factorielle

Les résultats de l'analyse factorielle réalisée pour les cinq variables V136, V138, V139, V145 et V146 nous montrent que deux facteurs peuvent être créés. Ils contiennent à eux deux, un peu plus de 50% de la variance initiale.

Composante	Valeurs propres initiales			Sommes extraites du carré des chargements			Sommes de rotation du carré des chargements		
	Total	% de la variance	% cumulé	Total	% de la variance	% cumulé	Total	% de la variance	% cumulé
1	1,535	30,690	30,690	1,535	30,690	30,690	1,433	28,665	28,665
2	1,123	22,469	53,159	1,123	22,469	53,159	1,225	24,495	53,159
3	.909	18,188	71,348						
4	.777	15,540	86,887						
5	.656	13,113	100,000						

Méthode d'extraction : Analyse en composantes principales.

Le tableau suivant, présentant les corrélations entre les facteurs et les variables initiales, après rotation, permet quant à lui de nommer les facteurs et identifier l'information qu'ils recouvrent.

Rotation de la matrice des composantes^a

	Composante	
	1	2
important in marriage: faithfulness (Q42A)	.099	.755
important in marriage: same social background (Q42C)	.792	-.053
important in marriage: shared religious beliefs (Q42D)	.779	.038
important in marriage: children (Q42J)	.431	.361
important in marriage: discuss problems (Q42K)	-.057	.722

Méthode d'extraction : Analyse en composantes principales.

Méthode de rotation : Varimax avec normalisation Kaiser.

a. Convergence de la rotation dans 3 itérations.

Le premier facteur est fortement corrélé aux variables « importance de provenir d'un même milieu social », « importance de partager des mêmes croyances religieuses » et, à moindre mesure, « importance d'avoir des enfants ». Il regroupe donc des variables qui traitent communément du partage d'un bagage socioculturel entre partenaires d'un même couple. Il regroupe aussi la variable « importance des enfants » mais qui est, a priori, moins corrélée avec le facteur.

Le second facteur regroupe donc des variables qui traitent de l'importance de la fidélité et de la discussion de problèmes, c'est-à-dire de soucis qui peuvent altérer la qualité de la relation matrimoniale (infidélité, absence de communication).

Résultats de l'analyse classificatoire

L'analyse classificatoire peut être réitérée mais dans ce cas-ci, avec les deux facteurs qui synthétisent les cinq variables initiales⁴⁰. Dans ce cas, pour la description des classes, nous n'utiliserons plus que des tableaux croisant la variable « classe » créée, avec les variables les plus corrélées à chacune des composantes utilisées, de même que quelques variables de contextualisation qui n'interviennent pas directement dans l'analyse factorielle (ex : âge, sexe, religion, etc.). Il vous est également possible de préférer une sortie de moyennes sur les facteurs.

⁴⁰ Vous aurez donc pris soin de renommer les variables FACx_1 et FACx_2, ainsi que de les labelliser.

Vous pouvez directement remarquer que l'interprétation du dendrogramme (Figure 53) est plus simple que celle du graphique de la manipulation précédente. Le résultat est cependant identique : nous conserverons trois classes.

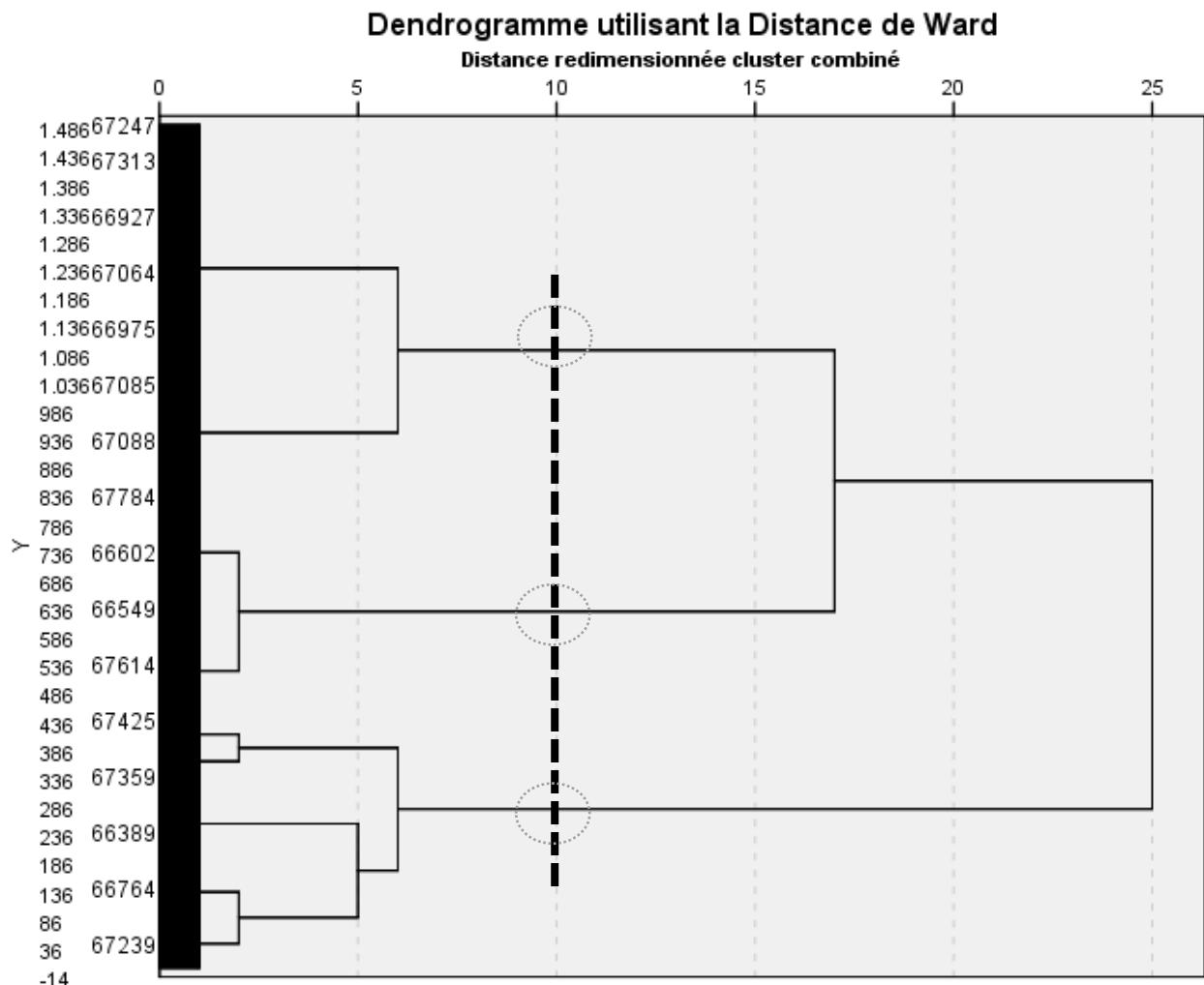


FIGURE 53 : DENDROGRAMME DU CLUSTERING APRÈS ACP

Ward Method * important in marriage: faithfulness (Q42A)
Tableau croisé

			important in marriage: faithfulness (Q42A)				Total
			very important	rather important	not very important		
Ward Method	1	Effectif	660	0	0	660	
		% dans Ward Method	100,0%	0,0%	0,0%	100,0%	
		% dans important in marriage: faithfulness	51,4%	0,0%	0,0%	44,2%	
	2	Effectif	216	185	23	424	
		% dans Ward Method	50,9%	43,6%	5,4%	100,0%	
		% dans important in marriage: faithfulness	16,8%	100,0%	100,0%	28,4%	
	3	Effectif	409	0	0	409	
		% dans Ward Method	100,0%	0,0%	0,0%	100,0%	
		% dans important in marriage: faithfulness	31,8%	0,0%	0,0%	27,4%	
Total		Effectif	1285	185	23	1493	
		% dans Ward Method	86,1%	12,4%	1,5%	100,0%	
		% dans important in marriage: faithfulness	100,0%	100,0%	100,0%	100,0%	

Ward Method * important in marriage: same social background (Q42C)

Tableau croisé

		important in marriage: same social background (Q42C)			Total
		very important	rather important	not very important	
Ward Method	1 Effectif	176	398	86	660
	% dans Ward Method	26,7%	60,3%	13,0%	100,0%
	% dans important in marriage: same social background (Q42C)	81,9%	57,3%	14,8%	44,2%
2	Effectif	39	215	170	424
	% dans Ward Method	9,2%	50,7%	40,1%	100,0%
	% dans important in marriage: same social background (Q42C)	18,1%	30,9%	29,2%	28,4%
3	Effectif	0	82	327	409
	% dans Ward Method	0,0%	20,0%	80,0%	100,0%
	% dans important in marriage: same social background (Q42C)	0,0%	11,8%	56,1%	27,4%
Total	Effectif	215	695	583	1493
	% dans Ward Method	14,4%	46,6%	39,0%	100,0%
	% dans important in marriage: same social background (Q42C)	100,0%	100,0%	100,0%	100,0%

Ward Method * are you a religious person (Q28)

Tableau croisé

		are you a religious person (Q28)			Total
		religious person	not religious person	convinced atheist	
Ward Method	1 Effectif	465	150	44	659
	% dans Ward Method	70,6%	22,8%	6,7%	100,0%
	% dans are you a religious person (Q28)	54,3%	32,7%	26,5%	44,5%
2	Effectif	223	139	57	419
	% dans Ward Method	53,2%	33,2%	13,6%	100,0%
	% dans are you a religious person (Q28)	26,1%	30,3%	34,3%	28,3%
3	Effectif	168	170	65	403
	% dans Ward Method	41,7%	42,2%	16,1%	100,0%
	% dans are you a religious person (Q28)	19,6%	37,0%	39,2%	27,2%
Total	Effectif	856	459	166	1481
	% dans Ward Method	57,8%	31,0%	11,2%	100,0%
	% dans are you a religious person (Q28)	100,0%	100,0%	100,0%	100,0%

Ward Method * sex respondent (Q86)

Tableau croisé

		sex respondent (Q86)		Total
		male	female	
Ward Method	1 Effectif	276	384	660
	% dans Ward Method	41,8%	58,2%	100,0%
	% dans sex respondent (Q86)	38,3%	49,7%	44,2%
2	Effectif	234	190	424
	% dans Ward Method	55,2%	44,8%	100,0%
	% dans sex respondent (Q86)	32,5%	24,6%	28,4%
3	Effectif	211	198	409
	% dans Ward Method	51,6%	48,4%	100,0%
	% dans sex respondent (Q86)	29,3%	25,6%	27,4%
Total	Effectif	721	772	1493
	% dans Ward Method	48,3%	51,7%	100,0%
	% dans sex respondent (Q86)	100,0%	100,0%	100,0%

Le tableau synthétique suivant permet de décrire les classes, en mettant l'accent à la fois sur ce qui les caractérise et sur ce qui les distingue des autres. Cette fois-ci, les classes sont plus exacerbées et discernables grâce au travail préalable de l'ACP.

		Classe 1	Classe 2	Classe 3
Importance de la fidélité	Très important	100.0	50.9	100.0
	Moyennement important	0.0	43.6	0.0
	Pas important	0.0	5.4	0.0
Importance d'avoir un background social identique	Très important	26.7	9.2	0.0
	Moyennement important	60.3	50.7	20.0
	Pas important	13.0	40.1	80.0
Appartenance religieuse du répondant	Personne religieuse	70.6	53.2	41.7
	Personne non religieuse	22.8	33.2	42.2
	Athée convaincu	6.7	13.6	16.1
Sexe du répondant	Homme	41.8	55.2	51.6
	Femmes	58.2	44.8	48.4

La classe 1 est donc l'ancienne classe 2 de la manipulation précédente : une classe très religieuse et un peu plus féminine, pour qui, la fidélité et le besoin de venir d'un même milieu est très important. La classe 2 est une classe plus mitigée sur l'importance de la fidélité et du besoin d'appartenir à un même milieu d'origine sociale, elle est moins religieuse également. Enfin la classe 3, est unanime sur la question de la fidélité, ne trouve pas que partager un même niveau d'origine social est important et n'est pas religieuse également. Elle ne se différencie pas grandement au niveau du genre.

Récapitulatif des méthodes d'exploration des données

Méthode	ACP	Cluster
Synthétise	des variables	Des individus
Variables utilisées	Ordinales / quantitatives	Tout type de variable
Génère des « méta »-variables	Quantitatives	Qualitatives
Condition ?	Indice KMO	//
Combien d'axes/modalités retenir ?	Méthode du coude de Catell / Critère de Kaiser	Repérer les sauts incrémentés de la distance de Ward dans la liste des agglomérations / Dendrogramme
Interprétation	Fortes corrélations des variables aux axes factoriels	Descriptives bivariées avec la variable de cluster retenue

Ressources supplémentaires

SPSS à l'UdeS (Université de Sherbrooke) – url : <http://spss.espaceweb.usherbrooke.ca/>

KINNEAR P., GRAY C. (2005). SPSS facile appliqué à la psychologie et aux sciences sociales. Maitriser le traitement des données, De Boeck, Bruxelles, 432 p.

Fox W. (1999). *Statistiques sociales* (3^e édition), De Boeck & Larcier s.a., Paris, Bruxelles, 374 p.

- Tilburg University, Gesis, Leibniz Institute for the Social Sciences (2010). *European Values Study. EVS 2008 – Belgium. Field Questionnaire*, 267 p., disponible sur i-campus.
- Tilburg University, Gesis, Leibniz Institute for the Social Sciences (2010). *European Values Study. EVS 2008 – Variable Report. Integrated Dataset*, 1280 p., disponible sur i-campus.
- Kinnear P., Gray C. (2005). *SPSS facile appliqué à la psychologie et aux sciences sociales. Maitriser le traitement des données*, De Boeck, Bruxelles, 432 p.
- Fox W. (1999). *Statistiques sociales* (3^e édition), De Boeck & Larcier s.a., Paris, Bruxelles, 374 p.
- Tenenhaus M. (2007). *Statistique. Méthodes pour décrire, expliquer et prévoir*, Dunod, Paris, 679 p.
- Masuy-Stroobant G. et Costa R. (dir.) (2013). *Analyser les données en sciences sociales. De la préparation des données à l'analyse multivariée*, P.I.E. Peter Lang s.a., Bruxelles, 301 p.
- Test de Brown-Forsythe, url : <http://www.wikilean.com/Articles/Analyse/9-Les-tests-d-hypotheses/Test-de-Brown-Forsythe>

<http://www.radisma.info/docannexe.php?id=522>

<http://webapps.fundp.ac.be/biostats/biostat/modules/module140/page7.html>

Wikipédia : décomposition de variances, test-t, anova, corrélation, brown-forsythe, f, t,
<http://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-des-multi>