Octobre 2008 **Numéro 41**

Cahiers de l'IMA

Fascicule SPSS

Ingrid Gilles
Eva Green
Paola Ricciardi-Joos
Régis Scheidegger
Chiara Storari
Thomas Tuescher
Pascal Wagner

Institut de Mathématiques Appliquées
Faculté des S.S.P.
Université de Lausanne
Anthropole
1015 Lausanne

UNIL | Université de Lausanne
Institut des sciences sociales
et pédagogiques

FASCICULE SPSS

Gilles, I., Green, E.T., Ricciardi-Joos, P., Scheidegger, R., Storari, C., Tuescher, T., Wagner, P.*

Université de Lausanne

* La suite des auteurs est définie par ordre alphabétique.

Historique du fascicule

Le fascicule SPSS que vous tenez entre les mains présente plusieurs particularités. La première est qu'il est un travail collectif. La deuxième caractéristique est qu'il a été modifié au cours des années sur une période qui va du 1997 au 2007 ! La troisième caractéristique, qui est la plus importante à notre avis, consiste dans le fait que ce fascicule est le produit de la pratique liée à un enseignement spécifique, celui de la recherche en psychologie sociale.

En effet, ce fascicule est née de la nécessité de rendre capables des étudiants en sciences sociales d'opérer des analyses statistiques simples avec le logiciel SPSS. Telle était la demande du séminaire « Psychologie sociale : recherche » de l'Université de Lausanne, dirigé pendant cette période par le Professeur (devenu ensuite Professeur Honoraire) Jean-Claude Deschamps.

Puisque le plus souvent les étudiants en sciences sociales de l'université de Lausanne adhèrent eux-mêmes au fâcheux stéréotype qui veut qu'ils ne soient pas doués en mathématiques, le fascicule en question n'a pas été conçu comme un manuel de statistique, mais plutôt comme un guide pratique pour l'application de certains tests statistiques aux sciences sociales. Ainsi, le lecteur avisé ne devrait pas être trop incommodé si des notions statistiques se trouvent amputées, simplifiées ou simplement elles ne sont pas abordées.

Il existe d'autres manuels d'utilisation spécifiques au logiciel SPSS. Cependant, ceux-ci sont souvent trop détaillés et d'utilisation peu pratique. Au fil du temps, les assistants du séminaire « psychologie sociale : recherche », qui sont également les auteurs de ce fascicule, ont essayé de créer un document facile à lire et à utiliser, en se fondant sur leur expérience directe des problèmes et des nécessités des étudiants. Ce document ne se veut pas limité à la psychologie sociale, mais à tout étudiant qui se trouve confronté pour la première fois avec SPSS et qui ne dispose que de quelques notions de base en statistique. Nous espérons qu'il sera utile à d'autres personnes comme il l'a été pour nous.

Les auteurs tiennent à remercier le Professeur Honoraire Jean-Claude Deschamps, les étudiants de l'Université de Lausanne qui ont participé à leurs enseignements, ainsi que les étudiants à venir, qui permettrons certainement d'améliorer ultérieurement ce document.
Nous remercions également l'Institut des Mathématiques Appliquées de l'Université de Lausanne, qui nous a permis de rendre ce document disponible à toute personne intéressée, et tout particulièrement Jean-Philippe Antonietti. Nous remercions également Karine Henchoz, André Berchtold et Dominque Joye pour leurs commentaires avisés.

TABLE DES MATIERES

1. Constitution d'une base de données SPSS

1.1. Lancer SPSS (PC ou MAC)

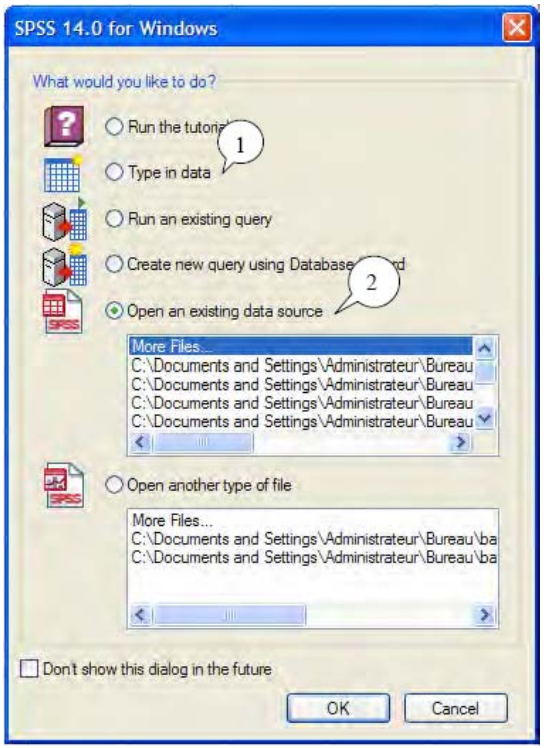
Sous PC, 2 façons :

1) Aller dans le menu **Démarrer** puis dans **Programmes**, choisir SPSS parmi la liste des programmes.
2) S'il existe un raccourci de l'application sur le bureau, double cliquez sur l'icône SPSS

Sous Mac :
Dans le menu **Applications,** vous trouverez le dossier SPSS et il suffit de cliquer sur l'icône de celui-ci.

1.1.2. Options de lancement de SPSS

Lorsque l'application est lancée, deux fenêtres s'ouvrent. En premier plan :



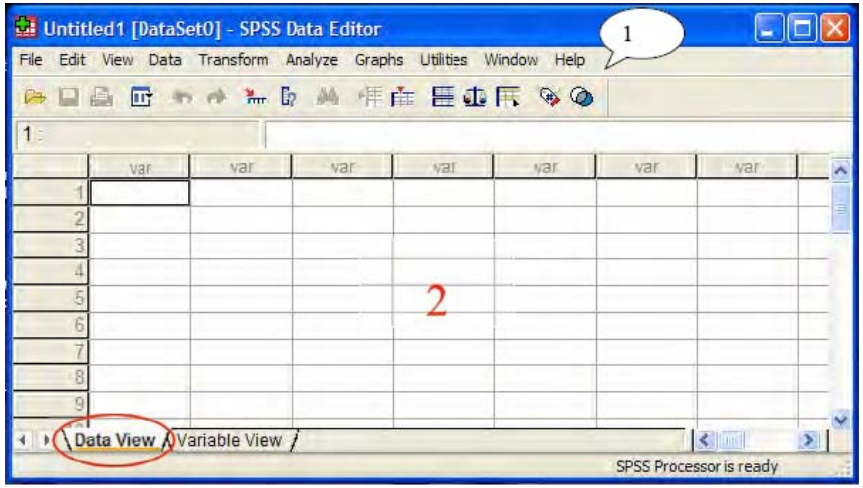
Cette fenêtre vous propose différentes options. La plupart du temps, on choisira :

1) L'option **Tape in data** pour faire la saisie des données récoltées à l'aide des questionnaires.
2) L'option **Open an existing data source** pour ouvrir un fichier déjà existant.

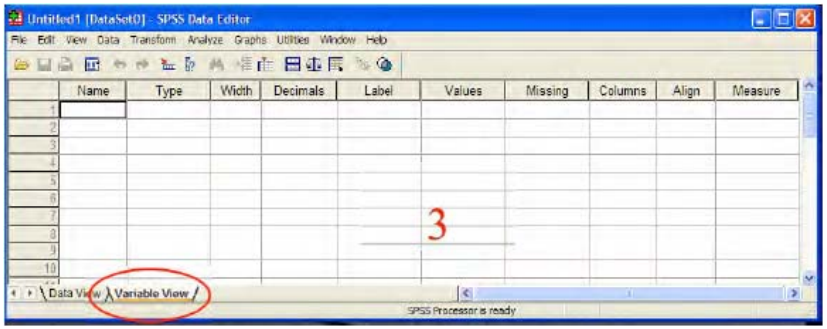
Lorsque vous sélectionnez l'option **Tape in data**, vous voyez apparaître au premier plan la fenêtre qui se trouvait en arrière plan et qui représente la fenêtre principale de SPSS.
Lorsque vous ouvrez une base de données déjà existante, il se peut qu'une autre fenêtre s'ouvre automatiquement : l'**output** (fenêtre d'édition des résultats). Pour l'instant nous ne nous occuperons que de la fenêtre principale.

1.2. Fenêtre principale

Elle se compose de plusieurs parties :



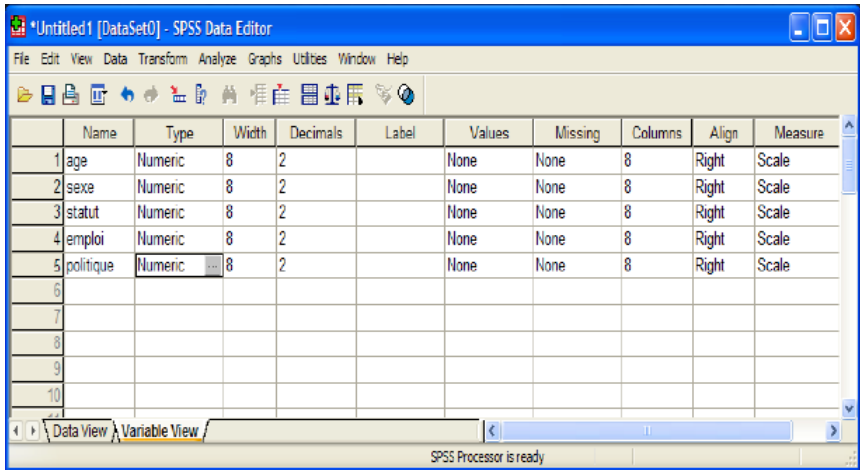
1) La barre des menus et des boutons de commande
2) La fenêtre principale de SPSS pour l'entrée et le traitement des données
3) La fenêtre de résumé des variables :



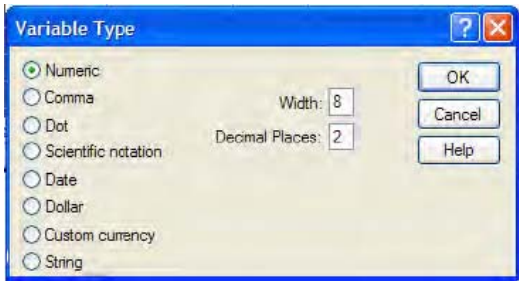
Pour la saisie des données, il faut dans un premier temps définir le nom des variables que l'on utilise.

1.3. Définition d'une variable et de ses propriétés

Cette opération s'effectue dans la fenêtre **Variable view**. Vous pouvez passer d'une fenêtre à l'autre en cliquant sur les onglets correspondants dans la barre en bas à gauche de la fenêtre.
Sous la colonne **Name** on indique le nom de la variable (nom sans accent, selon les versions du programme, vous disposez de 8 lettres et le seul trait reconnu est le souligné _), par exemple **age** pour l'âge des participants. On appuie ensuite sur **Enter** pour valider cette entrée. Des propriétés par défaut s'inscrivent alors sur la ligne qui concerne cette variable :



1) Type (par défaut, SPSS affiche **numeric**) :

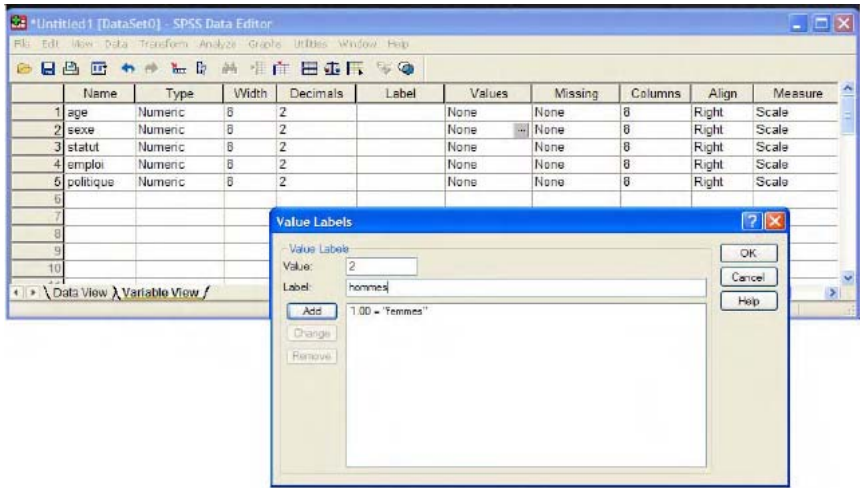


On obtient la fenêtre **variable type** en double cliquant sur le rectangle gris figurant dans la case **type** de chaque variable. Ici il s'agit d'indiquer quelle est la nature de la variable pour laquelle le type est défini.
Si vous entrez des chiffres ayant une valeur numérique → cliquez sur **Numeric**
Si vous entrez des chiffres sans valeur numérique, des lettres, des mots → cliquez sur **String**

Vous définissez également le nombre de caractères que vous pouvez inscrire (pour des variables nominales) et le nombre de décimales (pour les variables numériques).
Ces deux informations vont également s'inscrire dans les colonnes **width** et **decimals. Width** correspond à l'étendue de la case (le nombre de lettres ou de chiffres que vous pouvez inscrire). Par défaut, ces valeurs sont 8 et 2 respectivement. Vous pouvez les modifier à tout moment en changeant les chiffres qui apparaissent dans les cases correspondantes.

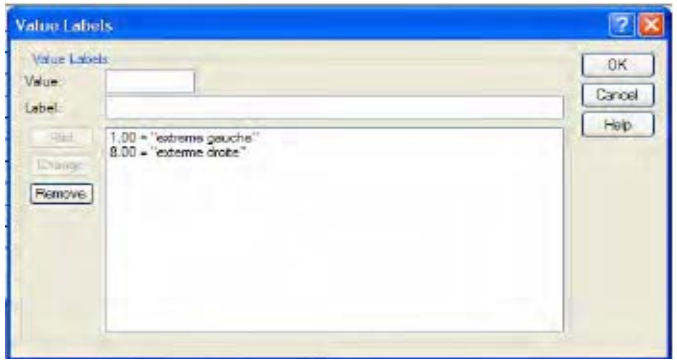
2) **Label** : le label correspond au descriptif de la variable. C'est un aide-mémoire. Le label est entré dans la case après avoir « double cliqué » dessus. Vous avez à disposition autant de caractères que vous le désirez. Pour agrandir physiquement cette case, sélectionnez la limite droite, gardez sélectionné et déplacez le curseur à droite ou gauche selon que vous voulez agrandir ou rétrécir la case.

3) **Value** : en général, les variables sont codées. Par exemple pour la variable sexe, on va attribuer le code 1 aux femmes et 2 aux hommes. Autre exemple, pour une question posée, on notera les réponses selon une échelle allant de 1 à 6 pour laquelle 1 correspondra à « pas d'accord » et 6 correspondra à « tout à fait d'accord ». Les codages peuvent être entrés sous **Values** afin de ne pas oublier quelles sont les valeurs qui leur sont attribuées.



Pour entrer une valeur il faut double cliquer sur le rectangle gris qui apparaît dans la case une fois celle-ci sélectionnée. La fenêtre ci-dessus s'affiche alors. Il suffit alors d'entrer dans l'onglet **Value** le premier code (par exemple *1* pour la variable sexe) et dans **Label** la modalité correspondante (par exemple *femme*). On clique alors sur **Add** pour valider la manœuvre. Attention, si vous ne cliquez pas sur **Add**, votre label ne sera pas pris en compte. Vous pouvez modifier labels et valeurs à tout moment. Pour tout nouveau label ajouté, répétez la procédure précédente. Pour enlever un label, cliquez sur la valeur et le label que vous voulez supprimer. L'onglet **Remove** apparaît alors. Cliquez dessus. Pour modifier un label, cliquez sur la valeur ou le label à modifier, effectuez les modifications dans les deux fenêtres supérieures et au lieu de cliquer sur **Add**, cliquez sur **Change**.

Pour une variable continue (par exemple la variable *politique* qui va de 1 = extrême gauche à 8 = extrême droite), vous pouvez n'indiquer que le code correspondant à extrême gauche et le code correspondant à extrême droite sans définir toutes les valeurs de l'échelle (cela vous permet de vous rappeler les bornes de l'échelle):



NB : les valeurs et labels indiqués dans cette fenêtre ne sont qu'indicatifs!!!! Modifier les labels et valeurs ne modifiera pas les données rentrées dans la fenêtre Data view! Il s'agit d'un aide-mémoire.

4) **Missing** : pour indiquer une non réponse (ou valeur manquante). Lorsque vous remplissez votre base de données, si des participants n'ont pas répondu à certaines question, vous pouvez laisser les casses correspondantes vides ou indiquer une valeur qui va être définie comme donnée manquante (par exemple 999). La première technique (ne rien mettre) ne vous permet pas de différencier entre elles plusieurs types de données manquantes (non réponse, n'a pas voulu répondre, ne sait pas), alors qu'avec le deuxième technique cela est possibles sans que les valeurs correspondantes soient prises en compte dans les calculs.

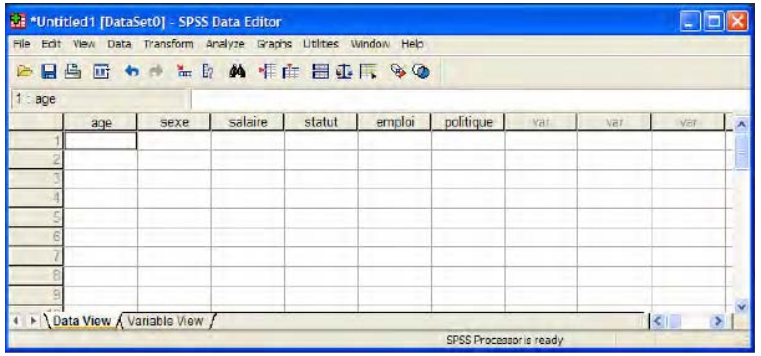
5) **Columns** : permet de contrôler la largeur physique de la case de la variable. Laissez 8 par défaut.

6) **Align :** permet de choisir si le texte dans les colonnes sera aligné à droite, à gauche ou centré. Laissez à droite par défaut.

7) **Measure** : permet d'indiquer si la variable que vous définissez est nominale (**Nominal**), ordinale (**Ordinal**) ou numérique (**Scale**). Laissez **scale** par défaut. Le type d'analyse que vous allez demander sur une variable va définir implicitement le type de variable que vous utilisez (si vous effectuez un chi carré, spss présume que les variables sont nominales). Cela ne sert pas à grande chose de les redéfinir ici. Dans la suite de ce fascicule, nous utilisons les termes **variables nominales** et **variables numériques**. Dans ce fascicule, nous avons décidé de regrouper les variables ordinales, les variables de rapport et les variables continues dans la catégorie des variables numériques. Cela ne signifie pas que nous considérons que ces différents types de variables présentent les mêmes caractéristiques statistiques (pour une définition des différents types de mesure et de variables, nous vous conseillons de faire référence à un manuel statistique, comme Howell, D.C. (2008). Méthodes statistiques en sciences sociales). Cependant, dans la pratique les chercheurs (du moins en psychologie

sociale) manipulent très souvent ces variables suivant les mêmes critères. C'est pourquoi nous avons décidé de les regrouper EXCEPTIONNELLEMENT dans la même catégorie des variables numériques.

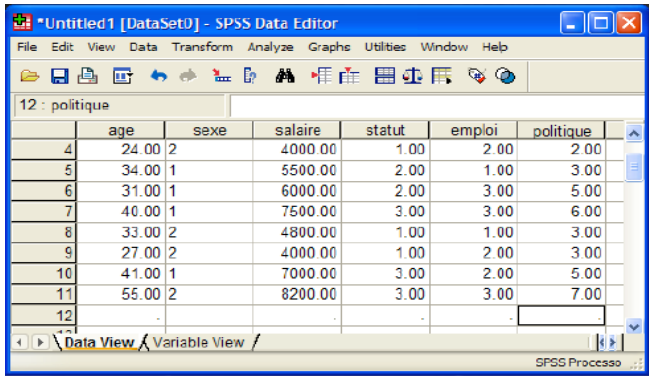
Si plusieurs colonnes doivent avoir exactement les mêmes caractéristiques, c'est-à-dire le même type, avec le même label, le même format de colonne, etc., utilisez dans **Variable view** les fonctions **Copy** et **Paste** du menu **Edit**. Le principe est le même que dans word : vous allez sur la case dont vous voulez copier le contenu, vous le copiez, vous allez sur la case dans laquelle vous désirez copier le contenu, et vous collez le contenu de la première case. Attention : vous devez avoir nommé une variable (la case à l'extrême gauche sur variable view) pour copier les autres cases. Au fur et à mesure que vous définissez les variables et leurs propriétés, la feuille des données (**data view**) se forme :



Les variables s'affichent en haut des colonnes et chaque ligne représente un participant.
Pour résumer:

1) Dans **Data view** une ligne = un participant et une colonne = une variable.
2) Dans **Variable view** une ligne = une variable et une colonne = les caractéristiques de cette variable.

Pour rentrer les données il faut donc remplir en ligne les informations correspondantes à chaque participant dans **Data view**:



1.4. Enregistrement des données

La première fois que vous enregistrez vos données (même principe que dans word) :

On enregistre les données en exécutant l'option **Save As** du menu **File**. Dans un premier temps, on choisit l'emplacement sur le quel on veut enregistrer le fichier (bouton **enregistrer dans**). Il est conseillé de sauver les données sur le bureau, vous les retrouverez plus simplement. Dans un deuxième temps, on entre un nom dans la fenêtre **Nom du fichier** ; ensuite on clique sur le bouton **Enregistrer**. L'extension d'un fichier de données SPSS est **« .sav »** et cette extension s'inscrit automatiquement à la suite du nom de votre fichier, il n'est donc pas nécessaire de la rajouter lorsque vous enregistrez vos données. Mais si vous voyez un fichier du type « nom.sav » sachez que c'est un fichier de données spss.

Pour ajouter des données à un fichier déjà existant (même principe que dans word) : lorsque votre document est ouvert, cliquez simplement sur **save** (!!pas « **save as** » !!). L'ordinateur va ajouter les nouvelles données à votre ancien dossier. Si vous cliquez « save as », le programme va créer une nouvelle base de données avec les nouvelles données ajoutées aux anciennes. Vous aurez alors deux bases de données. Si vous sauvez après chaque questionnaire saisi (ce qui est vivement conseillé) et que vous avez 100 participants, vous allez vous retrouver avec 100 bases de données. La dernière sera la bonne, mais comment s'en rappeler ?

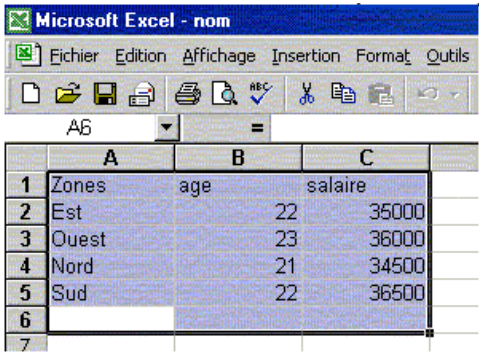
1.5. Quitter SPSS

Sur PC :
On exécute l'option **Exit** du menu **File** pour quitter SPSS. Le logiciel vous demande si vous voulez vraiment quitter l'application, cochez **yes**.
Sur MAC : On exécute l'option **Quit SPSS** du menu SPSS (qui se trouve à gauche du menu **file**). Le logiciel vous demande si vous voulez sauver les données contenues dans chaque fenêtre avant de quitter, répondez **yes**.

2. Importation d'un ficher de données à partir d'Excel

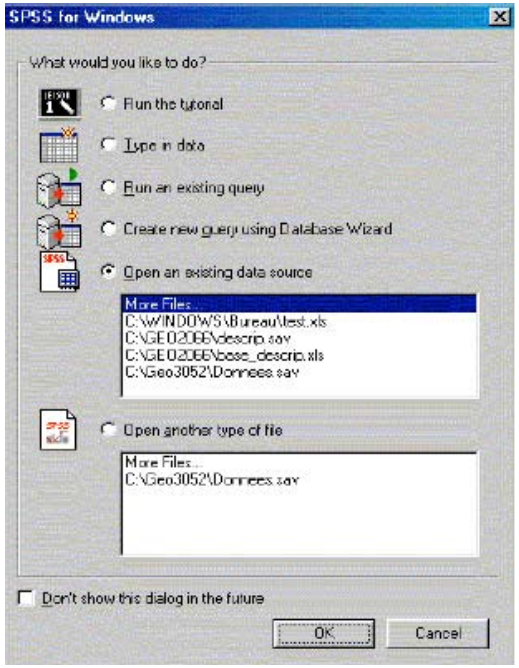
2.1. Fichier de données Excel

Le fichier aura comme extension .**xls**
Les variables correspondent aux colonnes et les participants aux lignes, comme dans SPSS.
Le nom des variables doit figurer sur la première ligne de la grille et doivent avoir 8 caractères au maximum (si vous voulez importer les données sous une version de SPSS antérieure à la 15). N'utilisez pas de caractères spéciaux ou d'accent.

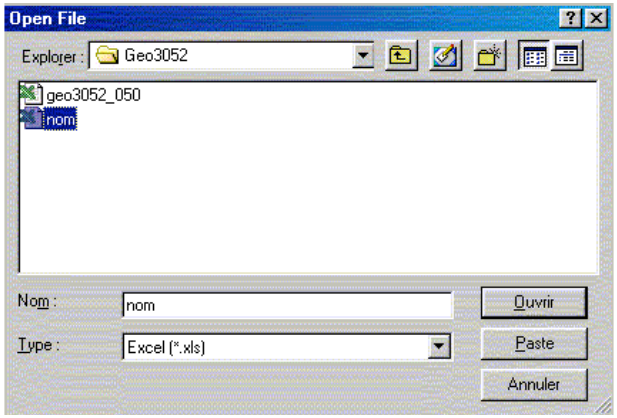


2.2. Importation des données

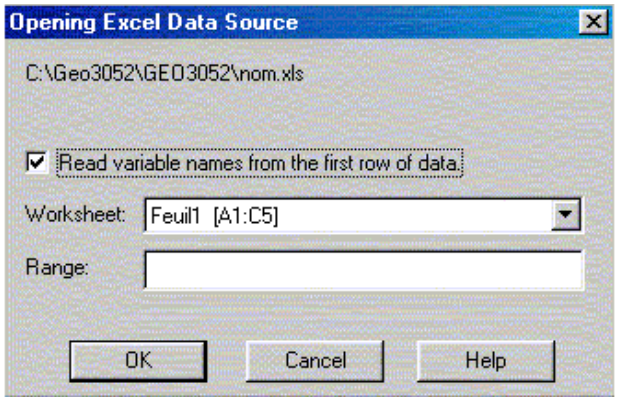
Dans SPSS, il faut suivre les étapes suivantes:
Spécifiez que vous désirez ouvrir un fichier déjà existant (**Open an existing data source**). Cliquez sur **More Files…** et ensuite, **OK**.



Indiquer dans la fenêtre **Open File** que le fichier à importer (**Fichier de type :**) est de type **Excel (.xls)**. Vous trouvez dans vos répertoires votre fichier Excel : sélectionnez-le et cliquez sur **Ouvrir.**



Cochez la case **Read variable names from the first row of data** dans la fenêtre **Opening Data Source** et cliquez sur **OK.**



Cette fonction permet au logiciel de reconnaître la première ligne de votre feuille excel comme étant le nom des variables. Attention les caractéristiques de vos variables sont définies par défaut par SPSS, puisque le logiciel ne dispose d'aucune information sur les données depuis Excel.

Le fichier de données s'ouvre alors dans SPSS et vous pouvez utiliser les fonctionnalités de SPSS. N'oubliez pas de sauvegarder le nouveau fichier.

Présentation des données, synthèse.
Après avoir saisi vos données, vous devez avoir un tableau des données avec :
1) Une ligne par participants (dans data view)
2) Une colonne par variables (dans data view)
3) Tous les résultats en chiffres (évitez les lettres)
4) Les caractéristiques de chaque variable définies

3. Fonctions de base de SPSS

Le logiciel SPSS travaille sur plusieurs feuilles (fenêtres) en même temps :

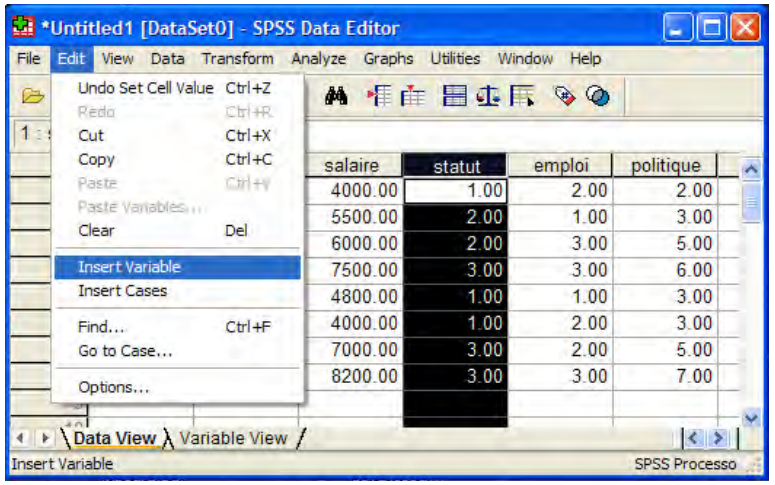
1) La feuille de données sur laquelle vous entrez vos données et dans laquelle elles sont définies. Cette feuille a une extension **.sav** (elle comprend la page **data view** et la page **variable view**). Vous pouvez ouvrir uniquement une fenêtre .sav en même temps (uniquement si vous utilisez la version 14 sur PC vous pouvez ouvrir plusieurs fenêtres .sav).

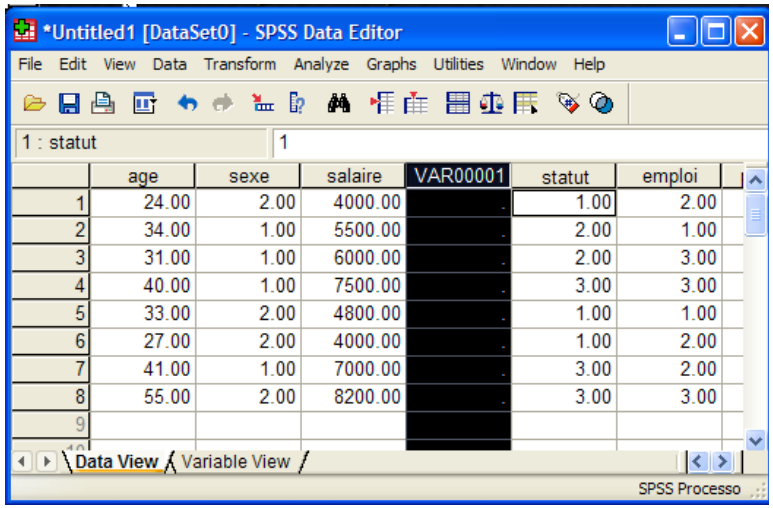
2) La (les) feuille(s) de sortie ou **output** sur laquelle apparaissent les résultats de vos analyses statistiques. Cette feuille a une extension **.spo**. Vous pouvez ouvrir plusieurs fenêtres de ce type en même temps. Sur ces fenêtres, le logiciel va vous présenter les résultats de vos analyses sous forme de tableaux ou de graphiques.

3) La (les) feuille(s) de syntaxe (**syntax)** sur laquelle vous pouvez programmer des analyses sans utiliser les menus. Cette feuille a une extension **.sps**. Vous pouvez ouvrir plusieurs fenêtres de ce type en même temps. Sur ces fenêtres, le logiciel va vous présenter les ordres que vous avez formulé à la machine sous forme écrite (programmation). Cette fenêtre vous permet de sauver les analyses sans devoir passer par l'output (qui est un fichier « lourd »). En plus, sauver la syntaxe peut être utile pour garder une trace des nouvelles variables que vous avez défini lors de vos analyses (voir chapitre 4).

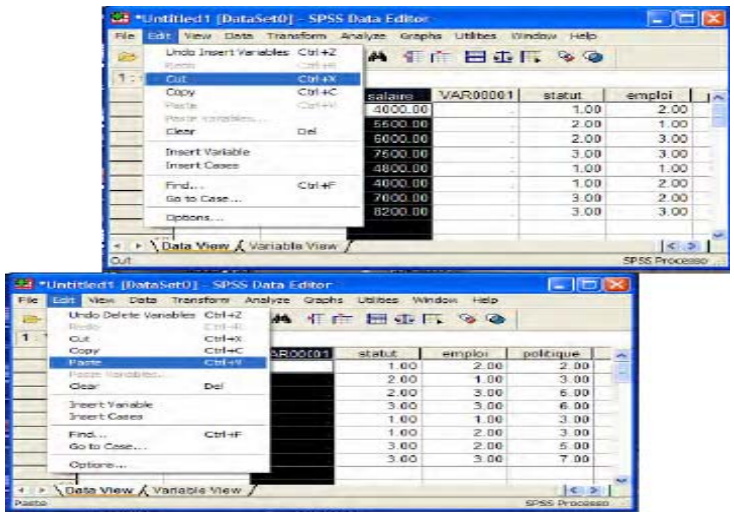
3.1. Insérer une nouvelle variable / déplacer une variable (sur la feuille de données)

Pour insérer une nouvelle variable dans une base de données existante, sélectionnez une colonne dans **data view** (ou une ligne dans **variable view**), puis suivez le chemin suivant : **Data => Insert Variable** (soit dans la fenêtre **data view** soit dans **variable view**)
Apparaît alors une nouvelle colonne vide dans **data view** ou une ligne dans **variable view**, précédant celle que vous avez sélectionnée, prédéfinie par défaut mais sans nom.





Pour déplacer une variable dans la base de données, insérez une colonne comme ci-dessus à l'endroit voulu, puis sélectionner la colonne à déplacer. Suivez alors le chemin suivant : **Edit => Cut**. Puis sélectionnez la colonne vide et **Edit => Paste**. Si vous voulez déplacer une variable qui contient des participants (par exemple vous avez entré 20 participants), vous devez sélectionner toute la colonne, c'est-à-dire toutes les valeurs entrées. Pour les copier, vous devez sélectionner un nombre de cases égal ou supérieur au nombre de participants (dans ce cas 20). Si vous en sélectionnez un nombre inférieur, spss va coller uniquement les cases « qui ont une place ». Par exemple les 15 premiers sujets, si vous n'avez sélectionné que 15 cases.

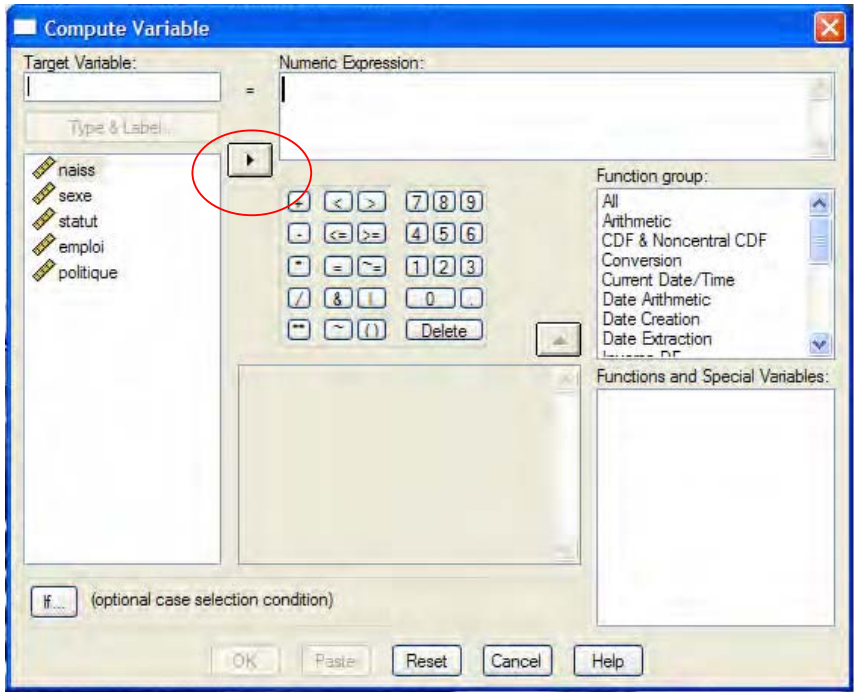


4. Manipulation des données

4.1. Créer une variable à partir d'une ou de plusieurs variables existantes

But : créer une nouvelle variable à partir de variables existantes, soit en les modifiant (par addition, soustraction, etc.), soit en les « agglomérant » (moyenne d'un ensemble de variables). L'aspect central de ces manipulations est que vous opérez des calculs sur vos variables (sous forme de simple équation ou sous forme d'une fonction). Il est très important de se rappeler de cet aspect. Normalement ces opérations sont effectuées sur des variables numériques. Exemple : vous avez codé la taille des participants en centimètres (nom variable : *tai_cen)* et vous voulez créer une variable nommée *tai_met,* qui représente leur taille en mètres.

Chemin : **Transform →Compute...**

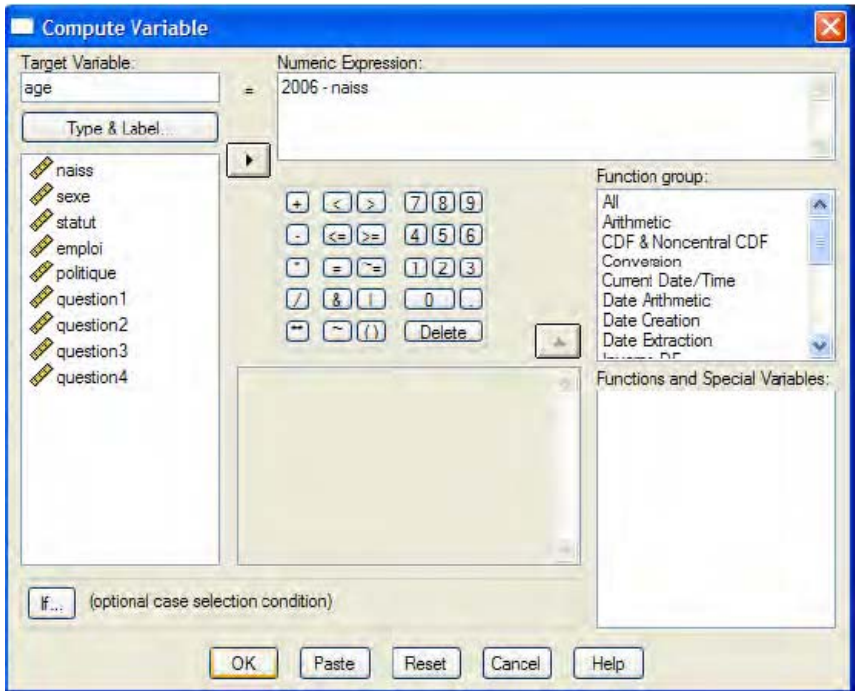


Dans **Target Variable** inscrire le nom de la nouvelle variable que vous voulez créer (par exemple *tai_met*). Dans **Numeric expression** inscrire, grâce au pavé numérique, la formule correspondante au calcul qui vous permet d'obtenir votre nouvelle variable (par exemple tai_cen/100). Vous avez deux possibilités : soit vous inscrivez vous mêmes le nom des variables dans l'expression, soit vous pouvez les chercher dans la fenêtre de gauche. Avec la flèche, vous les déplacez dans la fenêtre **Numeric expression** après les avoir sélectionné (une

à la fois). Pour déterminer les opérations, vous pouvez utiliser les symboles du pavé numérique qui se trouve sous la fenêtre **Numeric expression**. Dans la fenêtre **Functions**, comme dans Excel, vous disposez d'un certain nombre de fonctions préenregistrées. Par exemple, pour calculer la moyenne de plusieurs variables agrégées, cherchez **Mean**, sélectionnez cette fonction et faite-la glisser dans la fenêtre **Numeric expression** grâce à la flèche. Dans ce cas, le logiciel propose la fonction (par exemple **mean**) et des parenthèses. Dans les parenthèses, vous devez introduire les variables que vous voulez agréger, séparées par une virgule (les symboles les plus importants sont : *multiplication, / division, < plus petit que, <= plus petit ou égal, > plus grand que…). Cliquez ensuite sur **OK**.

Exemple 1 :

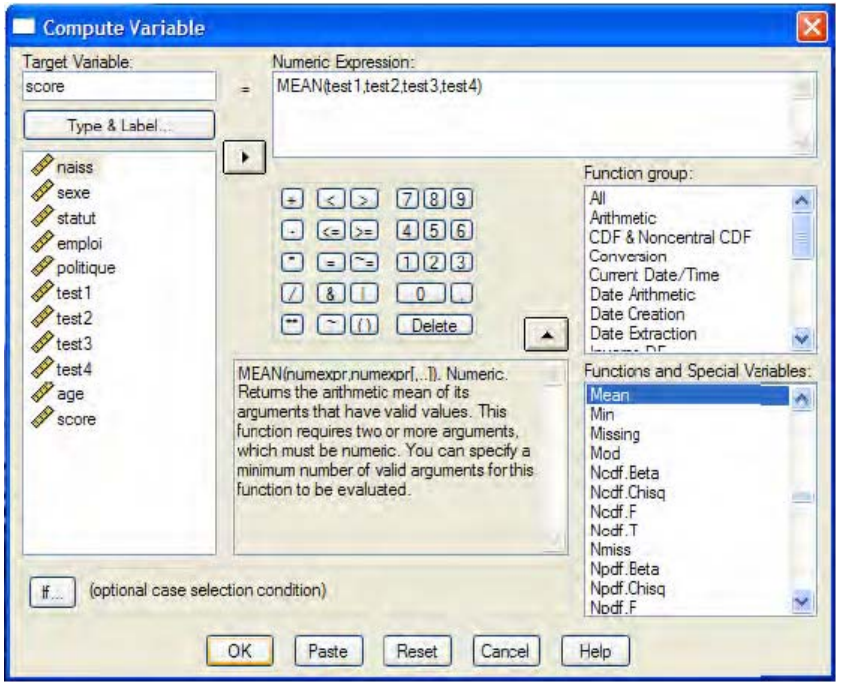
Dans notre base de données, il y a l'année de naissance des participants (variable *naiss* dans SPSS) mais pas leur âge. Nous voudrions travailler avec l'âge. Il faut donc créer cette variable. Pour obtenir l'âge à partir de la date de naissance, il faut écrire l'opération suivante : âge = 2006 –*naiss* (2006 représente l'année de l'étude). On reproduit donc ce calcul dans notre fenêtre SPSS :



Le variable âge ainsi créée va venir s'ajouter à la fin de la base de données, comme dernière variable.

Exemple 2 :

On a noté les points (scores) des participants à quatre tests que l'on a appelés test1, test2, test3 et test4. Ces tests mesurent l'aptitude au calcul des enfants, mais de manière différente (par équations, par des cubes de couleur, par calcul mentale ou par suite logique de chiffres). On veut calculer la moyenne de ces quatre tests pour créer une variable générale qui mesure l'aptitude générale au calcul. Toujours dans **transform** puis **compute**, on utilise la fonction **mean** comme indiqué ci-dessous :



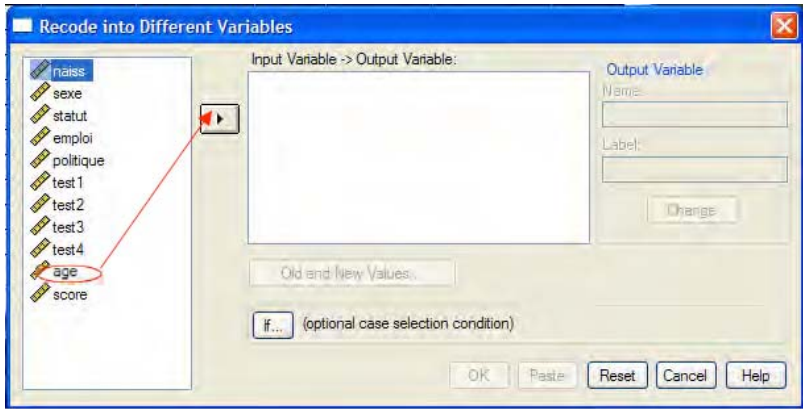
4.2. Recoder des variables

But : à partir de variables déjà existantes, créer d'autres variables. Dans ce cas, vous n'allez pas opérer des opérations arithmétiques sur vos données originales, mais vous allez plutôt recoder différemment les valeurs que vous avez donné à des catégories de réponse. Cela signifie que vous allez manipuler les étiquettes que vous avez déterminées pour vos données. Cette fonction vous permet aussi de créer des groupes : vous allez regrouper différemment des catégories de réponse pour en créer d'autres. Souvent, ce genre de manipulation porte sur des variables nominales, mais cela peut également concerner des variables numériques. Par exemple, si vous voulez former des classes d'âge (variable nominale) à partir de l'âge des participants (variable numérique). Vous devez utiliser cette fonction lorsque vous êtes amenés à inverser l'ordre de l'échelle utilisée pour mesurer une variable

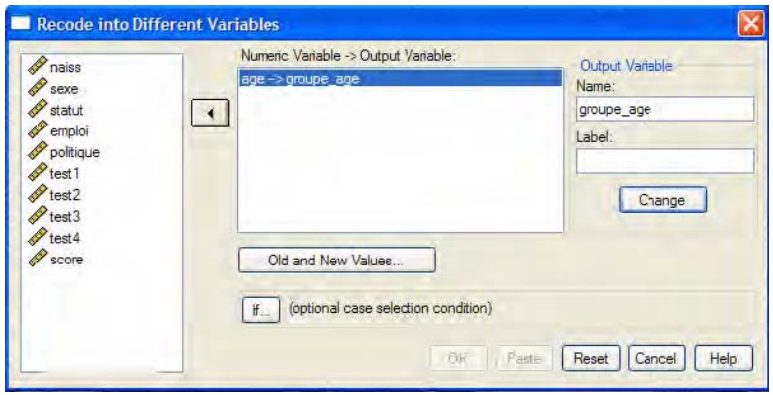
Chemin : **Transform → Recode...**

Une fenêtre apparaît alors et vous propose deux options. Le plus souvent sélectionnez : **Into Different Variable**. Cela signifie que le logiciel crée une nouvelle variable qui va s'ajouter à la fin de votre base de données. Si vous cliquez sur **Into same variable** le logiciel va modifier l'ancienne variable sans en modifier le nom. Le problème est que de cette manière, vous ne gardez pas de trace de cette transformation et il devient difficile de se rappeler si une variable est encore sous sa forme originale ou pas. Dans les deux cas, le principe est le même.

Pour la première option, la fenêtre apparaît :

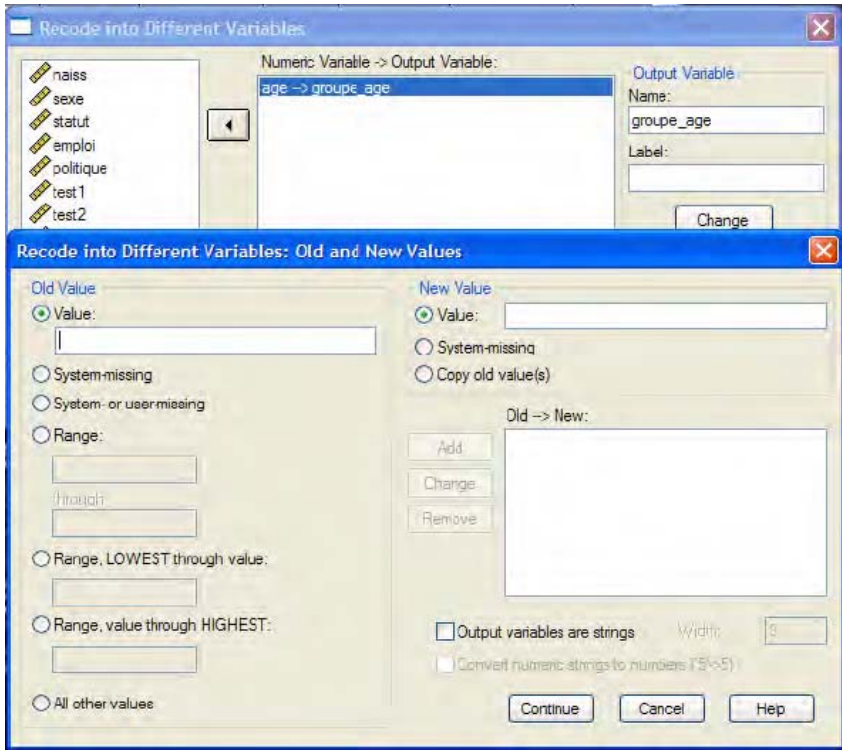


Sélectionnez dans le cadre de gauche le nom de la variable à recoder (*age*) et cliquez sur la flèche. Le nom de cette variable apparaît dans le cadre **Input Variable -> Output Variable**. Écrire dans le cadre de droite, sous **Name:** le nom de la future variable recodée (*groupe_age*). Cliquer sur **Change**. Si vous ne cliquez pas sur change, vous ne pouvez pas procéder à la création de la nouvelle variable. Un conseil: lorsque vous nommez la nouvelle variable, reprenez le nom de l'ancienne en lui rajoutant quelque chose qui vous rappelle qu'elle a été recodée. Par exemple, l'ancienne variable s'appelle *v30* et la nouvelle *v30_re*.



Cliquez ensuite sur **Old and New values…**

Une nouvelles boite de dialogue apparaît. Celle-ci vous permet de déterminer quelles nouvelles valeurs doivent être assignées aux anciennes.

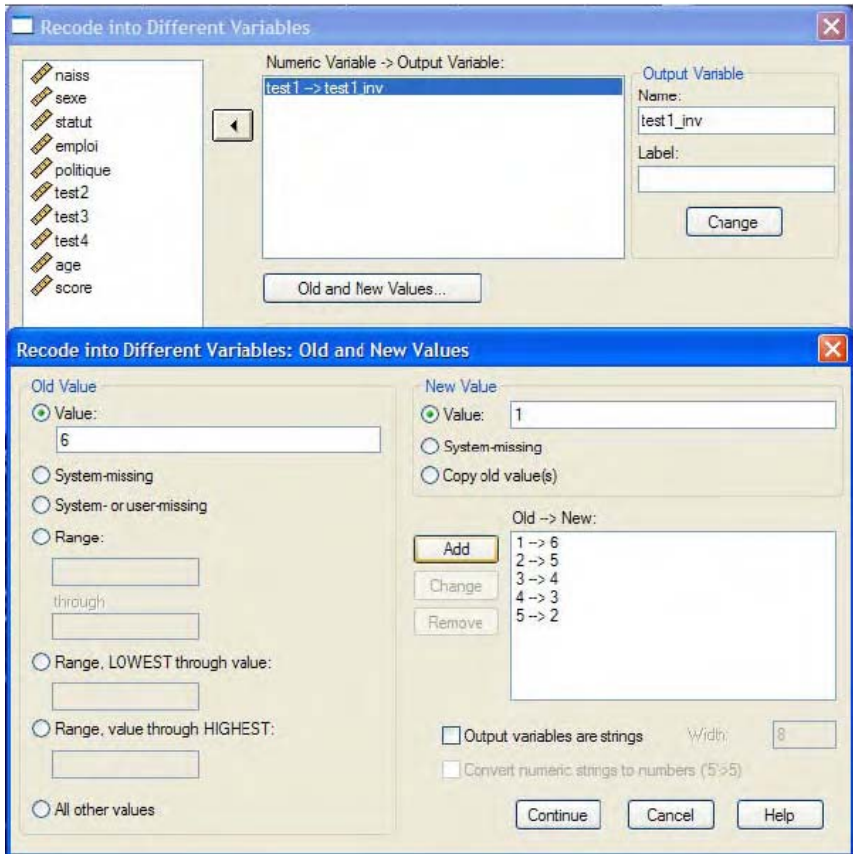


Dans le cadre **Old Value** et sous **Value**, écrire la valeur à recoder et à droite, dans le cadre **New Value** et sous **Value**, la nouvelle valeur recodée.
Si, par exemple, vous travaillez avec une variable numérique allant de 1 à 5 que vous voulez inverser, l'ancienne valeur 1 deviendra la « nouvelle » 5, l'ancienne valeur 2 deviendra 4, l'ancienne valeur 3 deviendra 3, l'ancienne valeur 4 deviendra 2, et l'ancienne valeur 5 deviendra 1.

Après chaque nouvelle valeur, cliquez sur **Add**; la transformation prévue s'affiche dans le cadre **Old →New**. Vous pouvez modifier ces valeurs, il suffit de sélectionner celle qui vous intéresse dans le cadre **Old →New** et de cliquer sur **change**. Par la suite, vous pouvez confirmer le changement en cliquant sur **Add**. Vous pouvez également éliminer un codage en cliquant sur **remove**. Une fois toutes les valeurs à modifier inscrites dans ce cadre, cliquez sur **Continue** puis cliquez sur **OK**. La première boîte de dialogue se ferme et le logiciel a créé votre nouvelle variable à la fin de la base de données.

Exemple 1 :

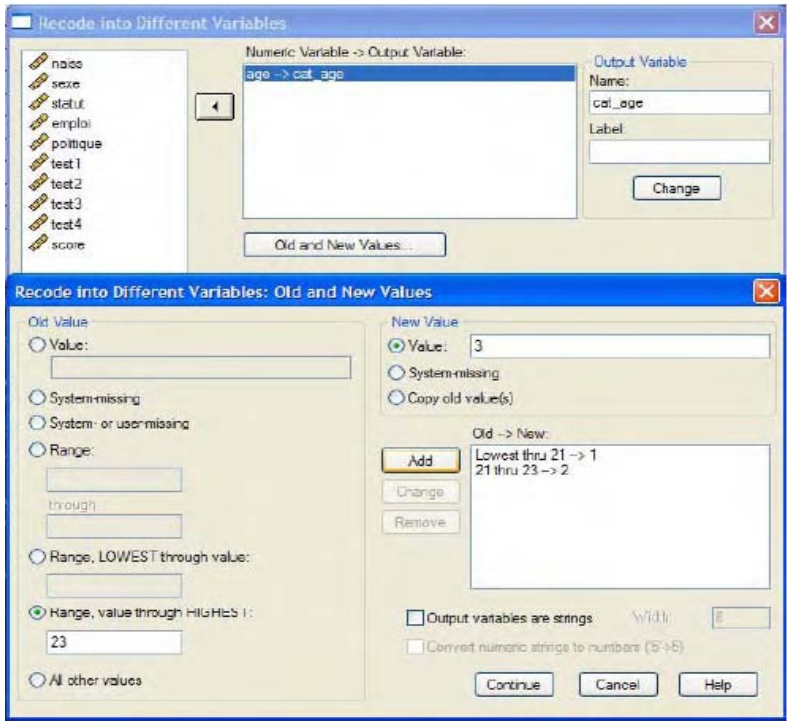
Le test1 a été mesuré à l'aide d'une échelle allant de 1 à 6, 1 = tout à fait réussi et 6 = pas du tout réussi, alors que le codage est inverse pour les trois autres tests (1 = pas du tout réussi et 6 = tout à fait réussi). Pour pouvoir calculer la moyenne de ces 4 variables (les tests), il faut que les échelles « aillent toutes dans le même sens ». Il faut donc inverser l'échelle du test1 de façon à ce que la valeur 1 = pas du tout réussi et que la valeur 6 = tout à fait réussi (bien entendu, il faut aussi recoder les valeurs intermédiaires). On suit la procédure décrite plus haut et on aboutit à la fenêtre suivante :



Après avoir appuyé sur **Add** pour ajouter le dernier recodage, on peut cliquer sur **Continue** puis **OK**. La nouvelle variable *Test1_inv* apparaîtra à la fin de la base de données.

Exemple 2 :

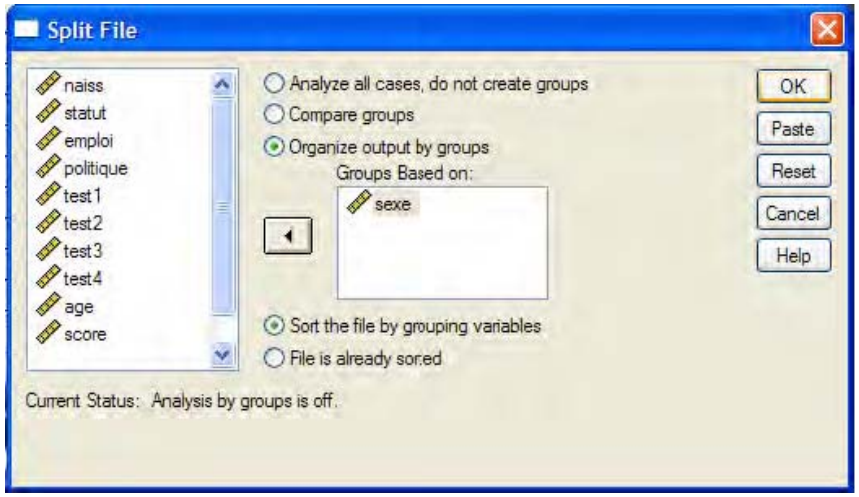
Dans les exemples traités jusqu'ici, il était question de faire correspondre un nouveau code à un ancien code. Il est également possible de créer des groupes, c'est-à-dire de faire correspondre un nouveau chiffre à un ensemble de chiffres. Par exemple, on veut créer 3 catégories d'âge (la catégorie 1 qui regroupe les individus qui ont moins de 21 ans, la catégorie 2 qui regroupe les individus qui ont entre 21 et 24 ans, et la catégorie 3 qui regroupe les individus qui ont plus de 24 ans) à partir de notre variable numérique *age*.
Dans les cadres **Old Value** et **New Value**, les catégories peuvent être définies de la manière suivante (pour ne pas recoder chaque valeur individuelle comme dans le cas précédent, puisqu'il serait trop laborieux ici) :
A gauche, **Range:, Lowest through** (plus petit que…les chiffres qui déterminent les limites de la catégorie sont comprises dans celle-ci) 21 et à droite, dans le cadre **New Value**, écrire **1**. puis cliquer sur **Add**. Ainsi toutes les valeurs inférieures à 21 (comprise) vont être regroupées dans une nouvelle catégorie 1.
À gauche, **Range, 21 through 23** et à droite, dans le cadre **New Value**, écrire 2. Cliquer sur **Add**. Les valeurs comprises entre 21 (non compris) et 23 (compris) vont être modifiées dans une nouvelle catégorie 2.
À gauche, **Range 23 through highest** et à droite, dans le cadre **New Value**, écrire 3. Cliquer sur **Add**. Ainsi, ceux qui ont plus de 24 ans (compris) seront regroupés dans une nouvelle catégorie 3. Une fois toutes les valeurs à modifier inscrites dans ce cadre, cliquez sur **Continue** puis **OK**.



4.3. Fragmenter la base de données (travailler uniquement sur une partie des données)

But : travailler sur l'ensemble des données, mais en obtenant des informations distinctes sur divers groupes en parallèle – comme les hommes et les femmes. Les résultats apparaissent alors séparément pour les groupes concernés dans l'output. Cette option vous permet de pouvoir considérer des groupes différents en même temps. C'est-à-dire que le logiciel, au lieu de travailler sur l'ensemble des individus, va faire la même analyse sur plusieurs groupes d'individus et va vous présenter les résultats pour chaque groupe séparément. Souvent la variable qui définit la fragmentation des données est une variable nominale (qui identifie des groupes de participants).

Chemin : **Data → Split File...**



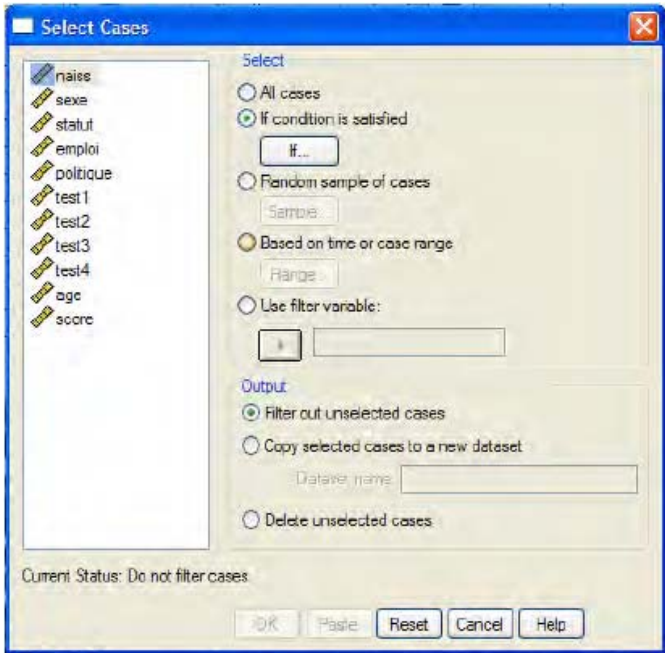
Cliquez sur **organize output by groups** (ou **compare groups** si vous désirez avoir l'information dans un seul tableau à la place de plusieurs. Le résultat est le même, uniquement la présentation change). Sélectionnez dans la fenêtre de gauche la variable en fonction de laquelle vous voulez séparer vos résultats (ici *sexe*) et faites-la glisser dans la fenêtre de droite (**Groups based on :**) à l'aide de la flèche.

N'oubliez pas de désactiver le **split file** quand vous n'en avez plus besoin. Il ne se désactive pas seul et le logiciel ne vous rappelle pas qu'il est activé. Cette fonction peut être désactivée par:
Data → Split file →Analyse all cases et **OK.**

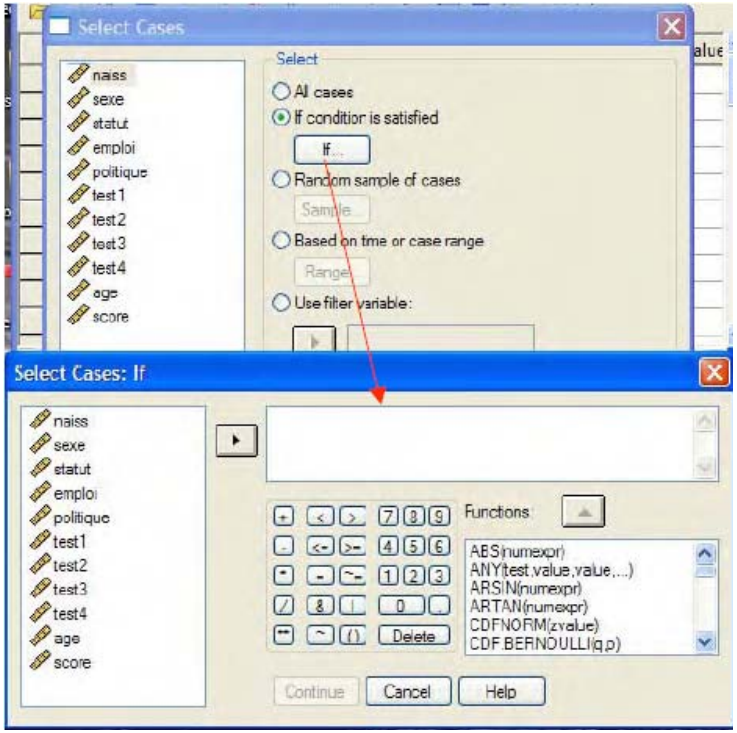
4.4. Travailler sur une partie des données

But : lorsqu'une seule partie de l'échantillon vous intéresse, vous pouvez demander de n'avoir les résultats que pour cette partie, pourvu qu'elle soit définie par des variables inclues dans la base de données (par exemple : résultats pour les femmes). Encore une fois, la variable qui vous permet de sélectionner les individus est, la plupart du temps, une variable nominale. La différence entre cette fonction et la précédente est qu'ici vous n'avez pas les résultats pour tous les groupes (par exemple les hommes et les femmes), mais uniquement pour une catégorie de participants (les femmes).

Chemin : **Data → Select cases…**



Cliquez sur **If condition is satisfied,** puis **If…**



Ecrire l'expression numérique nécessaire. Par exemple, pour traiter uniquement les données qui portent sur les femmes (codés 2), écrire : **sexe = 2**. Puis cliquer sur **Continue** et **OK.**

Cette sélection est active pour toute analyse faite par la suite jusqu'à ce que vous la désactiviez ou quittiez l'application. Pour désactiver :

Data → Select cases → All cases et **OK.**

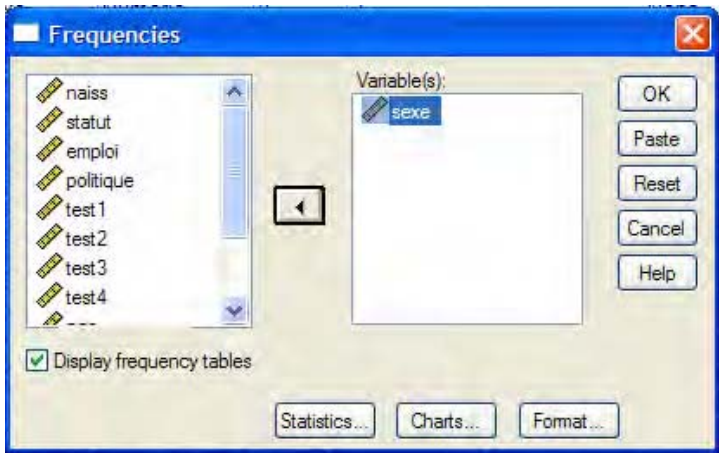
5. Statistiques Descriptives

Jusqu'à présent, il a été question de manipuler la base de données et les variables que celle-ci comprend. Nous passons maintenant à la phase de description des données.

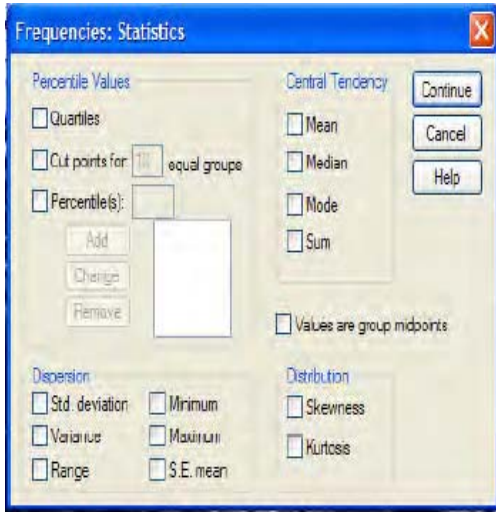
5.1. Calcul des fréquences (variables nominales)

But : calculer la fréquence (nombre et / ou pourcentage) des unités (participants) qui sont regroupées dans chaque modalité d'une variable nominale (par exemple, le sexe : N ou % de femmes et d'hommes).

Chemin : **Analyze → Descriptive Statistics → Frequencies...**



Sélectionnez dans la fenêtre de gauche les variables, puis passez-les à droite en utilisant la flèche.



Dans la fenêtre **Statistics**, on peut choisir l'information que l'on désire avoir sur les données. Ainsi vous pouvez avoir des indices de tendance centrale (moyenne, médiane et mode), vous pouvez obtenir la valeur pour laquelle l'échantillon sera divisé en x sous échantillons composés du même nombre (à peu près) de participants (**cut points for x equals groups**) et d'autres informations. Sélectionnez ce qui vous intéresse.

Cliquez sur **Continue** puis **OK** pour lancer votre analyse. Le logiciel va alors créer une nouvelle fenêtre, celle des résultats (**output**).

SPSS output :

Sexe

		Frequency	Percent	Percent valid	Cumulative percent
Valid	homme	76	43.9	43.9	43.9
	femme	97	56.1	56.1	100.0
	Total	173	100.0	100.0	

Dans cet échantillon, il y a 76 hommes (43.9% de l'échantillon) et 97 femmes (56.1% de l'échantillon). Vous remarquez qu'en ligne, vous avez les catégories qui composent votre variable (si vous leur avez donné des labels lors de la constitution des variables, ces noms apparaissent dans le tableau, autrement vous avez les chiffres). En colonne, vous voyez les fréquences (**Frequency** : le nombre de participants dans chaque catégorie). Les participants qui ne peuvent pas être classés sont indiqués comme **missing system.** Suivent (toujours de gauche à droite), les pourcentages des participants dans chaque catégorie (**percent**) : attention, parmi ces pourcentages se trouvent également les participants qui n'ont pas répondu. Si vous vouliez indiquer les pourcentages de participants dans les catégories qui ont répondu à la question, vous devez faire référence à la colonne qui suit (**valid percent**). La dernière colonne est la somme des pourcentages des individus à chaque nouvelle catégorie (**cumulative percent**).

5.2. Tableaux croisés ou tableaux de contingence (2 variables nominales)

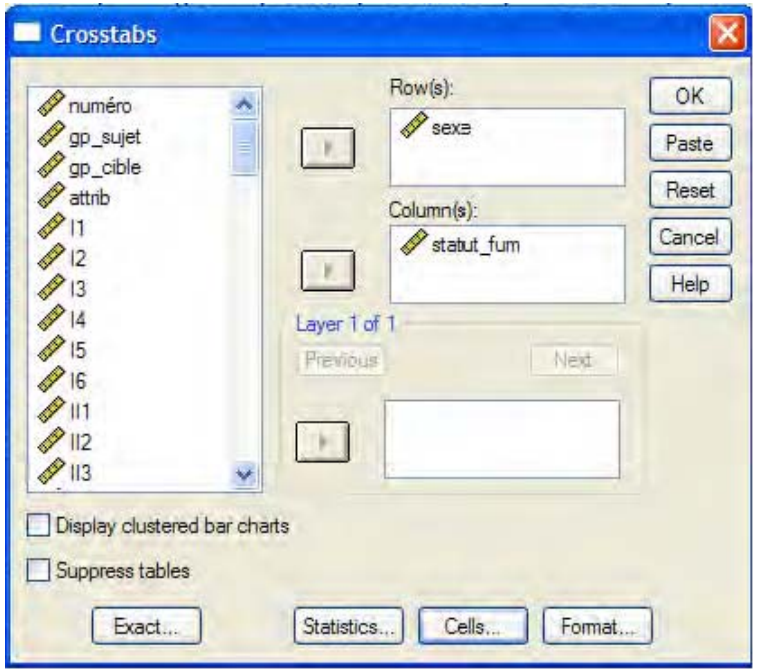
But: calculer la fréquence (nombres et/ou pourcentage) des participants qui sont compris dans les modalités d'une variable nominale, croisée avec les participants qui sont compris dans les modalités d'une seconde variable nominale (par exemple, nombre et pourcentage d'hommes et de femmes selon leur habitude à fumer). Attention : vous n'allez utiliser que des variables nominales pour cette fonction.

Chemin : **Analyze → Descriptive Statistics → Crosstabs...**

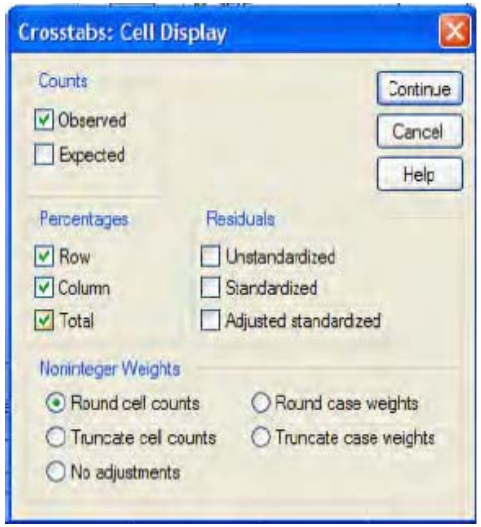
Entrer une première variable dans la fenêtre **Row(s):** et une deuxième variable sous **Column(s):** toujours en les sélectionnant et en cliquant sur la flèche qui sépare les deux fenêtres. L'ordre et la position que vous utilisez pour entrer les variables n'est pas important pour le résultat.

Exemple 1 :

Dans une étude sur la consommation de tabac, on croise le sexe (*sexe*) et la consommation de tabac des participants (*statut_fum* **:** 1 = fumeurs, 2 = fumeur occasionnel, 3 = ancien fumeur, 4 = non-fumeur).



Ensuite, cliquez sur **Cells…** pour demander les pourcentages (en ligne, en colonne et totaux).



SPSS output :

Si vous avez entré la variable *statut_fum* en colonne (**column**) cela signifie que les catégories fumeur, ancien fumeur, non fumeur et fumeur occasionnel seront sur le haut du tableau et définissent les colonnes de celui-ci. Les pourcentages en colonne se lisent alors comme le pourcentage des quatre groupes de fumeurs dans les catégories de la variable en ligne. Pour savoir quelle valeur regarder, faites attention à quels chiffres donnent une somme de 100% en colonne. Si vous avez entré la variable *sexe* en ligne (**row**) cela signifie que les catégories de sexe seront sur le côté gauche du tableau et définissent les lignes de celui-ci. Les pourcentages en ligne se lisent alors comme le pourcentage d'hommes ou de femmes dans les catégories de la variable en colonne. Pour savoir quelle valeur regarder, faites attention à quels chiffres donnent la somme de 100% en ligne.

Sexe*statut_fum Cross tabulation

			Statut_fum				
			Fumeur	Anc_fumeur	Non_fum	Fumeur_occ	Total
Sexe	Homme	Count	26	33	10	7	76
		% within sexe	34.2%	43.3%	13.2%	9.2%	100%
		% within statut_fum	43.3%	40.2%	47.6%	70.0%	43.9%
		% of Total	15.0%	19.1%	5.8%	4.0%	43.9%
	Femme	Count	34	49	11	3	97
		% within sexe	35.1%	50.5%	11.3%	3.1%	100%
		% within statut_fum	56.7%	59.8%	52.4%	30.0%	56.1%
		% of Total	19.7%	28.3%	6.4%	1.7%	56.1%
	Total	Count	60	82	21	10	173
		% within sexe	34.7%	47.4%	12.1%	5.8%	100%
		% within statut_fum	100%	100%	100%	100%	100%
		% of Total	34.7%	47.4%	12.1%	5.8%	100%

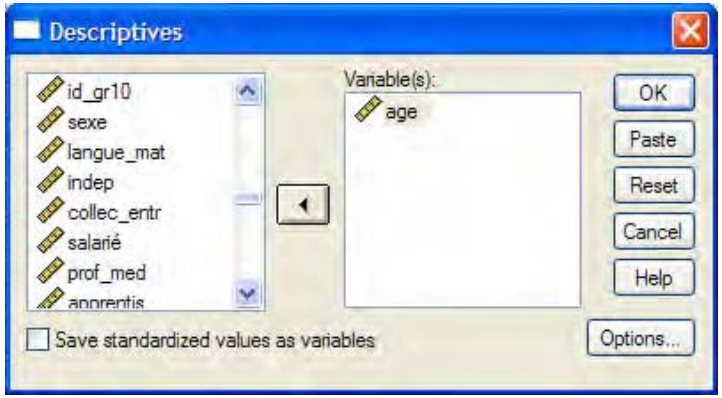
On peut voir que l'échantillon se compose de 34 femmes fumeuses et qu'elles représentent 35.1% des femmes de l'échantillon, 56.7% des fumeurs de l'échantillon et 19.7% du total de l'échantillon. Remarque : l'échantillon est l'ensemble des participants qui constituent la base des données.

5.3. Moyennes (variables numériques)

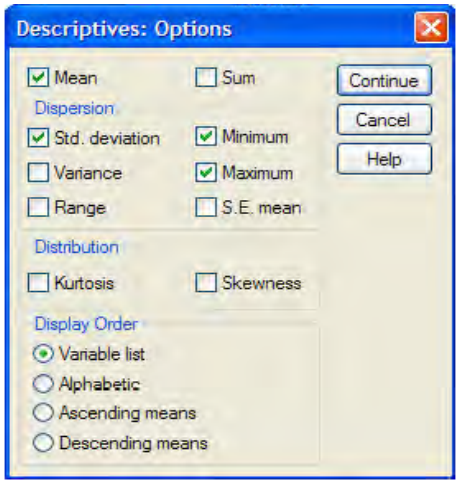
Obtenir la moyenne d'une ou de plusieurs variables.

Chemin: **Analyze → Descriptives statistics → Descriptives…**

Entrer la variable en question dans **Variable(s):**



Dans l'onglet **Option**, il est possible de demander plusieurs indices statistiques descriptifs concernant la variable sélectionnée: la moyenne (**mean**), l'écart-type (**std deviation**), la somme (**Sum**) et d'autres. Ce qui vous intéresse le plus souvent est la moyenne (**mean**), l'écart-type (**std deviation**), la valeur minimale (**minimum**) et la valeur maximale (**maximum**).



Exemple 1 :

Quelle est la moyenne d'âge des participants qui composent notre échantillon ?

SPSS output :

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	170	18.00	64.00	33.8294	11.64516
Valid N (liswise)	170				

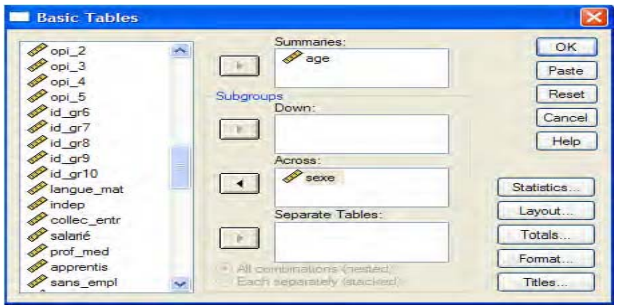
Le tableau montre que l'âge moyen des 170 participants est de 33.83 ans. Le plus jeune des participants est âgé de 18 ans et le plus âgé, de 64 ans. L'écart-type est de 11.64.

5.4. Obtention de la moyenne et de l'écart type de plusieurs sous-groupes de l'échantillon

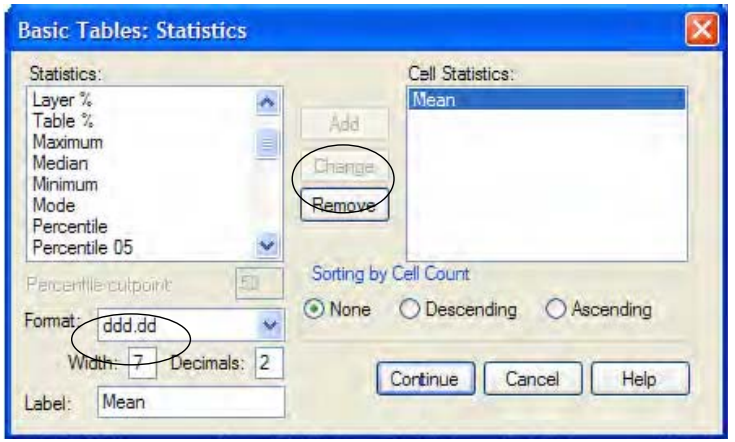
But: Obtenir la moyenne d'une variable numérique pour différents sous-groupes de participants, donc en fonction d'une variable nominale. Dans le chapitre 5.3, vous avez calculé la moyenne sur l'ensemble des participants qui constituent l'échantillon. Dans ce cas-ci, vous voulez calculer la moyenne de certains groupes : par exemple, quelle est la moyenne d'âge des femmes et des hommes. Vous pouvez procéder de plusieurs manières. Au chapitre 4.4, nous vous avons montré comment diviser vos participants en plusieurs sous-groupes (ou sélectionner seulement les participants qui vous intéressent). Vous pouvez utiliser cette méthode, mais il y a une autre manière de procéder (plus élégante), qui consiste à créer des tableaux avec l'information que vous cherchez. Attention : il est question d'une variable numérique et d'une variable nominale !!!!

Chemin : **Analyze → Tables → Basic Tables…**

Entrez la variable numérique dans la fenêtre **Summaries:** et la (ou les) variable(s) nominale(s) dans la fenêtre **Across** (vous pouvez également l'introduire dans la fenêtre **down** ou **separate tables**, le résultat est le même, uniquement la présentation est différente, à vous d'essayer).



Pour modifier la forme du tableau, différentes options peuvent être sélectionnées dans le menu **Layout** (à vous d'essayer). Ensuite, il est nécessaire de demander que les valeurs indiquées comprennent au moins deux chiffres après la virgule (autrement SPSS ne vous montre que le chiffre entier de la moyenne, par exemple 6 à la place de 6.13). Cliquer sur **Statistics**, une boîte de dialogue s'ouvre.



À gauche vous avez les différentes indications qu'il est possible d'obtenir (notamment la moyenne et écart-type). Sélectionnez ce que vous désirez connaître (par exemple **mean**) et déplacez-le dans la fenêtre de droite. Sélectionnez-le une fois déplacé et allez sous **Format**, sous la fenêtre de gauche. Choisissez le format **ddd.dd** et sous **Decimals** modifier le 0 en 2 (ou plus, comme vous désirez. C'est le nombre de décimaux après la virgule). Cliquez sur **change**. Si vous ne cliquez pas sur change, le logiciel ne tient pas compte de la modification). Cette opération est à exécuter pour chaque information demandée. C'est-à-dire que pour chaque élément que vous déplacez dans **cell statistics**, vous devez la sélectionner et en modifier le format. Cliquez sur **Continue** et **OK** une deuxième fois.

Exemple 1 :

Moyenne d'âge en fonction du sexe des participants.

SPSS output :

	Homme			Femme		
	Mean	Std. Deviation	Count	Mean	Std. Deviation	Count
	35.17	12.56	76.00	32.74	10.80	97.00

Les hommes de l'échantillon ont en moyenne 35.17 ans alors que les femmes ont en moyennes 32.74 ans.

6. Statistiques inférentielles

Dans le chapitre 5, il a été question de décrire vos données. Dans cette partie, il est question de tester des hypothèses. Avant de présenter les tests, nous proposons un petit tableau qui vous aidera à décider quelle analyse est la plus adéquate selon le type de variables (nominale ou numérique) dont il est question dans votre base de données et selon le rôle de chacune de celles-ci (variable dépendante ou indépendante).

Tableau de décision statistique : récapitulation du choix des analyses statistiques

VI \ VD	Variable nominale (exemple : sexe, section d'études)	Variable numérique (exemple : âge)
Variable nominale (exemple : sexe, section d'études)	Khi^2 (variables de même type: nominales)	ANOVA, t-test (variables de type différent : VI(s) nominale(s) et VD numérique)
Variable numérique (exemple : âge)	Nous ne traitons pas ce cas	Corrélation (deux variables de même type: numériques. Toutes les variables sont de même niveau, il n'y a pas de VI ou de VD)) Analyse de régression (variables de même type : VI(s) numérique(s) et VD numérique) Analyse factorielle en composantes principales (variables de même type : numériques. Toutes les variables sont de même niveau, il n'y a pas de VI ou de VD)

ATTENTION TRES IMPORTANT, concernant la logique des tests d'hypothèses et la notation officielle !!!!

Pour chaque test statistique que nous vous présentons, le logiciel associe une probabilité qui donne une indication du risque que vous prenez en considérant que votre hypothèse H0 est fausse. En effet, un test d'hypothèse en statistique consiste dans une démarche qui vous porte à rejeter ou à accepter une hypothèse nulle, sur la base d'un échantillon de données.

Une notion fondamentale concernant les tests statistiques est la probabilité que l'on a de se tromper. Dans l'idéal, on souhaiterait avoir un test qui renvoie toujours le "bon" résultat. Par exemple on aimerait avoir un test qui choisisse toujours l'hypothèse nulle lorsque celle ci est vraie et qui rejette tout le temps l'hypothèse nulle lorsque celle ci est fausse. Vous pouvez constater qu'un test statistique ne teste pas si H1 est vraie, mais si H0 peut être rejetée comme fausse.

Il y a deux façons de se tromper lors d'un test statistique:

1) Il y a la possibilité de rejeter H0 comme fausse alors qu'elle est vraie. On appelle ce risque le risque de première espèce et en général on note α la probabilité de se tromper dans ce sens.
2) Il y a la possibilité d'accepter H0 comme vraie alors qu'elle est fausse. On appelle ce risque le risque de deuxième espèce et en général on note β la probabilité de se tromper dans ce sens.

Dans l'idéal, on aimerait bien que ces deux erreurs soient nulles, malheureusement ce n'est pas possible, et il faut alors faire un choix. Ainsi, on peut décider de définir un seuil qui définit UN RISQUE RAISONNABLE DE SE TROMPER lorsqu'on rejette H0 comme fausse. La valeur *p* présentée par SPSS est calculée en tenant compte de ces deux types d'erreurs et la communauté scientifique (du moins pour ce qui concerne la psychologie sociale) s'est accordée pour considérer que lorsque la valeur du *p* est inférieure à .05 ($p < .05$), on peut rejeter H0 comme fausse et donc accepter H1 EN PRENANT UN RISQUE RAISONNABLE DE SE TROMPER AINSI FAISANT. Si $p < .05$ on dit que le test est **significatif**.

Par exemple, vous posez l'hypothèse selon laquelle 'il y a plus de femmes que d'hommes qui étudient la psychologie à l'université (Hypothèse 1, H1). L'hypothèse nulle, H0, sera qu'il y a la même proportion de femmes que d'hommes qui étudient la psychologie à l'université. SPSS vous donne une indication du risque que vous avez de vous tromper en rejetant H0 et en acceptant H1. Plus la valeur *p* est élevée (par exemple $p = .98$), plus vous avez de chances de vous tromper en rejetant H0 comme fausse (dans l'exemple, 98%).

Notations :

Si $p > .10$ (normalement 0,10 mais en notation scientifique .10), on note *ns* (Non Significatif). Le risque de rejeter H0 est trop élevé et votre hypothèse H1 n'est pas vérifiée.

Si $.10 > p > .05$, on note $p =$ (valeur donnée, par exemple .06) et on considère que le teste est tendanciel (pas significatif mais presque). Le risque de rejeter H0 est moins élevé que dans le cas précédent, mais il est toujours trop élevé pour les standards.

1) Si $.05 > p > .01$, on note $p < .05$.

2) Si $.01 > p > .001$, on note $p < .01$.

3) Si $p <. .001$, on note $p < .01$.

} Dans tous ces cas, on dit que le test est *significatif* : le risque de se tromper en rejetant H0 est raisonnablement petit selon les standards.

6.1. Le Khi carré

Principe du test : comparer une distribution observée (c'est-à-dire la distribution « réelle » des participants) à une distribution théorique (c'est-à-dire la distribution qu'il aurait fallu obtenir si l'hypothèse d'indépendance des variables, H0, était vraie). Le principe général consiste à analyser l'écart existant entre la distribution théorique et la distribution observée (c'est le calcul de la différence entre la valeur réelle et la valeur théorique pondérée). Cela revient à tester la relation entre deux variables nominales dans un tableau de contingence (un tableau des fréquences). Plus l'écart entre les deux distributions est grand, moins la valeur réelle est proche de la valeur de l'indépendance. Cela signifie qu'il y a dépendance entre les variables (H1 est acceptée).

Question posée par le test : Est-ce qu'il existe une vraie relation entre les variables X et Y (H1). Hypothèse nulle: il n'y a pas de relation entre les variables X et Y (H0).

Attention : Le *p* donne l'information uniquement sur la probabilité de l'existence d'une relation entre X et Y, mais pas sur la force de cette relation. C'est-à-dire que le *p* traité auparavant vous indique le RISQUE pris en rejetant H0 (qui suppose que le lien entre les deux variables est dû au hasard) comme fausse. Par contre, la valeur du *p* ne vous donne aucune indication de la FORCE du lien entre ces deux variables. Par exemple, un $p = .03$ signifie que vous pouvez considérer que le lien entre les deux variables n'est pas dû au hasard, mais cela ne signifie pas que le lien entre deux variable X et Y est plus fort que le lien existant entre deux variables Z et A pour qui $p = .05$.

Conditions de validité du test : Théoriquement, il faut au moins cinq participants dans chaque case du tableau croisé de la distribution théorique pour que l'analyse soit valable (le logiciel vous indique si c'est le cas).

Remarque

Dans le test du Khi carré, il n'y a pas vraiment de variable dépendante (VD) et de variable indépendante (VI) : les deux variables sont de même niveau. Le test questionne le lien entre elles.

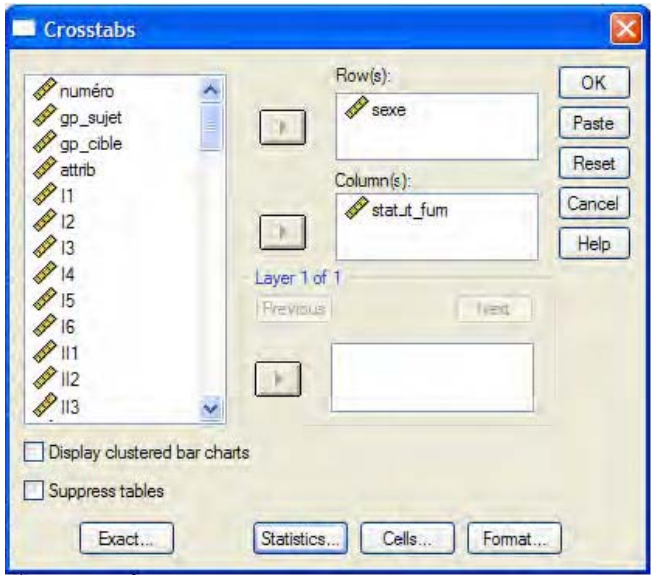
Exemple 1 :

Tester s'il existe une relation entre le fait d'être un homme ou une femme (variable X) et le fait de fumer (variable Y, voir chapitre 5.2). Exemple de H1 : les hommes sont plus souvent des fumeurs que les femmes.

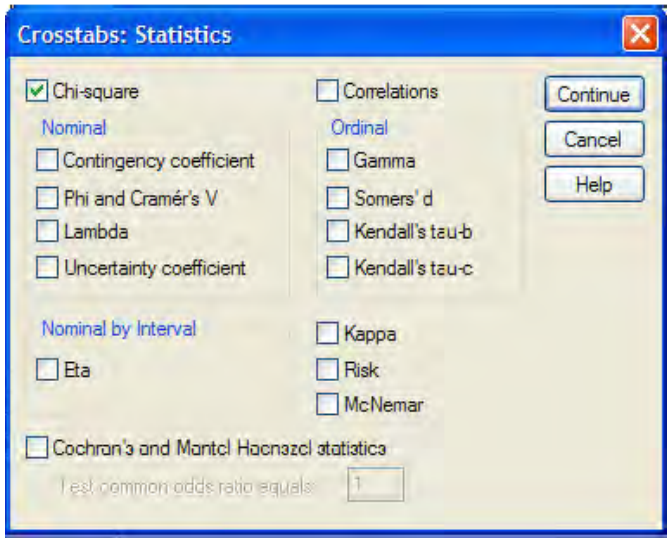
Chemin : **Analyze → Descriptive Statistics → Cross tabs...**

Dans **Rows**, ajouter une des variables (*sexe*).
Dans **Columns**, ajouter l'autre variable (*statut_fum*).
Que vous placiez le *sexe* dans columns et *statut_fum* dans row, ne change rien au résultat.
Dans **Cells** demandez les pourcentages (row, column and total).

CONSEIL : faites l'effort de comprendre la manière dont les pourcentages sont organisés par colonne et par ligne. Ce n'est pas difficile, mais c'est indispensable à l'interprétation des résultats !



Dans l'onglet **Statistics** cochez **Chi-square,** puis cliquez sur **Continue**, puis **OK.**



SPSS output:

			Statut_fum				
			Fumeur	Anc_fumeur	Non_fum	fumeur_occ	Total
Sexe	Homme	Count	26	33	10	7	76
		% within sexe	34.2%	43.3%	13.2%	9.2%	100%
		% within statut_fum	43.3%	40.2%	47.6%	70.0%	43.9%
		% of Total	15.0%	19.1%	5.8%	4.0%	43.9%
	Femme	Count	34	49	11	3	97
		% within sexe	35.1%	50.5%	11.3%	3.1%	100%
		% within statut_fum	56.7%	59.8%	52.4%	30.0%	56.1%
		% of Total	19.7%	28.3%	6.4%	1.7%	56.1%
	Total	Count	60	82	21	10	173
		% within sexe	34.7%	47.4%	12.1%	5.8%	100%
		% within statut_fum	100%	100%	100%	100%	100%
		% of Total	34.7%	47.4%	12.1%	5.8%	100%

On obtient donc le tableau croisé avec le nombre et le pourcentage de participants correspondant à chaque case.

χ^2

Valeur du *p* associé au χ^2

	Value	df	Asymp. Sig (2-sided)	Exact Sig. (2-sided)
Pearson Chi-square	3.336(a)	3	.343	.349
Likelihood Ratio	3.350	3	.341	.355
Fisher Exact Test	3.274			.353
Linear-by-Linear Association	1.367(b)	1	.242	.270
N of Valid Cases	173			

a 1 cell(12.5%) have expected count less than the minimum. Expected minimum count is 4.39
b The standardized statistic is -1.169

Condition de validité du test (a) : on peut voir qu'une case ne remplit pas la condition nécessaire au test. Le **Expected minimum** est inférieur à 5 (4.39). Le test risque de ne pas être fiable.

La valeur du χ^2 est de 3.34, le degré de liberté est de 3 et la probabilité associée à ce test est de .34, ce qui est supérieur au seuil de $p = .05$. Le test n'est donc pas significatif et il y a proportionnellement autant d'hommes que de femmes qui fument; $\chi^2(3) = 3.34$, n.s. Le risque de rejeter H0 en faveur d'H1 est trop élevé.

6.2. Corrélations de Pearson (variables numériques)

Principe du test : mesurer la force et la direction d'une relation linéaire entre deux variables numériques X et Y. Par exemple, on peut tester la corrélation entre l'âge X et l'attitude vis-à-vis de loisirs Y. L'hypothèse (H1a) est : plus les gens sont âgés et plus ils aiment les loisirs. H0 : pas de lien entre âge et le fait d'aimer ou pas les loisirs. Tout comme dans le cas du Khi^2, il n'y a pas de variable dépendante (VD) et de variable indépendante (VI) dans le calcul d'une corrélation. Ainsi, une hypothèse alternative est (H1b) : plus les gens aiment les loisirs, plus ils sont âgés. Le test est le même : la corrélation ne teste pas l'effet d'une variable sur une autre (la causalité), mais uniquement la relation entre elles. La corrélation vous donne une indication du fait que, « plus X alors plus Y » dans le cas d'une corrélation positive ou « plus X alors moins Y » dans le cas d'une corrélation négative.

Description : le coefficient de corrélation (*r*) peut être négatif ou positif, ainsi la relation est négative (si *r* a une valeur entre -1 et 0), positive (si *r* a une valeur entre 0 et 1) ou absente (si *r*= 0). Donc la valeur de *r* varie entre –1 et +1.

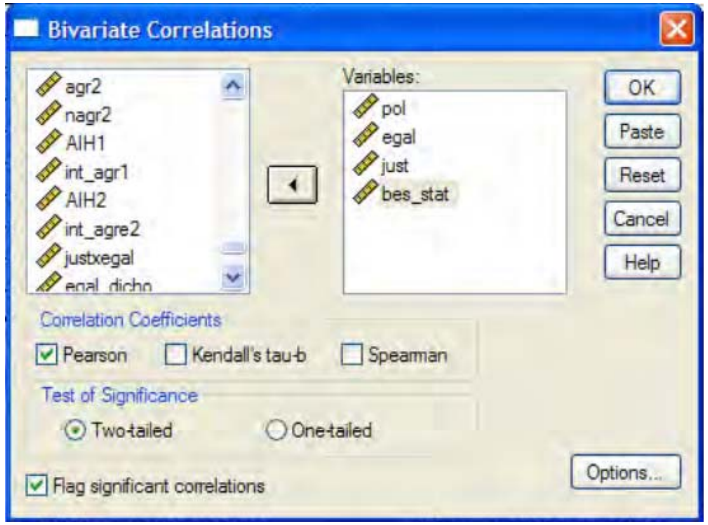
Concepts clés de l'interprétation des résultats :
1) Il n'y a pas de réel accord sur la nécessité d'évaluer la probabilité (le *p* d'une corrélation), mais nous allons partir du principe que c'est important. Si *r* est significatif ($p < .05$) alors la relation entre les deux variables n'est pas due au hasard et vous pouvez rejeter H0 (cela signifie qu'il n'y a pas de lien entre les deux variables).
2) Une fois que vous avez contrôlé que la relation entre les deux variables n'est pas due au hasard (si le risque de rejeter H0 est raisonnablement petit), vous allez vous intéresser à la force et la direction de la relation: la valeur du *r*. Il n'y a pas de règles fixe pour définir quand la taille du coefficient de corrélation montre un vrai lien. Les conventions considèrent qu'une corrélation en dessous de .30 est faible, de .30 à .50 moyenne et de .50 et plus, forte. Attention : la significativité (valeur de *p*) dépend fortement du nombre de sujets dans l'échantillon. Plus il y a de sujets, plus la corrélation devient facilement significative. Il se peut qu'une corrélation de $r = .10$ soit significative ($p < .05$), cela ne signifie pas qu'elle est forte ! Dans ce cas, on parle de force du lien, mais elle ne dépend pas de la valeur du *p* !

SPSS calcule le coefficient de corrélation (*r*), annonce l'effectif (le nombre de participants pris en compte dans le calcul, *n*) et teste la significativité du coefficient (*p*).

Exemple 1 :

Dans un questionnaire, des échelles ont été passées pour mesurer l'égalitarisme, le besoin de statut des individus, la façon dont ils justifient le système économique ainsi que leur position politique. H1 est : quelle est la relation entre les réponses des participants à ces questions ?
Attention : une corrélation teste toujours le lien existant entre deux variables à la fois uniquement !

Chemin : **Analyze → Correlate → Bivariate...**



Comme toujours, vous trouvez la liste des variables dans la fenêtre de gauche. Sélectionnez les variables qui vous intéressent et passez-les dans la fenêtre **Variables** à l'aide de la flèche qui se situe entre les deux fenêtres. Attention, vos variables doivent être indiquées comme étant numériques par SPSS. Si ce n'est pas le cas, vous ne pourrez pas les transférer. Par défaut, la case **Pearson** est cochée (**Kendall's tau-b** ou **Spearman** représentent d'autres manières de calculer vos coefficients). Nous allons toujours utiliser le test de **Pearson**, la différence avec les deux autres coefficients étant minime. Cliquez sur **OK**.

SPSS output :

Correlations

		pol	egal	just	bes_stat
pol	Pearson correlation	1	-.467**	.475**	.147
	Sig. (2-tailed)		.000	.000	.196
	N	79	79	79	79
egal	Pearson correlation	-.467**	1	-.556**	-.296**
	Sig. (2-tailed)	.000		.000	.008
	N	79	80	80	80
just	Pearson correlation	.475**	-.556**	1	.276*
	Sig. (2-tailed)	.000	.000		.013
	N	79	80	80	80
bes_stat	Pearson correlation	.147	-.296**	.276*	1
	Sig. (2-tailed)	.196	.008	.013	
	N	79	80	80	80

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

Valeur du *p* associée au *r*

Coefficient de corrélation

Nombre de participants – degrés de liberté du *r*

SPSS calcule toutes les corrélations possibles deux à deux et les présente sous forme de tableau croisé. On ne s'intéresse donc qu'aux résultats en dessus (ou en dessous) de la diagonale, puisque le tableau est symétrique. Vous remarquez que la diagonale présente des corrélations de valeur $r = 1$: une variable est parfaitement corrélée avec elle-même.

Dans le tableau, les corrélations significatives sont accompagnées d'étoiles *. Plus il y a d'étoiles, plus le *p* est petit. N'oubliez pas de regarder la valeur du *p* en tout cas (habituez-vous).

Commentons quelques corrélations :

1) On observe une corrélation positive significative entre la position politique (*pol*) et la justification du système (*just*) : $r = .475$, $p < .001$. Cette corrélation est assez forte et indique que plus les personnes se disent de droite (l'échelle de *pol* va de 1 = gauche à 8 = droite) plus elles justifient le système économique (l'échelle de justification du système va de 1 = ne pas justifier à 6 = justifier) et inversement, c'est-à-dire que plus les individus justifient le système, plus ils se disent de droite. En fait, le test ne permet pas de déterminer quelle variable est la conséquence de l'autre.

2) On observe une corrélation négative significative entre la justification du système et l'égalitarisme (*egal*) : $r = -.56$, $p < .001$. Cette corrélation est forte et indique que plus les personnes justifient le système économique, moins elles voudraient que tous les groupes soient égaux dans la société et inversement.

3) Le besoin de statut et la politique ne sont pas corrélés : $r = .15$, *ns*. Il n'a pas de relation entre le fait de se positionner à gauche ou à droite et le besoin de statut de l'individu.

Attention : une corrélation ne vous permet pas de dire, par exemple, que c'est parce que les participants justifient le système qu'ils disent être de droite. Elle vous permet uniquement de dire que les deux mesures sont liées. De plus, l'interprétation des corrélations dépend du sens de vos échelles ! Faites attention à ce que signifient sur vos échelles des valeurs élevées ou basses. Si, par exemple, l'âge est mesuré de 1 à 100 ans et le fait d'aimer les loisirs de 1 = aimer pas du tout à 10 = aimer tout à fait, une corrélation positive signifie que plus les individus sont âgés, plus ils aiment les loisirs. Si l'échelle des loisirs va de 1 = aimer tout à fait à 10 = aimer pas du tout, une corrélation positive signifie que plus les gens sont âgés, moins ils aiment les loisirs !!!!!

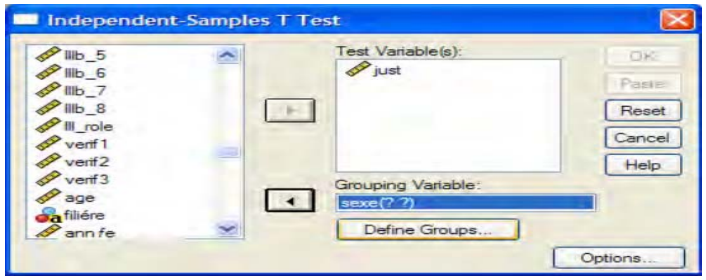
6.3. Test statistique de la différence entre deux moyennes

6.3.1. T-test avec 1 variable numérique et 1 variable nominale à deux modalités

But: tester si la différence observée entre deux moyennes est statistiquement significative, et donc si cette différence ne peut pas être expliquée par le hasard. On utilise en général le t-test lorsque l'on a une idée sur le type de différence entre les deux groupes de participants à l'étude.

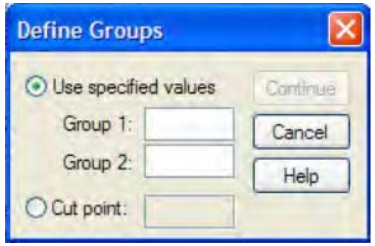
Par exemple, nous pouvons tester l'hypothèse suivante : les femmes justifient moins le système économique que les hommes (H1). Notre VI est donc le sexe (variable nominale à 2 modalités) des participants et notre VD est la justification du système (variable numérique).

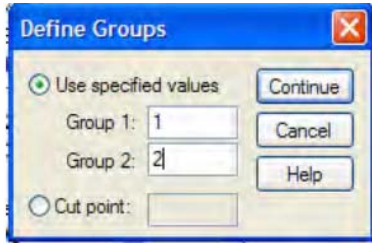
Chemin : **Analyze → compare means → independant samples T-test...**



Déplacez votre VD (pour notre exemple, *just*) de la liste de gauche vers l'onglet **Test Variable(s)** à l'aide de la petite flèche du haut. Déplacez ensuite votre VI (pour notre exemple, *sexe*) de la liste de gauche vers l'onglet **Grouping Variable(s)** à l'aide de la petite flèche du bas.

Le logiciel vous demande alors de définir les groupes de votre VI dont vous voulez comparer les moyennes. Cliquez sur **Define Groups…**





Une petite fenêtre apparaît alors et vous devez spécifier que le groupe 1 correspond à la modalité 1 de votre VI (les femmes) et que le groupe 2 correspond à la modalité 2 de votre VI (les hommes).

Cliquez sur **continue** puis **OK**

SPSS Output :

Independent Samples Test

		Levene's Test for Equality of Variance							95% Confidence Interval of the Difference	
		F	Sig.	t	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Just	Equal variances assumed	1.04	.31	1.73	78	.087	-.404	.285	-1.06	.074
	Equal variances not assumed			-1.87	48.85	.067	-.494	.263	-1.02	.035

Dans ce premier tableau se trouvent les informations qui concernent la significativité de la différence entre les deux moyennes. Deux cas sont possibles :
1) Dans la colonne **Levene's test for equality of variance** le F est non significatif (ici F = 1.04, p = .31, *ns*) et on regarde les résultats sur la colonne **Equal variances assumed**.
2) Dans la colonne **Levene's test for equality of variance** le F est significatif (il faudrait que $p < .05$, ce qui n'est pas le cas ici) et l'on regarde les résultats sur la colonne **Equal variances not assumed**.

Dans notre exemple, on considère la première ligne (**Equal variances assumed**). Dans la colonne ***t*** on peut lire la valeur -1.73 et dans la colonne **Sig (2-tailed)** on peut lire .087. Cela veut dire que la différence entre hommes et femmes concernant la justification du système n'est que tendanciellement significative. Les degrés de liberté sont notés dans la colonne **df**, donc on écrit $t(78) = -1.73$, $p = .09$.

Group Statistics

	sexe	N	Mean	Std. Deviation	Std. Error Mean
just	Femme	57	-.1420	1.21024	.16030
	Homme	23	.3520	1.00228	.20899

Dans ce second tableau, vous pouvez lire les effectifs pour chaque groupe (57 femmes et 23 hommes), ainsi que les moyennes correspondantes. On voit que les femmes justifient tendanciellement moins le système (M = -.14, SD = 1.21) que ne le font les hommes (M = .35, SD = 1.00).

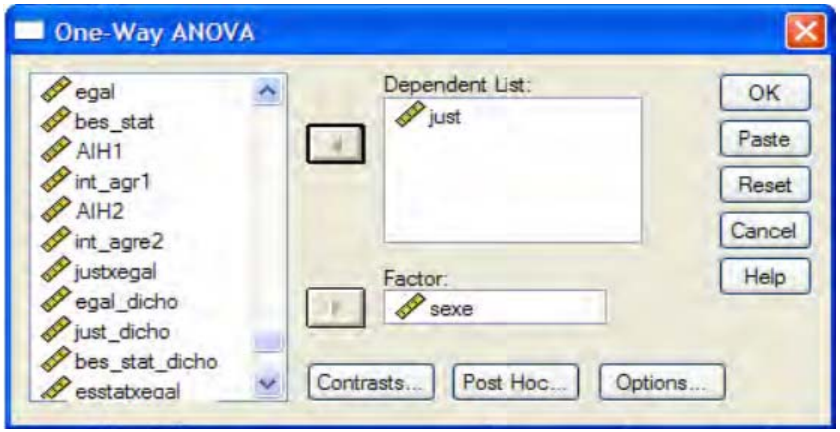
6.3.2. One Way ANOVA avec 1 variable numérique et 1 variable nominale à deux modalités

But: tester si la différence observée entre deux moyennes est statistiquement significative, et donc si cette différence ne peut pas être expliquée par le hasard. Pour cela, l'analyse de variance prend en compte non seulement l'importance de la différence des moyennes, mais également la dispersion des réponses des participants autour de la moyenne (l'écart-type). Si cette dispersion est forte, la différence entre les moyennes risque de ne pas être significative, tandis que si les réponses des participants sont proches de la moyenne (faible dispersion), la différence entre les moyennes a plus de chances d'être significative. Contrairement à l'analyse présentée au chapitre 6.3.1, ici H1 se limite à postuler l'existence d'une différence entre les groupes (les hommes et les femmes se différencient entre eux quant à leur adhésion au système), mais elle ne définit pas « dans quelle direction va » cette différence (par exemple, si les hommes adhèrent davantage au système que les femmes).

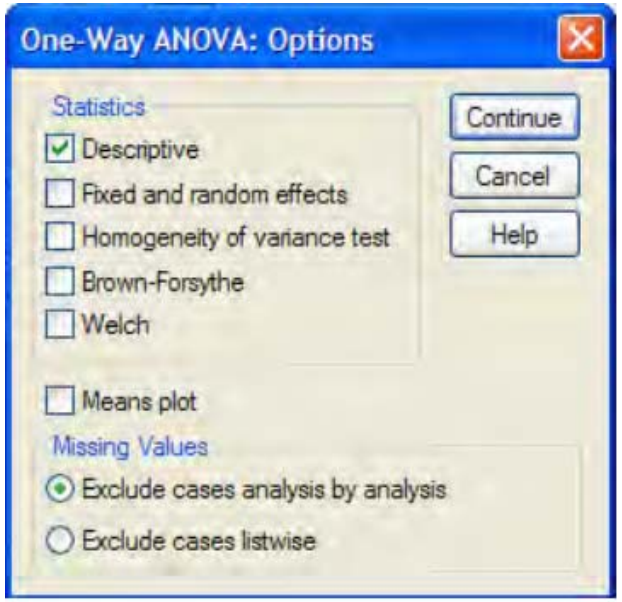
Exemple 1 :

H1: les femmes et les hommes justifient différemment le système économique. Notre VI est donc le sexe des participants (variable nominale à 2 modalités) et notre VD est la justification du système (variable numérique).

Chemin : **Analyze → Compare means → One Way ANOVA...**



Entrez la variable numérique dans la fenêtre **Dependent List:** et la variable nominale (dans ce cas : sexe) dans la fenêtre **Factor:**
Puis cliquez sur **Option**. Une nouvelle boîte de dialogue s'ouvre et vous pouvez demander que les statistiques descriptives apparaissent dans l'output.



Sélectionnez la case **Descriptive.** Puis **Continue** et la deuxième boîte de dialogue se ferme. Cliquez sur **OK**.

SPSS Output :

Degrés de liberté de l'effet du sexe

Valeur du *F* pour le sexe

p associé au *F* pour le sexe

ANOVA

Just

	Sum of Squares	df	Mean Squares	F	Sig
Between Groups	3.999	1	3.999	2.996	.087
Within Groups	104.123	78	1.335		
Total	108.122	79			

Degrés de liberté de l'erreur

Afin de comparer les deux moyennes à l'aide d'une ANOVA, SPSS calcule un indice appelé *F*. C'est la valeur de ce *F* et la probabilité *p* qui lui est associée qui vont nous permettre ou non de vérifier l'hypothèse H1. Le *F* qui nous intéresse dans le tableau est celui qui est associé à notre facteur intergroupe (**Between groups**). Dans notre exemple, il est de 2,99.

Par convention on note $F(1,78) = 2.99$, $p = .087$**.** Soit *F* (degrés de liberté de l'effet, degrés de liberté de l'erreur) = valeur du *F* pour le sexe, *p* = valeur de Sig dans tableau (ici *F* est tendanciel, donc il faut noter la valeur exacte du *p*).

Descriptives

Just

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
femme	57	-.1420	1.12024	.16030	-.4631	.1791	-2.62	2.20
homme	23	.3520	1.00228	.20899	-.0814	.7854	-1.56	1.85
Total	80	.0000	1.16989	.13080	.2603	.2603	-2.62	2.20

Dans le tableau, se trouvent les moyennes des hommes et des femmes sur la VD (dans notre exemple, la justification du système labellisé *Just*). Vous voyez que les femmes justifient tendanciellement moins le système ($M = -.14$, $SD = 1.21$) que ne le font les hommes ($M = .35$, $SD = 1.00$).

Dans un rapport, on écrit :

On trouve un effet tendanciel de l'appartenance sexuelle sur la justification du système : $F(1,78) = 2.99$, $p = .087$. Cet effet nous indique que les femmes justifient moins le système ($M = -.14$, $SD = 1.21$) que les hommes ($M = .35$, $SD = 1.00$) mais cette différence n'est que tendancielle.

Exemple 2 :

H1 : les femmes et les hommes se positionnent différemment sur l'échiquier politique. Notre VI est toujours le sexe des participants à l'étude, notre VD est la position politique.

On effectue l'analyse comme précisée plus haut:

SPSS output :

ANOVA

pol

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	14.965	1	14.965	7.339	.008
Within Groups	157.009	77	2.039		
Total	171.975	78			

Descriptives

pol

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
Femme	56	3.60	1.45	.19	3.22	3.99	1.00	8.00
Homme	23	4.56	1.38	.29	3.97	5.16	3.00	8.00
Total	79	3.88	1.48	.17	3.55	4.22	1.00	8.00

On peut dire qu'il y a un effet significatif de l'appartenance sexuelle des participants sur leur position politique : $F(1,77) = 7.34$, $p < .01$. Effectivement, le tableau des moyennes nous indique que les femmes disent voter plus à gauche ($M = 3.60$, $SD = 1.45$) que ne le font les hommes ($M = 4.56$, $SD = 1.38$). Nous rejetons H0 comme fausse en faveur de H1.

6.3.3. Test statistique de la différence entre plusieurs moyennes: ANOVA avec 1 variable numérique et 1 variable nominale à plus de 2 modalités

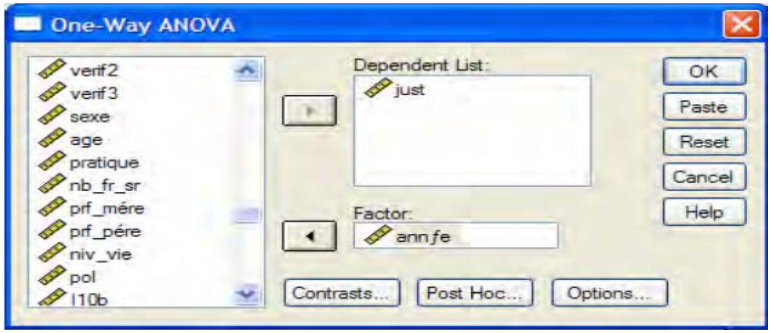
Le cadre expérimental est comparable à celui exposé dans le chapitre 6.3.2 (une VD numérique et une VI nominale).

But : parfois, la variable indépendante nominale a plus de deux modalités (on compare donc entre elles les moyennes de plusieurs groupes). L'ANOVA indique si ces moyennes sont globalement différentes entre elles, sans préciser exactement quelles moyennes sont différentes ou similaires entre elles. C'est-à-dire que le test peut signifier que la moyenne1 est différente de la moyenne2, mais que la moyenne 2 n'est pas différente de la moyenne3, alors que la moyenne1 est différente de la moyenne3. Mais il peut également signifier que la moyenne1 n'est pas différente de la moyenne2, qui, elle, est différente de la moyenne3. Par contre, la moyenne1 n'est pas différente de la moyenne3. Ainsi, une fois avoir vérifié que le test global (le *F* de l'ANOVA) est significatif, il faut tester les effets spécifiques.

Exemple :

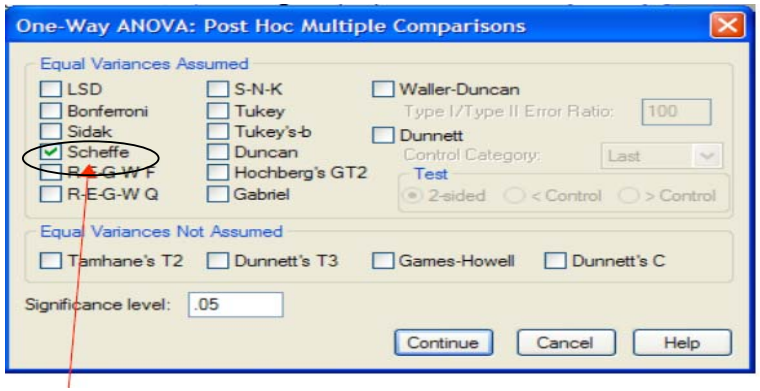
H1 : les étudiants universitaires justifient plus le système lorsqu'ils entrent à l'université que lorsqu'ils en sortent. Notre VI est donc l'année d'étude (3 modalités : 1 = première année bachelor, 2 = deuxième année bachelor, 3 = troisième année bachelor) et notre VD, la justification du système.
H0 : pas de différence entre les moyennes.
H1 : il y a une différence entre au moins deux des trois moyennes.

Chemin : **Analyze → Compare Means → One-Way ANOVA...**



Comme dans le paragraphe précédent, entrez votre VI dans **Factor:** et votre VD dans **Dependent list**

Pour le test des moyennes deux à deux, cliquez sur **Post Hoc…** Une nouvelle fenêtre s'ouvre:



Demandez le test de **Scheffe,** puis cliquez sur **Continue.**

Vous pouvez utiliser d'autres tests post-hoc disponibles dans SPSS, ce qui les différencie est la pondération des comparaisons entre les moyennes.

Sous **Option**, n'oubliez pas de demander les statistiques descriptives.

SPSS Output :

ANOVA

Just

	Sum of Squares	df	Mean Squares	F	Sig
Between Groups	4.830	2	2.415	1.793	.173
Within Groups	102.347	76	1.347		
Total	107.177	78			

Le tableau d'ANOVA montre que, globalement, la différence entre les trois groupes d'étudiants n'est pas significative : $F(2, 76) = 1.79$, *ns*. Ce qui signifie que les moyennes ne sont pas significativement différentes entre elles.

Descriptives

Just

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean Lower Bound	95% Confidence Interval for Mean Upper Bound	Minimum	Maximum
1	45	.1418	1.17339	.17492	-.2108	.4943	-2.50	2.20
2	11	.1641	1.06543	.32124	-.5516	.8799	-1.45	1.44
3	23	-.3978	1.17579	.24517	-.9063	.1106	-2.62	1.85
Total	19	-.0122	1.17220	.13188	-.2748	.2503	-2.62	2.20

Multiple Comparaisons

Dependent variable : just
Sheffe

(I) année	(J) année	Mean Difference (i – J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
1.00	2.00	-.02237	.39032	.998	-.9969	.9522
	3.00	.53962	.29745	.200	-.2031	1.2823
2.00	1.00	.02237	.39032	.998	-.9522	.9969
	3.00	.56199	.42541	.422	-.5002	1.6242
3.00	1.00	-.53962	.29745	.200	-1.2823	.2031
	2.00	-.56199	.42541	.422	-1.6242	.5002

Ce tableau (**Multiple Comparaisons**) nous permet de tester les différences entre les moyennes prises deux à deux. Par exemple, on voit qu'entre les 1ères années et les 2èmes années la différence entre les deux moyennes est de .022 et que cette différence n'est pas significative, puisque $p > .05$ (= .998). Puisque le test global n'est pas significatif, vous ne devez normalement pas vous intéresser à ce tableau. Ici, nous l'avons présenté comme exemple. Si le test global avait été significatif, il aurait été possible qu'une des 3 comparaisons (regardez bien, 3 d'entre elles sont des répétitions) ait été significative. Toutefois, le fait que le test global soit significatif, n'est pas une garantie d'observer une différence significative lors des comparaisons deux à deux. Ne paniquez pas : cela est dû au fait que le test post-hoc de Scheffe pondère les effets et que cette pondération peut ne pas mettre en lumière certains effets.

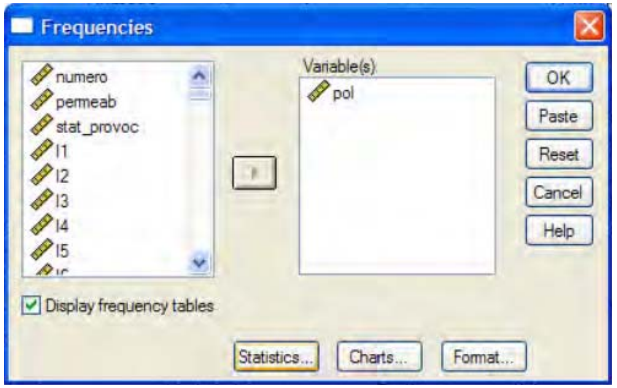
6.3.4. Test statistique de la différence entre plusieurs moyennes définies par plusieurs variables: ANOVA avec 1 variable numérique et plusieurs variables nominales

La variable dépendante (VD) de l'analyse est la variable numérique, alors que les variables indépendantes (VI) sont nominales. La différence par rapport au test présenté au chapitre 6.4 est que vous allez évaluer l'effet de plusieurs VI à la fois sur la même VD et non plus l'effet d'une seule VI !

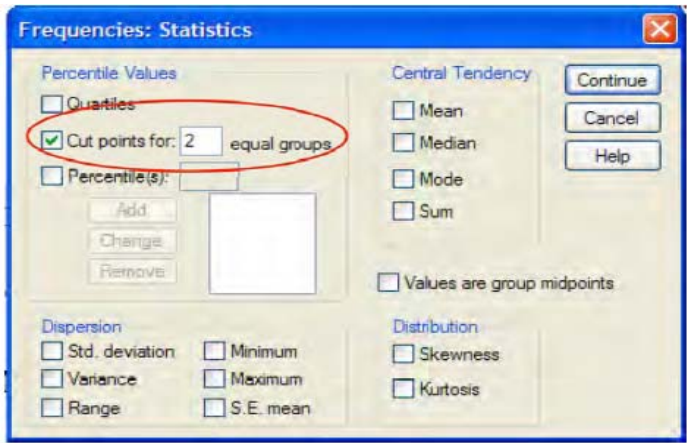
But: tester simultanément l'effet de plusieurs variables nominales, ainsi que celui de leurs interactions, sur la variable dépendante.

Par exemple, nous cherchons à savoir s'il existe un effet conjoint du sexe (VI1) et de la position politique (VI2, variable que l'on rendra dichotomique : les participants qui se disent de gauche et les participants qui se disent de droite) sur le fait de justifier le système. Nos deux VI sont donc le sexe et la position politique, toutes deux doivent être nominales. Notre VD est la justification du système, variable numérique.

En premier lieu il est nécessaire de recoder la variable *pol*. C'est un recodage et non une transformation mathématique ! Les étudiants SSP étant plutôt à gauche, on peut supposer qu'il n'y a pas vraiment de participants qui se déclarent de droite. Il est donc plus intéressant de comparer ceux qui sont plus à droite ou à gauche relativement à l'échantillon des participants plutôt que de créer les catégories de 1 à 4 et de 5 à 8, à priori selon l'échelle utilisée (de 1 = gauche à 8 = droite). Ce qui signifie que, si l'échelle va de 1 à 8, vous n'allez pas créer deux groupes sur la base de l'échelle (le groupe de gauche qui regroupe les gens qui ont répondu de 1 à 4 et le groupe de droite qui regroupe les gens qui ont répondu de 5 à 8), mais vous allez plutôt demander à spss de séparer les individus en deux groupes égaux (avec le même nombre de participants) sur la base des réponses de ceux-ci. On procède donc au découpage des participants en deux groupes. Pour obtenir deux groupes d'effectifs équivalents, allez sous **Descriptive statistics** puis **Frequencies**, sélectionnez la variable *pol* et déplacez-la dans la fenêtre de droite.



Cliquez ensuite sur **Statistics**, une nouvelle boîte de dialogue s'ouvre:



En haut à gauche se trouve la phrase « **cut point for X equal groups** ». Cochez-la et dans la case vide, ajoutez le nombre de groupes désiré (dans ce cas 2). Cliquez sur **Continue** et **OK**.

SPSS output :

Statistics

pol

N	Valid	79
	Missing	1
Percentiles	50	4.0000

Pol

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Gauche	3	3.8	3.8	3.8
	2	9	11.3	11.4	15.2
	3	23	28.8	29.1	44.3
	4	18	22.5	22.8	67.1
	5	17	21.3	21.5	88.6
	6	5	6.3	6.3	94.9
	7	2	2.5	2.5	97.5
	Droite	2	2.5	2.5	100.0
	Total	79	98.8	100.0	
Missing	System	1	1.3		

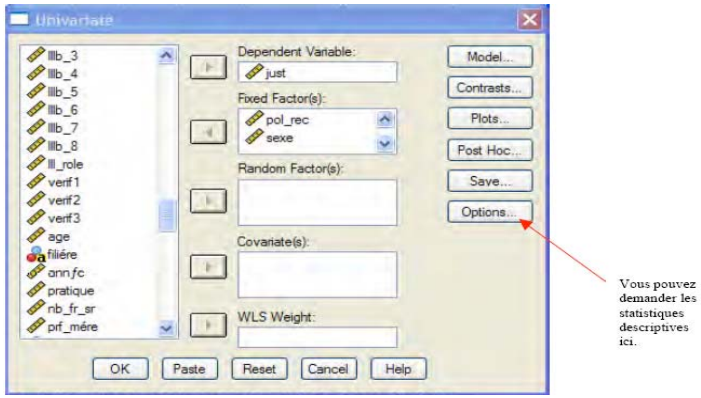
Le premier tableau indique le nombre de réponses valides et que la médiane se situe à la valeur 4. Le deuxième tableau montre que pour la valeur 4 on atteint un pourcentage cumulé de 67.1%. C'est-à-dire que 67.1% des participants a répondu en dessous du 4 (compris). Nous n'avons donc pas de valeur qui coupe l'échantillon exactement à 50%.
Nous allons donc subdiviser les participants sur la base de la valeur la plus proche de 50%, à savoir la valeur 3 et pour laquelle on atteint un pourcentage cumulé de 44.3%. Cela signifie que, pour obtenir deux groupes des participants plus ou moins égaux, il est nécessaire de regrouper les participants qui ont répondu 1, 2, ou 3 dans une catégorie (1 = participants de gauche), et les participants qui ont répondu 4, 5, 6, 7 et 8 dans une autre (2 = participants de droite). Bien entendu, ces deux catégories sont relatives à la distribution des réponses des participants à l'étude et elles ne sont pas absolues.

Pou créer la nouvelle variable politique dichotomique (voir chapitre 4.2) :

Chemin : **Transform → Recode → Into different variable...**

Maintenant vous pouvez procéder à l'analyse avec la variable *sexe* et la nouvelle variable gauche-droite (*pol_rec* dans l' exemple).

Chemin : **Analyze → General Linear Model → Univariate…**



Entrez la variable numérique (VD) dans la fenêtre **Dependant Variable:** et les variables nominales (VI) dans la fenêtre **Fixed Factor(s).** Demandez les **Descriptives** sous les **Options**.

L'analyse donne à la fois les effets principaux comme auparavant (différences éventuelles entre hommes et femmes, et entre les participants de gauche et les participants de droite) et, nouveauté, l'effet d'interaction entre les deux variables (par exemple différence éventuelle entre hommes et femmes, mais seulement parmi les participants de gauche). Un effet d'interaction représente le croisement des effets du sexe et du positionnement politique sur la VD. Par exemple, il se peut qu'il n'y ait pas de différence significative entre les hommes et les femmes, ni entre les participants de gauche et les participants de droite, mais que les femmes de gauche se différencient des femmes de droite, et les hommes de gauche ne se différencient pas des hommes de droite. Plusieurs cas de figure sont possibles. Vous pouvez formuler des hypothèses sur les effets principaux et sur les effets d'interaction, cela dépend de votre cadre théorique.

SPSS output :

Between-Subjects Factor

		Value Label	N
Sexe	1.00	Femme	56
	2.00	Homme	23
Pol_rec	1.00	Gauche	35
	2.00	Droite	44

Descriptive Statistics

Dependent Variable : Just

Sexe	Pol_rec	Mean	Std. Deviation	N
femme	1.00	-.5310	1.32216	28
	2.00	.2482	.98739	28
	Total	-.1414	1.29191	56
homme	1.00	-.6051	.79771	7
	2.00	.7707	.77734	16
	Total	.3520	1.00228	23
Total	1.00	-.5458	1.22532	35
	2.00	.4382	.94213	44
	Total	.0023	1.17719	79

Test of Between-Subjects Effects

Dependent Variable : Just

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	21.686 (a)	3	7.229	6.275	.001
Intercept	.050	1	.050	.043	.863
sexe	.726	1	.726	.630	.430
Pol_rec	16.779	1	16.779	14.564	.000
sexe*pol_rec	1.287	1	1.287	1.117	.294
Error	86.404	75	1.152		
Total	108.091	79			
Corrected Total	108.090	78			

A R Squared = .201 (Adjusted R Squared = . 169)

Les deux premiers tableaux donnent une indication des codes utilisés pour identifier les groupes et de leurs labels (homme ou femme, gauche ou droite) et aussi des moyennes de chaque condition prise en compte, les écarts-types et les effectifs (nombre de participants pas case). Ces informations ne sont pas essentielles pour interpréter l'analyse.

Le premier tableau que vous devez regarder est nommé **Tests of Between-Subjects Effects**. Dans ce tableau, on trouve les effets de chaque variable indépendante (effets principaux), puis de l'interaction (*sexe*pol_rec*). Donc sur la ligne *sexe*, vous avez l'effet du sexe sur la VD avec le *F* et le *p* associé, sur la ligne *pol_rec* vous avez l'effet de la position politique sur la VD (avec *F* et *p*), et sur la ligne *sexe*pol_rec*, vous avez l'effet des deux variables conjointes, auquel est associé un *F* et un *p*.

Dans les tableaux qui suivent vous avez les moyennes correspondant aux effets.

Tableau moyennes 1.sexe

Dependent Variable : just

Sexe	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Femme	-.141	.143	-.427	.144
homme	.083	.243	-.402	.567

Tableau moyennes 2.pol_rec

Dependent Variable : just

Pol_rec	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Gauche	-.568	.227	-1.020	-.116
Droite	.509	.168	.174	.845

Tableau moyennes 3.sexe*pol_rec

Dependent Variable : just

Sexe	Pol_rec	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Femme	Gauche	-.531	.203	-.935	.127
	Droite	.248	.203	-.156	.652
Homme	Gauche	-.605	.406	-1.413	.203
	Droite	.771	.268	.236	1.305

Dans le tableau **Tests of Between-Subjects Effects**, on voit que l'effet du sexe est non significatif : $F(1,75) = .63$, *ns*. Les hommes et les femmes ne diffèrent pas quant à leurs réponses sur l'échelle de justification. Même si les valeurs des moyennes ne sont pas exactement les mêmes, le test a mis en évidence que cette différence n'est pas significative.

Pour la politique, nous observons un effet significatif : $F(1, 75) = 14.56$, $p < .001$. Dans le **tableau moyennes 2.pol_rec,** on voit que les participants qui indiquent être de gauche justifient moins le système ($M = -.56$; $SD = .22$) que ceux qui indiquent être de droite ($M = .51$; $SD = .17$).

L'interaction entre les deux variables indépendantes n'est pas significative : $F(1,75) = 1.11$, *ns*. Donc il n'y a pas de différences significatives entre le fait d'être un homme qui indique être de droite, un homme qui indique être de gauche, une femme qui indique être de gauche ou une femme qui indique être de droite, dans les réponses données sur la justification. Si l'interaction avait été positive, le **tableau moyennes 3.sexe*pol_rec** nous aurait permis d'interpréter cette interaction. Attention, sur les tableaux qui présentent les moyennes ne figurent pas directement les écarts-types. Vous pouvez les obtenir en utilisant la fonction **Descriptives Statistics**.

Topo rapide sur les interactions

Parfois, il est plus simple de représenter des différences entre les groupes par des graphiques, plutôt que par des tableaux avec des moyennes. C'est souvent le cas lorsqu'il est question d'effets d'interaction. À partir de vos tableaux des moyennes (par exemple le **tableau moyennes 3.sexe*pol_rec**), vous pouvez réaliser les graphiques de vos interactions. Ceux-ci peuvent vous aider à interpréter vos effets et ils sont nécessaires pour la présentation de vos résultats. Nous allons vous proposer des allures de graphiques selon qu'il y a interaction ou pas.

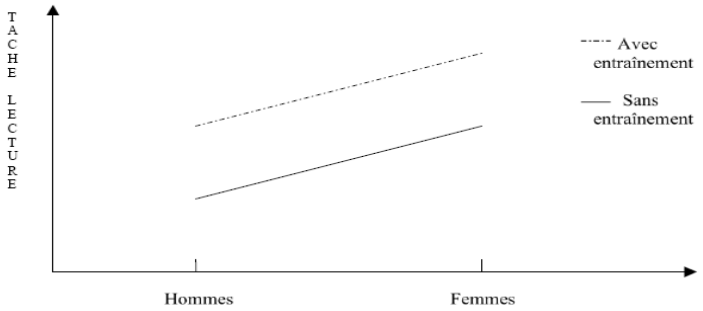
Pour illustrer nos graphiques, nous dirons que nous avons réalisé une expérience sur des hommes et sur des femmes, qui étaient soumis pour la moitié à un entraînement à une tâche verbale et pour l'autre moitié non. Nous avons regardé ensuite si cet entraînement était bénéfique, en fonction du sexe, sur une tâche de rapidité de lecture. Les variables indépendantes sont le sexe des participants (variable nominale à 2 modalités) et le fait d'avoir suivi un entraînent ou pas (variable nominale à deux modalités), la variable dépendante est le score à la tâche de rapidité à de lecture (variable numérique, échelle 1 = pas du tout rapide, 10 = très rapide).

Le plan expérimental se présente ainsi :

	Hommes	Femmes
Entraînement	Moyenne M1	Moyenne M2
Pas entraînement	Moyenne M3	Moyenne M4

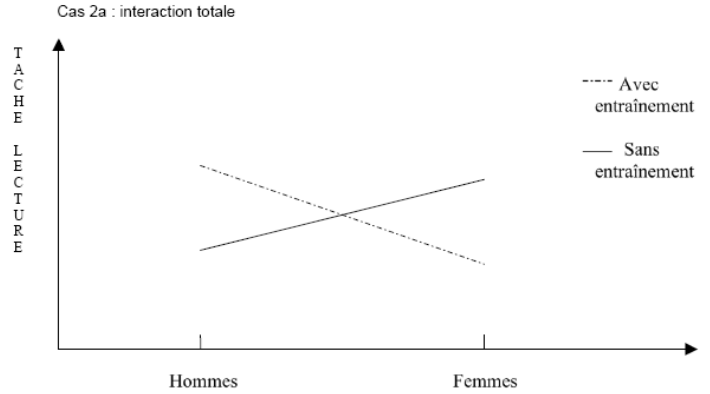
Trois effets (et donc trois types d'hypothèses) sont possibles :
1) Effet principal du sexe des participants (par exemple : les femmes lisent plus rapidement que les hommes)
2) Effet principal de l'entraînement (par exemple : les participants qui ont suivi un entraînement lisent plus rapidement des participants qui n'ont pas suivi cet entraînement).
3) Effet d'interaction (par exemple : l'effet de l'entraînement va améliorer la rapidité de lecture chez les femmes et empirer chez les hommes
Dans cet exemple, l'intérêt est de vous montrer comment représenter graphiquement l'effet significatif d'interaction ou l'effet non significatif d'interaction (selon les valeurs des moyennes dans le tableau)

- Cas 1 : pas d'interaction significative

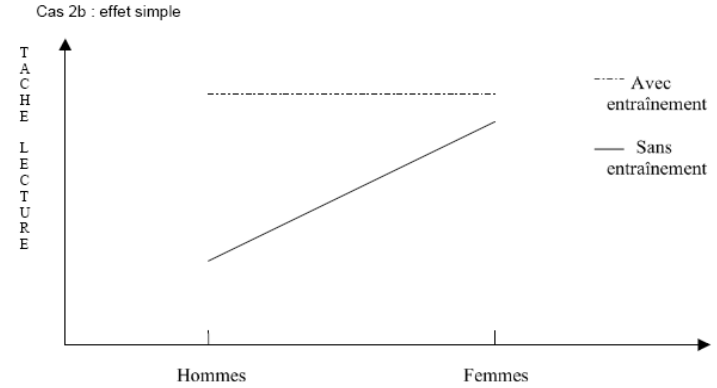


À partir du moment où les droites sont parallèles, il n'y a pas d'interaction. Le graphique montre les effets principaux du sexe des participants et de l'entraînement à la tâche de lecture. Si vous regardez les points qui représentent les femmes et les hommes, vous remarquez que les femmes sont plus rapides que les hommes à la tâche de lecture. En ce qui concerne l'effet de l'entraînement, les positions des points qui représentent ces groupes (et les lignes qui les unissent) montrent qu'en général lorsqu'il y a entraînement les performances des hommes et des femmes sont meilleures. Il va de soi que pour pouvoir dire qu'il y a différence ou pas, il faut que le test soit significatif !!!! Les deux droites pourraient aussi être confondues ou plates, ce qui indiquerait la disparition d'un des deux effets principaux (sexe ou entraînement).

- Cas 2 : Interaction significative, deux sortes



Dans ce cas, on voit que l'effet de l'entraînement n'a pas les mêmes répercussions sur les hommes et sur les femmes. Les femmes réussissent mieux la tâche lorsqu'elles ne reçoivent pas d'entraînement que lorsqu'elles en reçoivent un, alors que pour les hommes c'est exactement l'inverse. Les 4 conditions expérimentales diffèrent les unes des autres. Encore une fois, pour parler de différences, il faut que le test de l'interaction soit significatif.



Il s'agit encore d'une interaction, mais cette fois-ci l'effet de l'entraînement n'est bénéfique que pour les hommes et n'a pas d'effet sur les femmes. C'est un effet simple du sexe.

6.3.5. Test statistique de la différence entre deux ou plusieurs moyennes provenant des mêmes participants : ANOVA avec 2 variables numériques à mesures répétées (VD) et une variable nominale (VD)

Les variables dépendantes (VD) sont des variables numériques, alors que la variable indépendante (VI) est nominale. La spécificité de ce test est que vous allez évaluer l'effet

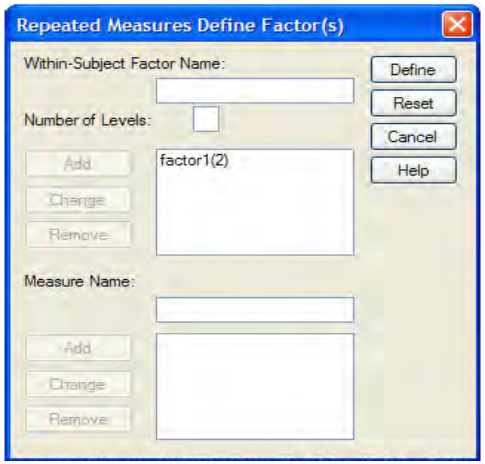
d'une VI sur la différence qui existe entre deux VD. C'est-à-dire qu'il n'est plus question de comparer des groupes sur une variable, mais de comparer des groupes sur la différence entre deux variables !

But : effectuer une analyse en croisant une mesure répétée (2 variables numériques ou plus mesurées sur les mêmes participants) et une variable indépendante nominale.

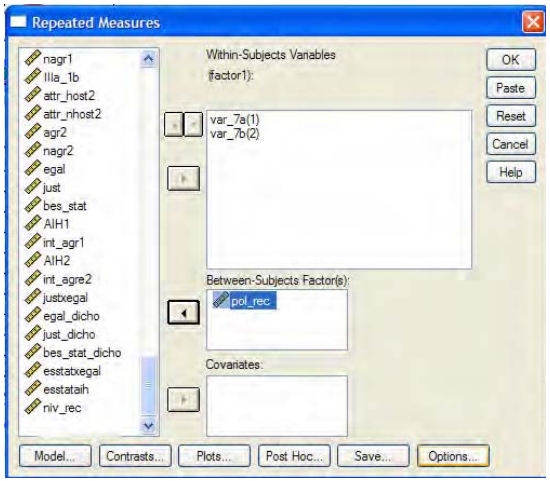
Par exemple, on pose deux questions aux participants sur leur opinion par rapport à la Suisse :
1) « Je suis fier/ère de la Suisses dans le domaine de sa réussite économique » (nom de la variable *var7a*, variable dépendante numérique, les individus ont répondu sur une échelle qui va de 1 = tout à fait d'accord à 5 = pas du tout d'accord
2) « Je suis fie/ère de la Suisse dans le domaine de ses réussites dans les arts et la littérature » (nom de la variable *var7b*, variable dépendante numérique, les individus ont répondu sur une échelle qui va de 1 = tout à fait d'accord à 5 = pas du tout d'accord ; c'est important que les échelles des VD soient les mêmes).
On se demande si la position politique (gauche ou droite, variable indépendante nominale, catégorie 1 = gauche et catégorie 2 = droite) a un impact sur les réponses à ces deux questions.
Les hypothèses peuvent être de différente nature, mais nous allons en proposer seulement un exemple. En premier lieu, il est possible de faire l'hypothèse d'une différence entre l'accord aux deux variables dépendantes.
Hypothèse 1a (H1a): les participants sont plus fiers de la Suisse dans le domaine de la réussite économique que dans le domaine des arts et de la littérature.
La deuxième hypothèse porte sur l'effet de la variable indépendante sur cette différence.
Hypothèse 1b (H1b) : les participants qu'indiquent être de gauche sont plus fiers de la Suisse dans le domaine de l'art et de la littérature que dans le domaine de sa réussite économique, alors que les participants qui indiquent être de droite sont plus fiers de la Suisse dans le domaine de sa réussite économique que dans le domaine des arts et de la littérature.

Pour tester l'hypothèse 1a, il faut créer ce que SPSS appelle un facteur, plus précisément un facteur intra-individuel. L'appellation *intra-individuel* signifie que ce sont les mêmes participants qui ont répondu aux VD et que l'analyse compare les participants avec eux-mêmes par rapport à leurs réponses à des variables différentes. C'est-à-dire que l'analyse compare la réponse du participant x à la question *var7a* à la réponse du même participant x à la question *var7b*. C'est pourquoi l'analyse crée une variable indépendante additionnelle qui s'appelle facteur intra-individuel. Nos VI sont: la position politique (gauche ou droite) et la position des participants sur les deux VD (*var7a* et *var7b*) qui est une mesure répétée puisque chaque participant a une valeur pour *var7a* et *var7b*.

Chemin: **Analyze → General Linear Model → Repeated Measures...**



Il faut d'abord définir un facteur, c'est-à-dire la combinaison des deux VD numériques (mesure répétée), en lui donnant un nom (on peut laisser le nom par défaut, **Factor 1**), puis définir le nombre de VD numériques qui constituent le facteur (2 ou plus, 2 dans l'exemple). Ensuite, cliquez **Add** afin que le facteur apparaisse dans la fenêtre en bas, puis **Define**. Une nouvelle fenêtre apparaît :



Entrez les deux ou plus VD numériques dans la fenêtre **Within-Subjects Variables**, puis la variable indépendante nominale dans **Between-Subjects Factor(S).** Cliquez sur **Options** pour demander les **Descriptive statistics** (les moyennes et les écarts-types), cliquez sur **Continue** puis **OK**.

SPSS produit plusieurs tableaux, mais seulement ceux qui sont commentés nous intéressent.

SPSS output :

Test of Within-Subjects Contrasts
Meausure : MEASURE_1

Source	Factor1	Type III Sum of Square	df	Mean Square	F	Sig.
Factor 1	Linear	3.675	1	3.675	8.118	.008
Factor 1 * G_D	Linear	1.008	1	1.008	2.227	.147
Error (Factor 1)	Linear	12.675	28	.453		

Tests of Between-Subjects Effects
Measure : MEASURE_1
Transformed Variable : Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	364.008	1	364.008	470.230	.000
G_D	8.333E-03	1	8.333E-03	.011	.918
Error	21.675	28	.774		

Descriptive Statistics

	Pol_rec	Mean	Std. Deviation	N
var7.a	1.00	2.50	.889	20
	2.00	2.20	.422	10
	Total	2.40	.770	30
vr7.b	1.00	2.75	.851	20
	2.00	3.00	.667	10
	Total	2.83	.791	30

Dans le premier tableau (**Test of Within-Subjects Contrasts**), spss présente les tests de la différence entre les deux V et entre les deux VD dans les deux groupes de participants, de gauche ou de droite. Dans le deuxième tableau (**Test of Between-Subjects Contrasts)**, est résumé le test de la différence dans les deux groupes de participants de la variable nominale si les deux VD sont agglomérées pour en constituer une seule (c'est la moyenne des réponses aux deux VD). Ce serait l'équivalent d'une nouvelle variable qui mesure la fierté des participants envers la Suisse de manière plus générale. Dans le troisième tableau (**Descriptive Statistics**) se trouvent résumées les moyennes des deux VD selon le groupe des participants.

Si on regarde le deuxième tableau, on peut remarquer que si l'on agglomère les variables *var7a* et *var7b*, il n'y pas de différence entre les participants qui indiquent être à gauche et les participants qui indiquent être à droite : $F(1, 28) = .011$, *ns*. Ce qui nous intéresse le plus dans

ce cas est le premier tableau qui met en évidence le fait que, oui, il y a une différence significative entre la manière de répondre à la *var7a* et à la *var7b* (les VD), puisque $F(1, 28) = 8.118$, $p < .01$. Le résultat de ce test se trouve sur la ligne **factor1**. Les moyennes présentées dans le troisième tableau montrent les participants sont globalement plus fiers de la Suisse par dans ses avancées en matière d'économie ($M = 2.40$; $SD = .77$) qu'en matière d'arts et culture ($M = 2.83$; $SD = .79$). Faites attention à l'échelle avec laquelle vous avez mesuré les variables.

En ce qui concerne l'effet d'interaction entre le facteur intra-individuel (les VD) et la variable indépendante (ce qui revient à dire : l'effet de la variable indépendante sur les différences entre les réponses aux deux variables dépendantes), ou encore la manière de répondre à ces deux variables dans les deux groupes de participants, il n'y a pas de différence significative, puisque : $F(1, 28) = 2.23$, *ns*. Ce qui signifie que, indépendamment du groupe des participants (gauche ou droite), tous les participants interrogés sont plus fiers des réussites de la Suisse dans le domaine économique que dans le domaine des arts et de la littérature.
Vous pouvez résumer vos résultats à l'aide d'un tableau comprenant les moyennes et écarts-types ou d'un graphique. Ainsi, H1a a été « confirmée », alors que cela n'a pas été le cas d'H1b.

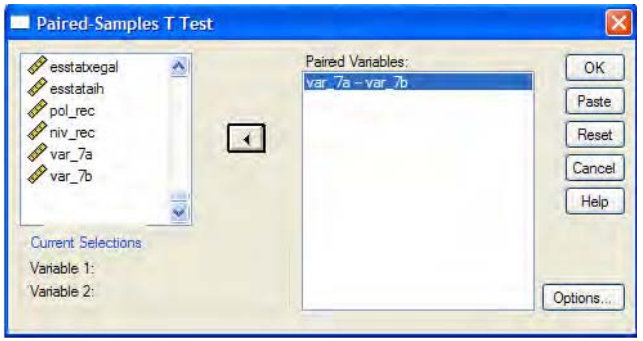
6.3.6. Test statistique de la différence entre deux moyennes provenant des mêmes participants : T-test avec 2 variables numériques à mesures répétées

Dans cette analyse, il est question de deux variables dépendantes (VD) numériques, sans aucune variable indépendante.

But: tester si la différence entre les réponses des mêmes participants à des variables différentes est statistiquement significative.

Nous prenons, par exemple, les réponses des participants aux questions *var7a* et *var7b* (voir chapitre 6.3.5 pour leur définition). Mais nous voulons seulement savoir si les participants ont répondus différemment à ces deux questions. C'est-à-dire que la seule hypothèse (H1a) qui vous intéresse est celle qui dit (toujours selon l'exemple) : les individus sont plus fiers des réussites de la Suisse dans le domaine de l'économie que dans le domaine de l'art et de la littérature.

Chemin : **Analyze → Compare Means →Paired-Sample T-Test...**



Cliquez sur la première des deux variables numériques (*var_7a*, par exemple), cliquez sur la seconde variable numérique (*var_7b*). Les introduire toutes deux dans la fenêtre **Paired Variables:** puis cliquez sur **OK**.

SPSS output:

T-test
Paired Samples Statistics

		Mean	N	Std. Deviation	Std Error Mean
Pair1	var7.a	2.42	33	.792	.138
	var7.b	2.82	33	.769	.134

Paired Sample Correlations

		N	Correlation	Sig.
Pair 1	var7.a & var7.b	33	.233	.191

Paired Sample T-test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	Lower Bound	Upper Bound			
Pair	Var7.a – var7.b	-.39	.966	.168	-.74	-.50	-2.342	32	.026

Toujours dans la même logique, vous devez avant tout vous intéresser à la valeur du test de vos variables (dans ce cas : le test de la différence entre les deux moyennes), donc au tableau **Paired Samples T-test**. La valeur du *t* est de -2.342, et la valeur du *p* associée est de .026 (écriture $t(32) = -2.34$, $p < .05$). Ce qui signifie que le test est significatif, et donc que la différence entre les deux VD est significative. A ce moment, vous allez regarder les moyennes dans le tableau pour savoir dans quel sens va la différence. Vous les trouvez dans le tableau **Paired Samples Statistics**. On voit que les sujets sont dans l'ensemble plus fiers de la Suisse dans le domaine économique ($M = 2.42$; $SD = .79$) que dans la littérature et dans l'art ($M = 2.82$; $SD = .77$).

6.4. Vérifier la fiabilité interne d'une échelle : alpha de Cronbach

But : lorsqu'on dispose de plusieurs questions dont on pense qu'elles sous-tendent la même idée, il est nécessaire de vérifier qu'elles peuvent être résumées en une seule variable. Sans cela, il serait erroné de les associer pour constituer une seule et même mesure. Généralement, cette opération porte sur des variables numériques. C'est-à-dire que, normalement, la cohérence interne mesure si un ensemble de variables « vont bien ensemble » et sous-tendent le même concept. Par exemple, il paraît qu'en Suisse le salaire gagné soit un sujet tabou. C'est pourquoi à la question « quel est votre salaire ? » il peut y avoir un taux élevé de personnes qui ne répondent pas. Pour dépasser ce problème, les chercheurs peuvent créer une série de

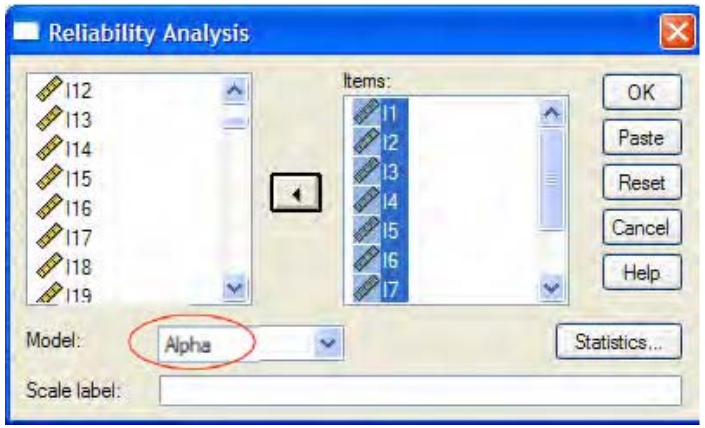
questions qui donnent des indications sur le revenu des personnes sans poser la question directement. Par exemple, ils peuvent demander aux participants d'indiquer le nombre de pièces dans lesquelles ils vivent, le nombre et le type des voitures du foyer, le budget pour les vacances. Ces mesures peuvent être « agrégées » pour en créer une unique variable qui va s'appeler « niveau de vie ». Mais, est-ce que le nombre et le type des voitures, le nombre de pièces et le budget pour les vacances sont bien des sous-dimensions qui indiquent le niveau de vie des participants?
L'alpha de Cronbach permet de répondre à cette question et d'évaluer dans quelle mesure ces trois variables vont « bien ensemble » et peuvent en constituer une seule.

La valeur de l'alpha varie entre 0 et 1 et on considère cet indice comme bon dès qu'il est de .80 et plus. Entre .60 et .80, il est satisfaisant. En deçà il devient risqué d'utiliser les variables pour en former une unique.

Exemple 1 :

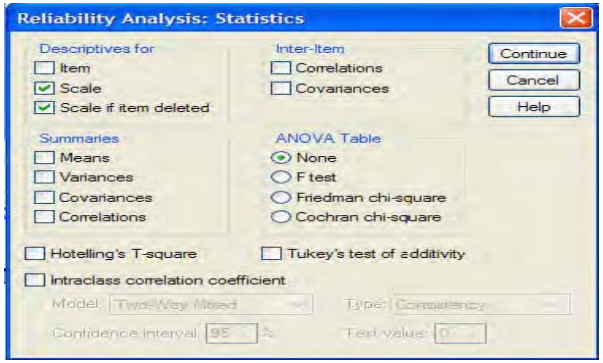
Dans un questionnaire, 7 variables sont censées mesurer l'attitude des participants envers une vision égalitaire de la société. Pour les participants en question, ces 7 variables vont-elles bien ensemble ? Est-ce que les participants les perçoivent comme relevant de la même sous-dimension où ils les considèrent comme indépendantes les unes des autres ?

Chemin : **Analyze → Scale → Reliability analysis...**



Sélectionnez dans la fenêtre de gauche les variables qui sont censées « aller ensemble » (mesurer la même sous-dimension) et faites-les glisser dans la fenêtre **Items** grâce à la flèche entre les deux fenêtres. Dans **Model**, le test alpha est sélectionné par défaut. Ne changez rien.

Cliquez sur l'onglet **Statistics….** Une nouvelle fenêtre s'ouvre :



Dans la rubrique **Descriptives for**, cochez les cases **Scale** (celle-ci vous donne l'indication de la valeur de l'alpha pour l'ensembles des variables que vous avez sélectionné) et **Scale if item deleted** (celle-ci vous donne l'indication de la valeur de l'alpha si vous enlevez une variables particulière de sa composition). Le fait de cocher cette dernière case vous permet d'évaluer la contribution de chaque variable à la mesure globale et de pouvoir éliminer, le cas échant, celle ou celles qui réduisent la valeur de l'alpha de la mesure globale.
Cliquez sur **Continue** puis **OK**.

SPSS output :

Reliability Statistics

Cronbach's Alpha	N of Items
.805	7

Ce premier tableau vous indique que l'alpha global des 7 variables est de .805, ce qui représente un bon alpha. On peut donc créer une variable unique en calculant la moyenne des 7 variables. Sur Mac et sur des versions plus récentes de SPSS, la configuration des tableaux est un peu différente et vous trouvez la valeur de l'alpha tout au fond de l'analyse plutôt qu'au début, mais la description est la même! Puisque la valeur de l'alpha est très satisfaisante, vous pouvez vous arrêter là et créer votre indice à partir des 7 variables. Pour l'exercice, nous considérons également si la valeur de l'alpha peut être augmentée (améliorée) en ne prenant pas en compte certaines variables.

Item-Total Statistics

	Scale Mean If Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha If Item Deleted
I1	45.3250	46.804	.669	.754
I2	45.3125	48.091	.643	.759
I3	44.1000	57.635	.474	.793
I4	44.6375	49.981	.712	.752
I5	44.1875	58.619	.433	.798
I6	45.6125	45.582	.568	.778
I7	45.7500	51.025	.397	.813

Ce second tableau indique que la variable qui contribue le moins à la mesure globale est la *I7*. Si elle n'avait pas été entrée dans le test, l'alpha aurait été de .813. Ainsi si notre alpha n'avait pas été bon, nous aurions pu ne pas inclure la variable *I7* pour le faire remonter. Ce qui veut dire que lorsqu'on créera la mesure globale on effectuera la moyenne sans la variable I7. Important: lorsque vous travaillez sur des échelles qui comportent plusieurs variables et qui ont déjà été validées par d'autres études, vous n'avez pas d'intérêt à enlever des variables de la mesure globale juste pour « remonter » la valeur de l'alpha. Il se peut que le fait d'enlever une variable modifie la signification du tout! Dans ces cas, une fois que la cohérence interne est satisfaisante (et si c'est une bonne échelle, cela est le cas), créez la mesure globale en comprenant toutes les variables. Autrement... posez-vous la question de la validité de l'échelle elle-même !

ATTENTION : lorsqu'on calcule un alpha, il faut que les échelles de mesure de toutes les variables aillent dans le même sens! Si votre alpha comporte des valeurs négatives, cela signifie qu'il y a des variables qui s'opposent à d'autres dans votre analyse et il faudra probablement inverser certaines d'entre elles à travers un recodage (voir chapitre 4.2).

6.5. Analyse en Composantes Principales Exploratoire (ACP)

Description : méthode d'analyse exploratoire (qui ne teste pas des hypothèses) permettant d'organiser et de synthétiser un ensemble de variables numériques en quelques dimensions (les agréger pour avoir moins de variables à étudier). L'analyse montre quelles variables décrivent et mesurent une même dimension (appelée facteur par l'analyse).

Fonction : détecter des dimensions (par exemple, lorsque vous avez créé une échelle qui mesure notion par plusieurs variables et vous vous demandez si les variables s'organisent comment vous pensez) ou étudier la fiabilité d'un modèle existant (par exemple, lorsque vous utilisez une échelle qui mesure plusieurs notions à l'aide de plusieurs variables et vous voulez contrôler que les variables de l'échelle mesurent bien des choses différentes). Attention : dans ce dernier cas rappelez-vous que vous ne testez pas un modèle, mais vous l'étudiez de manière exploratoire.

Fondement : le calcul se base sur la matrice des corrélations entre toutes les variables prises en compte dans l'analyse.

Condition nécessaire pour mener l'analyse: le nombre de participants doit être au moins 5 fois supérieur au nombre de variables inclues dans l'analyse et inclure au moins 100 participants.

Normalement, l'analyse factorielle exploratoire ne teste pas des différences entre les groupes de participants, mais s'intéresse à l'organisation des réponses de la totalité des participants.

Concepts clés de l'interprétation des résultats:
1) La relation entre une variable et un facteur est exprimée par un indice de saturation allant de -1 à 1. Les variables qui saturent fortement sur un facteur (saturation proche de 1 ou de -1) sont celles qui « résument » (représentent) le mieux ce facteur. Une saturation négative indique que la variable en question s'oppose aux autres variables qui saturent fortement sur le même facteur, mais avec une saturation de signe opposé (la signification d'un facteur est à interpréter intuitivement ou à l'aide d'une théorie). Normalement, ce qui vous intéresse ce n'est pas la variable individuelle qui sature sur un facteur, mais plutôt l'ensemble de variables qui saturent sur ce facteur; l'interprétation porte sur l'ensemble des variables qui saturent sur un facteur. Le facteur est une sous-dimension qui sous-tend les réponses des participants à ces variables et qui organise leurs réponses. Comme dans l'exemple du salaire, le niveau de vie est la sous-dimension qui sous-tend la réponse à ces variables.
2) Il y a trois types de facteurs : 1) facteur général : toutes les variables entrées dans l'analyse saturent sur ce facteur et la valeur des saturations est de même signe, 2) facteur unipolaire : toutes les variables qui saturent sur le facteur sont positives ou négatives, ou 3) facteur bipolaire : une partie des variables qui saturent sur le facteur présente une saturation positive, alors qu'une autre partie présente une saturation négative. Attention : la valeur de la saturation (positive ou négative) peut dépendre de l'échelle de mesure des variables !!!! Il est toujours préférable que les échelles de mesure des variables « aillent dans le même sens », cela simplifie l'interprétation des facteurs.
3) Communalité (**communality**): la variance relative à une variable qui est expliquée par les facteurs retenus et qui est de 1 au maximum (l'ensemble des facteurs mis en évidence par l'analyse expliquent le 100% de la variabilité des réponses des participants). Il faut qu'elle ne soient pas trop basses (< .5). Si c'est le cas, cela signifie que votre analyse explique très peu de cette variable.
4) Valeur-propre (**Eigenvalue**): la variance totale expliquée par le facteur. Si cette valeur est inférieure à 1, cela signifie que le facteur explique moins qu'une seule variable; il est donc sans intérêt.

Options : SPSS vous permet de limiter le nombre des facteurs que vous désirez retenir. Si vous ne limitez pas le nombre de facteurs, SPSS utilise un critère statistique comme la règle de Kaiser, pour ne pas retenir un nombre trop élevé de facteurs. La règle de Kaiser retient tous les facteurs qui ont une valeur propre supérieure à 1 (eigenvalue).

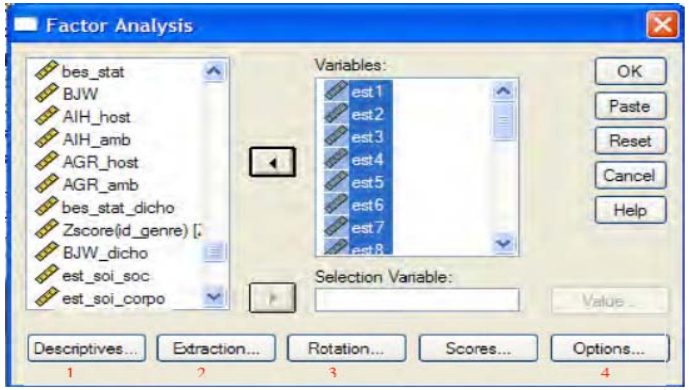
Logique des facteurs : le logiciel organise les variables retenues pour l'analyse en facteurs et indique le pourcentage de variance des réponses des participants qui est expliquée par chacun de ces facteurs. Le premier facteur sortant de l'analyse est celui qui explique le plus de variance dans les réponses des participants et le dernier est celui qui en explique le moins.

Pour faciliter l'interprétation des facteurs, SPSS vous donne la possibilité d'opérer une rotation sur vos résultats. Une rotation maximise les saturations de certaines variables sur les facteurs. Il existe différents types de rotation : 1) rotation orthogonale (par exemple VARIMAX) qui produit des facteurs indépendants les uns des autres (pas liés entre eux) et qui est la méthode la plus utilisée, 2) rotation oblique (par exemple OBLIMIN) qui produit des facteurs dépendants les uns des autres (liés entre eux).

Exemple 1:

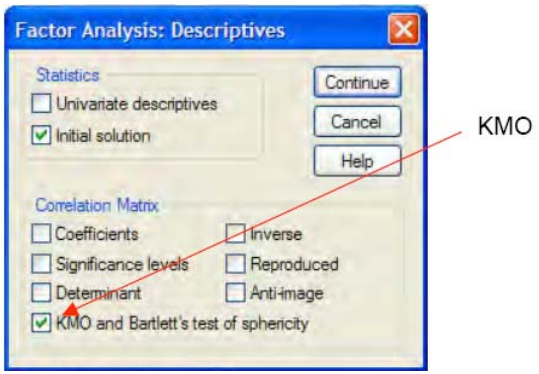
Nous avons utilisé une échelle d'estime soi qui est composée de 20 variables et se composant théoriquement 3 sous-dimension: 1) une sous-dimension « estime de soi sociale », 2) une sous-dimension « estime de soi performance » et 3) une sous-dimension « estime de soi physique » (échelle 1 = pas du tout, 6 = tout à fait). Est-ce que nous retrouvons ces trois dimensions dans les réponses de notre échantillon de participants?

Chemin : **Analyze → Data reduction → Factor…**



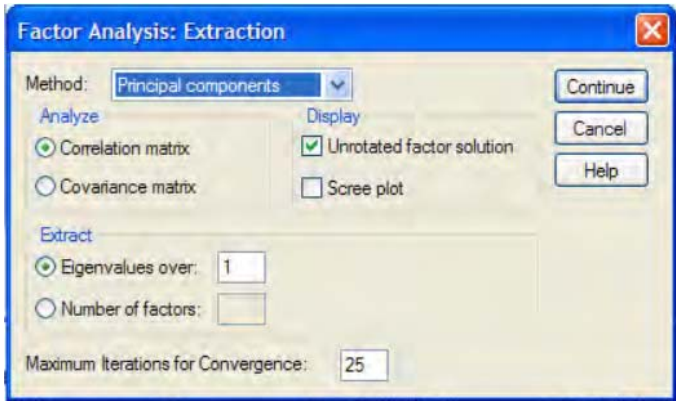
Sélectionnez dans la fenêtre de gauche toutes les variables de l'échelle (pour nous : de est1 à est20) et faites-les passer dans la fenêtre **Variable** à l'aide de la flèche du milieu.

Dans **Descriptives…**, sélectionnez la case **KMO.** Cliquez sur **Continue**.



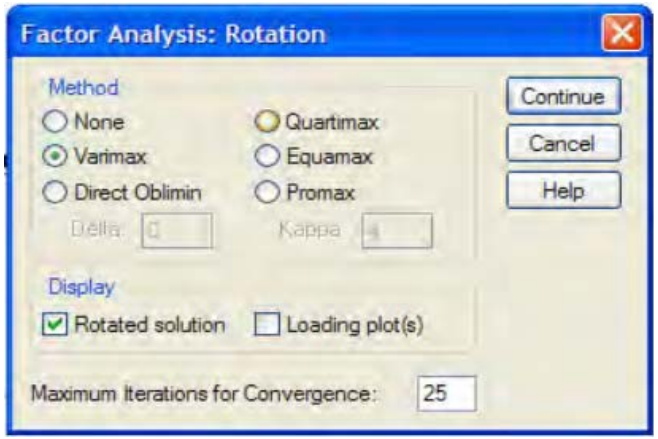
Le KMO vous indique si les résultats de l'ACP sont fiables. Il doit être supérieur à .600 pour que ce soit le cas, autrement cela signifie qu'il n'y a pas de réelle organisation dans les réponses des participants.

Dans **Extraction…**



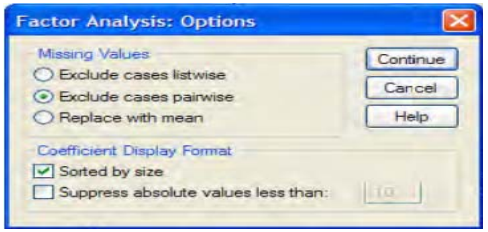
Dans l'onglet **Extract** : on peut choisir un nombre de facteur que l'analyse doit retenir en utilisant l'option **Number of factors** si l'on a une hypothèse/une idée de nombre de sous-dimensions. Si ce n'est pas le cas, on maintient le critère de sélection par défaut (Eigenvalue > 1). Dans l'exemple, vous pouvez sélectionner trois facteurs, puisque l'échelle utilisée comporte trois sous-dimensions. Laissez les autres cases enchangées. Cliquez sur **Continue**.

Dans **Rotation…**



Dans l'onglet **Method**, cochez **Varimax**. Cliquez sur **Continue**.

Dans **Options…**



Sélectionnez **Missing Values : exclude cases pairwise** (pour déterminer le traitement des données manquantes). Dans **Coefficient Display Format**, cochez **Sorted by size** : les variables seront organisées selon la valeur de leurs saturations sur chaque facteur en ordre décroissant, ce qui en facilite l'interprétation. Cliquez sur **Continue**, puis sur **OK**.

SPSS output:

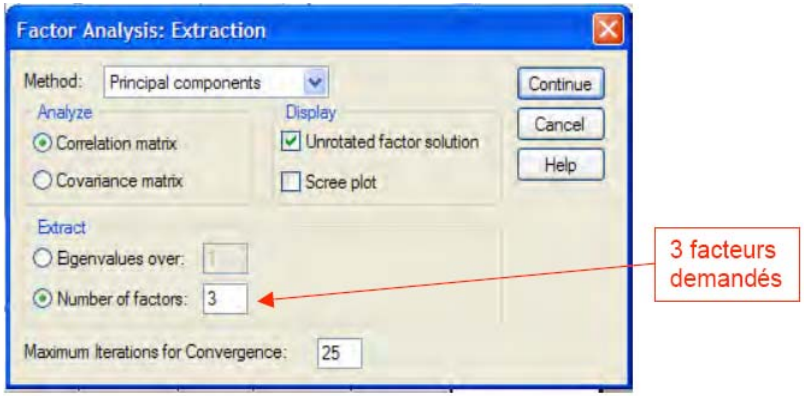
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.764
Barttlett's Test of Sphericity	Approx. Chi-Square	1006.168
	Df	190
	Sig.	.000

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.892	29.462	29.462	5.892	29.462	29.462	3.221	16.107	16.107
2	2.656	13.281	42.743	2.656	13.281	42.743	2.942	14.709	30.816
3	1.718	8.588	51.331	1.718	8.588	51.331	2.681	13.405	44.220
4	1.399	6.996	58.327	1.399	6.996	58.327	1.978	9.892	54.112
5	1.020	5.098	63.425	1.020	5.098	63.425	1.863	9.313	63.425
6	.946	4.728	68.153						
7	.842	4.211	72.364						
8	.820	4.100	76.464						
9	.710	3.550	80.014						
10	.672	3.359	83.373						
11	.626	3.131	86.504						
12	.547	2.737	89.241						
13	.417	2.084	91.325						
14	.393	1.964	93.289						
15	.339	1.697	94.986						
16	.307	1.536	96.522						
17	.232	1.158	97.680						
18	.211	1.057	98.737						
19	.130	.651	98.387						
20	.123	.613	100.00						

Le premier tableau vous indique la valeur du KMO. Ici, KMO =.76, l'analyse est donc fiable. Le tableau **Total Variance Explained**, montre que l'analyse retient 5 facteurs (si on n'introduit pas des limitations définies à priori et sur la base de la valeur propre). La première série de trois colonnes (de gauche à droite) met en évidence tous les facteurs qui sont sortis de l'analyse. Leur valeur propre se trouve tout à gauche (**Total**), suivie par le pourcentage de variance dans les réponses des participants qui est expliquée par chaque facteur (**% of variance**) et par la variance cumulée à chaque fois qu'un facteur s'ajoute aux précédents (**cumulative %**). La série de trois colonnes du milieu contient les mêmes informations, limitées aux facteurs retenus (ici, selon leur valeur propre supérieure à 1). Attention: ces valeurs font référence aux facteurs AVANT ROTATION. Si vous avez appliqué une rotation à vos données et que vous allez prendre en compte les résultats qui dérivent de celle-ci, vous devez vous intéresser à la série de trois colonnes de droite. Remarquez que chaque facteur explique de moins en moins de variance, mais complique de plus en plus l'interprétation des résultats. Il faut que vous tranchiez entre une interprétation qui résume le plus de variance possible et le fait de ne pas compliquer le modèle en incluant trop de facteur. Dans l'exemple, les 5 facteurs expliquent 63.42% de la variance des réponses des participants. Pour réduire le nombre de facteurs, nous pouvons refaire l'analyse et la forcer à ne mettre en évidence que trois facteurs, puisque théoriquement l'échelle comporte 3 sous-dimensions (il faut toujours faire référence à la théorie dans ces cas):



SPSS output:

Le premier tableau est strictement le même.

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.764
Barttlett's Test of Sphericity	Approx. Chi-Square	1006.168
	df	190
	Sig.	.000

Le tableau qui résume les informations concernant les facteurs est semblable à celui de la première analyse, mais il y a quelque différence :

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.892	29.462	29.462	5.892	29.462	29.462	3.699	18.495	18.495
2	2.656	13.281	42.743	2.656	13.281	42.743	3.470	17.350	35.845
3	1.718	8.588	51.331	1.718	8.588	51.331	3.097	15.486	51.331
4	1.399	6.996	58.327						
5	1.020	5.098	63.425						
6	.946	4.728	68.153						
7	.842	4.211	72.364						
8	.820	4.100	76.464						
9	.710	3.550	80.014						
10	.672	3.359	83.373						
11	.626	3.131	86.504						
12	.547	2.737	89.241						
13	.417	2.084	91.325						
14	.393	1.964	93.289						
15	.339	1.697	94.986						
16	.307	1.536	96.522						
17	.232	1.158	97.680						
18	.211	1.057	98.737						
19	.130	.651	98.387						
20	.123	.613	100.00						

Dans ce nouveau tableau, on retrouve bien 3 facteurs comme demandé. Ces trois facteurs expliquent à eux trois, 51.33% de la variance, ce qui est moins que lors de l'analyse précédente qui avait retenu 5 facteurs. Ainsi, en contraignant l'ACP à limiter la solution à 3 facteurs, on perd de l'information. Le précédent modèle était plus complexe, mais celui-ci explique moins de variance. Le choix d'un modèle plutôt qu'un autre se justifie sur la base de la théorie.

Le tableau montre que, après rotation VARIMAX, le premier facteur explique 18.5% de la variance, le second explique 17.3% et le dernier 15.5%.

Pour ce qui est de l'interprétation des facteurs, l'analyse produit d'autres tableaux.

Component Matrix(a)

	Component		
	1	2	3
est15	.776	.047	.152
est11	-.730	.165	.053
est12	-.703	.387	.368
est4	.699	.056	.267
est1	-.697	.270	-.011
est10	.602	.270	.181
est3	-.600	.326	.506
est19	.579	.183	.127
est16	.564	-.315	-.314
est20	.535	.364	.097
est14	-.511	.142	-.446
est8	.487	.233	.020
est6	-.322	.236	-.300
est17	.354	.776	-.251
est13	.395	.731	-.232
est2	.390	.650	-.018
est9	-.341	.367	-.039
est7	.425	-.238	-.569
est5	.508	-.216	.518
est18	.187	-.165	.217

Extraction Method : Principal Component Analysis.
a 3 components extracted

Ce tableau (**Component Matrix(a)**) donne le détail des variables qui composent chaque facteur avant la rotation. En général, si vous avez opéré une rotation vous ne consultez pas ce premier tableau, mais faites attentions de ne pas l'interpréter à la place de celui qui suit!

Rotated Component Matrix(a)

	Component		
	1	2	3
est17	.851	-.021	-.257
est13	.836	-.054	-.205
est2	.757	.050	-.017
est20	.599	-.096	.247
est10	.555	-.126	.380
est19	.473	-.187	.355
est8	.468	-.177	.205
est12	-.100	.844	-.238
est3	-.089	.842	-.055
est7	.074	-.736	-.119
est16	.077	-.692	.178
est11	-.282	.555	-.418
est1	-.174	.543	-.484
est9	.110	.352	-.342
est5	.086	-.076	.749
est14	-.152	.090	-.670
est4	.430	-.228	.571
est15	.472	-.354	.528
est6	.026	.112	-.486
est18	-.042	-.015	.365

Ce tableau (**rotated component matrix(a)**) est à lire comme indiqué sur celui qui suit (que nous avons remanié).

	Component		
	1	2	3
est17	.851		
est13	.836		
est2	.757		
est20	.599		
est10	.555		
est19	.473		
est8	.468		
est12		.844	
est3		.842	
est7		-.736	
est16		-.692	
est11		.555	
est1		.543	
est9		.352	
est5			.749
est14			-.670
est4			.571
est15			.528
est6			-.486
est18			.365

C'est-à-dire que les variables qui saturent le plus sur le premier facteur sont: *est17*, *est13*, *est2*, *est20*, *est10*, *est19* et *est8*. La variable *est12* sature davantage sur le 2ème que sur le 1er facteur, cela veut dire que l'on passe à la constitution du 2ème facteur. Les variables pour ce second facteur sont : *est12*, *est3*, *est7*, *est16*, *est11*, *est1* et *est9*. La variable *est5* sature davantage sur le 3ème facteur que sur le 2nd, donc on passe à la constitution du 3ème facteur composé des variables : *est5*, *est14*, *est4*, *est15*, *est6*, *est18*.

On peut voir que, pour les facteurs 2 et 3, les saturations des items sont positives et négatives : ce sont des facteurs bipolaires. Cela signifie par exemple que les variables *est7* et *est16* s'opposent aux variables *est12*, *est3*, *est11*, *est1* et *est9*. Dans cet exemple, l'analyse était menée sur une échelle qui fait référence à un modèle théorique, et il ne nous reste plus qu'à évaluer la concordance entre nos résultats et l'échelle originale. S'il n'y avait pas eu de modèle préalable (les trois sous-dimensions de l'échelle), il faudrait chercher à quoi se rapportent nos facteurs et leur donner un nom. Il n'y a pas de règle pour faire cela, le nom que vous donnez au facteur (qui fait référence à la sous-dimension que vous supposez qu'il mesure) dépend de votre théorie et du contenu (signification) des variables qui le constituent.

Remarques

1) Souvent, il est plus clair de présenter l'analyse en composantes principales sous forme de tableau (du même type que le tableau **rotated component matrix**)
2) Si possible, donnez la formulation complète des variables dans le tableau (ou au moins les tables), pour faciliter l'interprétation des facteurs

6.6. Analyse de régression linéaire

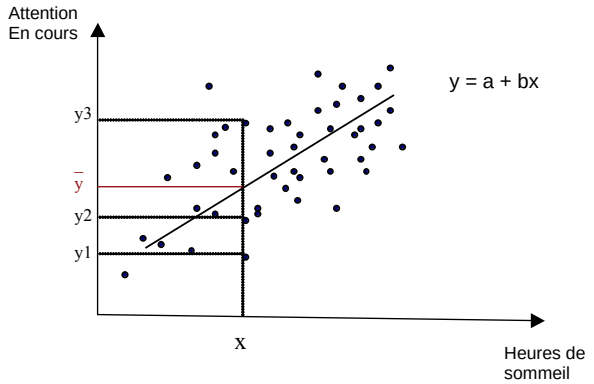
Description : l'analyse de régression linéaire fonctionne sur le même principe que l'analyse des corrélations. C'est-à-dire que cette analyse permet de quantifier le type (positif ou négatif) et la force du lien (valeur du *r*) existant entre deux ou plusieurs variables. La différence principale entre l'analyse de régression et l'analyse des corrélations est que, dans le cas de l'analyse de régression, il est question d'une variable dépendante (la VD, notée Y dans l'équation de régression) et de plusieurs variables indépendantes (les VI, notées Xi dans l'équation de régression). Ainsi, l'analyse de régression teste l'impact d'une ou plusieurs de VI sur une VD (on introduit donc l'idée de causalité).

Principe : à partir du nuage de points (voir figure 1) obtenu sur la base des corrélations entre les variables incluses dans l'analyse, SPSS définit une droite résumant le mieux possible ce nuage. Cette droite est appelée droite de régression. Par « résumant le mieux », on entend que la somme des distances, au carré, entre chaque point sur le graphique et la droite est la plus petite possible. Par exemple, si on étudie dans quelle mesure le nombre d'heures de sommeil en dehors des cours (VD ou X) prédit l'attention des étudiants en cours (VD ou Y), chaque étudiant interrogé sera représenté par un point sur un graphique regroupant ces deux informations (nombre d'heures de sommeil en dehors des cours et attention en cours). L'hypothèse H1 est alors que plus le nombre d'heures de sommeil en dehors des cours augmente, plus l'attention en cours augmente. La nouveauté de l'analyse de régression par rapport à l'analyse des corrélations est que la première permet aussi de tester le fait que ce sont les heures de sommeil en dehors des cours qui causent l'augmentation de l'attention en cours, et non pas l'inverse. Nous faisons donc l'hypothèse d'un lien linéaire positif entre ces deux variables, lien linéaire qui s'exprime par une droite de régression résumée par l'équation Y = a + b*X (a est une valeur constant calculée par l'analyse, c'est l'intercept).

Voir figure 1 pour la droite de régression). Ainsi, en termes de corrélations, cela signifierait qu'il y a une corrélation positive entre les deux variables. L'analyse de régression permet en addition de tester la causalité (quelle variable a un effet sur quelle variable), ce qui n'est pas le cas de l'analyse des corrélations.

Figure 1.

Représentation graphique du nuage de points et de la droite de régression dans le cas d'une variable X (VI) qui prédit une variable Y (VD)



Comme le montre la figure ci-dessous, à un nombre d'heures de sommeil en dehors des cours (qui représentent l'axe des x) correspondent des points qui n'ont pas la même ordonnée à l'origine, c'est-à-dire des valeurs y qui représentent l'attention en cours (y_1, y_2 et y_3). Par exemple : plusieurs étudiants dorment le même nombre d'heure en dehors des cours, mais ils n'ont pas le même niveau d'attention en cours. De même, à un nombre d'heures en dehors des cours x correspond une valeur d'attention sur la droite de régression ($y = a + b*X_i$).
Le rôle de SPSS sera de trouver la droite permettant de minimiser l'écart entre les trois points y_1, y_2 et y_3 et la droite elle-même. Cela signifie que la somme des différences y_1-y, y_2-y et y_3-y au carré est la plus petite possible. Dans cet exemple, nous avons utilisé uniquement deux variables qui définissent un espace à deux dimensions. L'analyse de régression s'intéresse à l'étude de n variables X (ou VI) qui prédisent la variable Y (ou VD) et qui définissent un espace à n + 1 dimensions, mais le principe présenté dans la figure 1 reste le même.

Concepts clés de l'interprétation des résultats:
1) Pour chaque VI, le logiciel calcule un coefficient β qui correspond à la pente de la droite de régression. Par exemple, si vous avez 3 VI, c'est-à-dire X_1, X_2 et X_3, l'équation de la droite de régression sera $Y = a + \beta_1*X_1 + \beta_2*X_2 + \beta_3*X_3$. β_1 est un indice de la force de la relation entre X_1 (une des VI introduites dans l'équation de régression) et Y (la VD), une fois que le lien entre Y (la VD) et les autres X_i (dans ce cas X_2 et X_3, les autres VI introduites dans l'équation de régression) est maintenu constant. Lorsque le coefficient β est standardisé, il varie entre -1 et +1, il est noté B et appelé Bêta. Les valeurs des coefficients non-standardisés s'expriment dans les unités originales dans lesquelles les variables ont été mesurées, alors que les valeurs des coefficients standardisés ne dépendent pas des unités de

mesure. Pour pouvoir comparer directement l'impact respectif des différentes VI sur la VD, nous vous proposons donc de considérer les coefficients standardisés Bêta
2) À chaque β (ou Bêta, si standardisé) correspond un *t* (qui correspond au *t* de student du chapitre 6.3.1) et qui indique si le X_i (ou VI) auquel il est associé prédit significativement Y (ou VD). Plus précisément, la probabilité *p* associée au *t* vous dit si le Bêta est significatif et donc si la VI (X) prédit la VD (Y). La valeur du Bêta vous donne alors une idée de la force du lien entre la VI et la VD. Le principe est le même que dans l'interprétation des corrélations (voir chapitre 6.2)
3) Une autre mesure est importante : l'indice associé à la puissance de l'analyse elle-même. Il s'agit du R^2 (appelé aussi **R square** par SPSS). Cet indice donne une indication du pourcentage de variance totale de la VD expliquée par la ou les X_i introduites dans la droite de régression (l'analyse). La valeur du R^2 dépend du nombre de participants inclus dans l'étude. Ainsi, il est préférable de prendre en compte le R^2 ajusté (appelé **adjusted R square** par SPSS). Plus la valeur du R^2 ajusté est élevée, plus les X_i sont pertinentes pour expliquer Y (la VD)
4) À la valeur du R^2 ajusté est associé un *F* (identique au *F* de Fischer du chapitre 6.3.2) qui indique si le R^2 ajusté, c'est-à-dire la proportion de variance expliquée par l'analyse, est significative. En d'autres termes, si la variance de Y qui est expliquée par l'analyse ne peut être pas être attribuée au hasard, mais aux X_i

Type de variables prises en compte: s'agissant du même principe que les corrélations, l'analyse de régression peut être appliquée uniquement à des variables numériques. Cependant, des variations sont possibles. Par exemple, moyennant certaines transformations, des variables nominales peuvent être introduites dans la droite de régression, mais les analyses spécifiques qui prennent en compte cette possibilité dépassent le but de ce fascicule.

Pour résumer :
1) L'analyse de régression porte sur une VD (Y) prédite par une ou plusieurs VI (X_i)
2) En principe, les variables introduites dans l'analyse de régression sont numériques.
3) Première étape de l'analyse : vérifier que les X_i introduites dans la droite de régression (l'analyse) prédisent de manière significative la variance de Y. C'est-à-dire vérifier que la valeur du *F* associée au R^2 ajusté soit significative. Si cela n'est pas le cas, la droite de régression ne prédit pas Y (aucune X_i ne prédit Y)
4) Seconde étape de l'analyse : vérifier la significativité du Bêta associé à chaque X_i (VI). Si la valeur du *t* est significative, examiner la valeur du coefficient Bêta, qui vous donne une indication de la direction (positif ou négatif) et de la force du lien entre X_i et Y une fois que la valeur des autres X_i est maintenue constante

Exemple[1] 1 :

Régression linéaire avec une X (VI)

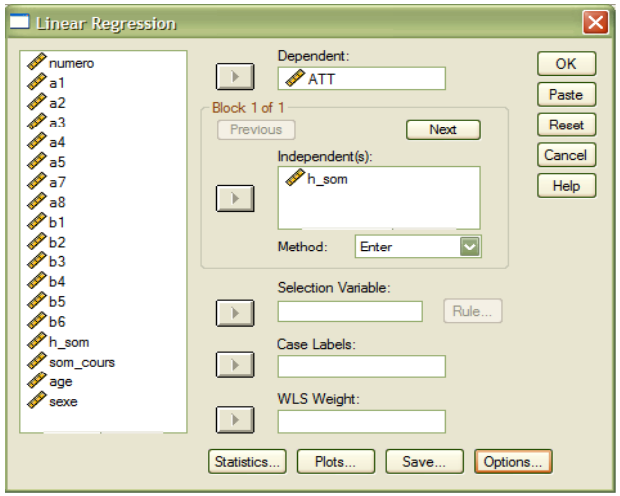
Afin de tester l'hypothèse (H1) selon laquelle le nombre d'heures de sommeil en dehors des cours (X) prédit positivement l'attention des étudiants en cours (Y), nous avons interrogé 100 étudiants. Dans ce cas, H0 est que le nombre d'heures de sommeil en dehors des cours ne prédit pas l'attention en cours. Nous avons demandé à ces 100 étudiants le nombre d'heures par jour qu'ils dorment en moyenne en dehors des cours (X ou VI numérique) et nous avons

[1] Les exemples proposés dans ce chapitre ont pris en compte des variables centrées. Pour ce faire, nous avons soustrait la moyenne générale de la variable spécifique des valeurs de celle-ci. Opération **compute**, équation **varx – Moyenne de x**. Pour cette raison, les chiffres sur les graphiques peuvent être négatifs.

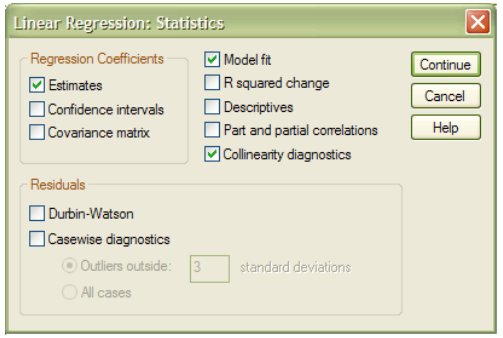
défini un indice d'attention en cours (Y ou VD numérique) qui a été mesuré sur une échelle qui a va de 1 = aucune attention à 6 = beaucoup d'attention. Dans SPSS, nous avons une variable *h_som* (qui représente une colonne dans la fenêtre **variables view**) qui correspond au nombre d'heures de sommeil en dehors des cours de chaque étudiants et une variable *ATT* et qui correspond à l'attention en cours du même étudiant. Ainsi, l'équation de la droite de régression est *ATT* = a + β**h_som*.

Procédure à suivre afin de réaliser l'analyse de régression linéaire dans SPSS :

Chemin : **Analyze → Regression →Linear…**



Sélectionnez dans la fenêtre de gauche la variable *ATT* et faites-la passer dans la fenêtre **Dependent** à droite (Y ou VD). Sélectionnez à gauche la variable *h_som* et faites-la passer dans la fenêtre à gauche **Independent(s)** (X ou VI). Cliquez sur l'onglet **statistics...**



Par défaut les cases **Estimates** et **Model fit** sont sélectionnées (gardez-les sélectionnées). Cochez la case **Collinearity diagnostics**.
Cliquez sur **Continue**. Vous retombez alors su la première fenêtre. Cliquez sur **OK.**

SPSS affiche alors la fenêtre d'output.

3 tableaux sont importants pour analyser les résultats :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.512(a)	.262	.254	1.18240

a Predictors: (Constant), h_som

Les heures de sommeil en dehors des cours expliquent le 25% de l'attention en cours (de la variance de Y ou VD).

Le R^2 peut être influencé par le nombre de participants inclus dans l'étude. Pour disposer d'une estimation de la variance de Y (ou VD) expliquée par X (la VI) qui prend en compte un nombre constant de participants, on considère le R^2 ajusté.

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48.629	1	48.629	34.783	.000(a)
	Residual	137.011	98	1.398		
	Total	185.640	99			

a Predictors: (Constant), h_som
b Dependent Variable: ATT

Valeur du *F*

Ce second tableau est le tableau d'ANOVA associé au R^2 ajusté. Il indique un *F* significatif: $F(1, 98) = 34.78$; $p < .001$ (voir chapitre 6.3.2). Cela signifie que la VI (X) introduite dans l'analyse (dans ce cas, les heures de sommeil en dehors des cours) prédit de manière significative Y (ou la VD ; dans ce cas, l'attention en cours). Ainsi, les 25% de part de variance de l'attention en cours qui est expliquée par le nombre d'heures de sommeil en dehors des cours ne sont pas attribuables au hasard. Le nombre d'heures de sommeil en dehors des cours est donc une variable pertinente pour expliquer l'attention en cours.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Constant	3.060	.118		25.880	.000		
	h_som	.391	.066	.512	5.898	.000	1.000	1.000

a Dependent Variable: ATT

Dans le troisième tableau, on trouve les coefficients de régression (B et Bêta). Celui qui nous intéresse concerne notre variable X h_som (dernière ligne du tableau). Pour cette variable, le B est égal à .39 et le Bêta (**standardisez Coefficients**) correspondant est de .51. Ces valeurs sont associées à un $t(98) = 5.90$; $p < .001$ (voir chapitre 6.3.1), qui indique que le nombre d'heures de sommeil en dehors des cours influence significativement l'attention pendant les cours. Le Bêta est positif, ce qui signifie que plus le nombre d'heures de sommeil en dehors des cours augmente, plus l'attention pendant les cours augmente. Le tableau indique aussi les tolérances pour chaque X (ou VI). La tolérance est un indicateur de la corrélation existante entre la X en question et les autres X_i. Pour que les résultats de l'analyse de régression soient fiables, il faut que les X_i soient peu corrélées entre elles (dans le cas contraire, il y a un problème de multicollinéarité). La valeur de la tolérance doit être supérieure à .60. Dans l'exemple, la valeur de la tolérance associée à h_som est 1 (tolérance parfaite) puisque l'analyse n'inclut qu'une seule X et elle ne peut pas être corrélée avec d'autres X_i.

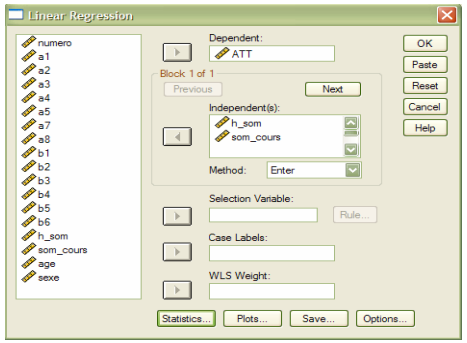
Les résultats peuvent être rédigés de la façon suivante: le nombre d'heure de sommeil en dehors des cours prédit de manière positive l'attention des étudiants pendant les cours (R^2 ajusté = .26 ; $F(1,98) = 34.78$, $p < .001$). Plus les étudiants dorment en dehors des cours, plus ils sont attentifs en cours (Bêta = .51; $t(98) = 5.90$, $p < .001$).

Exemple 2 :

Régression linéaire à 2 X_i (VI)

Le principe est exactement le même que lorsqu'il n'y a qu'une X. La différence se situe dans le fait que l'on rentre deux X dans la fenêtre **Indepentents** au lieu d'une. Admettons que l'on veuille tester l'impact des heures de sommeil en dehors des cours sur l'attention en cours, mais aussi celui des heures de sommeil effectuées en cours (variable *som_cours* dans la base de données), indépendamment des heures de sommeil en dehors des cours. Notre première hypothèse (H1a) est que le nombre d'heures de sommeil en dehors des cours (X_1) augmente l'attention en cours (Y). Notre deuxième hypothèse (H1b) est que le nombre d'heures de sommeil pendant les cours (X_2) diminuent l'attention en cours. Effectivement, on peut penser que plus les étudiants dorment en cours, moins ils sont attentifs.

Chemin : **Analyze → Regression →Linear…**



Rien ne change aux autres réglages par rapport à l'analyse présentée précédemment.

Après avoir effectué ces réglages et cliqué sur **ok** pour lancer l'analyse, on obtient l'output suivant:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.704(a)	.496	.485	.98941

a Predictors: (Constant), som_cours, h_som

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	91.501	2	45.750	46.735	.000(a)
	Residual	92.999	95	.979		
	Total	184.500	97			

a Predictors: (Constant), som_cours, h_som
b Dependent Variable: ATT

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.060	.100		30.618	.000		
	h_som	.478	.057	.625	8.349	.000	.947	1.056
	som_cours	-.596	.089	-.499	-6.665	.000	.947	1.056

a Dependent Variable: ATT

L'analyse des tableaux est identique, si ce n'est que deux effets sont à commenter. Ainsi, on observe que nos deux X_i expliquent 49% de la variance de Y et que cette proportion est significative, $F(2, 95) = 46.73$, $p < .001$. Comme précédemment, le nombre d'heures de sommeil en dehors des cours influence positivement l'attention en cours (Bêta = .63, $t(95) = 8.34$, $p < .001$). Conformément à notre hypothèse, nous trouvons également que les heures de sommeil pendant les cours diminuent l'attention en cours (Bêta = -.50, $t(95) = -6.66$, $p < .001$). Ainsi plus les étudiants dorment pendant les cours, moins ils sont attentifs pendant les cours. Nous observons que la tolérance est de .95 pour les deux X, ce qui est largement supérieur à .60. Les deux X ne sont donc pas trop liées entre elles et les résultats de l'analyse sont fiables.

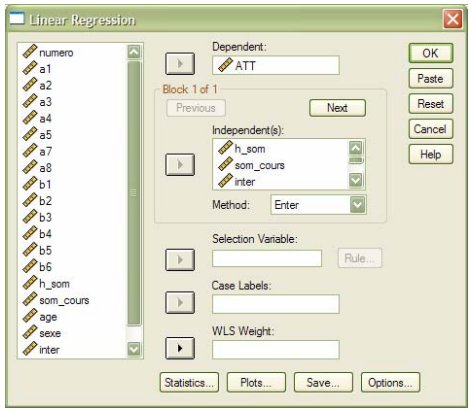
Exemple 3 :

Régression linéaire avec 2 X et l'interaction entre elles X (3 variables indépendantes sont incluses dans l'équation)

Contrairement à une analyse de variance ANOVA, l'interaction entre deux X_i n'est pas calculée automatiquement dans une analyse de régression (voir chapitre 6.5). Avant de procéder à l'analyse, il faut donc créer l'interaction entre les deux X_i. Pour cela, on crée une

nouvelle variable à l'aide du **compute**, qui correspond simplement à la multiplication des deux X_i (dans le **compute** rentrer dans la ligne de calcul: *h_som* * *som_cours*) et que nous appelons *inter*. Pour résumer, l'analyse comprend trois VI : *h_som*, *som_cours* et *inter*. Notre hypothèse concernant l'interaction entre ces deux variables est que le nombre d'heures de sommeil en dehors des cours influence davantage l'attention en cours lorsque la proportion d'heures de sommeil réalisées en cours est faible (H1c).

On réalise l'analyse de la même manière que dans les deux exemples précédents.
Dans la fenêtre principale, on rentre les deux X_i et l'interaction dans **Independents**. La VD (Y) est toujours introduite dans **Dependent.**



On obtient l'output suivant:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.796(a)	.634	.623	.84710

a Predictors: (Constant), inter, h_som, som_cours

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	117.048	3	39.016	54.372	.000(a)
	Residual	67.452	94	.718		
	Total	184.500	97			

a Predictors: (Constant), inter, h_som, som_cours
b Dependent Variable: ATT

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	3.174	.088		36.206	.000		
	h_som	.545	.050	.711	10.828	.000	.901	1.110
	som_cours	-.330	.089	-.276	-3.723	.000	.707	1.414
	inter	-.242	.041	-.452	-5.967	.000	.677	1.476

a Dependent Variable: ATT

L'analyse montre que les deux X_i et leur interaction expliquent 63% de la variance de l'attention en cours (Y, la VD) et que cette proportion de variance expliquée est significative: $F(3, 94) = 54.37$, $p < .001$. Les coefficients de régression standardisés (Bêta) indiquent que le nombre d'heures de sommeil en dehors des cours a une influence positive sur la VD (Bêta = .71 ; $t(94) = 10.83$, $p < .001$). De même, la proportion d'heures dormies en classe influence négativement la VD (Bêta = -.28; $t(94) = -3.72$, $p < .001$). Ces effets reproduisent les résultats de l'exemple 2. L'analyse montre que l'interaction entre les deux X_i a une influence significative sur la VD ($t(94) = -5.97$, $p < .001$) et que cette influence est négative (Bêta = -.45). Cette interaction nous apprend que, chez les étudiants qui dorment peu d'heures en dehors des cours, la proportion d'heures qu'ils passent à dormir en cours n'influence pas leur niveau d'attention en cours, qui reste assez bas. Par contre, quand le nombre d'heure de sommeil en dehors des cours augmente, on observe que plus la proportion d'heures de sommeil en cours diminue, plus l'attention aux cours augmente (pour l'interprétation des résultats, voir remarque et graphique 2). Notre hypothèse H1c est ainsi confirmée.

Exemple 4 :

Régression linéaire hiérarchique

Le principe des régressions hiérarchiques est un peu différent de celui de la régression linéaire simple dont nous avons traité jusqu'à maintenant. Il s'agit ici de tester les effets de chaque variable indépendante au cours de plusieurs étapes. À la première étape, on rentre une ou plusieurs X_i et on teste les effets de cette X ou ces X_i. À l'étape suivante, on rentre une ou plusieurs X_i qui vont s'ajouter à celles déjà présentes dans l'analyse. Ainsi, on teste l'effet de ces variables et si leur ajout augmente significativement la partie de la variance de la VD (Y) qui est expliquée par l'analyse. Le but principal de l'analyse de régression hiérarchique est de vérifier si l'ajout des nouvelles variables X_i à chaque étape augmente la variance de la VD (Y) expliquée et si cette augmentation est significative. Le nombre d'étapes et les variables ajoutées à l'analyse de régression de départ dépendent des hypothèses du chercheur.

La logique est ici davantage une logique de test de modèles (chaque droite de régression représente un modèle): on teste un modèle de base (la première étape) auquel on ajoute des variables X_i de façon à tester si le deuxième modèle (la deuxième étape) améliore de manière significative le premier modèle. Le but étant d'expliquer au mieux les variations de la VD (Y).

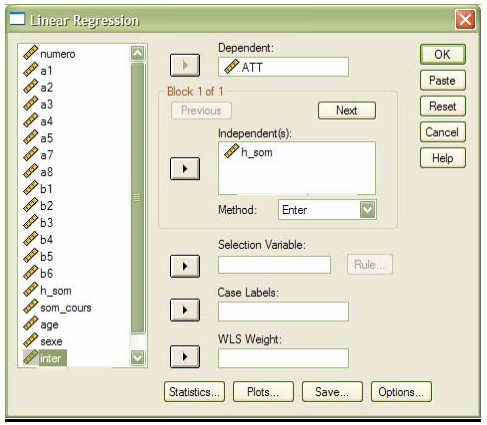
Nous allons réaliser la même analyse de l'exemple 3, mais avec une analyse hiérarchique. La régression de base (X_1 = h_som) testera l'impact du nombre d'heures de sommeil en dehors des cours sur l'attention en cours (Y ou VD). Nous ajouterons ensuite la proportion d'heures passée à dormir en cours (X_2 = som_cours), pour tester si le fait de tenir compte de cette

variable améliore la prédiction de l'attention en cours. Nous ajouterons finalement l'interaction entre les deux X_i (X_3 = inter), pour tester si le fait de tenir compte de l'effet différent des heures de sommeil en cours selon le nombre d'heures de sommeil en dehors des cours explique de manière plus précise l'attention en cours.

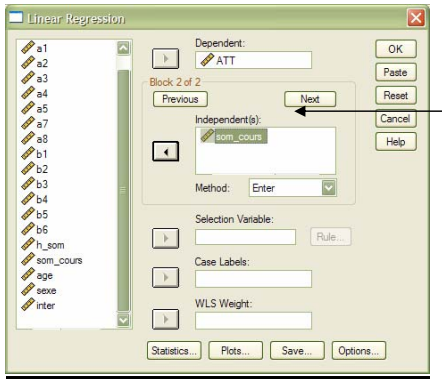
La démarche est la même que précédemment:

Analyze → regression → linear...

Dans la fenêtre **Dependent** glissez la variable *ATT*.
Dans la fenêtre **Independent** glissez la variable *h_som*. La droite de régression (le modèle) est $Y = a + \beta_1 * h\_som$.

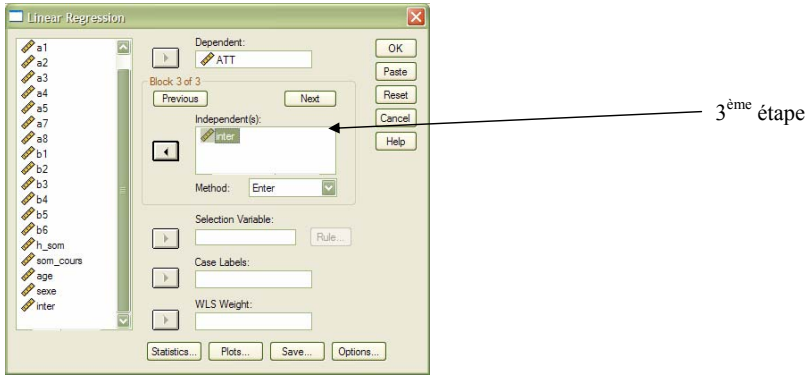


Cliquez sur l'onglet **Next** situé en haut à droite de la fenêtre **Independent.** Cette fenêtre « se vide ». Il s'agit en fait de la fenêtre **Independent** pour l'étape 2. Glissez alors la variable som_cours dans la fenêtre **Independent.** La variable h_som ne disparaît pas de l'analyse, simplement cela signifie que la variable som_cours s'ajoutera à celle-ci dans la deuxième droite de régression (le modèle), qui est $Y = a + B_1 * h\_som + B_2 * som\_cours$.

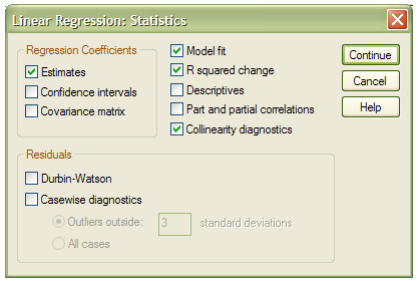


$2^{ème}$ étape

Cliquez à nouveau sur **Next**. La fenêtre **Independent «** se vide » à nouveau et vous pouvez alors y glisser la variable inter. La droite de régression (le modèle) est $Y = a + \beta_1 * h\_som + \beta_2 * som\_cours + \beta_3 * inter$.



Cliquez alors sur l'onglet **Statistics...** la fenêtre ci-dessous apparaît. Comme précisé plus haut, **Estimates** et **Model fit** sont cochés par défaut. Cochez **Collinearity diagnostics** et **R squared change** (cette dernière mesure permet de constater si l'ajout des variables de chaque étape améliore le modèle de manière significative par rapport à l'étape précédente).



Cliquez sur **Continue**. Vous revenez à la première fenêtre, cliquez alors sur **OK**.

L'output suivant apparaît :

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change
1	.510(a)	.260	.253	1.19235	.260	33.775	1	96	.000
2	.704(b)	.496	.485	.98941	.236	44.418	1	95	.000
3	.796(c)	.634	.623	.84710	.138	35.603	1	94	.000

a Predictors: (Constant), h_som
b Predictors: (Constant), h_som, som_cours
c Predictors: (Constant), h_som, som_cours, inter

Chaque tableau présente trois droites de régression (trois modèles), qui correspondent aux trois étapes (**Model**). Dans ce premier tableau, on voit que la première étape (**Model 1**, avec X_1 *h_som*) explique 26% de variance de la VD (Y), que la seconde étape (**Model 2**, les deux X_i *h_som* et *som_cours*) explique 49% et enfin que la troisième étape (**Model 3**, les deux X_i et l'interaction des deux) explique 63% de variance de la VD (Y). Ce sont les mêmes résultats des exemples précédents (exemple 1, 2 et 3).

Le **R Square Change** indique l'augmentation de R^2 entre les étapes. Par exemple l'ajout de l'interaction aux deux X_i lors de l'étape 3 augmente le R^2 de 14% par rapport à l'étape 2. Le F Change est le *F* associé au R change. Ainsi entre la 1ère et la 2ème étape, le R^2 augmente de 24%, ce qui est significatif ($F(1, 95) = 44.42$; $p < .001$). Entre la 2ème et la 3ème étape, le R^2 augmente de 14%, ce qui est significatif ($F(1, 94) = 35.60$; $p < .001$). Ainsi, l'ajout de X_2 (som_cours) à X_1 (h_som) et de l'interaction des deux (inter) est pertinent et améliore significativement la prédiction de l'attention en cours.

ANOVA(d)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	48.018	1	48.018	33.775	.000(a)
	Residual	136.482	96	1.422		
	Total	184.500	97			
2	Regression	91.501	2	45.750	46.735	.000(b)
	Residual	92.999	95	.979		
	Total	184.500	97			
3	Regression	117.048	3	39.016	54.372	.000(c)
	Residual	67.452	94	.718		
	Total	184.500	97			

a Predictors: (Constant), h_som
b Predictors: (Constant), h_som, som_cours
c Predictors: (Constant), h_som, som_cours, inter
d Dependent Variable: ATT

L'interprétation de ce second tableau ne change pas par rapport aux exemples précédents. Ce tableau montre que l'explication de la VD (Y) ne peut pas être attribuée au hasard lors de chaque étape (voir exemple 1 pour le **Model 1**, exemple 2 pour le **Model 2** et exemple 3 pour le **Model 3**).

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta	Tolerance	VIF	B	Std. Error
1	(Constant)	3.062	.120		25.424	.000		
	h_som	.391	.067	.510	5.812	.000	1.000	1.000
2	(Constant)	3.060	.100		30.618	.000		
	h_som	.478	.057	.625	8.349	.000	.947	1.056
	som_cours	-.596	.089	-.499	-6.665	.000	.947	1.056
3	(Constant)	3.174	.088		36.206	.000		
	h_som	.545	.050	.711	10.828	.000	.901	1.110
	som_cours	-.330	.089	-.276	-3.723	.000	.707	1.414
	inter	-.242	.041	-.452	-5.967	.000	.677	1.476

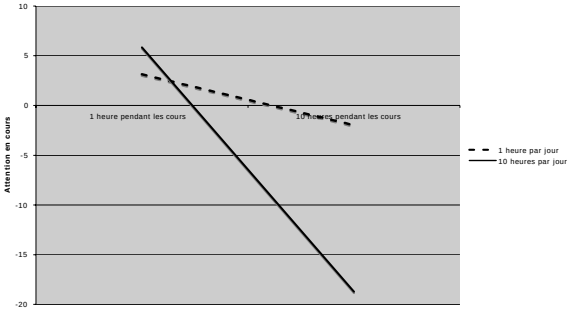
a Dependent Variable: ATT

Ce troisième tableau est interprété de la même manière que les tableaux de coefficients de régressions des trois exemples qui précèdent. Le modèle 1 indique que le nombre d'heures de sommeil en dehors des cours influence significativement l'attention pendant les cours. (voir exemple 1). Le modèle 2 indique qu'en plus de cet effet, on trouve un effet principal de la proportion d'heures passées à dormir en cours. (voir exemple 2). Enfin, le modèle 3 indique qu'à ces deux effets s'ajoute celui de l'interaction des deux X_i (voir exemple 3).

Remarques

1) Souvent, il est plus clair de présenter les résultats des analyses de régression sous forme de tableau (par exemple, avec les informations relatives au R^2 ajusté, sa significativité, les valeurs Bêta (ou B) et leur significativité, et le test de significativité du R square change lors de chaque étape s'il est question d'une analyse de régression hiérarchique).
2) L'interprétation des interactions n'est pas toujours évidente, puisqu'il ne s'agit pas ici de comparer des moyennes. Le graphique d'interaction peut être réalisé, mais SPSS ne le fournit pas. Le seul moyen de construire ce graphique est de résoudre l'équation de la droite de régression correspondant à l'analyse (dans ce cas, il convient d'utiliser les coefficients non-standardisés pour le calcul). Dans l'exemple 3, l'équation de la droite de régression est : $Y = a + \beta_1*X_1 + \beta_2*X_2 + \beta_3*X_1X_2$, dans laquelle $a = 3.17$, $\beta_1 = .54$, $\beta_2 = -.33$ et $\beta_3 = -.24$. L'équation devient, $Y = X_1* (\beta_1 + \beta_3*X_2) + \beta_2*X_2$. Maintenant, il faut définir X_1. Nous avons choisi de calculer les moyennes pour un étudiant qui dort une heure en dehors des cours ($X_1 = 1$) ou 10 heures ($X_1 = 10$). Ainsi, nous obtenons deux équations : $Y = 3.17 + (.54 + -.24*X_2) + -.33*X_2$ et $Y = 3.17 + 10*(.54 + -.24*X_2) + -.33*X_2$. Pour calculer les moyennes de ces deux variables, on doit encore définir la valeur de X_2. Nous avons décidé de considérer le cas d'un étudiant qui dort 1 heure pendant les cours et d'un autre qui dort 10 heures. Ainsi, nous obtenons 4 valeurs de Y:
- Étudiant qui dort 1 heure par jour en dehors des cours et 1 heure pendant les cours :
Y = 3.14 (attention en cours)
- Étudiant qui dort 1 heure par jour en dehors des cours et 10 heures pendant les cours :
Y = -1.99
- Étudiant qui dort 10 heures par jour en dehors des cours et 1 heure pendant les cours :
Y = 5.84
- Étudiant qui dort 10 heures par jour en dehors des cours et 10 heures pendant les cours :
Y = -18.73

Au niveaux graphique, cela donne (graphique 2):



La ligne pointillée représente les étudiants qui dorment 1 heure en dehors des cours et la ligne continue représente les étudiants qui dorment 10 heures en dehors des cours. Vous pouvez voir que la différence est bien plus marquée chez les seconds que chez les premiers (interaction).