Analyse de données quantitatives (LPOLS1221)
Séance 1 : Prise en main de SPSS et Data Management

Autres ressources :

- SPSS à l'UdeS (Université de Sherbrooke) – url : http://spss.espaceweb.usherbrooke.ca/
- KINNEAR P., GRAY C. (2005). *SPSS facile appliqué à la psychologie et aux sciences sociales. Maitriser le traitement des données*, De Boeck, Bruxelles, 432 p.
- FOX W. (1999). *Statistiques sociales* (3e édition), De Boeck & Larcier s.a., Paris, Bruxelles, 374 p.

Présentation du logiciel

3 types de fenêtres SPSS :

- ***Data editor :*** fenêtre principale avec deux onglets :
  - « *Data view* » : affichage brut des données (une ligne par unité d'observation et une colonne par variable)
  - « *Variable view* » : informations relatives à chacune des variables (une ligne par variable et les caractéristiques de chaque variable en colonnes)

- ***Syntax :*** fenêtre de commandes SPSS ;
  - soit on encode directement les commandes (exige une connaissance du langage de programmation SPSS)
  - soit on encode indirectement les commandes (par la fonction « *Paste* »). **Attention :** prendre le réflexe d'enregistrer la syntaxe pour *toute* opération effectuée plutôt que les résultats obtenus
    - automatiser la reproduction de vos procédures (gain de temps parfois non négligeable)
    - possibilité de retourner en arrière si une erreur s'est produite

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

- ***Output viewer :*** fenêtre présentant les résultats des analyses exécutées par SPSS

L'onglet Variable view du Data editor :

On y retrouve toutes les informations relatives aux variables.

- **Name (nom)**: code attribué (nom raccourci ou abréviation généralement) à la variable (pas d'espace autorisé).
  Utiliser le nom des variables rend l'utilisation du logiciel plus facile et plus rapide. N'hésitez pas à effectuer un clic droit sur une liste de variables dans un menu divers et cliquer sur « Display Variable names ».

- **Type** : format des données collectées. Utilisation fréquente des types « numérique », « chaine » (texte) et éventuellement « date ». Permet au logiciel de reconnaître et de pouvoir traiter les symboles contenus dans la base de données.

  - Si la variable contient uniquement des nombres : renseignez « *numeric* » ;
  - Si la variable contient des caractères : renseignez « *string* » ;
  - D'autres formats sont disponibles mais seront rarement utilisés dans le cadre du cours.

- **Width (largeur)** : nombre de caractères maximum autorisé au niveau de la base de données. Par défaut, affiche le nombre de caractères de la valeur/modalité la plus longue.

- **Decimals (décimales)** : nombre de décimales apparentes.

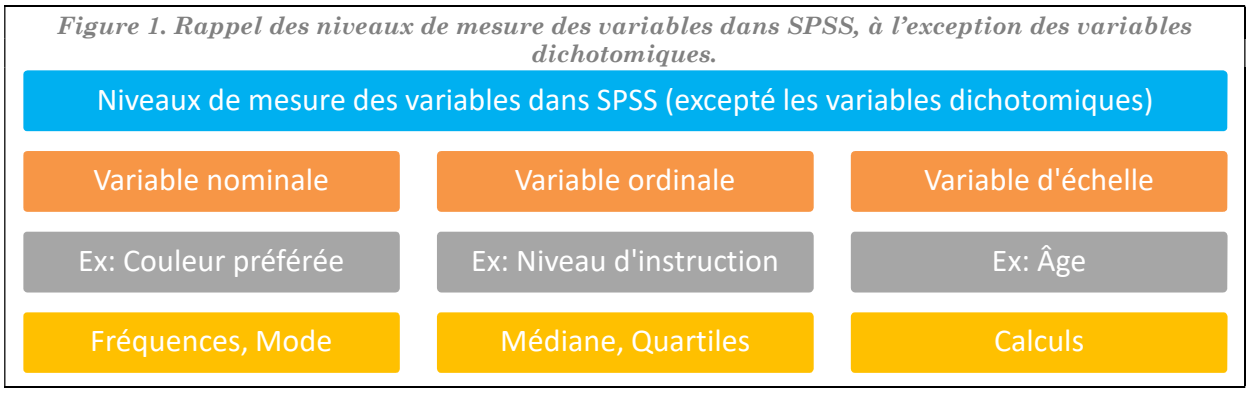
- **Label (libellé)** : descriptif ou « étiquette» de la variable. Lors du traitement des données, l'on utilise généralement le **nom** diminué de la variable pour des questions de rapidité. Cependant, lors de la production de rapport, il est préférable d'avoir des intitulés de variables complets. Lorsque vous produirez des résultats, vous verrez le libellé apparaître, et non pas le nom.

- **Values (valeurs)** : répertoire des valeurs/modalités possibles pour la variable et attribution d'un libellé aux différentes valeurs/modalités (pour des variables catégorielles). Grâce à cet outil, il est possible par exemple de dire que la valeur 1 enregistrée dans la base de données de la variable « Genre », va apparaître dans le rapport sous la forme de « Femme » et la valeur 2 sous la forme de « Homme ».
  Si vous avez oublié quel label est associé à quelle valeur, c'est dans l'onglet Values de la variable view qu'il faut se rendre.

- **Missing (Données Manquantes)** : codification de modalités comme valeurs manquantes.
  Si par exemple, lors d'une enquête, des personnes ont répondu « je ne sais pas » à la question posée, vous pouvez renseigner ce type de réponse comme donnée manquante. Elle ne sera alors pas prise en compte lors de l'analyse.

- **Measure (mesure)** : nature de la variable : (1) *nominal* : variable catégorielle nominale ; (2) *ordinal* : variable catégorielle ordinale ; (3) *scale* : variable quantitative (discrète ou continue).

Figure 1. Rappel des niveaux de mesure des variables dans SPSS, à l'exception des variables dichotomiques.



Les variables **nominales** sont les variables où il n'y a **aucun ordre intrinsèque entre les différentes modalités**, comme par exemple l'option politique ou la nationalité.

Les variables **ordinales** sont celles où il y a un ordre entre les modalités, mais **sans que pour autant il y ait une unité de mesure**. Par exemple, si les individus notent leur

soutien pour un parti politique sur une échelle allant de 1 à 5, ceci ne signifie pas que la distance entre 2 et 3 est la même que celle entre 4 et 5.

Les variables **quantitatives** permettent des opérations mathématiques, puisqu'**en plus d'un ordre entre les valeurs possibles, elles ont aussi une unité de mesure**. Si l'on mesure le revenu en euros, la différence absolue entre 1400 et 1600 est la même que la différence entre 1900 et 2100.

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

Principales fonctions de data management dans SPSS :

1. **Création de variables** : Encodage manuel des données et des caractéristiques de la variable ou importation d'une base de données

2. **Définir les caractéristiques d'une variable** : deux possibilités
   - Traitement manuel dans la fenêtre « Variable view »
   - « Data» → « Define Variable properties » → Permet d'utiliser la syntaxe et donc de garder une trace des opérations effectuées.

3. **Transformation et création de variables** (à partir de données existantes) : 5 possibilités
   - « Transform » → « **Recode into different variables** »
     Crée des variables de classes à partir de variables quantitatives ou qualitatives, en groupant une/des valeur(s) dans une nouvelle catégorie.
   - « Transform» → « **Compute variable** »
     Crée des variables sur base de calculs (opérations arithmétiques) sur d'autres variables. Utilisation courante des opérateurs suivants :
     + *(addition);* **-** *(soustraction);* **/** *(division);* ***** *(multiplication);* ****** *(exposant)*
   - « Transform » → « Recode into same variables » (**Prudence !!**)
     Ne crée pas une nouvelle variable mais écrase les données de la variables par de nouvelles valeurs.
   - « Transform » → « **Count values within cases** »
     Crée une variable de comptage sur base de plusieurs variables de départ, dont on décompte certaines valeurs/modalités cible.
   - « Transform » → « Visual Binning »
     Crée une variable sur base de prédécoupements (ne sera pas traité dans ce cours)

4. **Sélection d'unités d'analyse spécifiques** : « Data » → « Select cases » Utilisation des opérateurs logiques déjà présentés ci-dessus et également des opérateurs logiques du type : | *(« ou » logique);* & (« *et* » *logique)*

5. **Scinder les analyses et les résultats** pour des groupes particuliers de répondants : « Data » → « Split file ».

6. **Sélectionner une série de variables** : « Utilities » → « Define Variable sets » et ensuite « Use variable sets » pour afficher uniquement cette série de variables.

7. **Traitement des données manquantes**
   - Soit absence de données
   - Soit codification spécifique par des valeurs auxquelles on attribue une étiquette « donnée manquante »

   ⇒ Que signifie une donnée manquante ? Que faire des données manquantes ? Différents traitements possibles mais jamais une meilleure solution !
   - Première étape : vérifier la raison de la non-réponse :
     - Absence de données (refus/impossibilité/sans opinion/autre) ?
     - Question non applicable ?
   - Deuxième étape : prendre une décision quant au traitement des non-réponses (rem : différents traitements sont possibles mais il n'y a jamais une seule solution idéale ou recommandée !)
     - Exclure les données manquantes ?
     - Remplacer les données manquantes ? (moyenne, mode, réponse aléatoire et proportionnelle, etc.)

Application

Base de données (Exo-TP1) simplifiée : 20 individus et 10 variables de départ, avec le codebook suivant :

- « age » : Âge des individus (mois) au moment de l'enquête
- « sexe » : Sexe des individus (1 : Homme ; 2 : Femme)
- « taille » : Taille des individus (cm) au moment de l'enquête
- « poids » : Poids des individus (kg) au moment de l'enquête
- « option » : Option choisie par les étudiants (LG : Latin-Grec ; Sc : Sciences ; Eco : Économie ; SS : Sciences sociales)
- « note_frcs » : Note finale (sur 100) obtenue en cours de Français
- « note_math » : Note finale (sur 100) obtenue en cours de Mathématique
- « note_geohist » : Note finale (sur 100) obtenue en cours de Géographie-Histoire
- « note_option » : Note finale (sur 100) obtenue en cours à option
- « instr_père » : Niveau d'instruction le plus élevé obtenu par le père du répondant (1 : nul-primaire ; 2 : secondaire ; 3 : supérieur)

Importez cette base de données en SPSS, tenant compte du fait que la première ligne contient les noms des valeurs.

1. Créez deux groupes de variables :

   - Un groupe « constitution » reprenant les variables age, sexe, taille, poids

- Un groupe école reprenant les autres variables.

2. Utilisant le groupe de variables « constitution », effectuez les manipulations suivantes :
   - *Body Mass Index*[1] (BMI)
     - Créer une nouvelle variable qui illustre la valeur du BMI des répondants
     - Créer une nouvelle variable regroupant les individus par catégorie de BMI
       - < 18.5 : Insuffisance pondérale
       - 18.5-24.9 : Corpulence normale
       - 25-29.9 : Surpoids
       - > 29.9 : Obésité

       Tenez compte des unités de mesure et copiez la syntaxe de création de la variable.

     - Créer une nouvelle variable regroupant les individus en deux groupes : soit « Corpulence normale », soit « Corpulence "anormale" »
     - Comment interpréter les données manquantes ? Comment pourrait-on les traiter ? Quelle solution adopteriez-vous ? Discussions des avantages et inconvénients.

3. Résultats scolaires. Utilisant à nouveau l'entièreté de la base, effectuez les manipulations suivantes :
   - Créer une nouvelle variable calculant le résultat final pondéré (en %) des étudiants (Français : 5 ECTS ; Math : 5 ECTS ; Géo-Histoire : 2 ECTS ; Option : 4 ECTS)
   - Créer une nouvelle variable qui exprime le pourcentage obtenu par les étudiants en une note sur 20
   - Créer une nouvelle variable qui attribue un grade aux étudiants
     - < 10 : Ajournement
     - 10-13.9 : Satisfaction
     - 14-15.9 : Distinction
     - 16-17.9 : Grande distinction
     - ≥ 18 : La plus grande distinction
   - Créer une nouvelle variable qui répartit les étudiants en deux groupes : soit « Echec », soit « Réussite »
   - Créez une variable qui compte le nombre de dispenses que l'étudiant a obtenu (cote supérieure à 50).

[1] $BMI = \frac{poids\ (kg)}{taille\ (m)^2}$