Université de Montréal

Guide d'introduction au logiciel SPSS

Guide élaboré pour les étudiants du cours
CRI-1600 G : Initiation aux méthodes quantitatives

Certificat de criminologie
Faculté de l'éducation permanente

Automne 2010

TABLE DES MATIÈRES

LISTE DES DIFFÉRENTES PROCÉDURES PRÉSENTÉES

PRÉSENTATION DU LOGICIEL SPSS

1. L'ENVIRONNEMENT SPSS

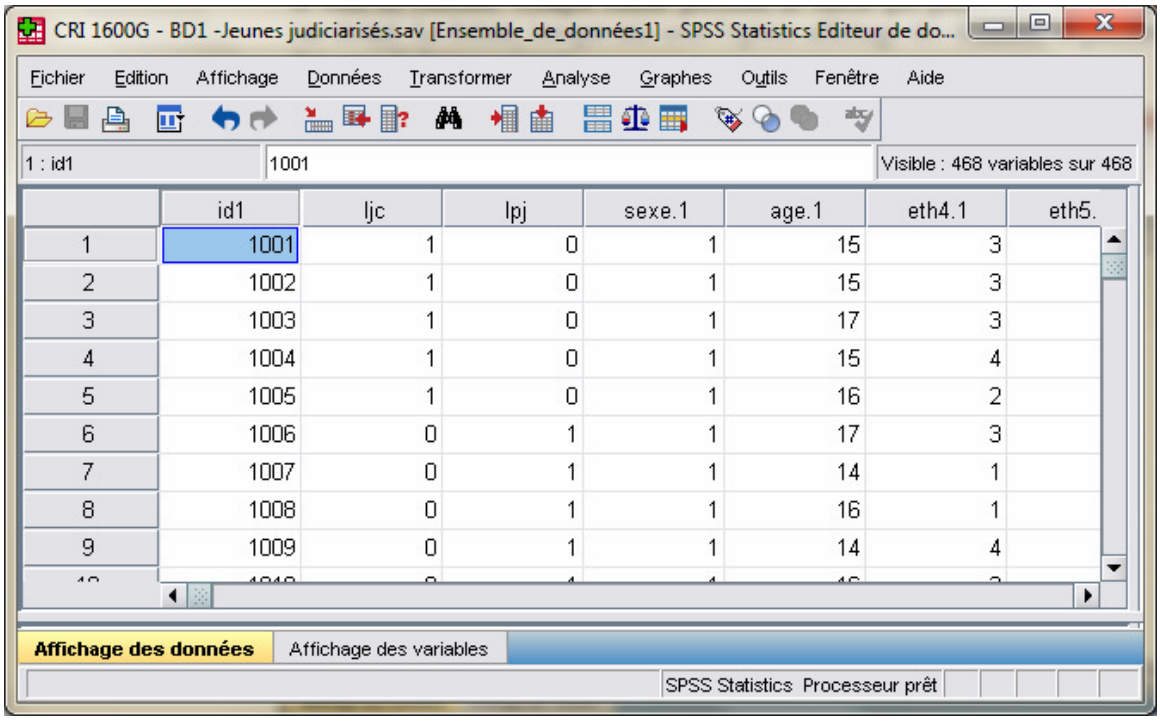
SPSS (*Statistical Package for Social Sciences*) est essentiellement un logiciel de traitement de données en vue d'analyses statistiques. Il lit les données, les traduit en format SPSS, les transforme - si demandé - et exécute des opérations mathématiques et statistiques. La version de SPSS utilisée dans le cadre du cours est la 17.0

1.1. LA FENÊTRE ÉDITEUR DE DONNÉES

La fenêtre *Éditeur de données* présente le contenu d'un fichier de données que vous avez préalablement sélectionné. Vous pouvez créer de nouvelles feuilles de données ou modifier des données préexistantes. Cette fenêtre comprend deux onglets :

L'ONGLET AFFICHAGE DES DONNÉES

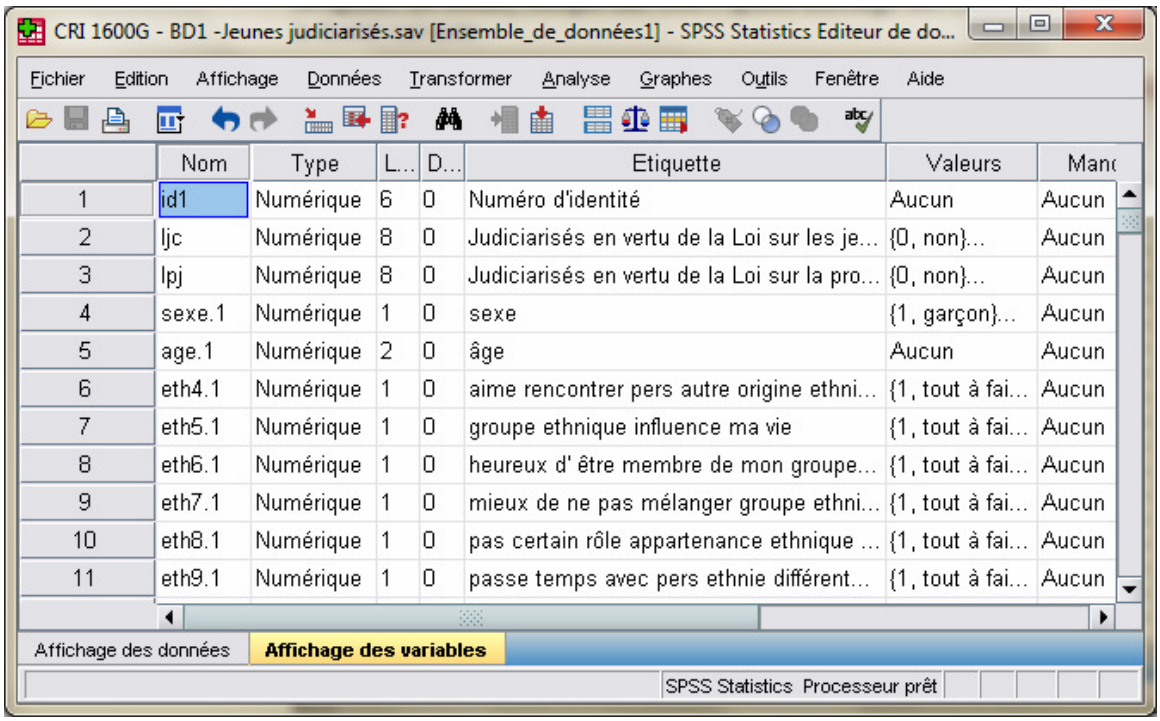
- Permet de voir la banque de données, où les cas sont présentés en lignes et les variables sont en colonne. Chaque cellule présente la valeur que prend une variable pour un cas donné. Le fichier, lorsque mis sous l'option affichage des données, se présente sous la forme suivante :



	id1	ljc	lpj	sexe.1	age.1	eth4.1	eth5.
1	1001	1	0	1	15	3	
2	1002	1	0	1	15	3	
3	1003	1	0	1	17	3	
4	1004	1	0	1	15	4	
5	1005	1	0	1	16	2	
6	1006	0	1	1	17	3	
7	1007	0	1	1	14	1	
8	1008	0	1	1	16	1	
9	1009	0	1	1	14	4	

L'ONGLET AFFICHAGE DES VARIABLES

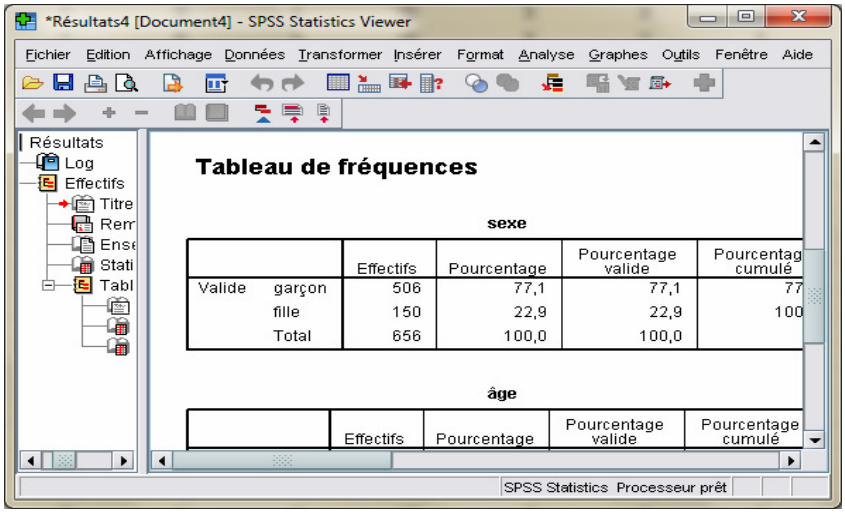
- Permet de voir toutes les variables présentes de la banque de données, leurs noms, ce qu'elles représentent, leurs valeurs manquantes, leurs valeurs possibles, les étiquettes qui les désignent. Le fichier, lorsque mis sous l'option affichage des variables, se présente sous la forme suivante :



	Nom	Type	L...	D...	Etiquette	Valeurs	Manc
1	id1	Numérique	6	0	Numéro d'identité	Aucun	Aucun
2	ljc	Numérique	8	0	Judiciarisés en vertu de la Loi sur les je...	{0, non}...	Aucun
3	lpj	Numérique	8	0	Judiciarisés en vertu de la Loi sur la pro...	{0, non}...	Aucun
4	sexe.1	Numérique	1	0	sexe	{1, garçon}...	Aucun
5	age.1	Numérique	2	0	âge	Aucun	Aucun
6	eth4.1	Numérique	1	0	aime rencontrer pers autre origine ethni...	{1, tout à fai...	Aucun
7	eth5.1	Numérique	1	0	groupe ethnique influence ma vie	{1, tout à fai...	Aucun
8	eth6.1	Numérique	1	0	heureux d' être membre de mon groupe...	{1, tout à fai...	Aucun
9	eth7.1	Numérique	1	0	mieux de ne pas mélanger groupe ethni...	{1, tout à fai...	Aucun
10	eth8.1	Numérique	1	0	pas certain rôle appartenance ethnique ...	{1, tout à fai...	Aucun
11	eth9.1	Numérique	1	0	passe temps avec pers ethnie différent...	{1, tout à fai...	Aucun

1.2. LA FENÊTRE DE RÉSULTATS VIEWER

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



Les résultats des analyses que vous ferez s'empilent les uns à la suite des autres, au fur et à mesure que vous effectuez de nouvelles opérations, à l'intérieur d'une même session de travail. Utilisez les *menus* de la fenêtre pour sélectionner des fichiers, des statistiques et des diagrammes. Vous comprendrez aussi rapidement la pertinence de « faire le ménage » au fur et à mesure dans les résultats que vous produisez, une fenêtre de résultats *Viewer* qui ne serait pas débarrassée des résultats erronés, ou éventuellement devenus inutiles, se faisant de plus en plus difficile à consulter.

2. LES FICHIERS DANS SPSS

2.1. LES FICHIERS DE DONNÉES

Les fichiers de données qui apparaissent dans la fenêtre *Éditeur de données* se reconnaissent par l'extension **.sav** suivant le nom. Les fichiers de données contiennent les *banques de données* avec lesquelles vous aurez à travailler. Ces fichiers se présentent sous la forme d'une feuille quadrillée remplie de chiffres ou de lettres répartis en colonnes et en rangées.

- Si vous désirez créer nouvelle une banque de données, il ne vous reste plus qu'à y entrer les informations pertinentes en respectant les normes qui vous auront été communiquées.
- Si vous devez plutôt travailler à partir d'une banque de données déjà constituée, vous devez alors ouvrir le fichier en question en suivant la procédure suivante :

PROCÉDURE POUR OUVRIR UN FICHIER DE DONNÉES

Cliquez sur **Fichier** dans le menu principal en entête de SPSS
↳ **Ouvrir**
↳ **Données** (la liste des fichiers de données lisibles pour SPSS apparaît)
cliquez sur le nom du fichier nécessaire à vos analyses (extension .sav) qui apparaît sous cette rubrique puis *cliquez* sur
↳ **Ouvrir**

Le nom du fichier de données que vous avez sélectionné apparaît en entête de la fenêtre. Vous pouvez dès lors effectuer les traitements que vous jugez nécessaires, qu'il s'agisse de modifications à l'intérieur même du fichier de données ou de traitements statistiques.

L'ouverture d'un fichier de données doit toujours précéder les instructions de traitement statistiques données à SPSS, autrement celui-ci ne pourra pas s'exécuter, ne sachant pas quelles données traiter. Si des modifications ont été apportées au fichier de données préalablement ouvert, SPSS demandera à l'utilisateur si ces modifications doivent être sauvegardées avant que le fichier soit fermé.

2.2. LES FICHIERS DE RÉSULTATS

Les fichiers de résultats se reconnaissent par l'extension **.spv** suivant le nom attribué au fichier par l'utilisateur. Ils apparaissent dans la fenêtre de résultats *Viewer* et consistent en divers résultats d'opérations précédentes qui ont été sauvegardés. Tant et aussi longtemps que l'utilisateur n'a pas sauvegardé le travail effectué en séance, le fichier se nomme RÉSULTATS1. Au moment de sauvegarder un fichier contenant les résultats qu'il désire conserver, l'utilisateur doit obligatoirement lui donner un nom. À partir de ce moment, le fichier peut être manipulé comme tout autre fichier, c'est-à-dire qu'il peut être ouvert, consulté, édité, à nouveau sauvegardé, etc.

Même si tout ce qui se retrouve dans le fichier résultat est généré automatiquement par SPSS, il est possible d'accéder à ces fichiers et de les éditer, c'est-à-dire d'en modifier le contenu dans le but, par exemple, de présenter intégralement les résultats dans le cadre d'un travail à remettre ou d'un article en production sans avoir à en refaire *manuellement* la présentation, ou encore pour « faire du ménage ». Il est possible d'effacer les opérations inutiles ou les résultats erronés, d'enlever les messages d'erreurs dont on a pris connaissance ou de produire une meilleure mise en page.

Pour ce faire, il s'agit de manipuler le fichier comme vous le feriez si vous étiez dans un traitement de texte simplifié, en coupant, en collant et en ajoutant du nouveau texte. Les capacités de SPSS en matière traitement de texte sont toutefois limitées. Nous verrons comment y pallier.

PROCÉDURE POUR OUVRIR UN FICHIER DE RÉSULTATS

Cliquez sur **Fichier** dans le menu principal en entête de SPSS
- **Ouvrir**
  - **Résultat** (*cliquez* sur le nom du fichier choisi (extension .**spv**) qui apparaît sous cette rubrique puis *cliquez* sur
    - **Ouvrir**

3. LES COMMANDES DANS SPSS

Il existe deux façons d'entrer des commandes dans SPSS. La première est assez simple puisqu'il s'agit de sélectionner, à l'aide de la souris, les éléments pertinents des différents *menus* qui vous sont présentés les uns à la suite des autres ou en cascade. Cette technique, très *user friendly* vous permet d'effectuer toutes les commandes à partir des *zones de dialogue*. C'est la technique qu'on vous enseignera dans le cadre du cours.

3.1. TRAVAILLER AVEC LES MENUS, SOUS-MENUS ET LA SOURIS

C'est à l'aide de la souris que vous vous déplacerez dans SPSS. Le curseur indique l'endroit où vous vous trouvez. La touche de gauche de la souris vous permet de sélectionner, en

cliquant, un élément du menu, ou encore une commande, une sous-commande ou une option qui lui sont associées.

3.2. LES COMMANDES, LES SOUS-COMMANDES ET LES OPTIONS

Les items qui apparaissent aux différents niveaux de menus sont des commandes, des sous-commandes ou des options que vous pouvez (ou dans certains cas devez) ou non sélectionner afin de compléter ou de préciser des analyses que vous envisagez demander à SPSS de mettre à exécution. SPSS prévoit ordinairement, par défaut, un certain nombre d'options minimales qu'il associe automatiquement à chaque commande ou sous-commande rencontrée sous SPSS, selon le cas. Dans la mesure où ces options vous satisfont, vous n'avez rien d'autre à spécifier que les commandes et sous-commandes pour que SPSS remplisse la mission de traitement des données que vous lui assignez.

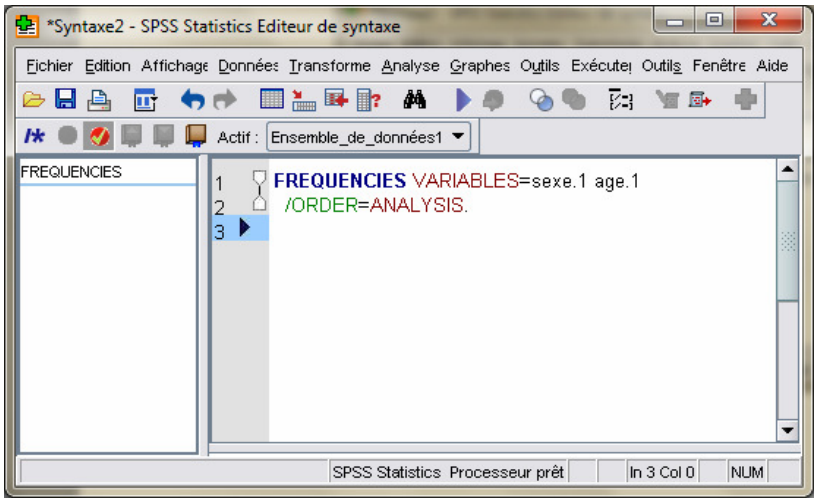
3.3. RÉALISER UNE OPÉRATION DANS SPSS

Rien n'est plus simple que de demander à SPSS d'exécuter une commande. En fait, il s'agit à chaque fois de sélectionner les opérations et les options que vous désirez voir exécuter, d'indiquer sur quelles variables les opérations doivent être faites et, finalement, de lancer la commande en cliquant sur **OK,** ce qui confirme à SPSS de se mettre en action.

Si l'opération que vous venez d'exécuter est une commande statistique, une fenêtre de résultats *Viewer* apparaîtra. Rappelez-vous aussi que tant et aussi longtemps que vous n'avez pas vous-même sauvegardé ce *fichier-résultat*, celui-ci est temporaire et s'effacera à la fin de la séance de travail en cours.

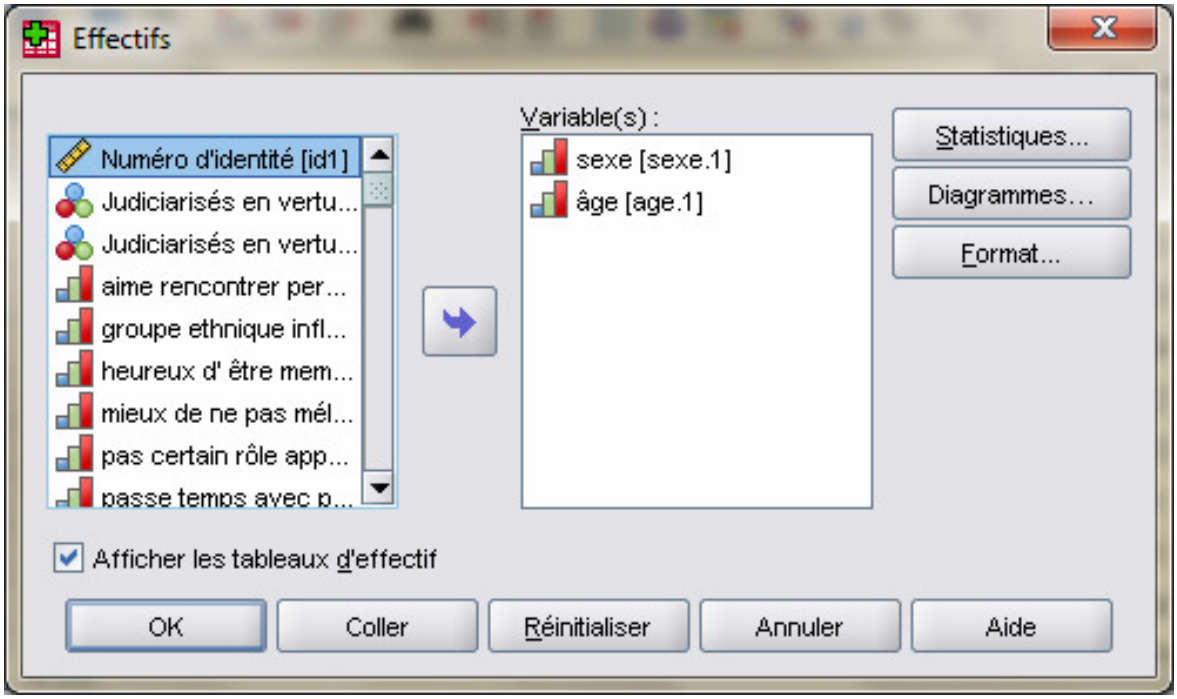
3.4. LA SYNTAXE DANS SPSS

La deuxième façon d'entrer des commandes dans SPSS est via la *syntaxe*. La syntaxe est essentiellement un fichier de texte simple contenant le *code* de toutes les commandes et sous-commandes demandées dans le cadre de vos analyses statistiques. Ce texte s'affiche dans la fenêtre *Éditeur de syntaxe* qui se présente sous la forme suivante :



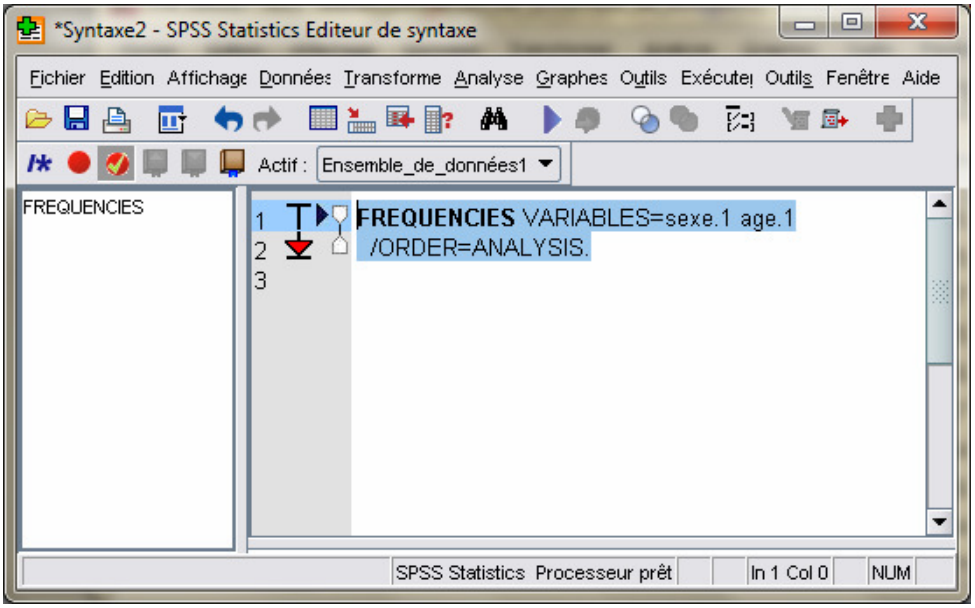
L'utilisation de la syntaxe présente plusieurs avantages. Parmi les plus importants, celui de pouvoir suivre la progression de vos travaux statistiques et celui de pouvoir « garder des traces » de toutes les commandes et sous-commandes que vous aurez demandé un logiciel d'exécuter. Ces particularités peuvent s'avérer fortes intéressantes lorsque vous essaierez plusieurs combinaisons de tests successivement lors d'une même séance de travail et que vous tenterez d'organiser vos analyses et vos résultats. Ces caractéristiques sont aussi particulièrement avantageuses lorsque vous êtes plusieurs à travailler sur les mêmes données, à effectuer quelque fois de la manipulation de variables, telles que des *recodages* ou alors que vous créez de nouvelles variables. Il est possible aussi d'inscrire des notes directement dans le fichier syntaxe, comme par exemple, *analyses bivariées en date du 1er septembre avec les données brutes*. Il s'agit simplement de mettre des astérisques * devant et après vos notes. Par ailleurs, ces notes se retrouveront dans vos *fichiers-résultats* et vous permettront de mieux situer vos tableaux. Un autre avantage des fichiers de syntaxe est qu'ils peuvent être traités comme un logiciel de traitement de texte simple et les fonctions *coupés, copiés, collés* sont possibles. Il est ainsi possible de faire plusieurs tests en recopiant seulement le nom des nouvelles variables, ce qui permet de sauver beaucoup de temps à *cliquer* et *recliquer* dans les *menus*. Au besoin, vous aurez simplement à ouvrir votre fichier de syntaxe et à la *rouler* à nouveau pour relancer les analyses, ce qui vous évitera de toujours sauvegarder ou imprimer vos résultats.

Dans les versions antérieures de SPSS, la syntaxe devait être rédigée *manuellement*. Les utilisateurs devaient donc apprendre le *code* des commandes et sous-commandes qu'ils désiraient effectuer. Dans la version 17.0 de SPSS, l'utilisation de la syntaxe est beaucoup plus *user friendly*. Il s'agit simplement d'utiliser les fonctions des *menus* et lorsque toutes les commandes et sous-commandes ont été vérifiées, *cliquer* sur le bouton **COLLER** situé au bas à gauche avant de *cliquer* sur **OK.** Une fenêtre *Éditeur de syntaxe* s'ouvrira automatiquement.



Dans la version actuelle de SPSS, le logiciel dispose d'un éditeur de syntaxe qui corrige automatiquement les erreurs de *code*, et propose des choix de commandes ou de sous-commandes au besoin par le biais d'un menu déroulant. Les erreurs sont plus facilement repérables puisque les fichiers sont maintenant en couleur et seuls les termes reconnus s'affichent en couleur (les commandes en bleu et les sous-commandes en vert). Un panneau d'erreur indiquant où les problèmes sont situés apparaît aussi automatiquement si nécessaire. Il est aussi très simple d'accéder à de l'aide supplémentaire en cliquant sur l'onglet **AIDE**.

Pour faire *rouler* une syntaxe, il s'agit simplement de sélectionner le paragraphe contenant le *code* désiré et de *cliquer* sur la flèche bleue.



La flèche avec le bout rouge située en marge à gauche indique les lignes de syntaxe qui ont été *roulées*. Notez qu'il est aussi possible de mettre des *points d'arrêt* en marge à gauche si on ne désire pas *rouler* toute une syntaxe dans un même fichier. Il s'agit simplement de *cliquer* en marge à gauche à la ligne désirée. Un point rouge apparaîtra ; *cliquez* à nouveau pour désélectionner.

Quoique cette technique ne sera pas beaucoup enseignée dans le cadre du cours, il importe de savoir que la syntaxe est un outil très pratique dans l'environnement SPSS et qu'elle est souvent employée pour détecter des problèmes de commandes et/ou de sous-commandes. Nous vous recommandons d'ailleurs de toujours *copier* vos commandes à partir des *menus* SPSS lorsque vous effectuerez vos travaux d'analyse.

4. SAUVEGARDER LES FICHIERS DANS SPSS

Une des opérations les plus importantes à connaître lorsqu'on travaille avec l'ordinateur, c'est sans aucun doute la sauvegarde des divers fichiers utilisés. Si vous ne sauvegardez pas régulièrement vos opérations, vous risquez de perdre votre travail, du moins en partie.Lorsque vous travaillez sur un ordinateur « privé » vous pouvez enregistrer votre

travail sur le disque dur ou sur une clé USB, à votre choix (préférablement les deux, pour plus de sûreté).

Par contre, **notez bien** que si vous travaillez sur les ordinateurs de l'Université,
vous devez absolument sauvegarder le fruit de votre travail
sur votre clé USB ou sur votre WebDépôt,
puisque vous n'avez pas accès à la mémoire centrale (disque dur) de l'Université.

PROCÉDURE POUR SAUVEGARDER UN FICHIER

Cliquez sur **Fichier** dans le menu principal en entête de SPSS

↳ **Enregistrer sous** (remplacez le nom par défault par un nom de votre choix; inutile d'ajouter l'extension .sav ou .spv qui sera ajouté automatiquement)

↳ **Enregistrer**

ATTENTION: Chaque nouvelle sauvegarde de votre travail à l'intérieur d'un fichier qui garde son nom **« écrase »** l'ancienne version !!! Soyez donc sûr de votre coup, particulièrement si vous avez fait plusieurs manipulations de variables, avant de sauvegarder un fichier nouveau ou modifié sous le nom d'un ancien fichier car vous risquez ainsi de perdre une partie du travail antérieurement réalisé, les nouvelles données prenant la place des anciennes.

Il peut être très utile de savoir sauvegarder des fichiers sous de nouveaux noms ou alors de s'en faire des copies ailleurs, comme par exemple sur un CD. Faites en sorte que les noms de fichiers que vous choisissez vous disent quelque chose. La gestion de vos fichiers (et de votre temps) en sera ainsi grandement facilitée. N'hésitez donc pas à multiplier les copies de vos fichiers, on ne sait jamais quelle catastrophe peut survenir et vous obliger à tout recommencer si vous n'avez pas de copies à jour de votre travail.

5. L'AIDE DANS SPSS

Vous vous sentez dépassé? Vous êtes mal pris? Vous paniquez? Demandez d'abord à SPSS de vous aider (avant de vous adresser aux chargées de cours ou aux assistantes). SPSS offre un assistant nommé **AIDE**, très pratique et très efficace. La plupart du temps, cette source d'information suffit à nous dépanner dans les pires circonstances.

Lorsque vous faites face à une difficulté qui vous semble insurmontable et pour laquelle le petit guide que vous avez en main et que vous avez parcouru en long et en large ne semble pas pouvoir répondre à vos questions, adressez-vous d'abord au **AIDE** de SPSS. Il se fera

un plaisir de tenter de vous répondre. Vous pouvez aussi consulter plusieurs ressources d'aide sur Internet, en *googlant* simplement votre question. Vous pouvez aussi consulter l'adresse suivante pour une formation interactive sur les principales fonctions SPSS : http://www.mapageweb.umontreal.ca/guayjea/

6. QUITTER SPSS

PROCÉDURE POUR QUITTER SPSS

Cliquez sur **Fichier** dans le menu principal en entête de SPSS
↳ *Cliquez* sur **Quitter**

Automatiquement, SPSS vous demandera successivement si vous voulez sauvegarder les modifications qui ont été faites aux divers fichiers que vous avez utilisés. En temps normal, on sauvegarde tous les fichiers utilisés, mais il arrive aussi qu'on veuille éviter enregistrer des fichiers ne contenant que des résultats erronés ou inutiles. On choisira alors de ne pas sauvegarder le fichier ou de faire le ménage avant de quitter de façon à ne conserver que les « bons résultats ». En d'autres mots, il vaut mieux prendre le temps de réfléchir avant de *cliquer* sur **OK** et de tout sauvegarder sans discernement. Ces quelques secondes de réflexion pourraient dans certains cas vous sauver beaucoup d'ennuis et de temps.

Par exemple, prenez le temps de vous demander si vous avez bel et bien effectué des transformations dans les données avant d'accepter de sauvegarder le fichier de données. Il arrive en effet qu'on « accroche » le clavier de l'ordinateur, par inadvertance, alors que la fenêtre des données est active. Cet « accrochage » pourrait éventuellement changer des données préalablement correctement inscrites dans le fichier de données.

Par contre, si vous êtes certains de vouloir sauvegarder le contenu du fichier qui vous est indiqué, alors vous n'avez qu'à répondre « oui » à l'invitation qui vous est faite de procéder de la sorte.

SPSS vous présente dès lors le nom du fichier à sauvegarder ainsi que le répertoire sur lequel il se trouve. Vous pouvez alors choisir de conserver le même nom de fichier. Dans ce cas, **SPSS écrase l'ancienne version du fichier pour la remplacer par la nouvelle** que vous venez de créer.

Vous pouvez aussi choisir de changer le nom du fichier. Il s'agit alors de remplacer la partie du nom qui précède le suffixe ***.sav*** ou ***.spv*** qui indique le type de fichier dont il s'agit, par un autre nom. Dans ce cas, l'information contenue dans le fichier original est conservée et l'information modifiée est sauvegardée sous le nouveau nom de fichier. Ceci est vrai (c'est-à-dire que l'information originelle est conservée) tant et aussi longtemps que nous n'effectuez pas une procédure de sauvegarde s'adressant au fichier original.

BANQUE DE DONNÉES DISPONIBLE POUR LES EXERCICES

BD EXERCICES : CRI 1600G - DUC

Données issues de la Déclaration uniforme de la criminalité (DUC). Il s'agit de données policières concernant tous les vols qualifiés rapportés à la police en 1999 au Québec. Les variables concernent les agresseurs, les victimes ainsi que les lieux et les montants des vols.

Nombre de variables : 17

Unités d'analyse : 8476 vols qualifiés

BANQUES DE DONNÉES DISPONIBLES POUR LES TRAVAUX

BD1 : CRI 1600G – JEUNES JUDICIARISÉS

Données issues d'une étude longitudinale québécoise effectuée auprès d'un échantillon de jeunes judiciarisés en vertu de la Loi sur les jeunes contrevenants (LJC) et/ou de la Loi sur la protection de la jeunesse (LPJ) en 1993-1994. Les données disponibles sont celles qui concernent le temps 1 seulement. Les variables portent sur l'école, le travail, la famille, les amis, la participation des jeunes à des activités délictueuses, la précocité de leur implication délinquante et la fréquence de leurs comportements déviants et délinquants.

Nombre de variables : 471
Unités d'analyse : 656 adolescents judiciarisés

BD2 : CRI 1600G – SENTENCES

Données provenant d'une recherche sur les décisions de justice. À l'été 2009, deux chercheurs criminologues on a mandaté la firme de sondage CROP pour qu'ils interrogent 200 citoyens de Montréal, choisi aléatoirement. Les citoyens sans connaissance juridique ont été interrogés sur leurs opinions sur une foule de sujet concernant la détermination de la juste peine et ils devaient évaluer trois causes criminelles (dont le descriptif se trouve dans le fichier intitulé *Complément pour la banque Sentences- Causes présentées*). Pour chaque cause, les répondants devaient faire une évaluation de la situation, ils devaient ensuite prendre une décision et opter pour la sentence de prison qu'ils souhaitaient imposer.

Nombre de variables : 203
Unités d'analyse : 200 citoyens de Montréal

LABORATOIRE 1 : CRÉER OU TRANSFORMER UN FICHIER DE DONNÉES

Objectifs d'apprentissages :

- Créer un fichier de données
- Ajouter une variable
- Nommer une variable et la définir
- Étiqueter les valeurs d'une variable
- Indiquer la présence de valeurs manquantes

La plupart du temps, lorsqu'on traite statistiquement des informations, on travaille à partir de *fichiers de données* déjà constitués. Mais il arrive aussi qu'il faille bâtir une nouvelle *banque de données*. L'opération est généralement très simple. Il faut toutefois avoir préalablement compris la façon dont se présente une *grille de données*.

Dans un *fichier de données*, chaque cas (individu, répondant, dossier...) est représenté par une *ligne* contenant les données pour un ensemble de *variables*. Les variables, pour leur part, figurent en *colonnes* (chaque variable occupe l'espace d'une colonne). La rencontre d'une variable et d'un cas se présentera donc sous forme d'une *cellule* dans laquelle l'information se trouve consignée. Exemple: si Paul a 21 ans, à l'intersection du cas « Paul » et de la variable « âge » se trouvera le chiffre « 21 ».

Par défaut, lorsque vous ouvrez SPSS, une grille de données vierge apparaît. Il s'agit alors simplement d'effectuer la *saisie des données* (ce qui veut tout simplement dire d'entrer les données dans la grille informatique prévue à cet effet). On aura alors soin de sauvegarder le fichier le plus tôt possible et de procéder à des sauvegardes régulières au fur et à mesure de l'entrée des données, de sorte qu'une panne soudaine n'aura que des effets limités (autrement désastreux).

Afin de vous déplacer dans un fichier de données, vous avez le choix procéder avec la souris 🖱, les flèches → ↑ ↓ ← se trouvant sur le sur le clavier ou alors grâce aux « ascenseurs » situés en bas et sur le côté de l'écran. Si vous voulez inscrire ou corriger des données, il s'agira alors simplement de placer le curseur dans la case appropriée et d'inscrire l'information en question. L'information apparaîtra dans la grille seulement après que vous ayez quitté cette case, soit en appuyant sur la touche **ENTER**, soit en vous déplaçant avec la souris ou les flèches vers une autre cellule ou à l'extérieur du tableau de données.

1. LA CODIFICATION DES DONNÉES

Le fichier de données comprendra uniquement des *données brutes*. Ces dernières peuvent être *numériques* (composées de chiffres ou de codes chiffrés) ou *alphanumériques* (combinaison de lettres seulement ou de lettres et de chiffres). On préférera toujours, dans la mesure du

possible, associer une *forme numérique* aux données. En effet, SPSS travaille beaucoup plus facilement avec des chiffres qu'avec des lettres.

Dans le cas des données de niveau de mesure nominal ou ordinal, il s'agit alors d'attribuer un *code numérique* aux données. Par exemple, pour indiquer le sexe d'un individu, on préférera coder les réponses de la façon suivante: garçon: "1"; fille: "2", plutôt que garçon: "G"; filles: "F". Bien entendu, pour certaines variables telles le nom, le prénom, les alias... il est quasi impossible d'imaginer une codification un tant soit peu sensée. Il s'agira donc d'une contrainte incontournable. Le traitement de telles données, dans le cadre d'analyses statistiques, est toutefois fort limité ce qui diminue les inconvénients que pose leur forme alphanumérique.

2. EXEMPLE D'UN FICHIER DE DONNÉES INFORMATIQUE

Prenons par exemple un *fichier de données* qui comprendrait les résultats obtenus au cours d'introduction aux méthodes quantitatives par un groupe de 10 étudiants. Chaque étudiant représente, en langage informatique, un *cas* ou un *enregistrement informatique*. Le fichier constitué comptera donc 10 cas ou 10 enregistrements. Le fichier comprendra, d'autre part, les données caractérisant chacun des dix cas pour neuf variables préétablies. Ces neuf variables sont les suivantes: 1) le **prénom** et 2) le **nom** de l'étudiant; 3) la **note obtenue au TP1**; 4) la **note obtenue au TP2**; 5) la **note obtenue à l'intra**; 6) la **note obtenue au TP3**; 7) la **note obtenue à l'examen final**; 8) le **sexe**; et finalement 9) la **date de naissance** de l'étudiant.

LES VARIABLES

Prénom	Nom	tp1	tp2	intra	tp3	final	sexe	date nai.
Annie	Léger	9,5	9	25	10	35	2	780116
Caroline	Pelletier	7	8	22	7	32	2	771221
Claude	Gagnon	8	8,5	24	8	34	1	760512
André	Perron	7	7,5	29	9	38	1	780721
Monique	gagné	6	6,5	29	7	33	2	681122
Marcel	Tremblay	9	9	27	6	36	1	790915
Roger	Piché	8,5	8	13	8	32	1	730508
Annie	Caron	7	7,5	19	8	39	2	800401
Michelle	Martin	8,5	7	17	8	35	2	770523
André	Séguin	6,5	9,5	19	7	30	1	580420

3. AJOUTER UN CAS OU UNE VARIABLE À UNE BANQUE DE DONNÉES DÉJÀ CONSTITUÉE

Il est toujours possible d'ajouter une *ligne* (un cas) ou une *colonne* (une variable) à un *fichier de données* déjà constitué. Pour ce faire:

PROCÉDURE POUR INSÉRER UN NOUVEAU CAS OU UNE NOUVELLE VARIABLE DANS UNE BANQUE DE DONNÉES

Placez le curseur soit sur la *ligne* qui se trouve immédiatement sous l'endroit où vous voulez que le nouveau cas apparaisse, soit sur la *colonne* située immédiatement après l'endroit où vous désirez que la nouvelle variable apparaisse.

- *Cliquez* sur **Édition** dans le menu principal
  - *Cliquez* sur **insérer une variable** ou **insérer des observations**

Vous pouvez dès lors insérer l'information se rapportant au nouveau cas ou à la nouvelle variable que vous ajoutez au tableau des données en remplissant chacune des cellules correspondant à l'emplacement où doit se trouver la nouvelle information.

4. NOMMER UNE VARIABLE

Le nom de la variable n'apparaît pas automatiquement en *entête de colonne*, il faut que vous l'inscriviez vous-même (par défaut, les variables sont appelées v00001, v00002, v00003... v0000z) ce qui complique leur reconnaissance, non pas pour la machine mais pour vous, surtout s'il vous arrive de devoir revenir au fichier de données quelques semaines, quelques mois voire quelques années plus tard. Vous préférerez alors renommer la variable de façon à pouvoir plus facilement la reconnaître.

Lorsque vous nommerez vos variables, évitez les caractères spéciaux du genre #, $, !..., (leur usage pouvant être réservé pour SPSS), et les espaces (on peut lier les deux parties du nom par un petit tiret « _ » comme dans sex_rep pour sexe du répondant).

PROCÉDURE POUR NOMMER PLUS PRÉCISÉMENT UNE VARIABLE

Mettez vous mode **affichage des variables**

- La liste de variables et de leurs caractéristiques, comme son nom, son type, sa largeur, les décimales prévues, son étiquette, ses valeurs, les valeurs manquantes, etc.
  - Vous n'avez plus qu'à compléter ou corriger chacune des cases afin de bien nommer et qualifier vos données.

5. ÉTIQUETER UNE VARIABLE

ÉTIQUETTE : Il arrive que vous vouliez joindre des *étiquettes* à vos variables afin qu'elles soient plus explicites. Par exemple, on peut vouloir signifier que la **note de l'intra** est calculée sur 30, on attachera donc cette information supplémentaire à la variable en

définissant une *étiquette*, spécifiant qu'il s'agit de la **Note de l'intra calculée sur 30**. Une *étiquette* peut prendre jusqu'à 60 caractères de toutes formes et comprendre autant d'espaces que nécessaire.

6. ÉTIQUETER LES VALEURS DE LA VARIABLE

VALEURS : Il est possible, comme dans l'exemple précédemment cité (variable SEXE), que vous ayez attribué une valeur numérique à des informations de type nominal ou ordinal (ainsi, garçon=1 et fille=2). Si c'est le cas, vous voudrez sans doute associer une étiquette à ces valeurs numériques de remplacement, histoire de vous souvenir à quelle valeur correspond en réalité chaque code numérique.

PROCÉDURE POUR ÉTIQUETTER LES VALEURS NUMÉRIQUES D'UNE VARIABLE NOMINALE OU ORDINALE[1]

Cliquez sur le petit carré gris apparaissant dans la case appropriée

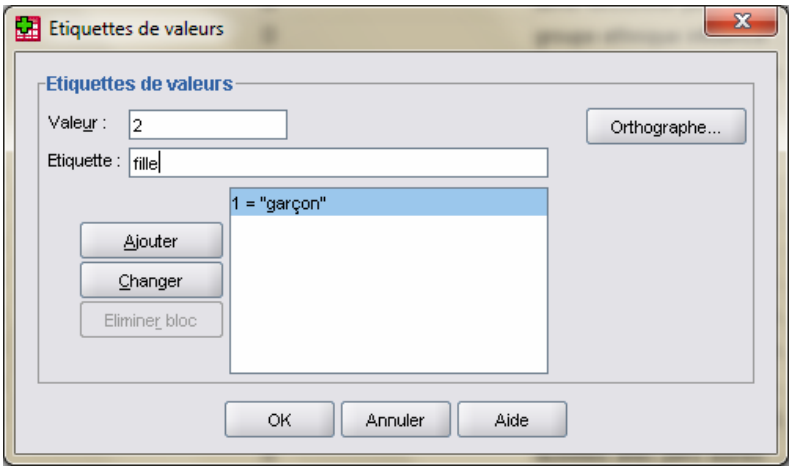
↳ Apparaît alors une *zone de dialogue* libellée étiquettes de valeurs.

↳ Il s'agit alors d'inscrire le code numérique dans le rectangle de **valeurs** et la valeur nominale correspondante dans le rectangle **étiquette de valeur**

↳ Une fois que vous avez associé la bonne valeur au chiffre correspondant, *cliquez* sur **ajouter**

Recommencez l'opération jusqu'à ce que toutes les valeurs numériques de la variable aient leurs étiquettes. N'oubliez pas d'appuyer sur **ajouter** après avoir précisé chaque étiquette.

↳ Lorsque toutes les valeurs de la variable ont reçu leurs étiquettes, *cliquez* sur **OK**



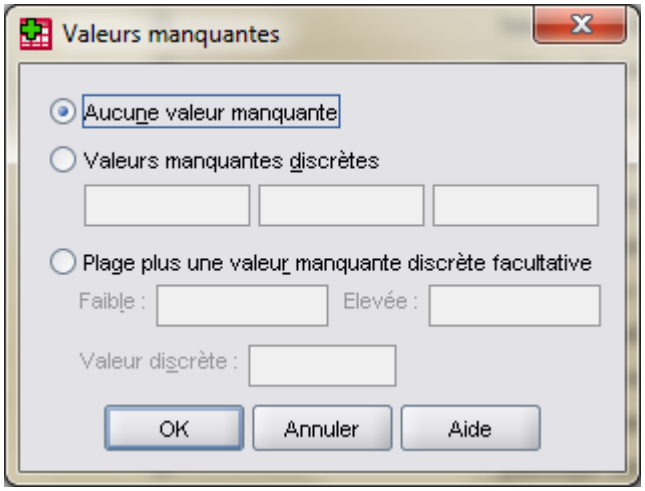
[1] Dans le cas de variables proportionnelles, comme par exemple l'âge du répondant ou le salaire familial brut calculé sur une base annuelle, il n'est pas utile de chercher à étiqueter les valeurs de la variable puisque les données portent *de facto* l'étiquette à laquelle elles correspondent (25 ans, 30 ans, 40 ans... 22,432$, 70,557$, 253,500$. L'apposition d'une étiquette précisant le nom de la variable demeure toutefois de mise.

7. INDIQUER LA PRÉSENCE DE VALEURS MANQUANTES

Lorsqu'il est question de traitements statistiques des données, une chose dont il faut s'assurer c'est que les valeurs manquantes (les répondants pour lesquels nous ne possédons pas l'information concernant une ou plusieurs variables) ne soient pas prises en compte. Ces données seront tout simplement considérées manquantes et retirées des analyses. Les valeurs manquantes sont définies et comptabilisées pour chacune des variables séparément (en effet, une donnée peut ne pas être disponible pour un répondant et pas pour un autre tout comme le fait de ne pas posséder l'information pour un répondant sur une variable ne signifie pas que l'information est manquante pour les autres variables concernant le même répondant).

PROCÉDURE POUR SIGNIFIER À SPSS LA PRÉSENCE DE VALEURS MANQUANTES POUR UNE VARIABLE

Cliquez sur le petit carré gris apparaissant dans la case appropriée
- ↳ Il apparaîtra alors une seconde *zone de dialogue*
  - ↳ SPSS vous soumet différentes options :
    - **Aucune valeur manquante**
    - **Valeurs manquantes discrètes**
    - **Plage plus une valeur manquante facultative**
    - **Valeur discrète**



- **Aucune valeur manquante** : signale qu'il n'y a pas de valeurs manquantes pour cette variable. Il s'agit de la valeur par défaut de cette sous-commande.

- **Valeurs manquantes discrètes** : trois rectangles sont ici prévus afin de permettre de définir jusqu'à trois valeurs manquantes pour une même variable. Très souvent, on choisi les valeurs 9, 99 ou 999 comme valeur manquante, à condition bien entendu que ces valeurs ne soient pas des valeurs possibles de la variable.

- **Plage plus une valeur manquante facultative** : vous permet de déclarer toutes les valeurs comprise entre une borne inférieure (*faible*) et une borne supérieure (*élevée*) comme étant autant de valeurs manquantes. Par exemple, on pourrait vouloir précises que toutes les valeurs comprises entre 75 et 99 pour la variable âge du répondant correspondent à autant de valeurs manquantes.

- **Valeur discrète** : reprend les propriétés de l'option précédente tout en permettant d'ajouter la déclaration d'une valeur discrète comme étant aussi manquante.

Une fois que vous avez constitué votre fichier de données, nommé vos variables, attribués des étiquettes plus explicites tant aux variables qu'aux catégories des variables de niveau de mesure nominal et ordinal, vous êtes prêts à passer aux traitements statistiques des données proprement dites. Les laboratoires suivants ont pour but de vous guider dans cette tâche, qu'il s'agisse de produire des distributions de fréquences, des tableaux croisés des statistiques univariées ou bivariées.

LABORATOIRE 2 : PRENDRE CONNAISSANCE DES DONNÉES ET APPRENDRE À LES MANIPULER

Objectifs d'apprentissages :

- Produire des tableaux de fréquences
- Observer les résultats
- Découvrir les valeurs manquantes et erreurs d'entrée de données
- Préparer le recodage
- Apprendre les techniques de manipulation de variables
- Éditer les résultats
- Imprimer ou transférer les résultats

Que vous ayez vous-même créé une banque de données ou que celle-ci vous ait été fournie, il s'agit maintenant d'explorer ces données, de vous les approprier suffisamment pour bien les comprendre et pour voir s'il n'y aurait pas quelques modifications à leur apporter pour qu'elles puissent être plus facilement analysées.

1. LES TABLEAUX DE FRÉQUENCES

Le tableau de fréquences est une bonne façon de présenter les informations que contient une banque de données à propos d'une variable. En effet, il indique, pour une variable donnée, toutes les valeurs que prend cette variable, le nombre de fois que chaque valeur ou catégorie de la variable apparaît et la proportion (%) que représentent les données correspondant à une catégorie d'une variable par rapport au nombre total d'observations prises en compte.

TABLEAU 1.1 EXEMPLE D'UN TABLEAU DE FRÉQUENCE ISSU DE SPSS

Judiciarisés en vertu de la Loi sur les jeunes contrevenants

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	non	341	52,0	52,1	52,1
	oui	314	47,9	47,9	100,0
	Total	655	99,8	100,0	
Manquante	Système manquant	1	,2		
	Total	656	100,0		

Dans le haut du tableau on retrouve l'étiquette de la variable pour laquelle on a demandé la production d'une distribution des fréquences (Judiciarisés en vertu de la Loi sur les jeunes contrevenants). Viennent ensuite les étiquettes attribuées à chacune des valeurs de la variable (non, oui).

Effectifs Le nombre de cas correspondant à chacune des valeurs de la variable.

Pourcentage Le pourcentages de cas que représente chacune des valeurs ou catégories de la variable par rapport à l'ensemble des cas, y compris les valeurs manquantes *(nombre de cas correspondant à une valeur ou catégorie donnée / nombre total de cas)*

Pourcentage valide Pourcentages des cas associés à chacune des valeurs ou catégories d'une variable par rapport au nombre de cas valides, c'est-à-dire à l'exclusion des valeurs manquantes *(nombre de cas correspondant à une valeur ou catégorie donnée / nombre total de cas — nombre de valeurs manquantes).*

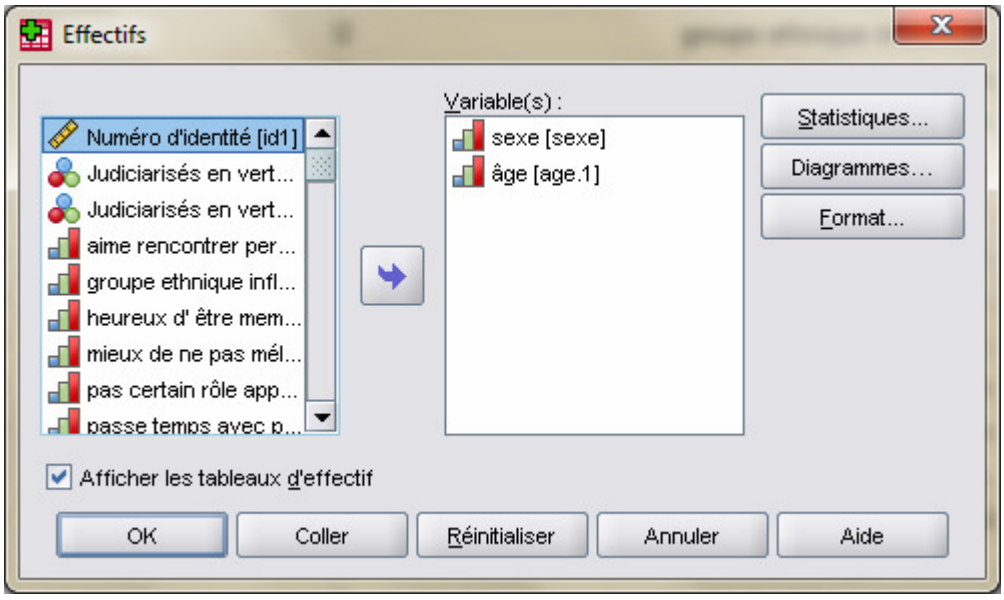
Pourcentage cumulé Pourcentage d'une catégorie d'une variable ajouté aux pourcentages que cumulent les catégories précédentes. Les pourcentages cumulés ne sont ordinairement pas pertinents lorsqu'il s'agit d'une variable de niveau de mesure nominal car, dans ce cas, il est impossible d'ordonner de quelque façon que ce soit les résultats. Par contre, dans le cas de la distribution d'une variable de niveau ordinal ou proportionnel, les pourcentages cumulés permettent de situer facilement l'emplacement d'une fraction des individus partageant une caractéristique donnée (ex: 75% des individus ont moins de 30 ans). Cette donnée ne sera toutefois intéressante que si le nombre de catégories que compte la variable à l'étude est suffisamment important.

PROCÉDURE POUR CRÉER UN TABLEAU DE FRÉQUENCES

Cliquez sur **Analyse** dans le menu principal

- *Cliquez* sur **Statistiques descriptives**
  - *Cliquez* sur **Effectifs**
    - *Sélectionnez* la variable pour laquelle vous désirez obtenir une distribution de fréquences en *cliquant* sur son nom tel qu'il apparaît dans le rectangle de gauche (lequel présente la liste des variables à l'étude, par ordre alphabétique).
      Si le nom de la variable n'apparaît pas à l'écran, déplacez-vous dans le rectangle grâce à « l'ascenseur » à droite de l'écran, la variable recherchée est certainement située plus haut ou plus bas dans la fenêtre.
      - *Cliquez* maintenant sur la flèche située entre les deux rectangles. Vous verrez alors le nom de la variable « sauter » du rectangle de droite au rectangle de gauche.
        - *Cliquez* sur **OK**

En format SPSS, l'écran de dialogue qui apparaît est le suivant, alors qu'on demande la distribution des fréquences de la variable « sexe » et de la variable « âge » laquelle est passée du rectangle de gauche, contenant les liste exhaustive des variables contenues dans le fichier de données, dans le rectangle de droite, lequel précise les variables pour lesquelles une distribution de fréquences est demandée.



N.B. Cette manière de sélectionner les variables sur lesquelles on veut effectuer des opérations (en les faisant passer d'un rectangle à l'autre) est la même partout dans SPSS, souvenez-vous-en.

Notez aussi que vous pouvez *sélectionner* plusieurs variables à la fois en retenant le **bouton Ctrl** et en déplaçant votre curseur vers le haut ou vers le bas de l'écran jusqu'à ce que l'ensemble des variables qu'on souhaite soumettre à l'analyse soient sélectionnées.

Mais surtout, notez qu'à chaque fois que vous effectuez une nouvelle opération, il faut commencer par RANGER LES ANCIENNES VARIABLES en les sélectionnant dans le rectangle de droite et en les envoyant dans le rectangle de gauche (ou en *cliquant* sur le bouton **Réinitialiser** ce qui a pour effet de ramener l'ensemble des variables à leur emplacement initial. Cela vous évitera de nombreux ennuis, dont celui d'accumuler inutilement des résultats reproduits plusieurs fois et évitera à SPSS de travailler pour rien.

2. COMMENT ÉDITER LES RÉSULTATS

Une fois que vous avez fait exécuter la commande de tableaux de fréquences en *cliquant* sur **OK,** la fenêtre*Résultats1 :Viewer SPSS* apparaît automatiquement devant vos yeux. Normalement, vous devriez avoir devant vous un fichier contenant le ou les tableaux de fréquences demandé(s).

Tout d'abord, il importe de savoir que la feuille des résultats des opérations que vous avez demandées à SPSS d'exécuter, se présente toujours par le début. Il s'agit donc d'utiliser « l'ascenseur » du côté droit de l'écran afin soit de se positionner n'importe où, selon le besoin.

Si vous voulez effacer ou déplacer une partie du texte se trouvant dans le *listing*, il faut d'abord le sélectionner.

3. COMMENT IMPRIMER DES RÉSULTATS

Une fois les résultats mis en forme à votre goût, certains d'entre vous préféreront travailler sur du papier et d'autres voudront peut-être transférer le tout dans un traitement de texte plus classique afin d'y intégrer des commentaires, de décrire les résultats ou encore de fignoler la présentation. Rappelez-vous que si SPSS est fort puissant pour ce qui est de calculer des statistiques, ses capacités en matière de traitement de texte sont par contre des plus limitées.

PROCÉDURE POUR IMPRIMER DIRECTEMENT LE FICHIER DE RÉSULTATS

À partir du menu d'entête de SPSS, choisir la commande
Cliquez sur **Fichier** dans le menu principal
↳ *Cliquez* sur **Imprimer** (choisir les options d'impression désirées)
↳ *Cliquez* sur **OK**

Parfois, il est plus intéressant de travailler à partir d'un logiciel de traitement de texte si vous voulez insérer du texte (l'interprétation des résultats par exemple), ou encore présenter vos résultats de manière différente. Pour ce faire, nous suggérons soit de faire une simple « copié-collé », soit de refaire entièrement vos tableaux sur *Word*, si vous désirez avoir encore davantage de latitude.

4. QUOI OBSERVER DANS LES TABLEAUX DE FRÉQUENCES

Une fois que vous avez les tableaux de fréquences en mains, prenez le temps de regarder rapidement chacun d'entre eux. De quelles variables s'agit-il? Pour combien de cas «valides» possédons-nous l'information? Quel est le niveau de mesure de la variable : est-elle nominale, ordinale ou quantitative? La distribution des valeurs de la variable vous paraît-elle conforme (y a-t-il des données qui apparaissent et qui ne devraient pas apparaître)?

Lors de cette première exploration des tableaux de fréquences (qui vous permet en quelque sorte de prendre le pouls de vos données et de vous rassurer sur leur « qualité » au moins apparente), il vous faudra procéder systématiquement en notant *manuellement*, sur une feuille à part, toutes les modifications que vous souhaiterez apporter à l'organisation des données : élimination de certaines variables, recodification d'un certain nombre d'autres, identification des valeurs étranges que prennent certaines d'entre elles, etc.

4.1. LES VALEURS MANQUANTES

Il arrive dans certains cas que toute l'information concernant tous les individus pour l'ensemble des variables ou encore pour une variable en particulier ne soit pas disponible. Dans ce cas, on dira que l'information est manquante. Un code sera alors utilisé pour en rendre compte. Ce code sert précisément à identifier les *valeurs manquantes*.

C'est à l'utilisateur de choisir le code qui lui convient pour identifier les *valeurs manquantes* à l'intérieur de sa banque de données. Par convention, le chiffre « 9 » (ou « 99 » ou « 999 », selon le cas) est le plus souvent utilisé.

On peut aussi prévoir différents types de valeurs manquantes et leur attribuer différents codes permettant de les distinguer. Par exemple, on attribuera la valeur « 7 » ou « NSP » à

ceux qui, en réponse à une question, indiquent « *ne sait pas* »; la valeur « 8 » à ceux qui *refusent de répondre*, et la valeur « 9 » pour toute *information vraiment manquante.*

Il arrive très souvent que l'on trouve des valeurs manquantes dans la distribution des données, mais que l'ordinateur ne le sache pas initialement. Il faut donc que vous observiez chaque tableau en vous demandant s'il y a des valeurs manquantes à signifier à SPSS. Vous reconnaîtrez les valeurs manquantes par leur code numérique (9, 99, 999,...), par l'étiquette qui leur est attachée (pas de réponse; NSP.; ne s'applique pas; indéterminé, ...) ou par l'absence d'annotation.

Il est important, au moment de traiter les fréquences des réponses, de faire le bilan des valeurs manquantes. Une variable qui présente un trop grand nombre de valeurs manquantes pourra être considérée non valide. On peut, au contraire, décider que le nombre de valeurs manquantes n'est pas suffisant pour invalider le traitement de la variable. Toutefois, on ne voudra pas tenir compte des valeurs manquantes dans les analyses. Il faudra alors indiquer à SPSS de ne pas tenir compte, au moment des analyses statistiques, des individus pour lesquels l'information n'est pas disponible. Nous verrons comment plus loin.

N.B.: La catégorie résiduelle « **autres** » ne correspond pas à des valeurs manquantes.
Elle indique plutôt que l'information est autre que celle prévue dans les catégories de réponses préétablies

4.2. LES ERREURS ÉVIDENTES

Il arrive aussi parfois que l'on trouve certaines absurdités dans les tableaux de fréquences et qu'elles soient dues à des erreurs dans la saisie des données. C'est le moment de noter la présence de ces *données erratiques*, lesquelles seront soit corrigées directement dans la banque de données, soit associées aux valeurs manquantes si on estime que cela ne remet nullement en cause la valeur des analyses.

Par exemple, il arrive que l'on se retrouve devant une valeur bizarre et incompréhensible, comme c'est le cas dans le tableau suivant pour la valeur « 6 » qui à tout a fait l'air d'être une erreur d'entrée de données. L'indice qu'il s'agit d'une *donnée erratique* est souvent la fréquence très peu élevée qui lui est associée. On devra s'inquiéter des *valeurs erratiques* qui apparaissent en trop grand nombre. Peut-être est-ce simplement qu'on a oublié d'associer une étiquette à une valeur qui, par ailleurs, représente vraiment un « choix de réponse » possible. Il s'agit de vérifier ce qu'il en est.

Statut de l'agresseur

	Valeur	Effectifs	%	% valide	% cumulé
Suspect	1	12	12.0	12.0	12.0
Accusé	2	22	22.0	22.0	36.0
Condamné	3	43	43.0	43.0	79.0
	6	**1**	**1.0**	**1.0**	**80.0**
N.S.P.	9	20	20.0	20.0	100.0
	Total	100	100.0	100.0	

Valeurs manquantes : 0

La logique aussi peut nous aider à détecter des erreurs, comme si dans un tableau représentant l'âge des meurtriers, on retrouve la valeur « 2 ». Puisqu'il est fort peu probable que l'un des tueurs apparaissant dans la distribution des données ait vraiment eu deux ans au moment de l'événement, on préférera inclure ce cas dans les valeurs manquantes ou, éventuellement, corriger l'information dans le fichier des données, pour peu qu'on puisse retrouver la bonne donnée.

4.3. LES CATÉGORIES À RETRAVAILLER

Une autre démarche à partir de la distribution de fréquences concerne des aménagements de catégories à faire pour que les données se présentent mieux. Il existe deux grandes familles de catégorisations : celles concernant les variables nominales ou ordinales et celles associées aux variables proportionnelles.

Les variables ordinales ou nominales

Ces variables sont déjà considérées comme des variables « catégorielles ». En effet, leurs valeurs se présentent sous forme de catégories, généralement établies par ceux qui ont fait le questionnaire. Normalement, dans ce cas, on ne devrait pas avoir à refaire la catégorisation. Mais il arrive qu'on veuille malgré tout procéder à quelques réorganisations de telles données. Il pourrait s'agir de regrouper des catégories qui semblent être parentes ou de limiter le nombre des catégories à analyser.

Par exemple, on pourrait vouloir faire seulement 3 catégories avec la liste de délits représentée au tableau suivant. On regrouperait alors les catégories « 1 » homicide + « 2 » agression sexuelle + « 3 » vol qualifié sous la grande rubrique: « 1 » délits contre la personne; les catégories « 4 » introduction par effraction + « 5 » vol de moins de 1000$ » + « 6 » vol de plus de 1000$ sous la grande rubrique: « 2 » délits contre la propriété; et finalement les catégories « 7 » possession de stupéfiants + « 8 » prostitution sous la grande rubrique: « 3 » délits sans victimes.

Délit dont est accusé le suspect			**Délit du suspect (3 catégories)***	
Étiquette	*Valeur*		*Étiquette*	*Valeur*
Homicide	1	➔	Délits contre la personne	1
Agression sexuelle	2			
Vol qualifié	4			
Introduction par effraction	5	➔	Délits contre la propriété	2
Vol de moins de 1000$	6			
Vol de plus de 1000$	7			
Possession de stupéfiants	8	➔	Délits "sans victimes"	3
Prostitution	9			

⇒ **Délits contre la personne** comprenant : homicide, agression sexuelle, vol qualifié;
⇒ **Délits contre les biens** comprenant: introduction par effraction, vol de plus et vol de moins de 1000$;
⇒ **Délits « sans victimes »** comprenant : possession de stupéfiants et prostitution**.**

Les variables proportionnelles

Les variables proportionnelles, pour leur part, présentent habituellement un tel nombre de valeurs qu'il faut, presque toujours, les regrouper en un nombre limité de catégories **si on veut en présenter un tableau de fréquences lisible.**

Vous constaterez en effet que le tableau de fréquences d'une variable proportionnelle est souvent extrêmement long, sa présentation peut occuper jusqu'à plusieurs pages, et qu'il est assez peu instructif, l'information se trouvant grandement éparpillée. Par contre, en créant des catégories plus larges, on règle souvent le problème de présentation et de consultation rapide du matériel.

RAPPEL : 4 types de regroupements sont possibles pour les variables proportionnelles

1. En classes de **largeurs égales**;
2. En classes **d'inégales largeurs** respectant **un principe conceptuel**;
3. En classes **d'inégales largeurs** constituées autour de **points de concentration** des données;
4. En classes **d'inégales largeurs** constituées pour **respecter un modèle préétabli** d'analyse.

Ces procédures de recodage ne sont par ailleurs pas nécessairement mutuellement exclusives. Il est en effet possible de combiner plusieurs d'entre elles. Ainsi, une catégorisation, même en classes d'égales largeurs, pourrait répondre à des caractéristiques conceptuelles... etc.

Dans tous les cas, la catégorisation d'une variable devra répondre à deux impératifs:

⇒ Les catégories doivent être **mutuellement exclusives** ET **exhaustives**.

DEUX REMARQUES IMPORTANTES:

⇒ Ne perdez pas votre temps à catégoriser des variables qui ne seront pas utiles pour la suite des analyses.

⇒ Rappelez-vous toujours que les analyses statistiques se pratiquent sur les données brutes (non regroupées). Le groupement en catégories sert essentiellement à des fins de présentation.

5. PRÉPARER LE RECODAGE

Vous devez maintenant préparer tous les changements à effectuer sur vos données avant de passer aux analyses définitives. Il s'agit tout simplement de noter, à propos de toutes les variables *pertinentes*, les transformations que vous désirez effectuer.

Demandez-vous:

⇒ Cette variable m'intéresse-t-elle?
⇒ Y a-t-il des valeurs manquantes non déclarées?
⇒ Y a-t-il des valeurs absurdes?
⇒ Faut-il organiser une nouvelle catégorisation (un recodage)?

C'est ici la base de tout travail d'analyse de qualité.

Voici à quoi pourrait ressembler l'utilisation d'une distribution de fréquences pour fins de nettoyage et de recodage (notification normalement faites manuellement):

Arme utilisée par l'agresseur

ARME	Valeur	Eff.	%	valid %	cum %	
Automatique	1	9	,6	,6	,6	
Canon scié	2	22	1,4	1,4	2,0	
Arme de poing	3	157	10,0	10,0	11,9	= 1:
Arme à feu						
Carabine/fusil	4	9	,6	,6	12,5	
Autre arme à feu	5	23	1,5	1,5	14,0	
Couteau	6	203	12,9	12,9	26,9	
Autre tranchant/pointu	7	26	1,7	1,7	28,5	
Objet contondant	8	90	5,7	5,7	34,2	=2:
Autre armes						
Explosifs	9	1	,1	,1	34,3	
Feu	10	3	,2	,2	34,5	
Force physique	11	989	62,8	62,8	97,3	= 3:
Pas autre arme	12	30	1,9	1,9	99,2	= 2
Menace	13	9	,5	,5	99,7	= 3
	43	1	,1	,1	99,8	erreur
Inconnu=missing	98	1	,1	,1	99,9	missing
N.S.P.	99	1	,1	,1	100,0	missing
		-------	-------	-------		
Total		1574	100,0	100,0		

Cas valides 1574 Cas manquants 0

Voici de quoi pourrait avoir l'air le *tableau de fréquence* de la même variable, une fois celle-ci retravaillée (ménage et recodage fait à partir de SPSS):

Arme utilisée par l'agresseur

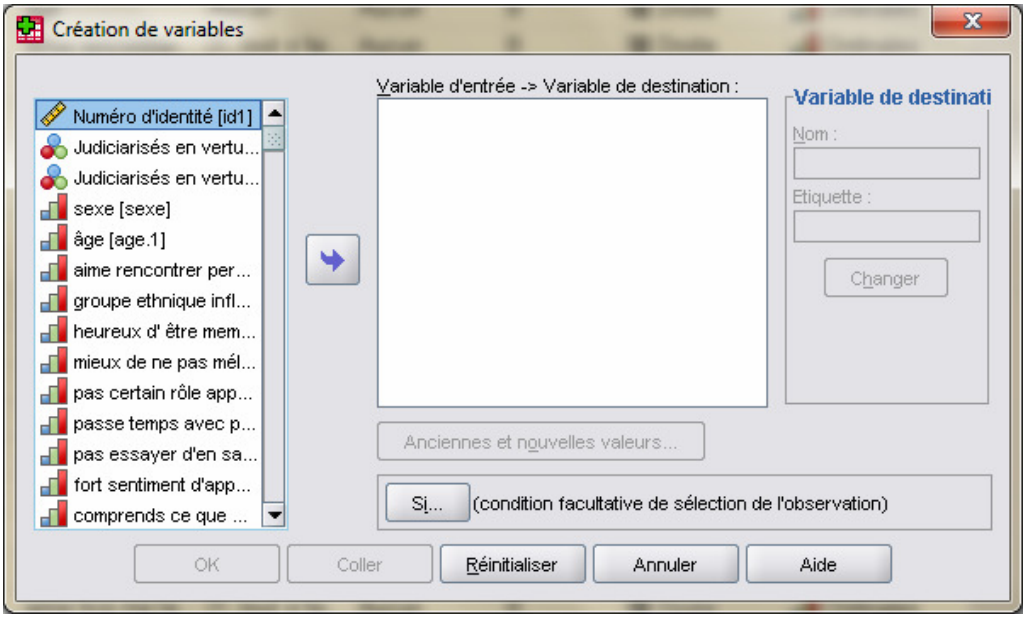
ARME	Valeur	Eff.	%	valid %	cum %
Armes à feu	1	220	14,0	14,0	14,0
Autres armes	2	353	22,5	22,5	36,5
Pas d'armes utilisées	3	998	63,5	63,5	100,0
		-------	-------	-------	
Total		1571	100,0		
Valeurs manquantes	3 (0,2%)				

6. RECODER OU TRANSFORMER LES VARIABLES

Il s'agit maintenant d'effectuer toutes les transformations sur les données, transformations que vous avez, bien entendu, minutieusement préparées! Voyons comment s'y prendre pour transformer la catégorisation initiale d'une variable, ou, en langage SPSS, comment "**recoder**" une variable.

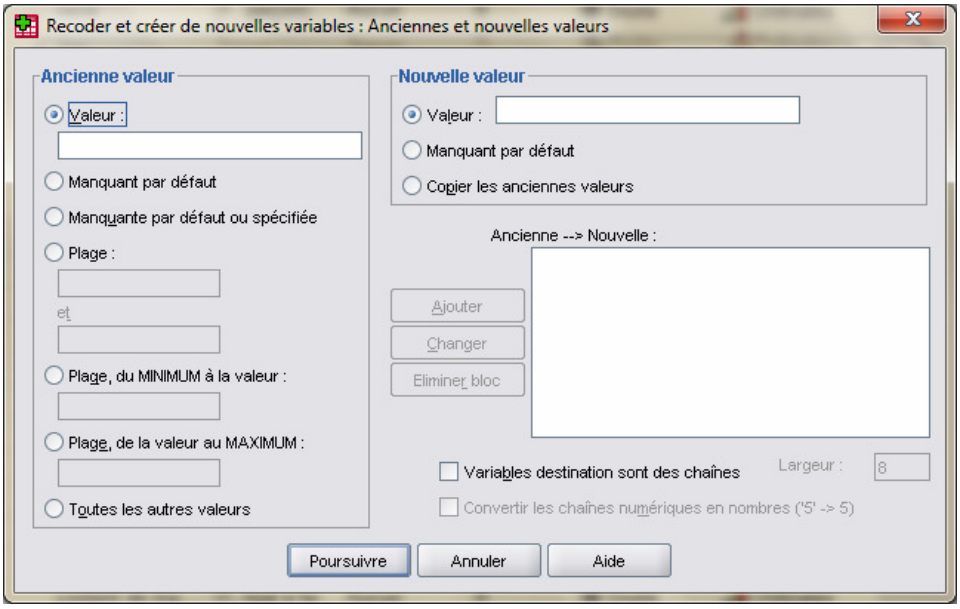
PROCÉDURE POUR CRÉER[2] UNE VARIABLE

Cliquez sur **Transformer** à partir du menu d'entête
- *Cliquez* sur **Création de variable**
  - *Sélectionnez* la variable à transformer dans le rectangle de gauche
    - *Cliquez* sur la flèche
      - Le nom de la variable sélectionnée apparaîtra dans le rectangle **Variable entrée --> Variable de destination**
        Inscrivez dans le rectangle de droite le nom de la nouvelle variable crée, ainsi que son étiquette (ex. infraction en trois catégories)
        - *Cliquez* sur **Réinitialiser**
          - *Cliquez* sur **Anciennes et nouvelles valeurs**



[2] L'option **RECODAGE DE VARIABLE** est aussi disponible. Toutefois, **nous suggérons fortement de ne jamais recoder la variable initiale en tant que telle, car une fois les transformations effectuées, il n'est plus possible de revenir en arrière et de retrouver les valeurs telles qu'initialement colligées**. Il vaut donc mieux garder une version intacte de la variable initiale et ne travailler qu'avec des « duplicata » pour tout ce qui touche les transformations des données.

Une nouvelle zone de dialogue apparaît :



⇒ *Inscrivez*, dans le rectangle de gauche, la ou les valeur(s) numérique(s) que vous voulez changer :

➔ Une valeur unique
➔ Une valeur manquante par défaut
➔ Une valeur manquante par défaut ou spécifiée
➔ Une plage de valeurs (ex: les valeurs 1 à 5----> 1 à 5)

⇒ *Associez*, dans le rectangle de droite, la nouvelle valeur (nouvelle: 1) à une ou plusieurs anciennes valeurs

↳ *Cliquez* sur **Ajouter.**

↳*Recommencez* les deux dernières opérations jusqu'à ce que vous ayez traité toutes les valeurs à recoder.

Attention, si certaines valeurs ne doivent pas être recodées, il n'est pas nécessaire de faire correspondre anciennes et nouvelles valeurs qui sont alors équivalentes. On doit toutefois préciser à SPSS que pour toutes ces valeurs *(toutes les autres valeurs* dans le rectangle de gauche en bas), la valeur de la variable demeure la même, ceci en lui précisant de les recopier (**copier les anciennes valeurs***)* dans le rectangle de droite en haut de l'écran).

Une fois toutes les opérations de transformation réalisées, terminez la procédure de la façon suivante:

↳ *Cliquez* sur **Ajouter** (pour enregistrer la dernière commande de modification)

↳ *Cliquez* sur **Poursuivre**

↳ *Cliquez* sur **OK**

MISE EN GARDE

- Appliquez la recodification d'une variable de préférence à la variable "nettoyée", i.e. celle où les valeurs manquantes sont déjà définies et les *données erratiques* corrigées ou éliminées en les associant aux valeurs manquantes, lorsque que vous procédez à la transformation d'une variable. Sinon, n'oubliez pas de redéfinir (ou rappeler) à SPSS la présence des valeurs manquantes.

- Une valeur ne peut être *recodée* plus d'une fois au cours d'une même procédure.

- Toutefois on peut appliquer à une même variable toute une série de recodages, en autant qu'on crée autant de nouvelles variables qu'il y a de recodages différents d'envisagés pour cette même variable.

- N'oubliez pas, immédiatement après le recodage, d'aller vérifier si les transformations ont effectivement eu lieu. Vous les verrez apparaître dans le fichier de données. Mais ne vous attendez pas à les voir apparaître dans un tableau de fréquences tant et aussi longtemps que vous n'avez pas commandé un tableau de fréquences pour regarder les nouvelles variables.

- Procédez immédiatement à l'étiquetage des valeurs des nouvelles variables. (labo 1).

7. CALCULER OU CRÉER DE NOUVELLES VARIABLES

La commande **Calculer** permet de créer une nouvelle variable, à partir d'une variable déjà existante, pour tous les cas d'une banque de données.

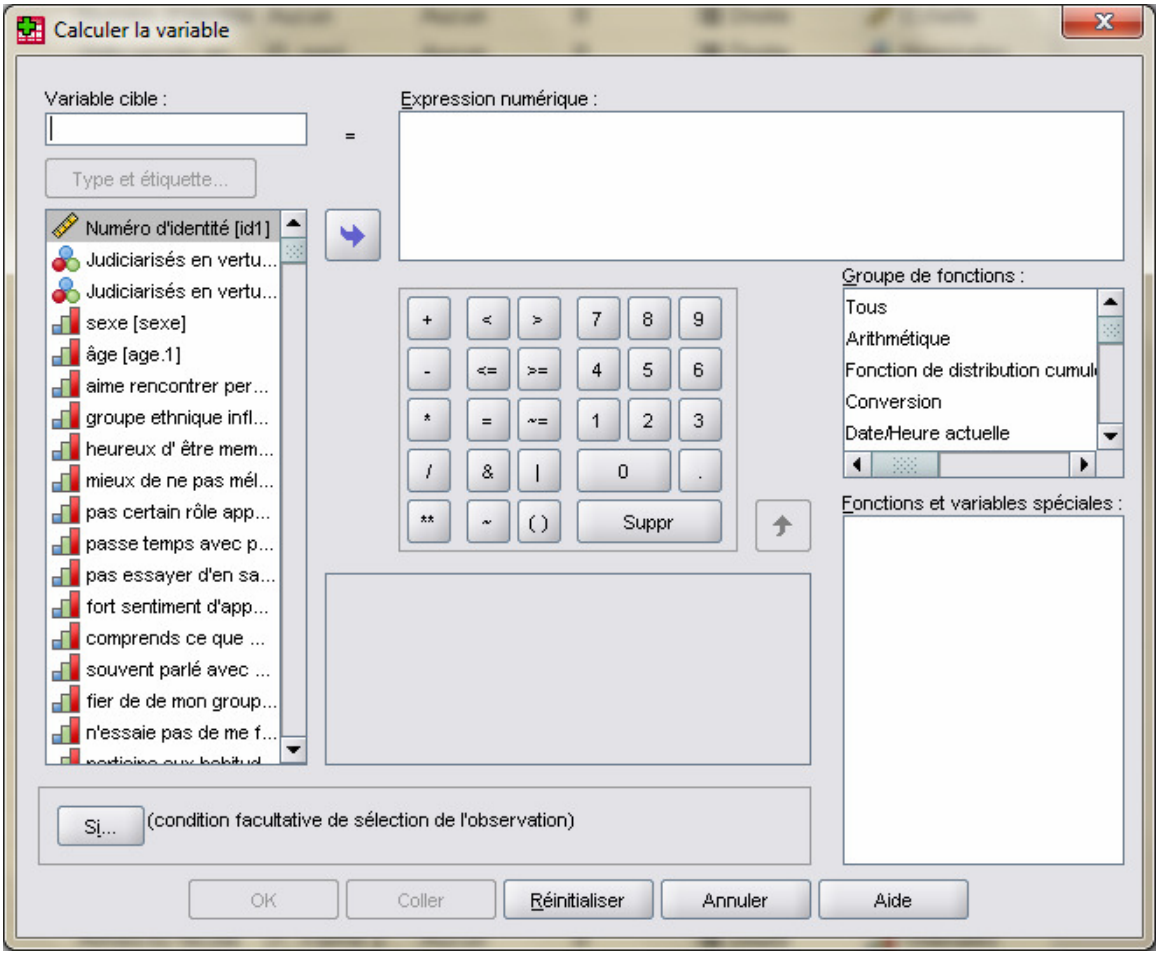
Elle permet, par exemple, de créer, à partir d'une banque de données incluant toutes les notes d'un étudiant dans un cours donné, une nouvelle variable qui correspond à la somme de toutes ces notes, donc son résultat final (resultat = tp1 + tp2).

Il s'agit là d'un exemple très simple de ce que l'on peut faire avec la commande **Calculer**. Dites-vous que la commande **Calculer** peut effectuer à peu près n'importe quelle opération mathématique que l'on souhaite appliquer aux données. Voyons brièvement comment s'y prendre dans les cas les plus simples.

PROCÉDURE POUR CRÉER UNE NOUVELLE VARIABLE À L'AIDE DE LA COMMANDE CALCULER

Cliquez sur **Transformer** à partir du menu d'entête
↳ *Cliquez* sur **Calculer la variable**

Un écran de dialogue apparaît :



⇒ *Inscrivez*, le nom de la nouvelle variable dans le rectangle (**Variable cible**) situé en haut à gauche de l'écran de dialogue.

⇒ *Cliquez* sur **Type et étiquette** pour apposer une étiquette à cette nouvelle variable.

⇒ *Écrivez*, à droite, **l'expression numérique** permettant la création de cette nouvelle variable (vous pouvez soit l'écrire grâce au clavier d'ordinateur, soit utiliser la zone de dialogue en sélectionnant les variables à sa gauche, les opérateurs et les fonctions en-dessous et en cliquant successivement sur les touches correspondant aux données nécessaires à la création de la nouvelle variable).

⇒ *Cliquez* sur **OK**

Il existe une infinité de possibilités pour créer de nouvelles variables à partir d'opérations plus ou moins complexes. Nous nous contenterons d'indiquer ici la façon d'inscrire les principaux opérateurs mathématiques ainsi que quelques fonctions de transformations parmi

les plus courantes, rappelez-vous seulement qu'opérateurs et fonctions peuvent être combinés à volonté.

Opérateurs:	**+** : addition	**<, >** :	plus petit, plus grand;
	- : soustraction	**<=, >=** :	plus petit ou égal, plus grand ou égal.
	***** : multiplication	**&** :	et
	/ : division	**|** :	ou
	** : exponentiel		

Quelques fonctions:	**Abs:**	valeur absolue
	Mean:	moyenne
	Squt:	racine carrée
	Rnd:	arrondir au nombre entier
	Sum:	faire la somme
	Yrmoda:	calculer la période entre deux dates

N.B. L'ordre d'exécution des opérations arithmétiques dans SPSS ne diffère pas de celui que vous avez appris dans vos cours de mathématiques:

- exponentiels en premier lieu;
- multiplications et divisions en second lieu;
- additions et soustractions en troisième lieu.

Il est important de se le rappeler sinon on risque de ne pas obtenir les résultats souhaités. Seules les parenthèses peuvent modifier l'ordre d'exécution des opérations.

L'OPTION *SI*:

L'option **SI**, qui se présente sous la forme d'un rectangle situé juste en dessous des opérations et des fonctions de plusieurs *commandes*, permet de créer, de recoder ou d'associer une valeur à une variable **SOUS CERTAINES CONDITIONS**.

On retrouve l'option **SI** pour les commandes *recoder, création de variable* ou *calculer* et bien d'autres (que vous verrez ultérieurement).

La procédure est toujours la même ou fort approchée :

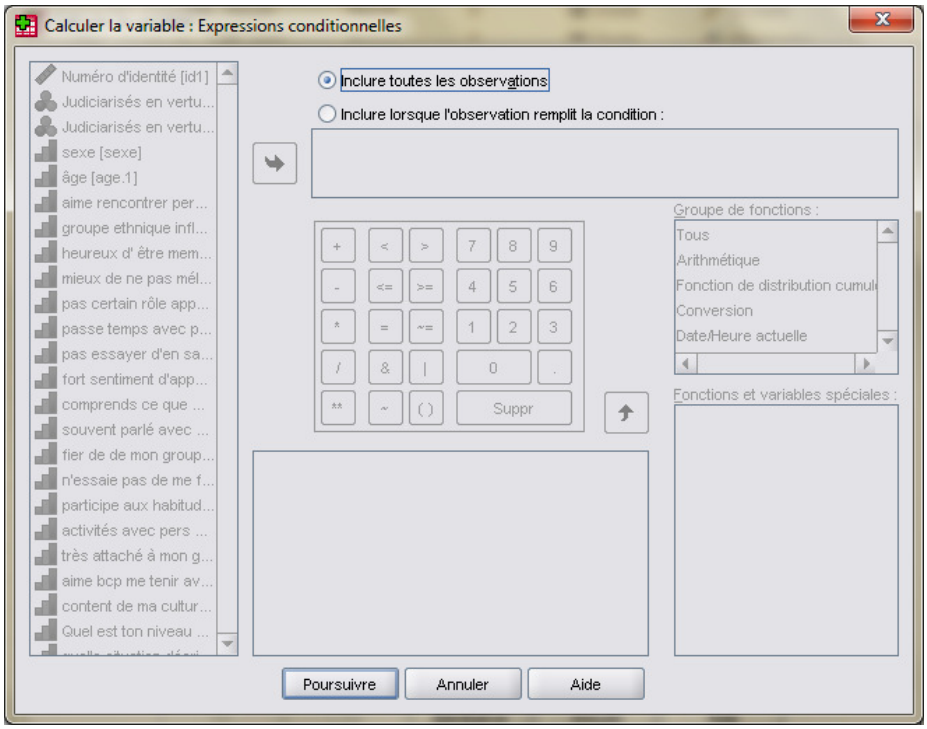
PROCÉDURE POUR IMPOSER UNE CONDITION À UNE COMMANDE

Atteindre la commande pour laquelle on veut imposer une condition d'exécution (recode, calculer, etc.)

↳ *Repérez* l'option **SI** dans l'écran de dialogue correspondant

↳ *Cliquez* sur **SI**

Un nouvel écran de dialogue apparaît lequel revêt plus ou moins la forme suivante vous permettant de préciser les conditions à appliquer à la commande (recode, calculer...) en cours:



⇒ *Choisir* l'option ⊙ Inclure lorsque l'observation remplit la condition :

Le rectange gris devient actif et vous permet d'entrer, *manuellement*, en vous servant du clavier de l'ordinateur, ou à l'aide des clés présentant les opérateurs mathématiques ou les fonctions nécessaires, la définition de la condition imposée.

Celle-ci peut-être fort simple (ex: **recode SI** SEXE_VIC=1)
ou fort complexe (ex: **Calculer** = (TP1+TP2+TP3 + EXAM1 + EXAM2)
SI TP1+TP2+TP3*.50) >=50 | (EXAM1 + EXAM2/2)**2 >= (TP1+TP2+TP3)

↳ *Cliquez* sur **Poursuivre**

↳ *Cliquez* sur **OK**

LABORATOIRE 3 : TRAVAILLER SUR UN ENSEMBLE SÉLECTIONNÉ DE DONNÉES

Objectifs d'apprentissages :

- Sélectionner des observations
- Créer un échantillon

1. SÉLECTIONNER UNE SOUS-POPULATION

La commande **Sélectionner des observations** permet d'effectuer une sélection temporaire ou permanente de certaines données sur lesquelles doivent porter plus spécifiquement une partie des analyses. **Une fois qu'un sous-ensemble de la population ou de l'échantillon de départ a été sélectionné, toutes les commandes suivantes ne concernent que ce sous-ensemble de données et ce, jusqu'à ce qu'une nouvelle commande indique de revenir à la population ou à l'échantillon entier**.

Cette commande sert soit à effectuer des analyses uniquement sur une partie spécifique des cas issus de l'ensemble initial des données, c'est-à-dire une sous-population ou un sous-échantillon selon que l'ensemble de départ constitue une population ou un échantillon; ou alors quand on désire utiliser un critère comme *variable contrôle*.

Exemple 1 :
Recherche s'appliquant entièrement à un sous-ensemble de données

Prenons, par exemple, une recherche s'intéressant aux facteurs qui influencent la criminalité des mineurs. Admettons que la banque de données à laquelle vous avez accès initialement contienne à la fois des données concernant des adultes et des mineurs. Vous ne vous intéressez qu'aux mineurs contenus dans la banque de données. Il suffit alors de demander à SPSS d'écarter définitivement tous les cas qui ont 18 ans ou plus, à l'aide de l'option ***Supprimées*** (voir plus bas), ce qui permet de créer une nouvelle banque de données uniquement composée des mineurs du groupe initial.

La condition imposée serait alors **Selon une condition logique SI : AGE < 18**
(la variable AGE représentant l'âge des délinquants)

ATTENTION

On prendra la précaution de sauvegarder le fichier constitué uniquement des données concernant les mineurs sous un **nouveau nom**. Ceci a pour effet de conserver le fichier de données initial intact et de créer un nouveau fichier de données contenant uniquement celles pertinentes à l'étude en cours, ici une étude portant sur la criminalité des mineurs. Ceci fait en sorte qu'advenant qu'on veuille reproduire l'étude pour les adultes, ou encore qu'on souhaite comparer ce qui se passe dans le cas des adultes par rapport aux mineurs, on possède toujours l'information nécessaire dans le fichier de données d'origine.

Exemple 2 :
Recherche qui utilise un sous-groupe de la population ou de l'échantillon initial pour une partie seulement des analyses

L'étude que vous entreprenez consiste à dresser le portrait des populations en institutions carcérales. Vient le moment d'analyser les données concernant la nature (peine pour non-paiement d'amende, sentence de juridiction provinciale ou sentence de juridiction fédérale) et la durée (en nombre de jours) des peines imposées. On sait que les prisons du Québec accueillent à la fois des prévenus (des suspects en attente de verdict ou de sentence) et des condamnés (individus reconnus coupables qui se voient imposer une peine de prison).

La question de la durée des sentences concerne uniquement les individus condamnés (les autres ne connaissant pas encore le verdict ou la sentence dans leur cause). Aussi, cette partie des analyses, spécifiquement, ne portera que sur ceux déjà condamnés. Il s'agit alors en quelque sorte de neutraliser la partie de la population ou de l'échantillon de départ qui ne doit pas être prise en compte pour cette partie précise des analyses. On effectuera alors une sélection temporaire (option **Filtrées**, voir plus bas) des cas devant être pris en compte à cette étape de l'analyse.

REMARQUE

Au moment d'interpréter les résultats, il faudra prendre la précaution de préciser que les analyses portent sur une partie seulement des données, celles concernées par la variable à l'étude. Ainsi, par exemple, on écrira que: « pour les personnes admises dans les institutions carcérales du Québec et condamnées à une peine de prison, la durée moyenne de la peine est de ... ».

Exemple 3 :
Recherche qui utilise une variable contrôle

On vous demande de vérifier l'hypothèse selon laquelle le mariage serait pour les femmes un facteur important dans la décision de cesser une carrière criminelle mais pas pour les hommes (*variable contrôle*: sexe de l'individu). Il s'agit alors d'utiliser une banque de données comprenant à la fois des hommes et des femmes néanmoins analysés séparément. Ici, il s'agit, dans un premier temps, d'effectuer les analyses sur le sous-groupe (sous-population ou sous-échantillon) composé uniquement des femmes pour ensuite refaire ces analyses sur le sous-groupe (sous-population ou sous-échantillon) regroupant spécifiquement les hommes.

Schéma d'analyse:

1ère condition imposée : *Sélectionner selon une condition logique SI* SEXE = 2
(Si la valeur "1" représente les femmes)

1ère série de traitements statistiques concernant le sous-groupe des femmes.

2ème condition imposée : *Sélectionner selon une condition logique SI* SEXE = 1
(Si la valeur "2" représente les hommes)

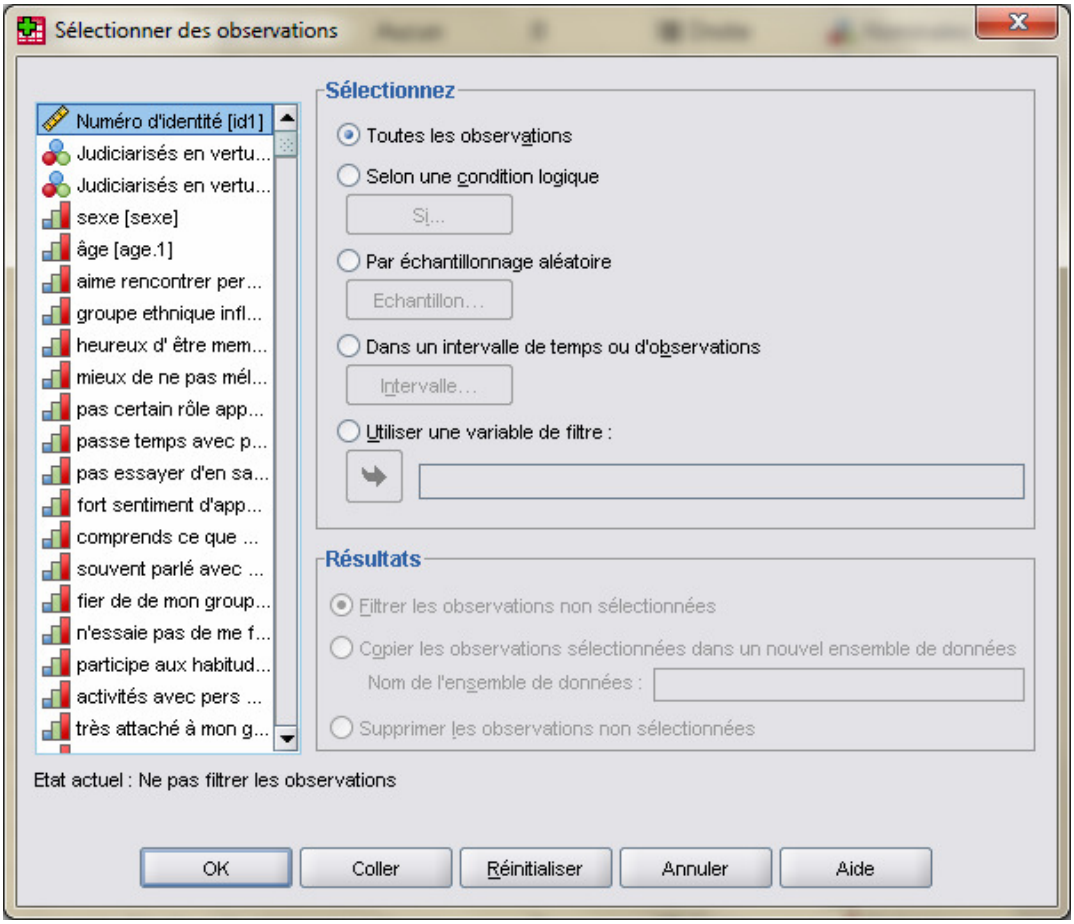
2ième série de traitements statistiques (équivalents aux premiers) concernant cette fois le sous-groupe des hommes.

Comparaison des données obtenues dans les deux cas.

PROCÉDURE POUR SÉLECTIONNER UN SOUS-GROUPE (TIRÉ D'UNE POPULATION OU D'UN ÉCHANTILLON INITIAL)

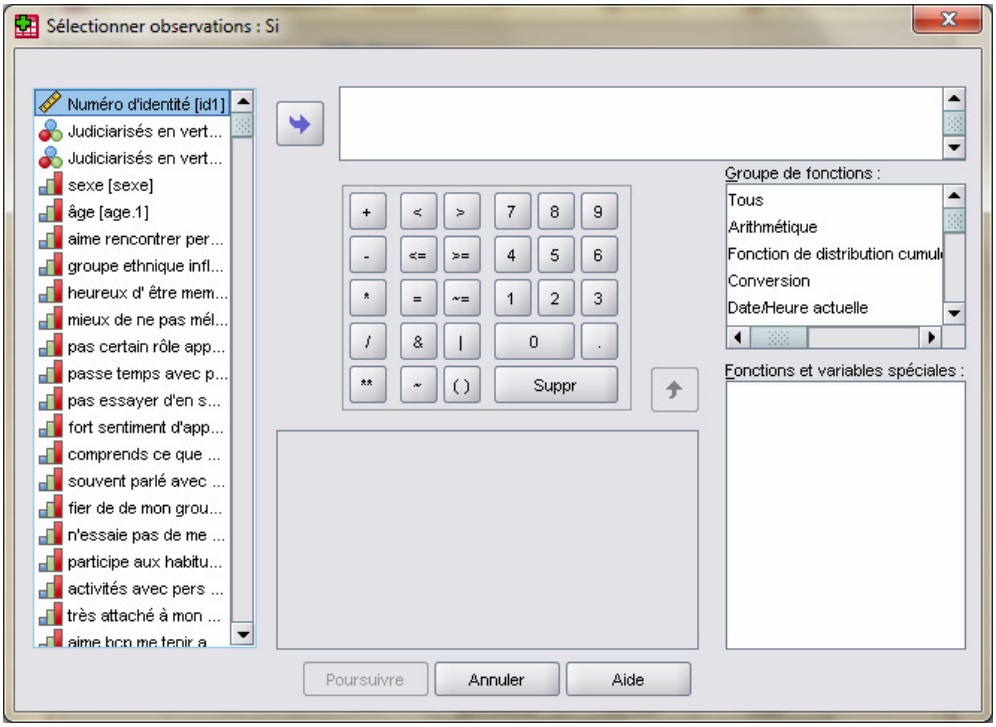
Cliquez sur **Données** à partir du menu d'entête de SPSS
↳ *Cliquez* sur **Sélectionner des observations**

Un écran de dialogue prenant la forme suivante apparaît:



Choisissez **l'option**
↳ ⊙ **Selon une condition logique**
↳ *Cliquez* sur le rectangle **SI...**

Un écran de dialogue prenant la forme suivante apparaît :



Écrivez dans le rectangle de droite, la condition permettant la création du sous-groupe (sous-population ou du sous-échantillon) requis.

↳ *Cliquez* sur **Poursuivre**

REMARQUE: Pour éliminer DÉFINITIVEMENT, c'est-à-dire pour l'ensemble des analyses réalisées dans le cadre de l'étude en cours, les cas qui ne répondent pas aux conditions du groupe soumis aux analyses:

choisir l'option ↳ ⊙ **Suppr**

sinon laisser sur ↳ ⊙ **Filtrées** (*option par défaut*) la sélection est alors temporaire et il est possible de retourner aisément à l'ensemble de la population ou de l'échantillon de départ, ceci en sélectionnant:

↳ *Cliquez* sur **OK**

Remarquez que lorsqu'une sélection temporaire d'une sous-population ou d'un sous-Èchantillon (par l'option **Filtrées**) est effectuée pour une partie des analyses, les cas non-sélectionnés sont facilement identifiables dans la fenêtre **Données** puisque le numéro d'identification des lignes correspondantes apparaît comme étant barré.

2. SÉLECTIONNER UN ÉCHANTILLON

La commande **Sélectionner des observations**…**Par échantillon aléatoire** permet de créer un échantillon aléatoire temporaire ou permanent à partir d'un fichier de données. Une fois que vous avez créé un échantillon, toutes les commandes suivantes sont exécutées à partir de cet échantillon, jusqu'à ce qu'une nouvelle commande indique à SPSS de revenir à la population ou à l'échantillon de départ.

PROCÉDURE POUR SÉLECTIONNER UN ÉCHANTILLON ALÉATOIRE PARMI L'ENSEMBLE DES CAS

Cliquez sur **Données** à partir du menu d'entête de SPSS
- *Cliquez* sur **Sélectionner des observations**
  - ⊙ Par échantillon aléatoire
    - *Cliquez* sur **Échantillon** (**soit** *indiquez* le % des cas devant être gardés dans l'échantillon, **soit** *indiquez* le nombre de cas que l'ordinateur doit sélectionner. Suivant cette procédure, *indiquez* dans le second rectangle le nombre de cas que contient initialement la banque de données.
      - *Cliquez* sur **Poursuivre** (indique à SPSS de procéder à la sélection des cas)
        - *Choisissez* soit :
          ⊙ Filtrer les observations non sélectionnées
          ou ⊙ Supprimer les observations non sélectionnées
          selon que vous vouliez que l'échantillonnage soit temporaire ou permanent, suivant en cela la même logique que celle précédemment présentée.

PROCÉDURE POUR REVENIR À LA POPULATION ENTIÈRE

Après un **SI** ou un **Échantillon temporaire**
- *Cliquez* sur **Données**
  - *Cliquez* sur **Sélectionner des observations**
    - **⊙ Toutes les observations**
      - *Cliquez* sur **OK**

LABORATOIRE 4 : MESURES DE TENDANCES CENTRALES ET DE DISPERSION

Objectifs d'apprentissages :

- Obtenir les mesures de tendances centrales et de dispersion sur un ensemble de données
- Obtenir et choisir les graphiques correspondants

1. LES MESURES DE TENDANCES CENTRALES ET DE DISPERSION

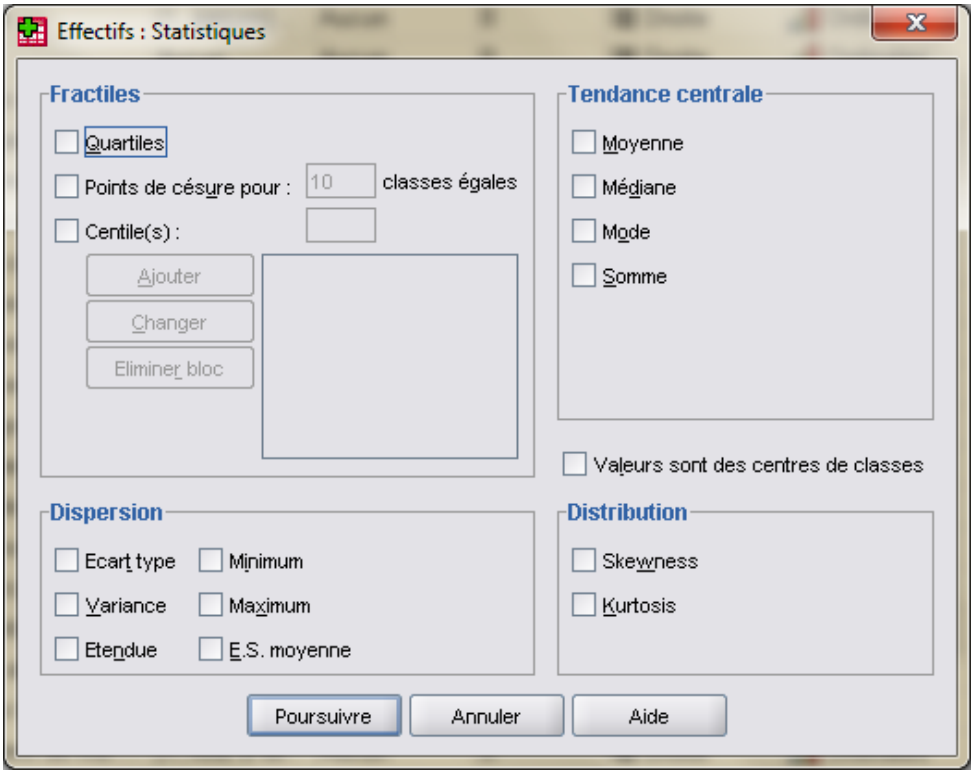
Les mesures de tendances centrales et de dispersion sont tout particulièrement pertinentes lorsqu'il s'agit de décrire des variables de niveau de mesure proportionnelle ou intervalle.

PROCÉDURE POUR OBTENIR LES MESURES DE TENDANCES CENTRALES ET DE DISPERSION

Cliquez sur **Analyse** à partir du menu d'entête de SPSS

- *Cliquez* sur **Statistiques descriptives**
  - *Cliquez* sur **Effectifs**
    - *Sélectionnez* la variable voulue (la version proportionnelle ou brute débarrassée des valeurs manquantes : il s'agit de la seule valide pour effectuer des statistiques)
      - *Cliquez* dans le carré **Afficher les tableaux** d'effectifs afin d'enlever le **X** qui s'y trouve (s'il s'y trouve) car un **X** dans ce petit carré indique à SPSS qu'on ne souhaite pas que le tableau de fréquence soit fourni. Par contre, lorsqu'il s'agit de produire des statistiques sur un tableau de fréquences concernant une variable de niveau interval ou proportionnel, on peut indiquer à SPSS les statistiques souhaitées tout en lui précisant qu'on ne désire pas voir apparaître le tableau de fréquences correspondant, lequel s'étendrait probablement sur plusieurs pages. À vous de choisir la bonne option.
        - *Cliquez* sur l'option **Statistiques**

L'écran de dialogue suivant apparaît :



⇒ *Choisir* les mesures de tendances centrales ou de dispersion dont vous avez besoin en cliquant dans les carrés précédant la statistique en question.

↳ *Cliquez* sur **Poursuivre ...**

↳ *Cliquez* sur **OK**

LES MESURES DE TENDANCES CENTRALES

MOYENNE : Mesure de tendance centrale la plus utilisée. Il s'agit de la somme des valeurs de toutes les observations, divisée par le nombre d'observations.

MÉDIANE : Valeur qui occupe la place du milieu dans le rangement ascendant ou descendant des valeurs de la variable. Autrement dit, c'est la valeur de la variable qui divise la distribution de telle sorte que 50% des valeurs se trouvent au-dessus d'elle et 50% des valeurs se rencontrent en-dessous d'elle.

MODE : Valeur la plus fréquemment rencontrée dans une série de données.

SOMME : Somme de toutes les valeurs d'une série de données.

LES MESURES DE DISPERSION

ÉCART-TYPE : Mesure la dispersion des observations autour de la moyenne. Un écart-type qui est grand par rapport à la moyenne indique la présence de données dispersées autour de la moyenne donc hétérogènes, alors qu'un écart-type petit par rapport à la moyenne indique la présence de données concentrées autour de la moyenne donc relativement homogènes.

VARIANCE : Écart-type élevé au carré. S'interprète en termes d'unités carrées.

ÉTENDUE : Différence entre la plus grande valeur et la plus petite valeur d'une série d'observations.

MINIMUM/MAXIMUM : Plus petite et plus grande valeurs rencontrées dans la distribution.

ÉCART-MOYEN : Distance moyenne (en valeur absolue) séparant les observations de la moyenne.

LES MESURES QUI RENDENT COMPTE DE LA FORME DE LA DISTRIBUTION

APLATISSEMENT : Indique le degré de voussure de la distribution (varie en -1 qui indique la présence d'une distribution leptokurtique et 1 qui indique la présence d'une distribution platikurtique en passant par 0, distribution mésokurtique).

ASYMÉTRIE : Indique le degré de symétrie de la distribution (varie entre -1 qui indique la présence d'une distribution négativement dissymétrique ou étendue vers la gauche et 1 indiquant la présence d'une distribution positivement dissymétrique ou étendue vers la droite en passant pas zéro, distribution symétrique).

Généralement, on ne présente pas les mesures de tendances centrales et de dispersion sous forme de tableau, on préfère en discuter dans le texte en les y intégrant. Toutefois, il peut arriver, pour une raison ou pour une autre, qu'on veuille aussi présenter ces résultats plus schématiquement. La présentation des mesures de tendances centrales et de dispersion pourrait alors se faire de la manière suivante:

TABLEAU 1.2. DISTRIBUTION DE LA DURÉE EN JOURS DES SENTENCES DE PROBATION (N = 12 369)

MOYENNE : 890	ÉCART-TYPE : 366.6
MÉDIANE : 730	MINIMUM : 90
MODE : 730	MAXIMUM : 3650
ASYMÉTRIE : 0,122	APLATISSEMENT : 0,336

Valeurs manquantes: 502 (4,1%)

RAPPEL : Attention, les statistiques concernant la distribution des valeurs d'une variable: mesures de tendances centrales, de dispersion ou de position se calculent sur les données brutes nettoyées et à l'exclusion des valeurs manquantes.

2. LES REPRÉSENTATIONS GRAPHIQUES

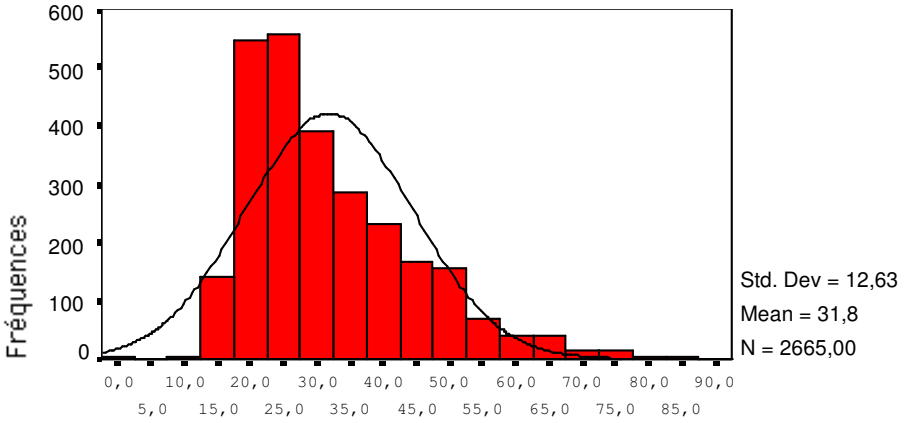
La forme de présentation la plus visuelle concernant la distribution d'une population en fonction d'une variable donnée est sans doute le graphique. SPSS vous offre le choix de plusieurs types de graphiques, allant du « graphique en pointes de tarte » (ou pointe de camembert pour les français!) jusqu'au graphique linéaire en passant par le diagramme en bâtons. Vous trouverez la liste des types de graphique que permet de réaliser SPSS sous l'item **Graphe** du menu principal.

Il n'y a pas de règle qui établisse clairement quand et comment utiliser une forme de graphique plutôt qu'une autre, notez toutefois que les **histogrammes** (pour les données de niveau de mesure intervalle ou proportionnel) et les **pointes de tarte** (pour les données de niveau de mesure nominal ou ordinal) **sont parmi les plus utilisés**.

EXEMPLE D'HISTOGRAMME

FIGURE 5.1

Age du meurtrier (SPCUM, 1989)

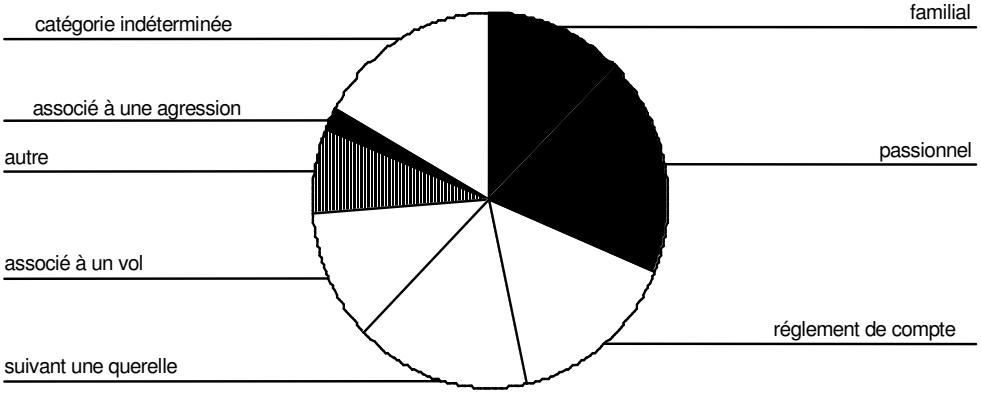


Age du meutrier

EXEMPLE D'UNE POINTE DE TARTE OU DE CAMENBERT

Figure 4.2

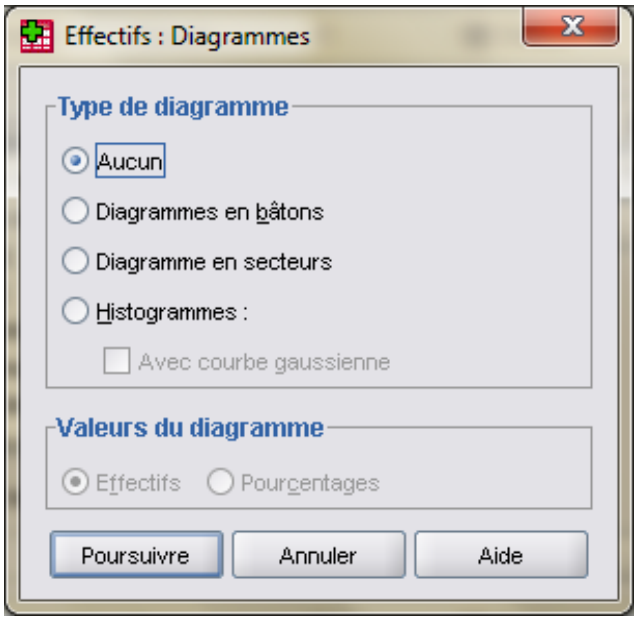
Catégories de meurtre (SPCUM, 1989)



PROCÉDURE POUR CRÉER UN GRAPHIQUE

Cliquez sur **Analyse** à partir du menu d'entête de SPSS
↳ *Cliquez* sur **Statistiques descriptives**
↳ *Cliquez* sur **Effectifs** (l'écran de dialogue que vous connaissez apparaîtra).
↳ *Sélectionnez* la variable pour laquelle vous désirez une représentation graphique.
↳ *Cliquez* sur l'option **Diagrammes**

L'écran de dialogue suivant apparaîtra :



↳ *Sélectionnez* alors le diagramme de votre choix
↳ *Cliquez* sur **OK**

N.B. Pour produire d'autres formes de diagrammes ou alors pour en faire de type plus élaboré, choisissez plutôt l'option **Graphes** du menu principal.

Conseil : Les graphiques ont l'avantage d'être éloquents et amusants à faire, par contre, ils prennent beaucoup de place et présentent des informations relativement limitées, n'en abusez donc pas. En outre, une représentation graphique, tout comme une présentation des données sous forme de tableau de fréquences, appelle un commentaire de la part du chercheur qui présente les données. Ne vous illusionnez pas : de toute manière, vous n'y échapperez pas !

PROCÉDURE POUR ÉDITER UN GRAPHIQUE

Une fois le diagramme apparus dans le fenêtre de résultats :

- *Sélectionnez* le diagramme en *cliquant* deux fois dessus (La fenêtre *Éditeur de diagramme* SPSS devrait maintenant présenter le graphique, lequel peut alors être modifié à votre guise.
  - *Ajoutez* un titre, une légende, ou ajustez n'importe quel élément du graphique (son orientation, la taille ou la police des caractères), soit en *double-cliquant* sur l'élément à transformer, soit en utilisant les menus qui vous présentent les différentes options de transformations possibles à cette étape de procédures de création et d'édition d'un graphique.

LABORATOIRE 5 : TESTS DE COMPARAISON DE MOYENNES

Objectifs d'apprentissages :

- Effectuer des tests de comparaison de moyenne
- Comprendre les différents résultats obtenus par SPSS
- Rapporter les résultats dans le cadre d'un travail pratique
- Interpréter les résultats

Le test d'hypothèse sur deux moyennes détermine si une différence entre deux statistiques échantillonnales est significative ou si elle est simplement due à une variation d'échantillonnage (on dira plus simplement due au hasard). En d'autre mots, on utilise le test de moyenne quant on veut savoir si la différence de moyennes entre deux groupes, pour une variable donnée, est significative ou pas.

N.B. Le test de différence de moyennes peut se faire lorsqu'on est face à:

1 variable DÉPENDANTE de type PROPORTIONNELLE (QUANTITATIVE)

et 1 variable INDÉPENDANTE de type NOMINALE DICHOTOMIQUE
(c'est-à-dire présentant seulement deux modalités ou catégories)

Exemple:

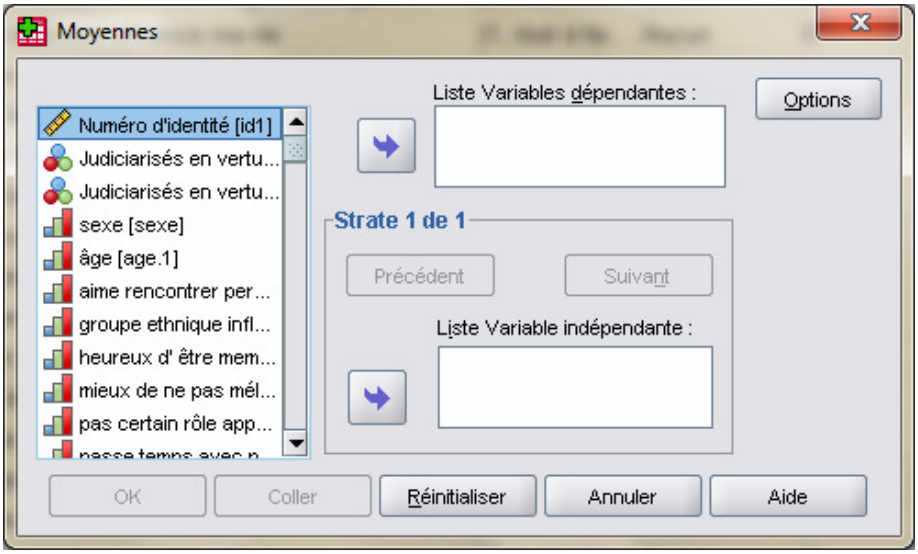
On veut savoir si, parmi les délinquants qui commettent des vols qualifiés, la différence dans la moyenne des gains est significative si on compare ceux qui s'attaquent aux dépanneurs et ceux qui s'attaquent aux garages? On fait alors un test de moyennes :

⇒ la variable dépendante (celle qui est appelée à varier) est le montant d'argent volé
⇒ la variable indépendante (celle qui pourrait faire varier la variable dépendante) est le lieu où se commet le délit (dépanneur ou garage). Cette dernière se présente sous forme dichotomique.

PROCÉDURE POUR EFFECTUER UN TEST DE COMPARAISON DE MOYENNES

Cliquez sur **Analyse** à partir su menu principal
↳ *Sélectionnez* **Comparer les moyennes**
↳ *Sélectionnez* **Moyennes**

L'écran de dialogue suivant apparaîtra :

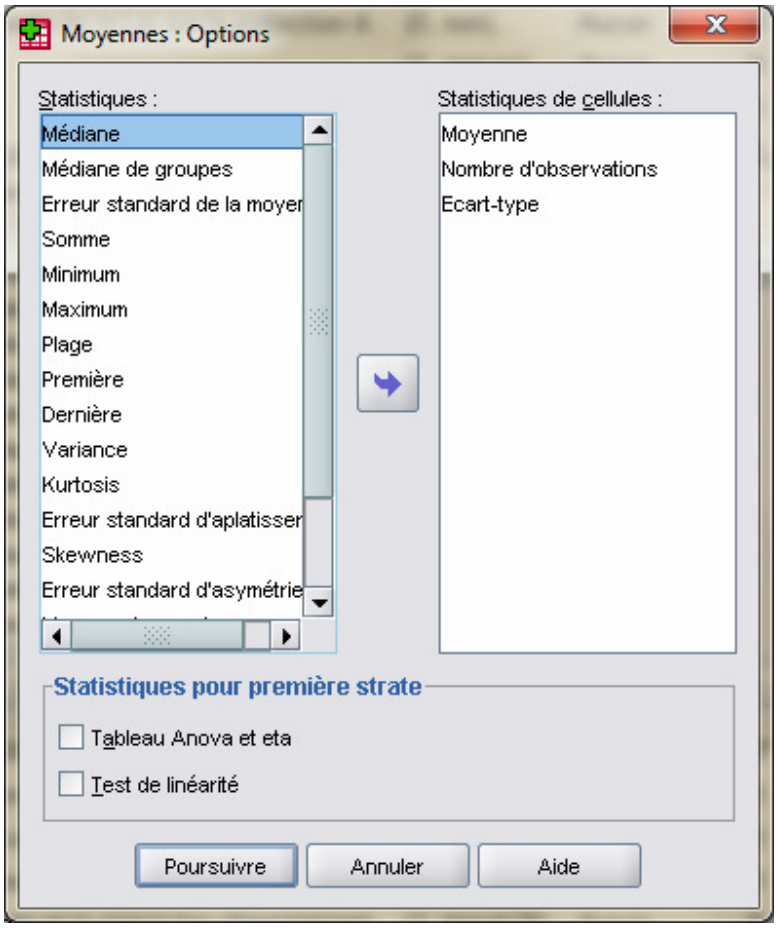


↳ *Sélectionnez* la **variable dépendante** (qui devrait toujours être de niveau de mesure proportionnel) dans le rectangle de gauche et faites-la passer, en usant de l'espace fléché prévu à cet effet, dans le rectangle intitulé **Liste variables dépendante.** On peut choisir plus d'une variable dépendante si plusieurs tests de moyennes doivent être réalisés en fonction d'une même variable indépendante (par exemple on cherche à établir si le nombre d'heures passées à étudier, le nombre d'heures passées à travailler, et les résultats scolaires varient en fonction du fait qu'on vit encore ou non chez ses parents).

↳ *Sélectionnez* la **variable indépendante** (laquelle devrait toujours se présenter sous une forme dichotomique, ce qui peut vouloir dire un certain travail de recodage) dans le rectangle de gauche présentant l'ensemble des variables contenues dans la banque de données et faites-la passer, en usant de l'espace fléché prévu à cet effet, dans le rectangle ***Liste variable indépendante.*** On peut choisir plus d'une variable indépendante, si celles-ci sont toutes susceptibles d'influencer la moyenne d'une ou de plusieurs variables dépendantes. Celles-ci doivent toutes être testées en regard de la ou des mêmes variables dépendantes (par exemple, on calcule la différence de moyennes dans la durée des procédures et la durée de la sentence en fonction du sexe - masculin / féminin - du statut - libre / détenu - de la situation d'emploi - possède un emploi oui / non -)

↳ *Cliquez* sur **Option**

Vous obtenez la boite de dialogue suivante vous permettant de choisir un certain nombre d'options selon qu'elles vous semblent plus ou moins à propos compte tenu des analyses que vous voulez effectuer :



Assurez-vous que les options **Moyenne, écart-type et nombre d'observations** (nombre de données valides sur lesquelles s'appuient les analyses) sont sélectionnées. Ce sont les données dont vous avez minimalement besoin pour procéder à votre analyse. Les options **Variance** et **Somme** (qui fait la somme de toutes les données disponibles) sont jugées moins utiles et servent surtout de base à d'autres analyses.

↳ *Sélectionnez* l'option statistique **Table Anova et eta** (qui vous donnera deux statistiques indispensables : Table et Eta)

↳ *Cliquez* sur **Poursuivre**

↳ **OK.**

RÈGLE DE DÉCISION: COMMENT DÉTERMINER SI LA DIFFÉRENCE DE MOYENNES OBSERVÉE EST STATISTIQUEMENT SIGNIFICATIVE?

Si on accepte un **pourcentage d'erreur de 5%** (donc qu'on veut être sûr à 95% de ne pas se tromper en disant qu'il existe une différence significative entre les moyennes observées et que, par conséquent, cette différence n'est pas simplement due au hasard échantillonnal):

Si P (ou Sig) inférieur à 0.05, la différence est jugée statistiquement significative.
Si P (ou Sig) est supérieur à 0.05, alors la différence n'est pas statistiquement significative.

Si on se fixe un **pourcentage d'erreur de 1%** (donc on veut être sûr à 99% de ne pas se tromper en disant qu'il existe une différence significative entre les moyennes observées et que, par conséquent, cette différence n'est pas simplement dûe au hasard échantillonnal):

Si P (ou Sig) est inférieur à 0.01, alors la différence est jugée statistiquement significative.
Si P (ou Sig) est supérieur à 0.01, alors la différence n'est pas statistiquement significative.

En sciences humaines on accepte généralement un pourcentage d'erreur maximum de 10%, 5% ou de 1%, selon les visées de l'étude et le degré de certitude ou de précision requis. C'est pourquoi on compare P à l'un des seuils suivants, soit: 0.1, 0.05 ou 0.01

1. EXEMPLE D'UN LISTING SPSS D'UN TEST DE COMPARAISON DE MOYENNES

Voici de quoi aurait l'air le **résultat** d'un test de différences de moyennes permettant de savoir si la valeur des vols dépendent de l'endroit ou a été perpétré le délit :

Observation Calculer Récapituler

	Observations					
	Inclus		Exclu(s)		Total	
	N	Pourcentage	N	Pourcentage	N	Pourcentage
VALVOLFR Valeur du vol ou de la fraude * DEP_BANQ dépanneur ou banque	1366	76,7%	416	23,3%	1782	100,0%

Tableau de bord

VALVOLFR Valeur du vol ou de la fraude

DEP_BANQ	Moyenne	N	Ecart-type
5,00 Dépanneur	422,8920	1093	1592,3371
6,00 Banque	9393,5128	273	28697,03
Total	2215,7028	1366	13379,35

Tableau ANOVA

			Somme des carrés	df	Moyenne des carrés	F	Signification
VALVOLFR Valeur du vol ou de la fraude * DEP_BANQ dépanneur ou banque	Inter-groupes	Combiné	1,758E+10	1	1,758E+10	105,734	,000
	Intra-classe		2,268E+11	1364	166250827		
	Total		2,443E+11	1365			

Mesures des associations

	Eta	Eta carré
VALVOLFR Valeur du vol ou de la fraude * DEP_BANQ dépanneur ou banque	,268	,072

2. EXEMPLE D'INTERPRÉTATION DE TEST DE COMPARAISON DE MOYENNE

Aménagés, pour être rendus dans le cadre d'un travail ou d'un article, les résultats d'un test de moyenne pourraient être présentés ainsi:

Tableau 1.1. Durée du procès en fonction des antécédents judiciaires du suspect

	Montant du vol	**Écart-type**	**Nombre de cas**
Dans des dépanneurs	422,89	1592,34	1093
Dans des banques	9393,51	28697,03	273
Au total	2215,70	13379,35	1 366

P = .000
Êta carré : 0,072
Valeurs manquantes : **416 ou 23,3%**

Quant à **l'interprétation**, on pourrait simplement dire qu'il existe une différence de moyennes statistiquement significative dans le montant des vols selon l'endroit où ils ont été commis. En effet, on peut dire que ceux qui s'attaquent aux dépanneurs font moins d'argent (en moyenne, 422,89$) que ceux qui volent des banques (en moyenne, 9393,51$), puisque $p<.05$. La mesure d'association êta carré nous permet d'affirmer que 7,2% de la variation des montants des vols pourrait être expliqué par l'endroit ou celui-ci a été commis.

LABORATOIRE 6 : TABLEAUX CROISÉS

Objectifs d'apprentissages :

- Créer des tableaux croisés
- Comprendre les différents résultats obtenus par SPSS
- Rapporter les résultats dans le cadre d'un travail pratique
- Interpréter les résultats

La commande **Tableaux croisés** permet de présenter sous forme de tableau croisé la mise en relation de 2 ou plusieurs variables catégorisées. Le tableau croisé se présente sous forme de rangées et de colonnes à la croisée desquelles on retrouve des cellules représentant le nombre de cas prenant simultanément une valeur donnée pour la variable indépendante et une valeur donnée pour la variable dépendante. Chacune des cellules du tableau représente donc une seule combinaison de valeurs croisant une valeur de la variable dépendante avec une valeur de la variable indépendante.

Par exemple, le tableau croisant la variable CRIME DE VIOLENCE (laquelle présente 2 catégories: crime violent vs crime non violent) avec la variable EFFET DE L'ALCOOL (2 catégories: sous l'effet d'alcool vs sobre) donnerait un tableau présentant quatre cellules constituant respectivement les quatre combinaisons suivantes:

Sous l'effet de l'alcool / Crime violent	Sobre / Crime violent
Sous l'effet de l'alcool/ Crime non-violent	Sobre / Crime non-violent

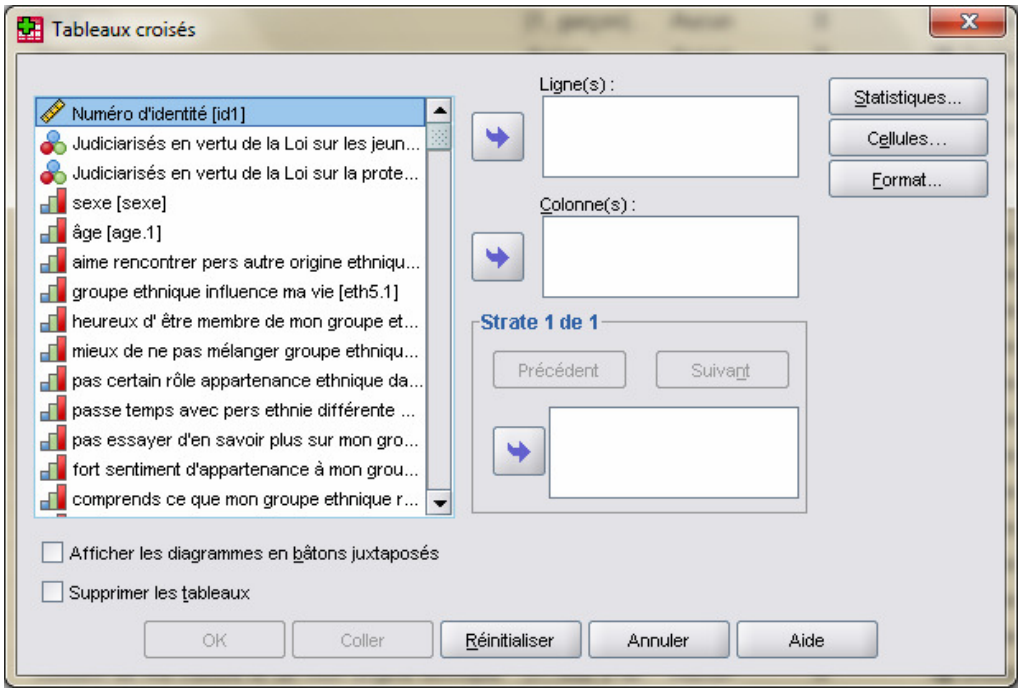
N.B. Il est généralement préférable de présenter **les valeurs de la variable indépendante en colonnes et celles de la variable dépendante en rangées**. En effet, par convention implicite on a convenu de standardiser la présentation des tableaux sous cette forme, la lecture des données s'en trouvant facilitée.

Toutefois, par souci d'esthétisme ou pour d'autres raisons, il est possible de passer outre cette règle. Par exemple, dans le cas où la variable dépendante présente un grand nombre de catégories alors que la variable indépendante n'en affiche pour sa part qu'un très petit nombre, il peut paraître préférable de présenter le tableau croisant les deux variables sur le sens de la *largeur* plutôt que sur celui de la *longueur*. Une chose demeure **capitale**, quelle que soit le présentation que l'on choisi de retenir, **il faut référer aux pourcentages appropriés au moment de procéder aux analyses et de formuler son commentaire sur les données**.

PROCÉDURE POUR CRÉER UN TABLEAU CROISÉ

Cliquez sur **Analyse** à partir su menu principal
↳ *Sélectionnez* **Statistiques descriptives**
↳ *Sélectionnez* **Tableaux croisés**

L'écran de dialogue suivant apparaîtra :



↳ *Sélectionnez* dans le rectangle de gauche, présentant l'ensemble des variables contenues dans la banque de données, la ou les variables **dépendantes** que vous souhaitez soumettre à l'analyse et faites-les passer, à l'aide des rectangles fléchés prévus à cet effet, dans le rectangle **Lignes(s).**

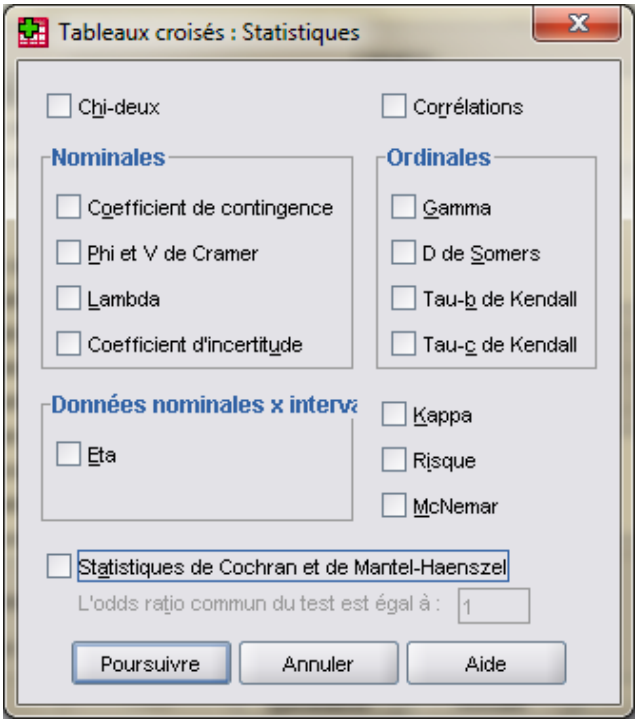
↳ Faire de même pour la ou les variables **indépendantes** de sorte qu'elles s'inscrivent, cette fois, dans le rectangle **Colonne(s)**.

↳ Vous constatez ainsi que vous pouvez soumettre au cours d'une même commande, plusieurs variables indépendantes et plusieurs variables dépendantes qui seront mises en relations les unes avec les autres. Sachez toutefois que vous multipliez de ce fait le nombre de tableaux croisés produits. Ainsi, le fait de préciser deux variables dépendantes et deux variables indépendantes, comme dans l'exemple fourni plus haut, donnera lieu à 4 tableaux (sexe des victimes et catégories d'homicide en fonction de l'âge de l'agresseur et sexe des victimes et catégories d'homicide en fonction du sexe de l'agresseur.

L'option **Strates 1 de 1** présent dans le même écran de dialogue permet pour sa part de faire intervenir des *variables contrôles*. De la même façon que pour les procédures précédentes, vous devez choisir dans le rectangle de gauche une ou des variables que vous voudriez faire intervenir en tant que variables contrôles. Ainsi, dans l'exemple précédent, le fait de faire intervenir la variable RESOLU (affaire résolue ou non), comme variable contrôle, double le nombre de tableaux produits, les quatre tableaux d'origine étant produits d'une part pour les homicides non résolus (4 premiers tableaux) et d'autre part pour les homicides résolus (4 autres tableaux).

L'option **Supprimer les tableaux** vous permet d'obtenir les statistiques concernant la mise en relation de deux variables, sans que les tableaux croisés correspondant ne soient produits. Ceci peut-être pratique si la seule chose qui vous est utile ce sont les statistiques, ou encore, si après avoir produit un tableau de fréquences vous vous apercevez que vous avez besoin des statistiques et que vous avez oublié de demander à SPSS de les produire. Si la mise en oeuvre de cette option n'est pas précisée (il s'agit pour la sélectionner de *cliquer* dans le petit carré précédant le nom de l'option afin d'y faire apparaître un x qui signale que l'option est en vigueur), SPSS produit par défaut les tableaux croisés demandés.

L'option **Statistiques** permet de sélectionner les coefficients dont vous avez besoin afin de préciser la relation entre les deux variables (sa significativité, dans certains cas sa force, dans certains cas encore sa direction).



Voici les statistiques les plus couramment utilisées :

- **Khi-deux** (Khi-carré, χ^2) **: permet de déterminer s'il existe une relation statistiquement significative** entre les variables. Seul coefficient qui, avec ses dérivés (phi, C de contingence, V de Cramer), s'applique pour les variables de niveau de mesure nominal, et par conséquent s'applique aussi pour toutes les variables de niveaux de mesure supérieur. SPSS nomme le χ^2 tel que nous avons appris à le calculer le Khi-deux de Pearson. Les autres données afférentes au χ^2 sont des variantes que nous n'apprendrons pas à interpréter.

- **Coefficient de contingence :** Dérivé du χ^2 qui **mesure la force de la relation**. S'utilise le plus souvent dans le cas de variables de niveau nominal mais s'applique aussi dans le cas de variables de niveaux de mesure supérieurs. Utilisé lorsque le tableau **comprend le même nombre de colonnes et de rangées**.

- **Phi et V de Cramer :** Dérivés du χ^2 qui **mesurent la force de la relation**. S'utilisent le plus souvent dans le cas de variables de niveau nominal mais s'appliquent aussi dans le cas de variables de niveaux de mesures supérieurs. **Phi** s'applique dans le cas de tableaux de dimension **2** (colonnes) **X 2** (rangées). **V de Cramer** s'applique **quelle que soit la dimension** du tableau.

L'option **Cellules** permet pour sa part de sélectionner les informations que vous désirez voir apparaître dans les cellules du tableau croisé.

Dans l'encadré libellé **Effectifs**

- **Observé :** présente le nombre de cas, la fréquence de deux caractéristiques (l'une concernant la variables dépendante l'autre la variable indépendante) prises simultanément.

- **Attendu** : présente les fréquences espérées (ou attendues ou théoriques) de chacun des points de rencontre (cellule) entre une valeur de la variable indépendante et une valeur de la variable dépendante si les variables mises en relation étaient parfaitement indépendantes (première étape du calcul du χ^2).

Dans l'encadré libellé **Pourcentages**

- **Ligne :** donne le pourcentage calculé en ligne pour chacune des rangées
  (Nb de cas de la cellule / Nb de cas (fréquence) de la rangée) * 100

- **Position :** donne le pourcentage calculé en colonne pour chacune des colonnes
  (Nb de cas de la cellule / Nb de cas (fréquence) de la colonne) * 100

- **Total :** donne le pourcentage de cas qui appartient à chaque cellule en rapport avec le nombre de cas total à l'étude (Nb de cas de la cellule / Nb total de cas) * 100

1. EXEMPLE DE TABLEAU CROISÉ

Voici un exemple de tableau croisé où la variable dépendante est le fait d'avoir commis un crime violent ou au contraire un délit ne présentant pas de violence et où la variable indépendante représente le fait d'avoir consommé ou non de l'alcool juste avant la commission du délit.

Ici, à la sous-commande **Statistiques**, on a sélectionné les options

- **Khi-deux**
- **Phi et V de Cramer**

Tandis qu'à la sous-commande **Cellules** les options retenues étaient

- **Effectif > Attendu**

Et à la sous-commande

- **Pourcentage > Position**

Tableau 1.1. Relation entre la consommation d'alcool juste avant de délit et le fait d'avoir commis un crime violent ou non

	Alcool	Sobre	Total
Crime violent	19 70,3%	5 21,7%	24 48,0%
Non-violent	8 29,6%	18 78,3%	26 52,0
Total	27 54,0%	23 46,0%	50 100,0%

Aucune valeur manquante

Khi-Deux	Valeur	ddl	Signification
Pearson	**11,7681**	**1**	**,00060**
Phi	**,48514**		

1.2. EXEMPLE D'INTERPRÉTATION DE TABLEAU CROISÉ

Un questionnaire passé à 50 détenus d'un pénitencier québécois tend à montrer l'existence d'une relation entre la consommation d'alcool et la violence des délits. En effet, malgré qu'au total il y ait environ le même nombre de crimes violent (24) et non-violent (26), on remarque que 70,3% des répondants ayant consommé de l'alcool juste avant la commission du délit ont commis des crimes violents, comparativement à seulement 21,7% des détenus affirmant qu'ils étaient sobres au moment du passage à l'acte.

Un test du Khi carré a permis de confirmer la présence d'une relation statistiquement significative entre les deux variables ($P < .05$). Le coefficient Phi (0.48) permet pour sa part d'affirmer que la relation entre la consommation d'alcool et les crimes violents semble être relativement forte. Notons toutefois que ces tests ne nous permettent absolument pas d'affirmer que l'alcool pousse les gens à faire usage de violence. On pourrait, en effet, tout aussi bien dire que les individus boivent souvent avant de commettre un crime violent, soit pour se donner du courage, soit pour se donner une excuse.

LABORATOIRE 7 : CORRÉLATION

Objectifs d'apprentissages :

- Le R2 de Pearson (coefficient de détermination)
- Le nuage de points

1. LE R2 DE PEARSON

Le R^2 de Pearson, appelé aussi le COEFFICIENT DE DÉTERMINATION, est un nombre qui mesure l'intensité de la **liaison linéaire entre deux variables de niveau de mesure quantitatives.**

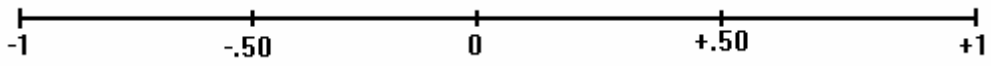
Le coefficient **R^2** varie toujours entre **-1 et 1**

- **"-1" représentant une relation linéaire négative parfaite** (on dira aussi que les variables sont liées par une relation linéaire parfaitement inversement proportionnelle);
- **"0" représentant l'absence totale de relation linéaire entre les variables;**
- **"1" représentant une relation linéaire positive parfaite (ou encore le fait que les variables sont liées par une relation linéaire parfaitement directement proportionnelle).**

Entre ces valeurs « extêmes », le coefficient de détermination indique si les variables sont plus ou moins bien linéairement liées entre elles, dans un sens ou dans l'autre. Ainsi, **plus on s'approchera de « -1 » ou « 1 » plus on conclura que la force de la relation (inversement ou directement proportionnelle) est importante.**

À l'inverse, **plus le coefficient de détermination tend vers zéro, plus on s'approche d'une situation d'absence de relation linéaire entre les variables à l'étude.**

- Plus précisément encore, le coefficient de détermination se lit comme étant **le pourcentage de variance de la variable dépendante expliquée par sa mise en relation avec la variable indépendante.** On détermine donc ainsi, si la variable indépendante constitue un plus ou moins **bon prédicteur** des valeurs que prendra la variable dépendante.

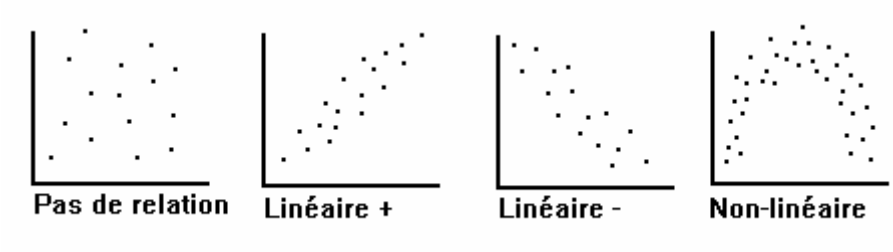


Notez que les relations linéaires parfaites ou très élevées (R^2 >.80) doivent être questionnées : il est possible qu'elles soient **dues au fait qu'on mesure deux fois le même phénomène**. En d'autres mots, il est possible que les variables ne soient pas réellement distinctes.

Par exemple, on mesure la peur du crime dans le quartier et en général et on met les deux variables en relation. On trouve une relation quasi parfaite entre les deux. C'est l'indication, qu'en fait, la peur du crime dans le quartier et la peur du crime en général ne font qu'un. On avait toutefois raison, dans un premier temps, de vouloir s'en assurer « scientifiquement ». Toutefois, pour la suite de l'étude, on sera avisé de ne conserver qu'un seul des deux indicateurs puisqu'on vient d'établir que les deux mesuraient finalement la même chose.

N'oubliez pas qu'il peut y avoir une forte relation entre deux variables, sans que cette relation soit linéaire. Pour que le coefficient de détermination rende justice à la relation, il faut absolument que la relation entre les variables soit positivement ou négativement linéaire.

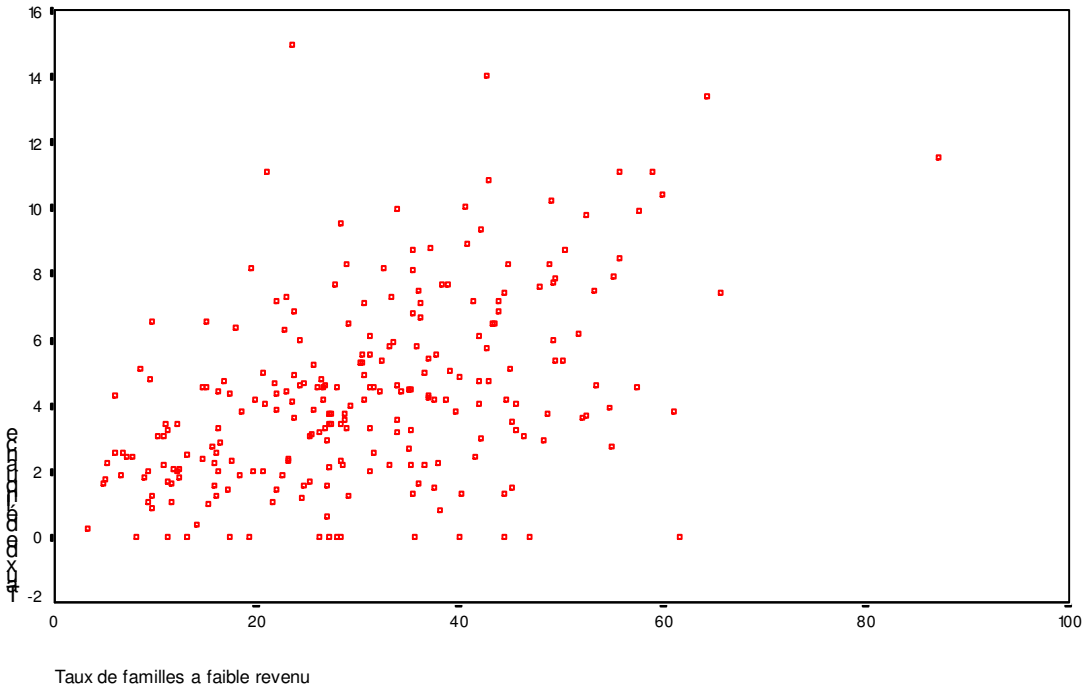
1.1 QUELQUES EXEMPLES DE RELATIONS ENTRE VARIABLES ILLUSTRÉS À L'AIDE DE DIAGRAMMES DE DISPERSION (OU NUAGE DE POINTS)



2. LE DIAGRAMME DE DISPERSION OU LE NUAGE DE POINTS

Le *diagramme de dispersion, ou le nuage de points*, constitue la **représentation graphique de la relation** entre deux variables de niveau de mesure proportionnel. **Chaque point représente un couple d'observations** (x,y) rapporté sur un graphique en prenant pour **abscisse** (axe horizontal) la **variable indépendante (x)** et pour **ordonnée** (axe vertical) la **variable dépendante (y).**

Graphique 1.1. Relation entre le taux de famille vivant sous le seuil de pauvreté et le taux de délinquance dans les secteurs de recensement de Montréal.

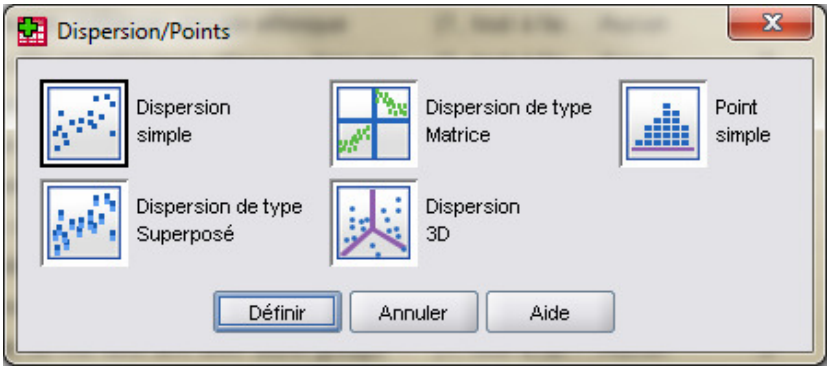


Le diagramme de dispersion permet de déceler à l'oeil que ces deux variables varient dans le même sens, dans une relation linéaire directement proportionnelle (ou positive). On doit néanmoins **faire appel au calcul des coefficients de détermination et de corrélation afin d'établir précisément dans quelle mesure (avec quelle force) ces deux variables sont inter-reliées, de manière linéaire ou non.**

PROCÉDURE POUR LA PRODUCTION D'UN DIAGRAMME DE DISPERSION (OU NUAGE DE POINTS) ILLUSTRANT LA RELATION ENTRE DEUX VARIABLES DE NIVEAU PROPORTIONNEL

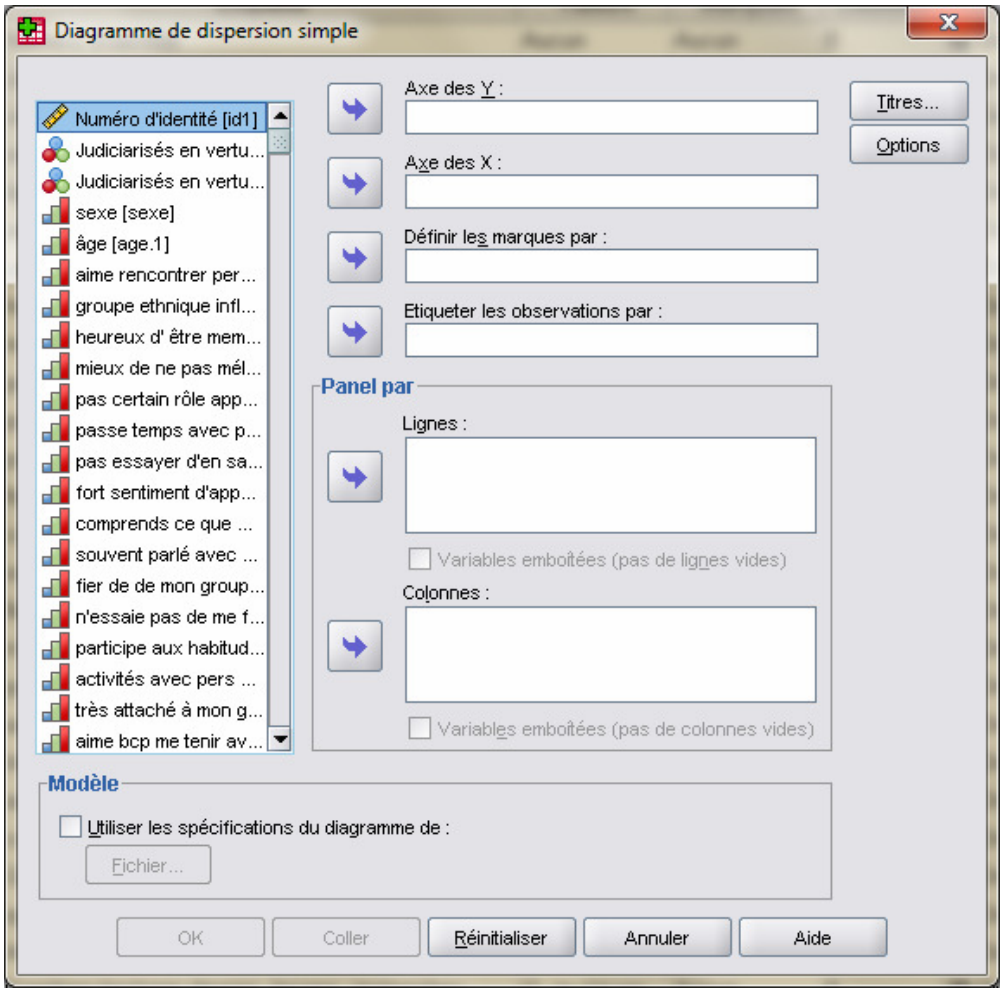
Cliquez sur **Graphes** à partir su menu principal
↳ *Sélectionnez* **Boîte de dialogue ancienne version**
↳ *Sélectionnez* **Dispersion/points**

L'écran de dialogue suivant apparaîtra :



↳ *Sélectionnez* **Dispersion simple**
↳ *Cliquez* sur **Définir**

Un nouvel écran de dialogue apparaîtra :



↳ *Sélectionnez* **la variable dépendante** soumise à l'analyse dans le rectangle de gauche qui vous présente l'ensemble des variables contenues dans la banque de données et l'inscrire, en la faisant passer à l'aide du rectangle fléché prévu à cet effet, dans le rectangle correspondant à **l'Axe des Y** (axe vertical), lequel doit représenter les valeurs de la variable dépendante.

↳ Répétez l'opération pour ce qui est de la **variable indépendante** laquelle sera pour sa part portée dans le rectangle représentant **l'Axe de X** (axe horizontal dans la représentation graphique).

Notez bien : Vous pouvez dès cette étape préciser les titres des axes et du graphique (à partir de l'option TITRES... présentée en bas de la boîte de dialogue, mais cette opportunité vous sera aussi donnée, une fois le graphique constitué, et il est alors un peu plus aisé de procéder. Alors attendre avant de procéder à ces réaménagements esthétiques.

↳ *Cliquez* sur **OK** - Le graphique se crée et apparaît dans la fenêtre *Viewer*. Vous n'avez plus qu'à cliquer afin d'aménager le graphique à votre goût !

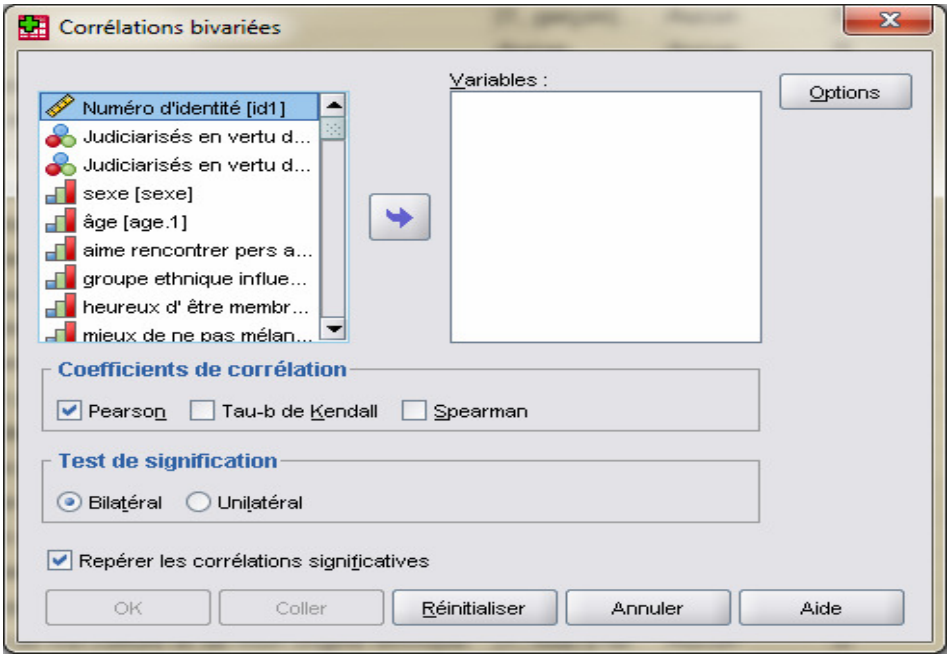
PROCÉDURE POUR L'ANALYSE DE CORRÉLATION : R DE PEARSON

Cliquez sur **Analyse** à partir su menu principal

↳ *Sélectionnez* **Corrélation**

↳ *Sélectionnez* **Bivariée**

L'écran de dialogue suivant apparaîtra :



↳ *Sélectionnez* les variables pour lesquelles vous désirez mesurer la corrélation. SPSS en traitera deux par deux.

À cette étape, SPSS ne vous demande pas de préciser quelle est la variable dépendante et quelle est celle indépendante parce que, quel que soit le sens dans lequel il effectue le calcul du coefficient de corrélation, celui-ci étant symétrique, il donnera le même résultat.

↳ S'assurer que le coefficient **Pearson** est bien sélectionné ☒

↳ *Cliquez* sur **OK**

2.1. EXEMPLE D'UN RÉSULTAT DE CORRÉLATION ET INTERPRÉTATION R DE PEARSON:

Notez bien que les coefficients de Pearson sont souvent présentés sous forme de matrice, donc qu'il n'y a que les chiffres en gras qui peuvent réellement servir.

TABLEAU 1.1: CORRÉLATION ENTRE LE TAUX DE FAMILLE PAUVRE ET LE TAUX DE DÉLINQUANCE DANS 245 SECTEURS DE RECENSEMENT DE L'ÎLE DE MONTRÉAL

	TAUX_DEL	**PAUVRES**	
TAUX_DEL	1,0000	**,4707**	(r de Pearson)
	(245)	**(244)**	(nombre de cas ayant servi au calcul)
	P= ,	**P= ,000**	(niveau de significativité)
PAUVRES	,4707	1,0000	
	(244)	(245)	
	P= ,000	P= ,	

INTERPRÉTATION POSSIBLE DU TABLEAU 1.1. :
Il existe une relation statistiquement significative (p<.05) et directement proportionnelle (puisque le signe du coefficient n'est pas négatif) entre le taux de familles vivant sous le seuil de pauvreté et le taux de délinquance juvénile. De fait, on constate que plus il y a de familles pauvres dans un secteur de recensement donné, plus le taux de délinquance y est élevé. Ceci est confirmé par un coefficient de corrélation *R de Pearson* de .47 (p - 0,000) indiquant une relation statistiquement significative positive assez forte entre les deux variables.

Notez bien : Dans le cadre d'un article scientifique, lorsque seulement deux variables sont mises en relation, on se contentera de « glisser » la valeur du coefficient et le degré de significativité dans le texte à l'appui de notre analyse. Inutile de présenter la matrice des résultats de l'analyse de corrélation.

LABORATOIRE 8 : RÉGRESSION

Objectifs d'apprentissages :

- La régression
- Le R de Pearson (coefficient de corrélation)

1. LA RÉGRESSION

1.1. LA DROITE DE RÉGRESSION SIMPLE

La droite de régression est la droite qui s'ajuste le mieux possible aux points d'un diagramme de dispersion mettant en relation deux variables.

FORMULE DE LA DROITE DE RÉGRESSION :

$$y = a + bx$$

où
"y" = variable dépendante,
"x" = variable indépendante
"a" = l'ordonné à l'origine (qu'on nommera aussi constante)
"b" = la pente de la droite.

Exemple de la relation entre le nombre de victimisations subies et la peur du crime, telle que mesurée sur une échelle croissante de 1 à 7, 7 représentant le plus haut degré de peur que peut indiquer un répondant.

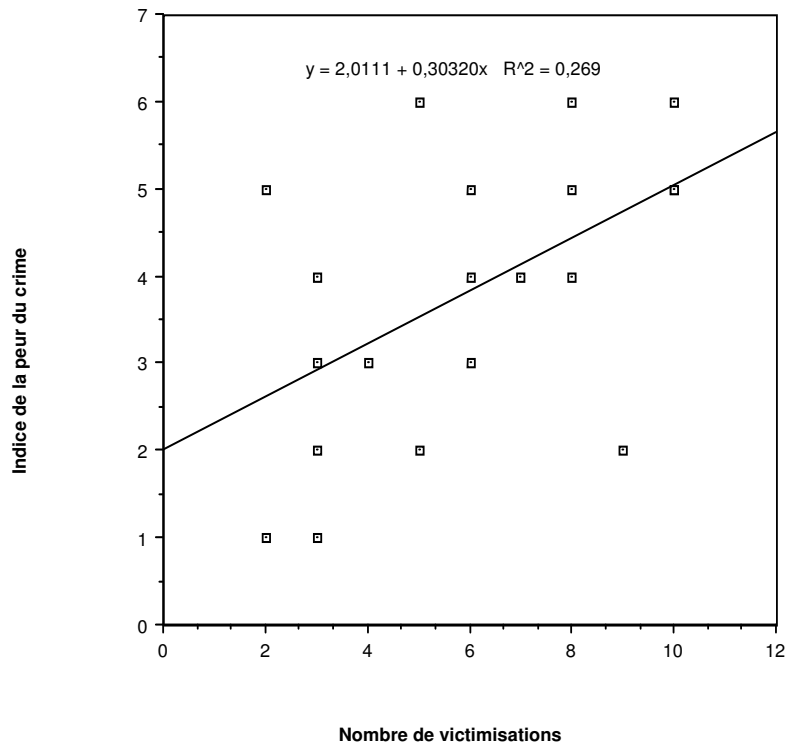
Cas no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
VICTIM	8	10	2	9	6	6	2	3	4	10	5	3	10	7	3	3	6	8	8	5
PEUR	4	6	1	2	5	3	5	1	3	5	2	2	5	4	3	4	4	5	6	6

Ici, la droite de régression s'inscrit :

Indice de peur du crime = 2,01 + 0,30 (x = nombre de victimisations)

Le R^2 de Pearson est ici de 0,27, c'est-à-dire que le nombre de victimisations subies permet d'expliquer 27% de la variance observée en ce qui a trait au niveau de peur du crime exprimé par les répondants.

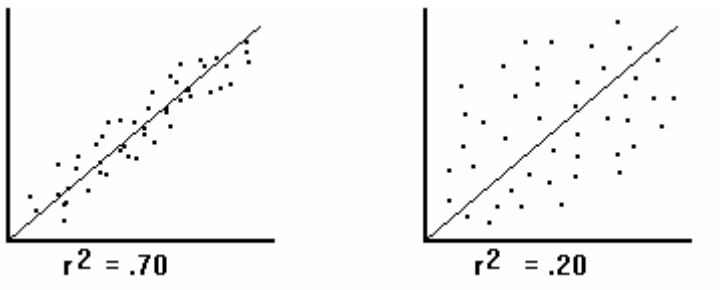
Tableau 8.2
Indice de la peur du crime en fonction du nombre d'expériences de victimisation subies par les répondants



1.2. LE COEFFICIENT DE DÉTERMINATION: R^2

Le coefficient de détermination, le **R^2 de Pearson**, représente **la proportion de la variation totale que subie la variable dépendante qui est expliquée par la droite de régression, lorsqu'une variable indépendante quelconque est prise en compte**. Il s'exprime comme le rapport de la variation expliquée (par la présence de la variable indépendante) ‡ la variation totale (lorsque la variable indépendante n'est pas tenue en compte). En d'autres mots, il s'agit **d'un indice de l'ajustement de la droite au nuage de points découlant de la mise en relation de deux variables.**

Ainsi, on peut voir que le premier diagramme, qui représente une relation plus forte entre les deux variables, présente un coefficient de détermination plus élevé que le second diagramme.

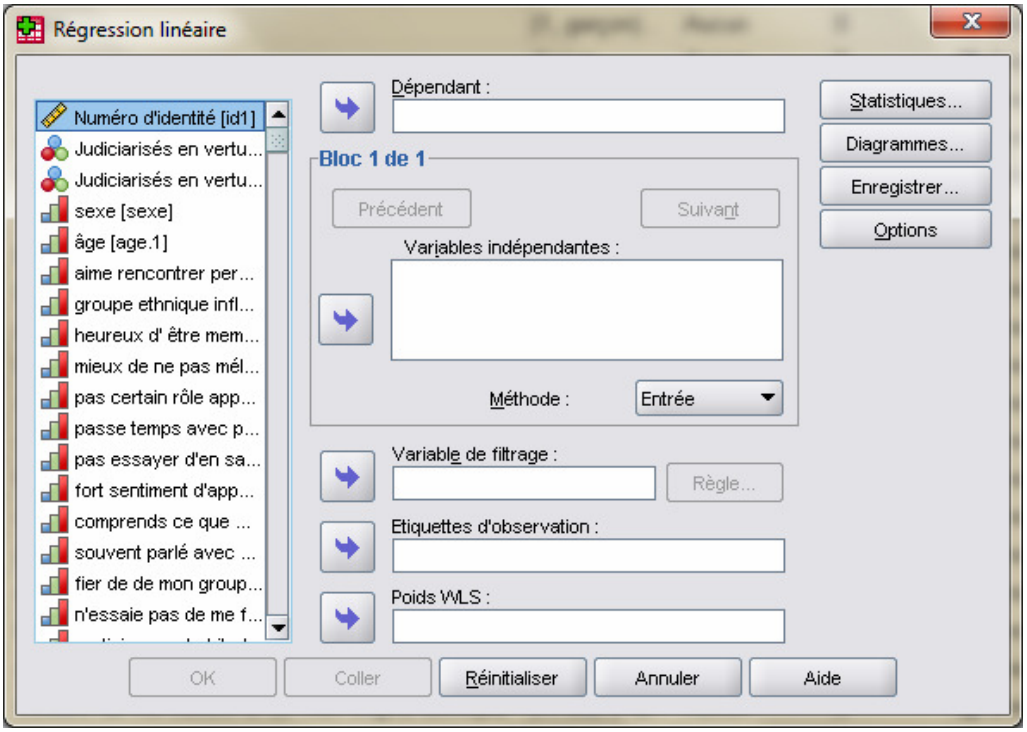


La procédure de régression réalisée par SPSS sert donc **à déterminer la présence d'une relation statistiquement significative**, c'est-à-dire qui ne serait pas due au hasard, entre les variables (donné par le niveau de significativité; la direction de cette relation (donnée par la pente positive ou néative); ainsi que la proportion de variance de la variable dépendante expliquée par la présence de la variable indépendante avec laquelle elle est mise en relation (R^2).

PROCÉDURE AFIN D'EFFECTUER UN TEST DE RÉGRESSION

Cliquez sur **Analyse** à partir su menu principal
↳ *Sélectionnez* **Régression**
↳ *Sélectionnez* **Linéaire**

L'écran de dialogue suivant apparaîtra :



↳ *Sélectionnez* les variables dépendantes et indépendantes que vous souhaitez soumettre à l'analyse. Ici il est important de bien déterminer et préciser à SPSS quelles sont les variables indépendante et dépendante.
↳ *Cliquez* sur **OK**

1.3. EXEMPLE D'UN LISTING SPSS D'UNE ANALYSE DE RÉGRESSION :

Régression

Variables introduites/éliminées[b]

Modèle	Variables introduites	Variables éliminées	Méthode
1	PEUR Indice de la peur du crime[a]	,	Introduire

a. Toutes variables requises introduites

b. Variable dépendante : VICTIM Nombre de victimes

Récapitulatif du modèle

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,537[a]	,288	,249	1,4506

a. Valeurs prédites : (constantes), PEUR Indice de la peur du crime

ANOVA[b]

Modèle		Somme des carrés	ddl	Carré moyen	F	Signification
1	Régression	2878,398	1	15,325	7,283	,015[a]
	Résidu	37,875	18	1193,624		

a. Valeurs prédites : (constantes), PEUR Indice de la peur du crime

b. Variable dépendante : VICTIM Nombre de victimisation

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification
		B	Erreur standard	Bêta		
1	(constante)	34,780	18,432		1,887	,015
	VICTIM Nombre de victimisation	1,494	,962	,169	1,553	

a. Variable dépendante : VICTIM Nombre de victimisation

1.4. INTERPRÉTATION POSSIBLE DES COEFFICIENTS ET/OU DE LA DROITE DE RÉGRESSION ET/OU DU NUAGE DE POINT:

Nous constatons l'existence d'une relation statistiquement significative (p = .01) et directement proportionnelle (signe de la pente = positif) entre le nombre de victimisations subies et la peur du crime telle que mesurée par une échelle allant de 1 à 7, où 7 indique le niveau le plus élevé de peur. Ainsi, plus une personne aura été victimisée, plus elle aura tendance à craindre la criminalité (R de Pearson: 0,54).

La droite de régression nous permet même de dire que la variable indépendante « nombre de victimisations » permet d'expliquer près de 30% de la variance observée au sujet de la variable dépendante « peur du crime » exprimée par les répondants ($R^2 = 0{,}288$).