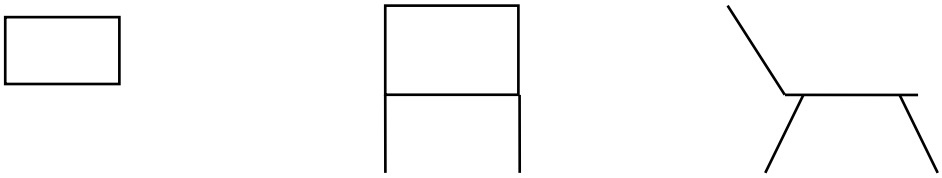
Table de matières

I. Introduction

L'analyse des données est une technique relativement ancienne 1930. Elle a connu cependant des développements récents 1960-1970 du fait de l'expansion de l'informatique. L'informatique est importante car cette technique nécessite le brassage de beaucoup de données par beaucoup de calculs pour en tirer des représentations graphiques. Elle apporte rapidité et fiabilité.

L'analyse des données est une technique d'analyse statistique d'ensemble de données. Elle cherche à décrire des tableaux et d'en déduire des relations pertinentes.

En effet, il faut en particulier réduire les dimensions de la réalité, c'est-à-dire ne pas considérer certaines variables tout en cherchant à conserver le maximum de sens. Cela revient à effectuer une projection.



Ces trois schémas représentent une chaise dessinée dans le plan. Ils permettent une plus ou moins bonne identification selon l'axe de projection.

L'analyse de données entend se démarquer des statistiques paramétriques. La statistique paramétrique effectue des mesures quantitatives et utilise le théorème central limite qui ramène à la loi de LAPLACE-GAUSS. La contestation par l'approche non paramétrique cherche d'une part, à se débarrasser de l'obligation de passer par la loi normale et donc des contraintes sur la taille des échantillons, et d'autre part, à s'intéresser aux données qualitatives.

Il n'y a pas très longtemps, on ne pouvait pas traiter un tableau de 3000 lignes et 300 colonnes. L'analyse des données est un ensemble de techniques qui permet de découvrir la structure d'un tableau de données à plusieurs dimensions et de la résumer à une structure plus simple. Les méthodes d'analyse de données ont commencées à être développées dans les années 50, elles sont appliquées dans tous les domaines de recherche qui doivent gérer de grande quantité de données.

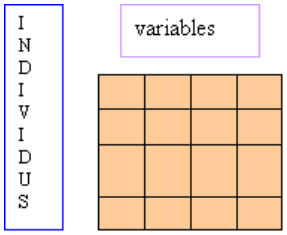
Ce cours d'analyse de données vise à préparer l'étudiant à utiliser les techniques statistiques. Toutefois, dans le court terme, un tel cours se veut aussi une précieuse aide pour l'étudiant qui réalise un mémoire de fin d'études qui fait intervenir une analyse plus approfondie de données.

Le cours comprendra une partie théorique, qui présentera les techniques modernes de l'analyse de grands ensembles de données, et une série de travaux pratiques réalisés sur ordinateur en salle informatique avec le support du logiciel SPSS14, qui développera les outils de base de l'analyse de données. A l'issue de ce cours, l'étudiant sera capable de :

- Traiter et décrire l'information contenue dans de grands tableaux de données ;
- Comprendre les mécanismes qui justifient l'emploi de telle ou telle méthode ;
- Interpréter correctement les graphiques et résultats fournis par les logiciels ;

II. Structure d'une base de données

Les données se présentent généralement sous la forme d'un tableau rectangulaire, dont les lignes correspondent à des individus ou unités statistiques et les colonnes à des variables appelées caractères ou caractéristiques.



Les analyses statistiques appliquées aux variables provenant des réponses aux questionnaires, vont dépendre du type de ces variables. Globalement, les variables peuvent être scindées en deux groupes principaux : les variables discrètes et les variables continues.

II.1. Types de variables

Une variable est une caractéristique étudiée pour une population donnée. Le sexe, la couleur préférée, ou encore l'âge sont des variables.

Il existe 2 types de variables :

Les variables **qualitatives**: sont des variables représentées par des qualités, telles que le sexe, le programme d'études ou encore l'état civil. Les variables qualitatives s'expriment en modalités. Les modalités sont comme des choix de réponses aux variables étudiées.

Les variables **quantitatives**: sont quant à elles des variables représentées par des quantités telles que l'âge, le poids et la taille. Elles s'expriment en valeurs. Les valeurs représentent les choix de réponses aux variables quantitatives.

Les variables discrètes ou qualitatives, ont des codes qui indiquent l'appartenance à un groupe ou une catégorie définie.

Les variables continues ou quantitatives ont des valeurs qui correspondent à des mesures provenant d'une échelle continue ordonnée.

II.2. Echelles de mesures

Pour les variables qualitatives, il y a deux échelles différentes :

Nominales ou catégorielles : permet d'affecter les individus à des catégories sans relation hiérarchique (nationalité, sexe, appartenance à un parti politique…).

Ordinales ou de rangements : permet d'affecter les individus à des catégories avec une relation hiérarchique qui rend possible leur comparaison (petit –moyen - grand).

Il en va de même pour les variables quantitatives, il existe 2 types d'échelles :

Echelle d'intervalle : permet de tenir compte de la différence entre deux valeurs d'une variable. Les valeurs observables sont ici numériques. Exemple: échelle de température Celsius. Le choix de l'origine (le zéro) de l'échelle est arbitraire. Le rapport entre les valeurs n'a pas de sens en soi.

Echelle de rapport : semblable à l'échelle d'intervalle mais avec l'existence d'une origine significative (zéro « naturel » ou « absolu »). Le rapport entre les valeurs a un sens précis. Exemple: durée de vie, taille, vitesse…

«Bien plus que la nature de la variable, c'est l'échelle de mesure utilisée qui importe puisque c'est elle qui détermine le degré de précision des résultats.».

Le choix d'une échelle n'est pas toujours unique : il est dicté par le point de vue sous lequel on considère la variable mesurée. Ex: la variable "âge" peut être mesurée sur une échelle : de rapports (l'âge exact), ordinale (enfance, adolescence), nominal (actif, passif).

III. Analyse de données

L'analyse de données est l'ensemble de méthodes descriptives ayant pour objectif de résumer et visualiser l'information pertinente dans un grand tableau de données.

La statistique : ensemble des méthodes permettant de collecter des données, de les organiser, les représenter, les décrire, les synthétiser, et de les analyser afin d'en tirer de l'information utile.

La statistique descriptive se compose de 3 domaines distincts :

1. Statistique univariée: répartition d'une population selon une variable (la taille, le poids…)
2. Statistique bivariée: étudie la relation qui peut exister entre deux variables (entre la taille et le poids, par exemple…)
3. Statistique multivariée: étudie les relations entre plusieurs variables.

III.1. Statistique descriptive univariée

La statistique descriptive univariée vise à explorer et décrire les données résultant de l'observation d'une variable x sur n individus.

III.1.1 Effectifs et fréquences cumulés

Le tableau d'effectifs donne des informations sur le nombre d'occurrences des valeurs prises par une variable catégorielle dans la base de données. Elle permet donc de constater la fréquence à laquelle les participants ont donné chacune des réponses possibles à la variable choisie.

Par exemple, si on veut savoir combien d'hommes et de femmes se retrouvent dans une base de données, on exécuterait la procédure de fréquences sur la variable nominale SEXE (ayant comme étiquettes 1=Homme, 2=Femme). On obtiendrait alors le nombre d'occurrences (ou l'effectif) de la valeur « 1 » (donc le nombre d'hommes) et de la valeur « 2 » parmi les répondants.

III.1.2 Représentations graphiques

Un graphique doit être un outil de communication; il a pour objectif de montrer des données de façon claire et adéquate :

- Le graphique doit contenir le maximum d'informations utiles (légendes,..); il doit être compris par lui-même, sans que l'on soit obligé de recourir à la lecture d'un texte explicatif
- Il n'est pas nécessaire de vouloir représenter des situations simples par des graphiques sophistiqués
- La façon de présenter graphiquement un phénomène doit mettre en évidence ses caractéristiques essentielles
- Il ne faut comparer des graphiques que si l'on a choisi des unités communes sur les axes

III.1.3 Mesures de position, de dispersion et de forme

But de ces mesures : résumer la distribution observée au moyen de valeurs typiques (ou statistiques caractéristiques); faciliter la comparaison entre des séries statistiques distinctes.

a) les mesures de position (le mode, et les moyennes)
b) Les mesures de dispersion : valeurs caractérisant globalement les écarts entre les observations ou encore leur dispersion autour d'une valeur centrale (l'étendue, la variance, l'écart-type, les écarts interquartile et interdécile, le box-plot, le coefficient de variation, les écarts moyen et médian absolus).
c) Les mesures de forme : valeurs caractérisant la forme (symétrie, aplatissement) d'un diagramme en bâtons ou d'un histogramme (les coefficients de Pearson, de Yule et Kendall, et de Fisher).

A) Les mesures de position :

Le mode : est la valeur observée qui apparaît le plus souvent. Le mode est une mesure de tendance centrale uniquement dans le cas d'une distribution unimodale en forme de cloche, car il peut y avoir des distributions bimodale ou plurimodale, des distributions sans mode.

Remarque : le mode peut parfois être une mesure de tendance centrale importante (ex: pour un commerçant qui veut voir quel produit spécifique est le plus vendu), parfois il n'est pas représentatif de l'ensemble des observations (ex: pour un commerçant, le mode de la distribution des tailles des clients quant aux habits est une information insuffisante).

La moyenne :

- ✓ Chaque observation intervient avec le même poids (1/n)
- ✓ Il y a d'autres types de moyennes (géométrique, pondérée, harmonique) que la moyenne arithmétique
- ✓ Uniquement si variable mesurée sur une échelle d'intervalles ou de rapports
- ✓ La moyenne arithmétique est unique
- ✓ La moyenne arithmétique est rarement une valeur observée

Attention, la moyenne arithmétique a une grande sensibilité face à la présence de valeurs extrêmes (valeurs aberrantes)

Remarque: La moyenne est un outil de comparaison limité : -Elle ne reflète pas l'hétérogénéité des résultats. -Elle peut être biaisée "vers le haut" ou "vers le bas" à cause de valeurs importantes. En présence de valeurs extrêmes, elle perd donc sa validité, car elle n'est pas représentative de l'ensemble.

A la différence de la moyenne, la médiane n'est pas sensible à la présence de valeurs extrêmes, et peut donc rendre mieux compte de la tendance centrale d'une série statistique que la moyenne.

Remarques sur les mesures de position :

• Le choix d'une mesure de position est fonction de l'objectif poursuivi, et le choix d'une mesure de tendance centrale est fonction de la nature de la variable étudiée (pour une variable nominale, le concept de centralité n'a pas de sens; pour une variable ordinale, la médiane s'impose; pour une variable quantitative d'intervalles ou de rapports, plusieurs choix sont possibles)

• Il faut toujours examiner la série statistique observée et se poser quelques questions préalables: y a-t-il présomption d'existence de valeurs extrêmes? Les mesures sont-elles précises ou s'agit-il d'appréciations plus floues ?

B) Les mesures de dispersion :

L'étendue d'un ensemble de données est la différence entre la plus grande et la plus petite valeur observée.

Attention, l'étendue ne tient pas compte de toutes les observations (idéalement, pour que l'étendue soit significative, il faut que les observations se répartissent de façon régulière entre la plus petite et la plus grande observation); L'étendue est de plus particulièrement sensible à la présence de valeurs extrêmes.

Les écarts interquartile et interdécile :

- L'écart interquartile mesure la dispersion des 50 % d'observations centrales.
- L'écart interdécile mesure la dispersion des 80 % d'observations centrales.

-> Qualité : Les écarts interquartile et interdécile ne sont pas sensibles à la présence éventuelle de valeurs extrêmes

-> Défaut : Ils ne mesurent la dispersion que des 50 ou 80 % d'observations centrales et négligent donc l'influence que jouent sur la dispersion de la série les 50 ou 20 % d'observations les plus grandes et les plus petites.

La variance : Elle mesure combien la variable varie en calculant les écarts par rapport à la moyenne. Elle correspond à la moyenne des carrés des différences entre les observations et leur moyenne arithmétique. Plus elle est importante, plus le groupe est hétérogène. Elle n'est valable que pour des variables quantitatives mesurées sur une échelle d'intervalles ou de rapports. Elle vaut 0 uniquement si toutes les observations ont la même valeur.

Attention, elle est très sensible à la présence de valeurs extrêmes (bien plus encore que la moyenne).

L'écart-type : C'est la racine carré de la variance, donc il s'exprime dans les mêmes unités que les observations (la variance le faisait aussi mais au carré).

! Ne pas interpréter les écarts-types trop influencés par les valeurs extrêmes.

! Ne comparer les écarts-types de deux séries statistiques que si elles ont un domaine de variation similaire et donnent lieu à des moyennes du même ordre de grandeur ! Comment calculer leur similarité ? En calculant le coefficient de variation (rapport de l'écart-type à la moyenne).

Car deux séries statistiques peuvent avoir le même écart-type mais si on ne regarde pas la moyenne, en vue de calculer le coefficient de variation, l'information fournie par le seul écart-type peut être trompeuse (cfr. exercice).

Le coefficient de variation :

C'est le rapport de l'écart-type à la moyenne. Plus sa valeur est élevée, plus la dispersion autour de la moyenne est grande. Il fournit l'information permettant de savoir si on peut ou non calculer deux écarts-types de séries statistiques différentes (car on met en rapport leurs écarts-types et moyennes respectifs, et pas seulement leurs écarts-types pris isolément.

C) Les mesures de forme :

Le coefficient empirique de Pearson - Le coefficient (empirique) de Yule et Kendall :

- ✓ Lorsqu'ils valent 0, on a affaire à une symétrie
- ✓ Lorsqu'ils sont supérieurs à zéro, on a une asymétrie de la distribution à gauche (la bosse, qui représente le mode, est vers la gauche du graphique)
- ✓ Lorsqu'ils sont inférieurs à zéro, on a une asymétrie de la distribution à droite

! Une distribution observée presque symétrique peut fournir des coefficients de Perason et de Yule et Kendall de signes contraires. Ils ne peuvent donc être considérés que comme des outils d'appréciation, simples à obtenir, mais pouvant parfois être contradictoires.

Remarque : on peut aussi analyser la symétrie d'une distribution à l'aide de l'allure du diagramme en bâtons, de l'histogramme,

III.2.Statistique descriptive bivariée

Observation de 2 variables x et y sur n éléments, afin d'Analyser des valeurs observées pour x d'une part et pour y d'autre part analyser le lien éventuel (association, dépendance) entre les valeurs prises par x et celles prises par y.

Si les données sont quantitatives (mesurées sur une échelle d'intervalles ou de rapports), on peut les représenter sur un graphique de dispersion (un scatter plot), où elles forment un nuage de points.

Attention pour les graphiques de dispersion à bien choisir les unités le long des 2 axes.

-> Il faut observer la forme du nuage de points :

- Est-il concentre/dispersé ?
- Présente-t-il une structure particulière ?
- Détecte-t-on la présence de valeurs "aberrantes" ?

On peut construire un tableau de contingence et/ou un tableau des fréquences si on observer plusieurs fois les mêmes couples de valeurs (sinon il serait inutile de faire une distribution observée, on se contenterait de la série statistique) (c'est fréquent si n est élevé, si les variables sont qualitatives, quantitatives discrètes, on continues donnant lieu à des arrondis importants)

III.2.1. Association entre deux variables quantitatives

La comparaison des profils-colonnes (ou lignes) entre eux et avec le profil-colonne (ou ligne) marginal permet de déceler s'il existe une association entre x et y, et d'analyser la nature de ce lien (ce type d'analyse peut aussi être réalisé lorsque x et/ou y sont des variables qualitatives (nominales ou ordinales)).

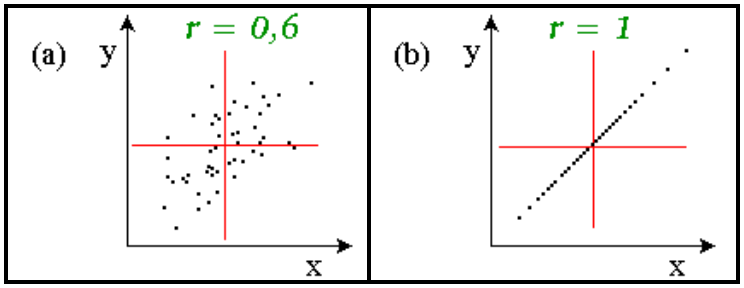
-Le coefficient de corrélation de Bravais-Pearson :

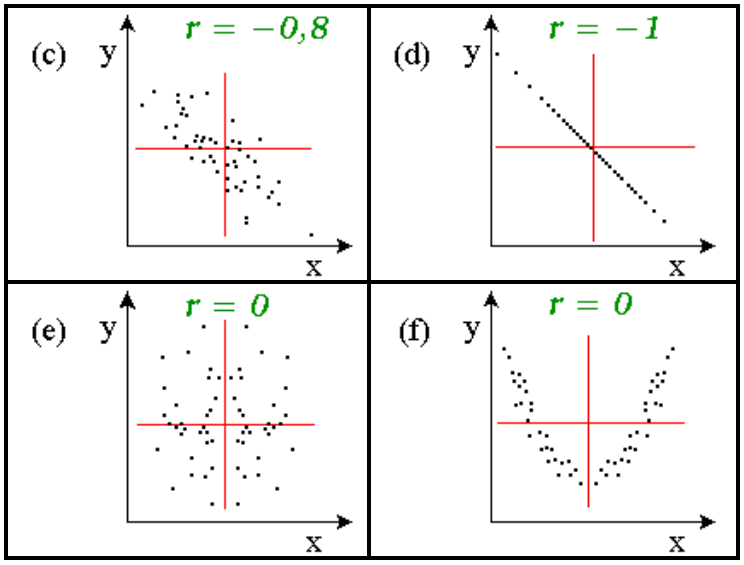
- Existe-t-il une association entre les 2 variables ? (-> coefficient de corrélation)
- Quelle est l'intensité de cette association ? (-> coefficient de corrélation) (associations linéaire positive, linéaire négative, ou absence d'association)
- Comment peut-on modéliser/représenter cette association ? (-> régression linéaire)
- Peut-on prédire la valeur d'une variable à partir de la valeur prise par l'autre variable ? (-> régression linéaire)

-> D'abord il faut calculer la covariance, en changeant éventuellement d'unité et d'origine; ensuite on peut calculer le coefficient de corrélation de Bravais-Pearson (r, nombre sans unité; compris entre -1 et +1 : si proche de 1, association linéaire positive, si = 0, pas d'association, si < 1, association linéaire négative).

La qualité de la corrélation peut être mesurée par le coefficient de corrélation r. Le coefficient de corrélation est compris entre -1 et +1. Plus il s'éloigne de zéro, meilleure est la corrélation

Quelques exemples de corrélation





Remarques :

1. Corrélation n'implique pas causalité : une forte corrélation n'implique par automatiquement une relation directe de cause à effet; elle peut simplement être due au fait que les 2 variables sont soumises à des influences communes !
2. Ex: Ce n'est pas parce que le prix de l'essence augmente que celui du mazout augmente aussi : ils évoluent dans le même sens parce qu'ils sont tous deux liés au prix du pétrole, qui est le même facteur sous-jacent (une variable, déterminante).
3. Le type de dépendance concerné par le coefficient de corrélation est exclusivement linéaire ! L'usage d'une droite de régression ne doit pas être "automatique" dès que r a une valeur proche de 1 ou -1, car il peut y avoir une forte association, un lien fort, mais qui ne soit pas linéaire. -> Une étude préalable du graphique de dispersion est donc indispensable !

III.2.2. Corrélation entre deux variables ordinales

Rappel : échelle ordinale : les concepts de moyenne et d'écarts des observations à leur moyenne n'ont pas de sens, seul l'ordre des observations est exploitable.

-> De fait, on va regarder les rangs des observations associés aux valeurs et non elles-mêmes.

Remarque : s'il y a plusieurs variables x prenant la même valeur, on établira des rangs moyens.

a) Le coefficient de corrélation de rangs de Spearman : c'est le coefficient de corrélation de Bravais-Perason de la série statistique bivariée où l'on considère les rangs des valeurs ordinales.

b) Le coefficient de corrélation de Kendall repose sur la notion de paries d'observations concordantes et discordantes et permet d'en calculer le nombre.

III.2.3. Test Chi carrée ou Chi deux

Il est nécessaire de recourir au test chi carré ou chi deux lorsque l'analyse porte sur une relation bi variée comprenant deux variables non métriques (nominales et/ou ordinales) Cette analyse s'effectue à l'aide de fréquence conjointe (c à d voir tableau de contingence ou tableau croisé).

X^2 est un calcul statistique qui permet de trancher la question de savoir si la relation entre les deux variables est significative ou non. Plus précisément c'est une procédure qui permet de trancher la proposition suivante : il n'y a pas de relation entre les deux variables.

Cette proposition est appelée hypothèse nulle. Pratiquement pour montrer qu'il existe une relation, on cherche à montrer que l'inexistence de la relation à une faible probabilité de se réaliser.

Le principe du test est simple :

Il suffit de comparer la répartition des observations entre les cases du tableau à une distribution idéale qui correspond exactement à une situation de liaison nulle entre les deux variables. Il nous faut donc deux tableaux : le tableau des effectifs observés, et le tableau des effectifs théoriques (correspondant à l'absence de liaison) et notre Khi deux ; il mesure l'écart des deux tableaux.

Plus l'écart sera grand et, plus faible sera la chance que le tableau observé est semblable au tableau théorique, autrement dit que l'hypothèse nulle se réalise, et par conséquent que la liaison soit significative.

Quand et pourquoi faut-il faire un Khi Carré ?

Quand ? Si votre recherche comporte deux groupes (deux mesures) et que votre variable dépendante est qualitative.

Pourquoi faire un test khi carré ? Pour comparer les fréquences de ces deux groupes afin d'inférer une relation entre X (ex: sexe) et Y (Ex : les réponses - oui ou non - à la question. Les tests statistiques comme le khi carré permettent au chercheur de rejeter ou non l'hypothèse nulle, donc de prendre une décision.

Avant de procéder à ce test, il faut formuler vos hypothèses statistiques (Ho et H1). Formuler les Hypothèses statistiques d'un Khi Carré : Dans la logique d'un test d'hypothèse, il y a toujours deux hypothèses statistiques.

La première - l'hypothèse nulle ou Ho - est, comme son nom l'indique, une hypothèse qui postule qu'il n'y a pas de différence entre les fréquences ou les proportions des deux groupes; notez : Groupe 1 = Groupe 2.

La seconde - l'hypothèse alternative ou H1 - correspond habituellement à l'hypothèse de votre recherche. Contrairement à l'hypothèse nulle, cette hypothèse postule qu'il existe une différence entre les fréquences des deux groupes, différence qui ne serait pas due au hasard.

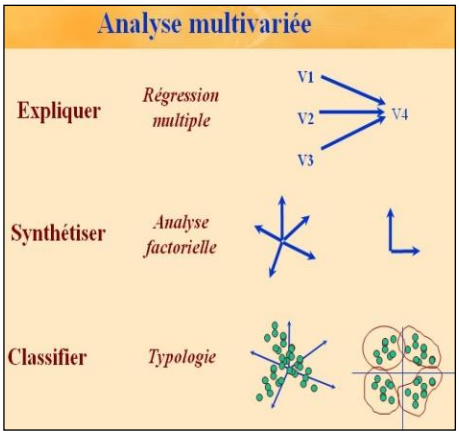
L'existence de cette différence permet d'inférer que X est bel et bien la cause de Y. Le test de Chi-deux est utilisé pour tester l'hypothèse nulle d'absence de relation

entre deux variables catégorielles. On peut également dire que ce test vérifie l'hypothèse d'indépendance de ces variables.

Si deux variables dépendent l'une de l'autre, elles partagent quelque chose, la variation de l'une influence la variation de l'autre…

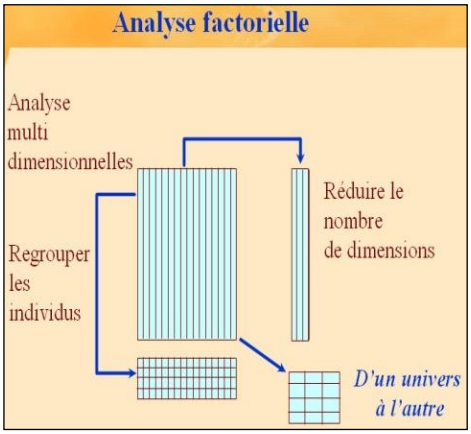
III.3. Analyse multi -variée

L'analyse multi -variée recouvre un ensemble de méthodes destinées à synthétiser l'information issue de plusieurs variables, pour mieux l'expliquer.



Il existe deux grandes catégories de méthodes : les méthodes descriptives et les méthodes explicatives.

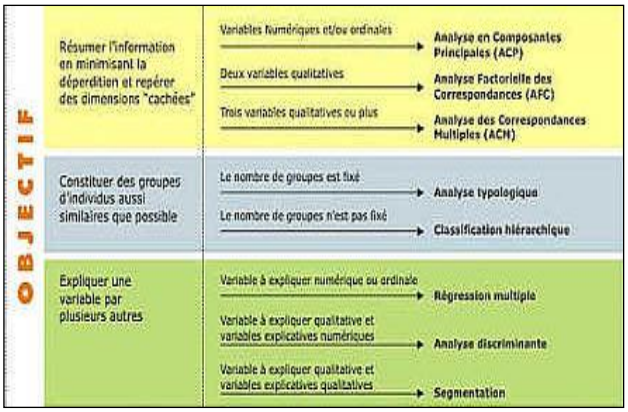
Dans ce cours, nous intéresserons aux méthodes descriptives. Ces méthodes visent à structurer et simplifier les données issues de plusieurs variables, sans privilégier l'une d'entre elles en particulier.



Les méthodes les plus utilisées dans le traitement des enquêtes sont

- l'analyse en composantes principales (ACP),
- l'analyse factorielle des correspondances (AFC),
- l'analyse des correspondances multiples (ACM),
- la typologie et les méthodes de classification.

Les différentes méthodes d'analyse multi variée permettent de répondre à des problématiques variées. Le choix d'une méthode dépend de l'objectif initial, des types de variables manipulées mais aussi, de la forme des résultats obtenus qui peuvent être plus ou moins faciles à présenter et à expliquer.



III.3.1 Analyse en composantes principales(ACP)

L'ACP s'applique à un ensemble de variables numériques. Elle permet de positionner les individus sur un plan en deux dimensions, en fonction de la proximité de leurs réponses aux questions sélectionnées. Les variables sont également représentées sur le mapping, mais de manière indépendante des points -individus.

L'ACP permet ainsi de mettre en évidence la structuration des réponses en montrant le regroupement des individus selon des combinaisons de réponses aux questions prises en compte.

Les axes du mapping ne correspondent généralement pas à l'une ou l'autre des variables mais à un regroupement optimal de plusieurs variables (ex : revenu et niveau d'études peuvent participer ensemble à la formation d'un axe dans la mesure où elles peuvent être fortement corrélées).

Le tableau de départ de l'ACP comporte les individus en ligne et les variables en colonne, avec, dans chaque case, la réponse numérique de l'individu à la question correspondante. Les questions qualitatives ordinales, c'est-à-dire, celles dont les réponses peuvent être ordonnées entre elles (échelles, fréquences...) peuvent être re-codifiées pour entrer dans le tableau de l'ACP.

Cette recodification doit être généralement préparée à l'avance. Toutefois, certains logiciels d'analyse statistique permettent de réaliser cette recodification en direct, lors du choix des variables à faire entrer dans l'ACP. L'algorithme de l'ACP effectue sur la matrice individus/variables différentes opérations (centrage- réduction des données, diagonalisation de la matrice, extraction de valeurs propres et de vecteurs propres…), en vue de passer du nombre de variables initial à un petit nombre de variables obtenues par combinaison des premières.

Ces nouvelles composantes forment les axes du mapping. La première composante est celle qui résume le mieux les informations contenues dans le tableau. La deuxième apporte un pourcentage inférieur mais complémentaire d'information, et ainsi de suite.

Le mapping d'ACP représente d'abord la première composante (axe horizontal) et la seconde (axe vertical). La somme des pourcentages d'explication des deux composantes renseigne sur le taux de déperdition d'information à partir des données de base. Ainsi, si la première composante résume 62% du tableau et la seconde 21%,

l'information représentée sur le mapping est de 83%. L'information « perdue » est donc de 17%.

Les points -individus sont représentés sur le mapping en fonction de leurs coordonnées sur les facteurs. Les points proches correspondent à des individus ayant des profils proches, a priori, quant aux réponses aux variables prises en compte dans l'analyse. Les points- variables sont également représentés sur le mapping, mais de manière complètement indépendante des individus.

Leur représentation indique leur corrélation avec les facteurs, à l'intérieur d'un cercle de rayon 1 défini avec une échelle arbitraire (qui peut être changée à loisir sans affecter la représentation des points-individus). Ces points variables renseignent sur le sens à donner aux axes. Ainsi, une variable proche du cercle de corrélation (corrélation forte) et proche d'un axe participe beaucoup à la formation de cet axe.

III.3.2 Analyse des correspondances

L'analyse factorielle des correspondances s'applique à deux variables qualitatives (nominales). Elle permet de positionner sur un mapping les modalités de réponses des deux questions. L'analyse des correspondances Multiples (ACM) généralise l'AFC à un nombre quelconque de variables et permet donc de représenter sur le même mapping les modalités de réponses de plus de deux variables.

Comme pour l'ACP, le but de ces analyses est de dégager des dimensions cachées contenues dans les réponses aux variables sélectionnées, pour faciliter l'interprétation de tableaux pas toujours lisibles au départ.

Le tableau de départ de l'AFC simple est un tableau croisé (tableau de contingence) présentant la ventilation d'une population sur les modalités de réponses de deux questions qualitatives (ex : diplôme et profession).

Comme en ACP, les deux premiers axes fournissent une partie généralement importante de l'information contenue dans le tableau initial (l'axe horizontal étant, par convention, le plus significatif).

La proximité des points renseigne, a priori, sur leurs associations. La disposition des modalités de chaque variable les unes par rapport aux autres aide à donner un sens à chaque axe (Ce qui n'est pas toujours évident, à la seule observation du graphique).

III.3.3 Les méthodes de classification (analyse en cluster ou clustering)

La classification est une méthode de regroupement des individus selon leurs ressemblances ou en groupe homogène.

La différence est que le nombre de groupes n'est pas à fixer a priori et que le résultat est représenté sous la forme d'un arbre de classification. L'élaboration de cet arbre peut être ascendante (méthode la plus fréquemment utilisée), par regroupements successifs des individus ou descendante, par divisions successives.

L'arbre de classification relie un individu à un autre ou à un sous-groupe d'individu issus eux-mêmes de regroupements. Lorsque l'on coupe l'arbre au niveau

du dernier regroupement, on obtient deux groupes d'individus. Si la division est effectuée au niveau de l'avant-dernier regroupement, on obtient trois groupes.

Un ensemble de clusters s'obtient en coupant le dendrogramme à un certain niveau d'agglomération.

Plus simplement, L'utilisateur cherche essentiellement à synthétiser, par la classification, son information en structurant la population par « groupes homogènes ». Pour la structuration, on peut citer toutes les techniques de classification automatique :

A) La classification hiérarchique ascendante

La méthode consiste à :

Sélectionne les deux classes les plus proches (au sens de la mesure de similarité entre classes choisie) et les réunit selon 3 types de lien ;

- ✓ Lien simple : Agglutine à chaque étape les deux clusters ayant la plus petite distance entre leurs voisins les plus proches. Tend à constituer rapidement des gros clusters et isole mal les clusters qui sont mal séparés.
- ✓ Lien complet : Agglutine à chaque étape les deux clusters ayant la plus petite distance entre leurs voisins les plus éloignés. Tend à constituer des petits clusters compacts.
- ✓ Lien moyen de groupe : Agglutine à chaque étape les deux clusters dont les moyennes de distances entre voisins sont les plus faibles. Produit des clusters dont la taille est intermédiaire entre celles des clusters produites par les deux méthodes précédentes.

A) La classification hiérarchique descendante

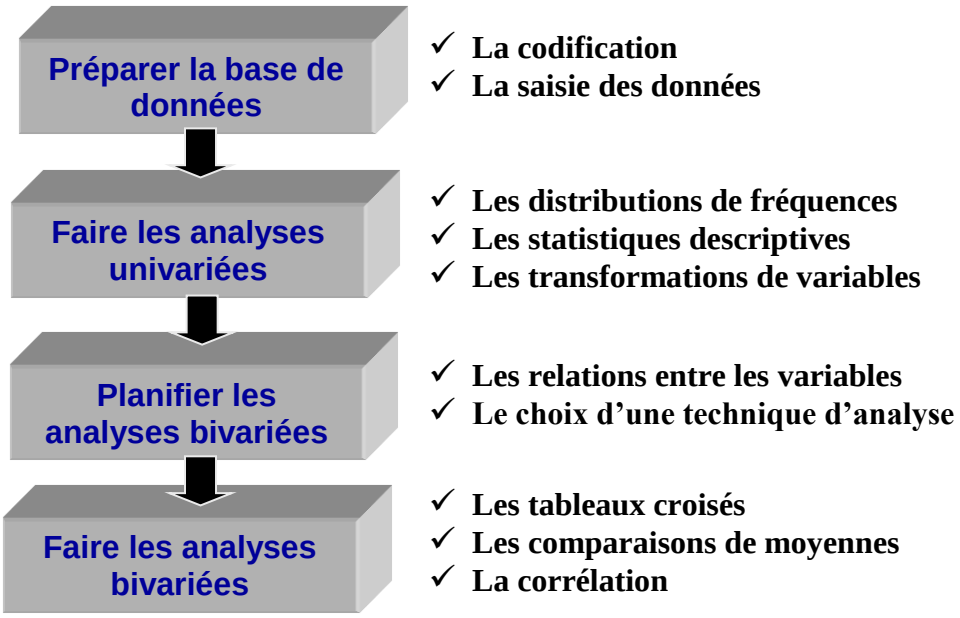
Sélectionne la classe la moins cohérente (au sens de la mesure de similarité entre classes choisie) et la subdivise. Pour réaliser la subdivision, la plupart des algorithmes de classification descendante ont besoin d'une classification ascendante. En raison de cette récursivité, ils sont donc moins attractifs que les algorithmes ascendants.

IV Analyse de données par SPSS

Actuellement, parmi les outils informatique capables de gérer de grandes bases de données, faire de la statistique de haut niveau et proposant un outil intégré d'analyse de données est le système SPSS. Il est largement répandu dans les grandes entreprises. SPSS, dont le sigle anglais signifie Statistical Package for the Social Sciences, est un programme informatique d'analyse de données statistiques. Il permet de saisir des données, d'en faire des présentations résumées (tableaux, graphiques), de les organiser et surtout de les analyser (**TP0** annexe).

Dans ce qui suit nous apprendrons ce que sont ces fichiers et comment les utiliser. Le programme SPSS est constamment mis à jour et amélioré, mais les différences entre les versions ne sont pas toujours majeures. Nous utiliserons ici la version 17.

La démarche générale pour le traitement de données sous SPSS est schématisée comme le montre la figure ci-dessous :



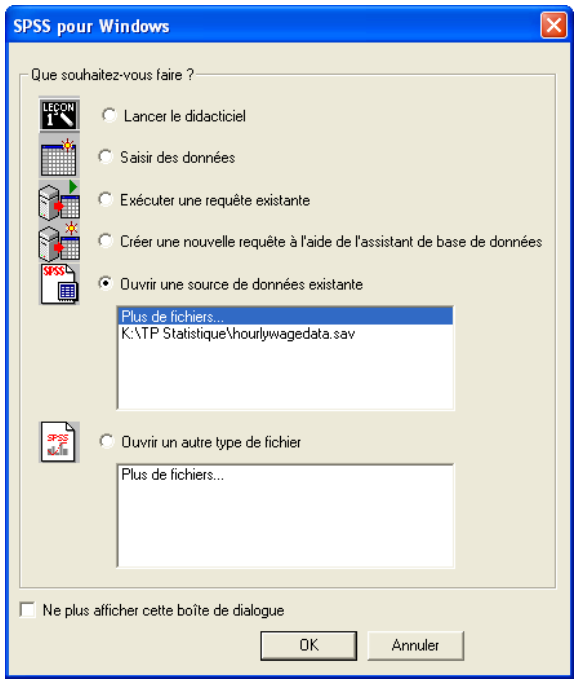
IV.1. Préparation de la base de données SPSS

Cette section porte principalement sur la préparation de votre base de données SPSS. Elle présente les étapes pour créer une base de données à partir d'un questionnaire en précisant tous les aspects à documenter pour chaque variable. De plus, elle indique les techniques pour ajouter des variables ou des cas à un fichier existant.

IV.1.1 Lancer SPSS sous Windows

Pour utiliser SPSS sous Windows, cliquer sur **Démarrer > Programmes > SPSS Inc > Statistics 17.0 >SPSS Statistics 17.0**. On obtient soit un fichier de données vide, ou alors la figure ci-contre que l'usager a le choix de faire apparaître ou non au démarrage de SPSS.

Cette boîte de dialogue vous donne plusieurs choix : ouvrir un fichier existant si vous choisissez l'option ouvrir une source de données existante et puis plus de fichiers… qui est sélectionnée par défaut, et que vous cliquez OK. Vous pouvez aussi créer un nouveau tableau de données et demander à utiliser le didacticiel qui vous permettra d'apprendre SPSS, comme vous pouvez ne plus avoir l'affichage de cet assistant.



Dès le lancement, deux fenêtres apparaissent :

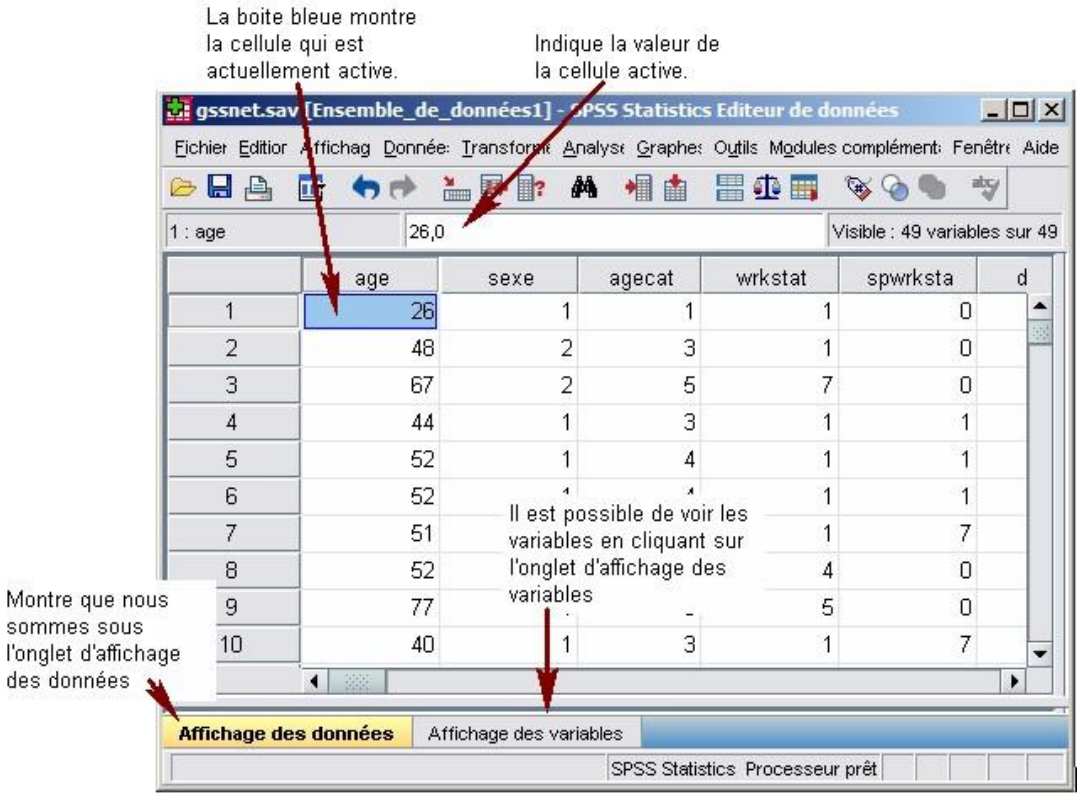
1. la fenêtre de l'éditeur de données qui vous permettra de saisir le tableau des données
2. la fenêtre de l'éditeur de syntaxe qui recueillera vos instructions d'analyse.

IV.1.2 Création de variables

La première étape de tout traitement des données doit nécessairement être celle de la préparation du fichier de données.

Lorsque la fenêtre Affichage de données est sur votre écran figure ci dessous, notez la forme rectangulaire de la fenêtre :

- Les lignes seront des cas (on peut entrer un nombre impressionnant de lignes);
- Les colonnes, des variables (on peut entrer un nombre astronomique de variables);
- Les cellules contiendront les valeurs des données.



	age	sexe	agecat	wrkstat	spwrksta	d
1	26	1	1	1	0	
2	48	2	3	1	0	
3	67	2	5	7	0	
4	44	1	3	1	1	
5	52	1	4	1	1	
6	52	1	4	1	1	
7	51	2	3	1	7	
8	52	2	4	4	0	
9	77	.	.	5	0	
10	40	1	3	1	7	

L'ensemble des lignes utilisées (observations, sujets ou cas) et des colonnes définies (variables) se nomme matrice de données. Cette matrice peut être vide (aucune donnée d'entrée), complète (toutes les cellules ont une valeur) ou incomplète (une proportion des cellules sont vides, c'est-à-dire qu'elles contiennent des données manquantes).

Il y a deux principales façons de créer un fichier de données de format SPSS :

1. Directement (de manière interactive) à l'aide de l'éditeur de données. Pour ce faire, vous devez entrer les variables dans l'onglet Affichage des variables, puis remplir les colonnes en indiquant les valeurs de chaque variable pour chaque cas (exemple de fichier SAV).
2. Par importation de fichiers de données
   - ✓ fichiers texte (RTF, Tab delimited) fichiers SPSS de versions antérieures
   - ✓ tabulateurs (Excel, Lotus 1-2-3)
   - ✓ base de données (Dbase, fichiers ODBC)
   - ✓ fichiers SAS… (autres logiciels d'analyses statistiques).

IV.1.3 Paramètres des variables

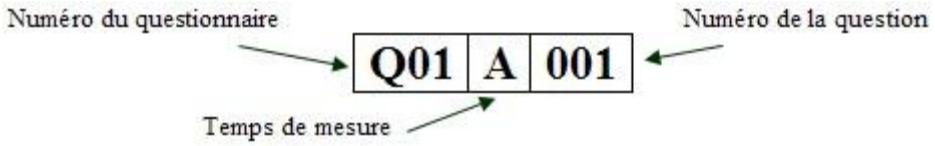
Avant de saisir les données, il faut définir chaque variable de la matrice de données à l'aide de la fenêtre qui correspond à l'onglet Affichage des variables de l'Éditeur de données.

Contrairement à la facette Affichage des données, cette facette de l'Éditeur de données présente les variables horizontalement (lignes) et les colonnes correspondent aux paramètres essentiels qui définissent chaque variable de la matrice de données que nous voulons créer. La facette Affichage des variables sert uniquement à la création et à la modification des variables de la matrice de données.

La création de toute variable nécessite la définition de ces paramètres essentiels :

1. **NOM** : Contient le NOM de la variable (maximum 8 caractères).

Habituellement, on tente de codifier le nom des variables pour les reconnaître facilement. Lorsqu'il y a quelques questionnaires différents, on adopte un code qui peut ressembler à ceci.



2. **TYPE** : Décrit le format de la variable (de la donnée recueillie)
3. **Numérique** : donnée numérique, nombre (décimale)
4. **Date** : donnée sous forme de date
5. **Chaîne** : donnée alpha-numérique (texte).
6. **LARGEUR** : Représente le nombre de caractères maximum de la donnée.
7. **DÉCIMALES** : Indique le nombre de décimales désirées.
8. **ÉTIQUETTE** : Ce champ très important est l'étiquette du nom de la variable (sa description).
9. **VALEURS** : Contient la déclaration des valeurs possibles d'une variable catégorielle et leurs étiquettes respectives.
10. **MANQUANT** : Contient la déclaration des valeurs possibles des données manquantes.
11. **COLONNES** : Décrit la largeur graphique de la colonne (en nb de caractères). Habituellement, on laisse la valeur par défaut de 8 caractères.
12. **ALIGN** : Aligne la donnée à l'intérieur de la colonne (droite, gauche, etc.).
13. **MESURE** : Définit la nature de la variable

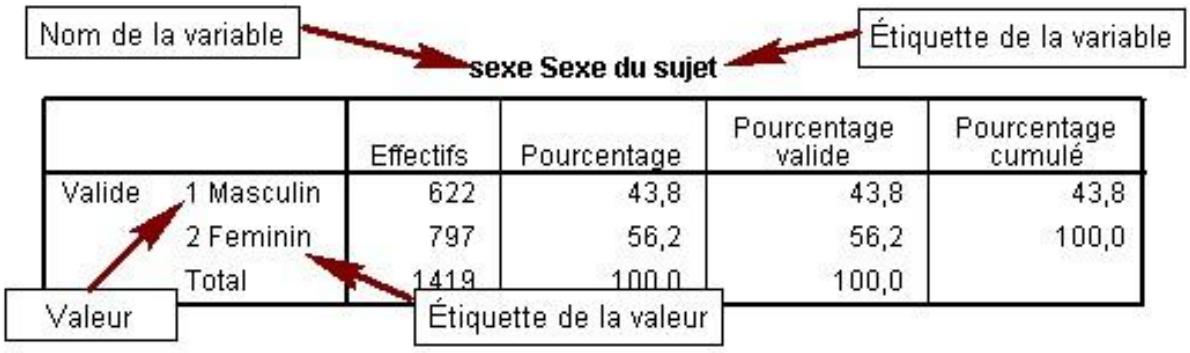
    Échelle : variable continue (intervalle ou de rapport) (ex :l'âge)

    Ordinale : variable catégorielle ordinale (ex : scolarité)

    Nominale : variable catégorielle nominale (ex : sexe).

Il faut toujours garder à l'esprit que SPSS utilise un vocabulaire très précis afin d'identifier, pour chaque variable comprise dans un fichier, son nom, son étiquette, ses

valeurs et les étiquettes des valeurs. Le tableau d'effectifs, ci-dessous, illustre les différentes composantes qui servent à la nomenclature d'une variable.



		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	1 Masculin	622	43,8	43,8	43,8
	2 Feminin	797	56,2	56,2	100,0
	Total	1419	100,0	100,0	

Exemple de fichier .SAV

Dans le but de bien comprendre le processus de création d'une variable, nous allons procéder à un premier TP1. Nous allons tenter de créer un fichier de données à partir d'un questionnaire qui supporte le **TP1** (voir annexes). Il s'agit d'un court questionnaire portant sur la satisfaction des clients d'un Carrefour Jeunesse Emploi (CJE) quelconque au Québec.

IV.1.4 Ajout de cas/variables

Il est possible d'ajouter de nouvelles observations (cas) ou variables en ajoutant des lignes supplémentaires (observations) ou des colonnes supplémentaires (variables) à un fichier de données existant.

A) Nouvelles observations

Celles-ci peuvent être ajoutées :

1. À la fin de la matrice en entrant tout simplement les données de ces nouvelles observations sur de nouvelles lignes;
2. Au sein des lignes déjà existantes en utilisant la commande « **Insérer les observations** » du menu **Édition** de l'Éditeur de données ou, plus simplement, en appuyant sur le bouton  .

B) Nouvelles variables

Elles peuvent être ajoutées :

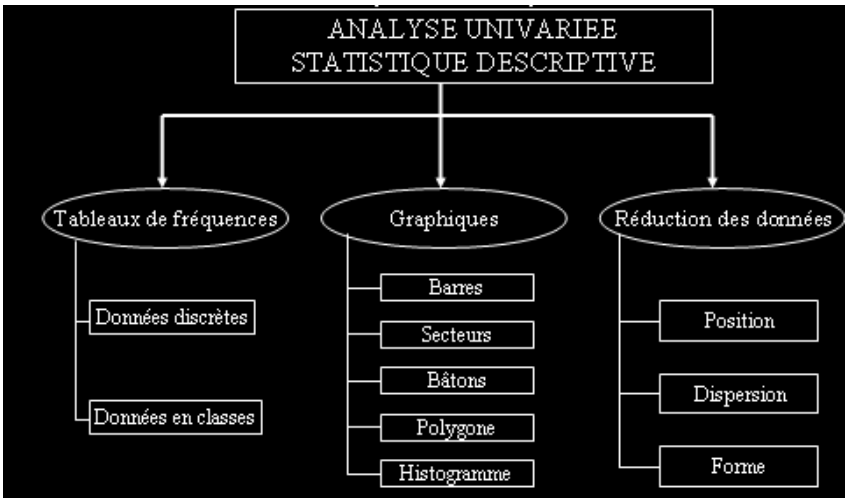
1. À la droite de la matrice en définissant de nouvelles variables sous l'onglet Affichage de variables;
2. Entre des variables existantes en utilisant la commande « **Insérer une variable** » du menu Édition de l'Éditeur de données ou le bouton  .

Cependant, il est aussi possible d'ajouter de nouvelles observations ou variables contenues dans d'autres fichiers SPSS à l'aide des commandes Ajouter des variables et Ajouter des observations à partir de la fonction **Fusionner des fichiers** du menu **Données**. Dans ce cas, les nouvelles observations ou variables sont contenues dans des fichiers SPSS distincts créés avant de procéder à la fusion (**TP2** Annexe).

IV.2. Analyse univarié

L'analyse univarié revient à décrire et synthétiser les résultats de la recherche en analysant les variables une à la fois.

Pour choisir la procédure convenable en SPSS à votre analyse univarié, il faut faire attention à l'échelle de mesure de la variable étudiée : il ne sert à rien de calculer une moyenne quand la variable est qualitative, par exemple. Les tableaux de fréquences sont appropriés quand on a un nombre restreint de catégories, et qu'on veut mesurer leur importance relative ou absolue. Par contre, cette même procédure offre de nombreuses options intéressantes pour les variables quantitatives. Les procédures Descriptives… et Explore… ne sont applicables que pour les variables quantitatives.

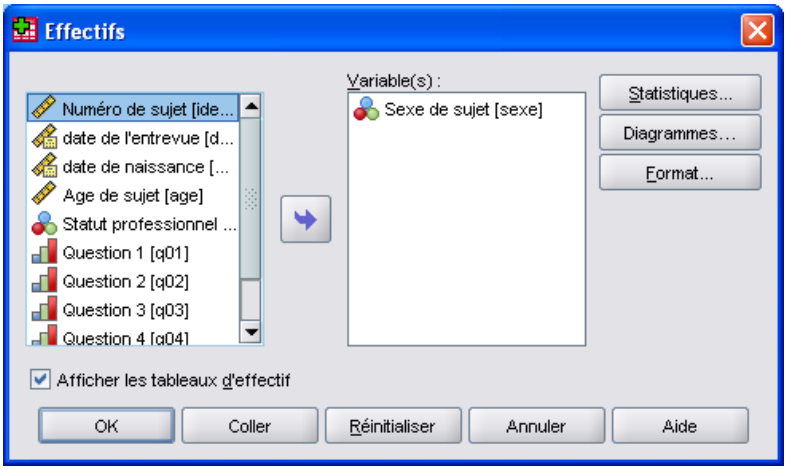


IV.2.1 Variable qualitative

Dans le cas de variables non métriques, on utilise des distributions de fréquences. La procédure SPSS est la suivante :

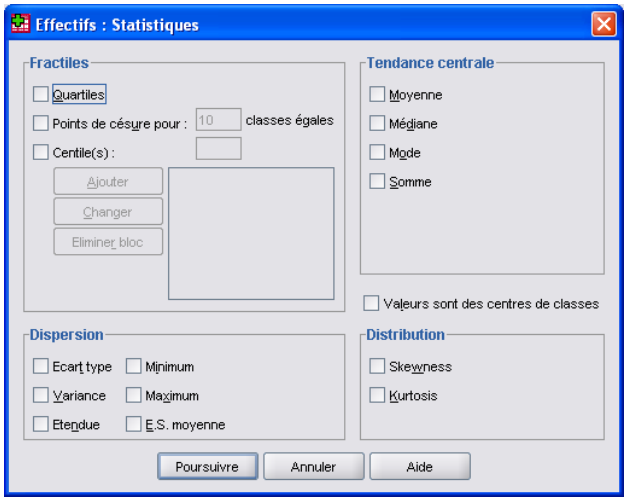
1. Sélectionnez la commande **effectifs** montrée ci-haut

Vous obtenez la boîte de dialogue ci-dessous. Cette procédure est utile quand les variables sont qualitatives, mais elles sont aussi très utiles quand la variable est quantitative mais qu'elle a été regroupée en un nombre restreint de catégories, comme par exemple pour la variable Age Catégories (enfance, adolescence …).



Toutes les variables du fichier sont inscrites dans la partie gauche de cette boîte de dialogue. Pour obtenir le tableau de fréquence d'une variable, il faut la sélectionner, puis la placer dans l'espace prévu à droite en cliquant sur le bouton contenant une mini-flèche. Remarquez qu'il y a plusieurs boutons permettant de spécifier des options.

2. Sélectionnez les variables Status et sexe et placez-les dans l'espace prévu à droite. Laissez le petit carré de l'option **Afficher les tableaux d'effectif** sélectionné.
3. Cliquez maintenant sur le bouton Statistiques….Vous obtenez la boîte de dialogue suivante :



Il y a quatre sections dans cette boîte de dialogue, chacune permettant un type de mesure descriptives : des mesures de position telles que les quartiles ou les percentiles, des mesures de tendance centrale, des mesures de dispersion, et des mesures qui décrivent la distribution dans son ensemble. Revoyez les définitions de ces termes vues au début du cours. Si la variable est qualitative, seul le Mode sera utile parmi ces mesures.

4. Cliquez sur poursuivre, vous reviendrez à la boîte de dialogue Effectifs.
5. Cliquez sur le bouton Diagrammes… . Vous permet de réaliser des graphiques sans passer par le menu Graphes à travers la boite de dialogue suivante :

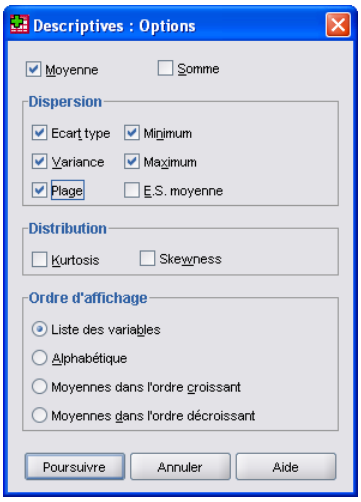
On a le choix entre plusieurs types de graphiques. Choisissez bâtons, secteurs, Histogramme et cliquez poursuivre.

IV.2.2 Variable quantitative

Dans le cas de variables métriques, on utilise les statistiques descriptives (mesures de tendance centrale et de dispersion), la procédure SPSS est la suivante :

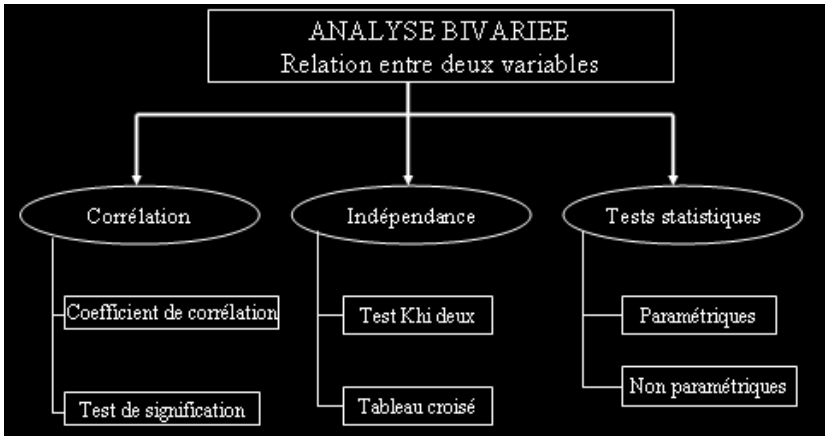
Cette commande n'est appropriée que pour les variables quantitatives, préférablement mesurées avec une échelle d'intervalle ou de ratio (scale dans la terminologie de SPSS). Nous allons l'illustrer avec un exemple que vous êtes invité à exécuter sur votre poste de travail (TP3 Annexe).

1. Après avoir Sélectionnez la commande Descriptives
2. Placez les variables que vous voulez analyser dans l'espace désigné par Variables du côté droit de la boîte de dialogue obtenue. Nous allons le faire pour la variable Age.
3. Cliquez sur le bouton Options pour spécifier les statistiques que vous souhaitez obtenir. Vous obtenez la boîte de dialogue ci-dessous. Remarquez que vous ne pouvez pas obtenir de graphiques par l'entremise de cette commande.

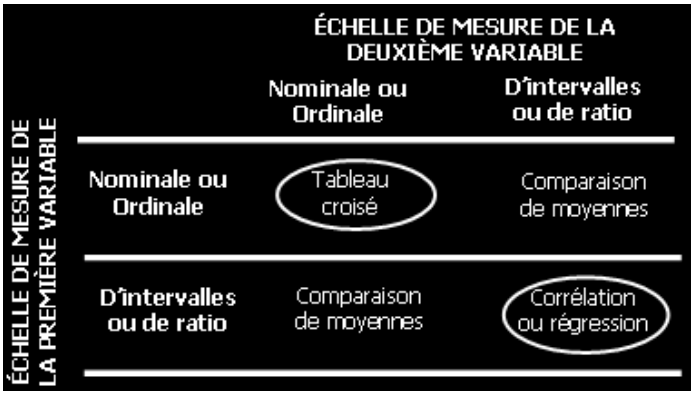


IV.3. Analyse bivarié

On s'intéresse dans ce type d'analyse aux relations qui existent entre des variables prises deux à la fois. On distingue deux types de relation bivariée : les relations de dépendance (plus fréquentes) et les relations d'interdépendance.



Le choix de la technique d'analyse appropriée dépond de l'échelle de mesure des variables comme illustré sur le tableau suivant.



IV.3.1 Variables quantitatives

Quand les deux variables sont quantitatives, l'association statistique entre elles prend la forme de la corrélation. Ce terme est synonyme du terme : association statistique entre variables quantitatives.

Nous allons effectuer deux opérations distinctes. L'une consiste à produire les coefficients de corrélation entre des variables, et l'autre à dessiner le nuage de points et à obtenir la ligne de régression. SPSS peut produire les coefficients de corrélation entre plusieurs variables prises deux à deux, d'un seul coup. On général on fait cette opération dans un premier temps pour explorer la situation, afin de déceler les relations significatives, puis on analyse avec plus de détails ces relations.

Ce type d'analyse est employé lorsque les deux variables sont mesurées avec des échelles métriques.

On cherche à établir si l'augmentation des valeurs d'une des deux variables entraîne systématiquement l'augmentation ou la diminution des valeurs de l'autre variable.

Un coefficient de corrélation r est une mesure d'interdépendance ou d'association entre deux variables métriques. Elle mesure l'intensité de la co-variation entre les deux variables. Cette mesure ne dépend pas de l'unité utilisée pour chaque variable, elle est comprise entre deux valeurs extrêmes -1 et 1.

Plus le coefficient est proche de 1 en valeur absolue, plus les valeurs sont dites corrélées

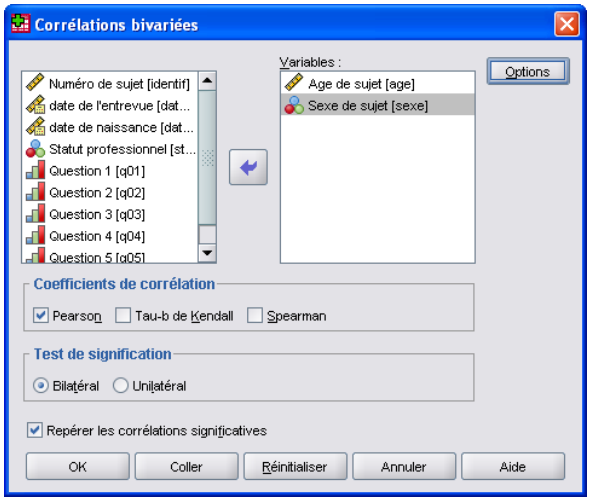
- Si r est proche de +1, ceci signifie que les deux variables varient dans le même sens
- Si r est proche de -1, ceci signifie que les deux variables varient en sens inverse l'une de l'autre
- Plus r est proche de 0, moins les variables sont corrélées. 0 signifie l'absence de corrélation entre les deux variables

A) Calcul des coefficients de corrélation

Pour obtenir les coefficients de corrélation avec SPSS, nous allons exécuter les étapes suivantes.

A) Ouvrez le fichier .**sav**.

B) Sélectionnez : **Analyse** > **Corrélation** > **Bivariée**. Vous obtenez la boîte de dialogue suivante :



C) Prenez le temps d'examiner les variables présentes dans le fichier, et surtout leur échelle de mesure. Et sélectionner les variables métriques à étudier.

D) Placez les variables sélectionnées dans l'espace prévu à cet effet à droite.

E) Cliquez OK (TP4 Annexe).

Vous aurez sans doute remarqué qu'en plus de donner le coefficient de corrélation (appelé coefficient de Pearson), le tableau vous donne aussi un niveau de signification, et le nombre de cas qui ont été inclus dans le calcul. Le niveau de signification nous dit quel risque de se tromper on prend si on prétend que la relation observée est valable pour l'ensemble de la population étudiée en supposant évidemment que les données que l'on a constituent un échantillon représentatif). Le nombre de cas utilisé est important car il se peut qu'il y ait des données manquantes.

Remarquez aussi qu'il y a une certaine redondance dans le tableau. La corrélation d'une variable avec elle-même est toujours 1. De plus, la corrélation entre x et y est la même qu'entre y et x. Donc, une partie du tableau aurait pu être omise, et certaines versions de SPSS omettent effectivement certaines des cellules redondantes.

B) Illustration graphique et ligne de régression

Les étapes suivantes vont nous permettre d'illustrer la situation des corrélations.

1. Cliquez Graphes Interactif diagramme de dispersion.
2. Dans la boîte de dialogue qui en résulte, faites glisser la variable x vers l'axe horizontal du graphique, et la variable y vers l'axe vertical.
3. Cliquez sur l'onglet dénommé **Fit** dans la partie supérieure de la boîte de dialogue. Vous obtenez une nouvelle boîte de dialogue : assurez-vous que l'option **Régression** a bien été choisie, et que la petite boîte correspondant au mot **Moyenne** n'a pas été sélectionnée.
4. L'onglet **Options** vous permet de choisir plusieurs styles de diagrammes pour le nuage de points.
5. Cliquez OK. Vous devriez obtenir le diagramme

IV.3.1 Variables qualitatives

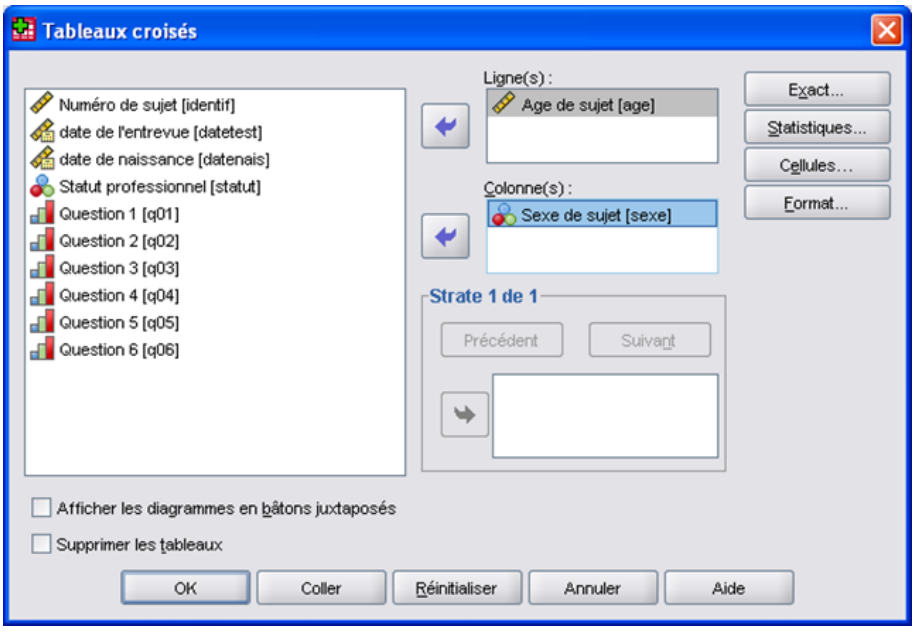
L'analyse de deux variables non métriques s'effectue à l'aide de fréquences conjointes (tableau de contingence). Il est nécessaire de recourir au test d'indépendance du Khi-deux lorsque l'analyse porte sur une relation bivariée comprenant deux variables non métriques (nominales ou/et ordinales).

A) Tableau croisé

Le tableau croisé contient les fréquences correspondant au croisement des caractéristiques qui définissent les deux variables. Les tableaux croisés à deux ou plusieurs modalités sont en général complétés par des mesures d'association qui permettent de démontrer la signification statistique d'une association observée entre les variables. Ces tests seront développés dans la section suivante.

Les tris croisés ont pour objet de rassembler dans un tableau unique les distributions de fréquences ou d'effectifs de deux ou plusieurs variables. Ce premier outil d'analyse examine la relation entre deux variables catégorielles. Il décrit comment subdiviser les fréquences de proportion d'une variable catégorielle avec une autre variable catégorielle.

Pour obtenir un tableau croisé (TP4 Annexe), il faut sélectionner : **Analyse** > **statistiques descriptives** > **Tableaux croisés** la boite de dialogue suivante apparait à l'écran :



Il faut sélectionner les deux variables à croiser puis cliquer sur cellules pour obtenir la boite de dialogue : cocher ligne, position et total dans les pourcentages puis poursuivre; dans statistiques il faut sélectionner Khi-deux puis poursuivre.

Le problème des tris croisés est qu'ils ne permettent pas de donner une conclusion ferme sur l'existence et/ou la force d'une relation entre les variables. Pour mesurer véritablement la relation entre les variables, il est nécessaire de mettre en place des tests de signification statistique de l'association. Dans la section suivante on s'intéressera au test du khi-deux pour vérifier l'association de deux variables qualitatives.

B) Test du khi-deux (χ2)

Le test du khi-deux (χ2) est couramment utilisé. Il cherche à tester si deux variables qualitatives (nominales ou ordinales) sont significativement associées. En réalité, c'est l'indépendance des variables qualitatives, présentées dans un tableau croisé, qui est testée. On cherche à vérifier si l'association des deux variables est suffisamment forte pour que l'hypothèse de leur indépendance puisse être rejetée.

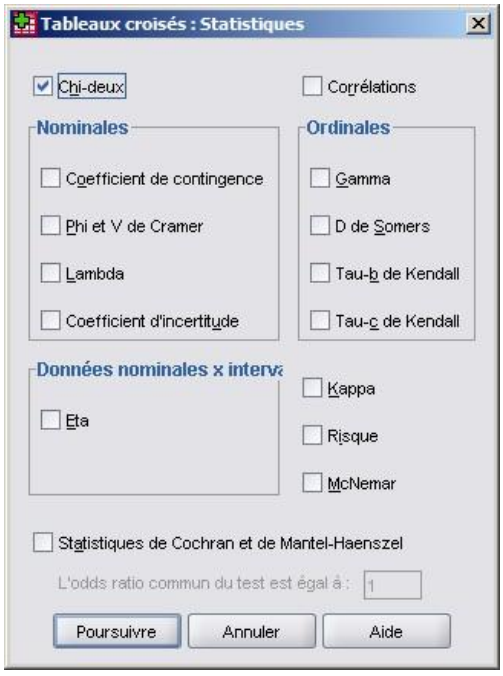
Le test du Chi-deux est un test de validation d'hypothèses. Il s'applique aux tableaux croisés. On calcule une statistique qui mesure l'écart entre une situation théorique où il n'y aurait pas d'association statistique, et une situation observée. Cette statistique suit une distribution connue, qui dépend du nombre de catégories des variables étudiées.

On pose donc :

- ✓ H0 : Il n'y a pas d'association statistique entre les variables.
- ✓ H1 : Il y en a.

On peut effectuer les calculs qui vont nous amener à accepter l'hypothèse nulle ou à la rejeter (TP4 Annexe). SPSS va en effet nous donner la probabilité qu'on obtienne la valeur du Chi-deux observée sur l'échantillon s'il n'y avait aucune association statistique au niveau de la population. On rejette l'hypothèse nulle si la probabilité est plus petite que le seuil qu'on s'est fixé. SPSS, donnent une signification ou p-value, s'interprétant comme le niveau risque de se tromper en rejetant H0. Ainsi, si elle est inférieure à 5 %, on rejette l'hypothèse d'indépendance entre les deux variables, qui sont alors significativement associées.

L'option Statistiques de la boite de dialogue **tableaux croisés** (figure ci-dessus) est utilisée pour les statistiques inférentielles. Elle permet de choisir le type de test que vous désirez utiliser pour évaluer s'il y a des différences significatives entre les groupes de variables. Pour obtenir le khi carré et le V de Cramer : Sélectionné les tests Chi-deux et Phi et V de Cramer comme sur la figure suivante :

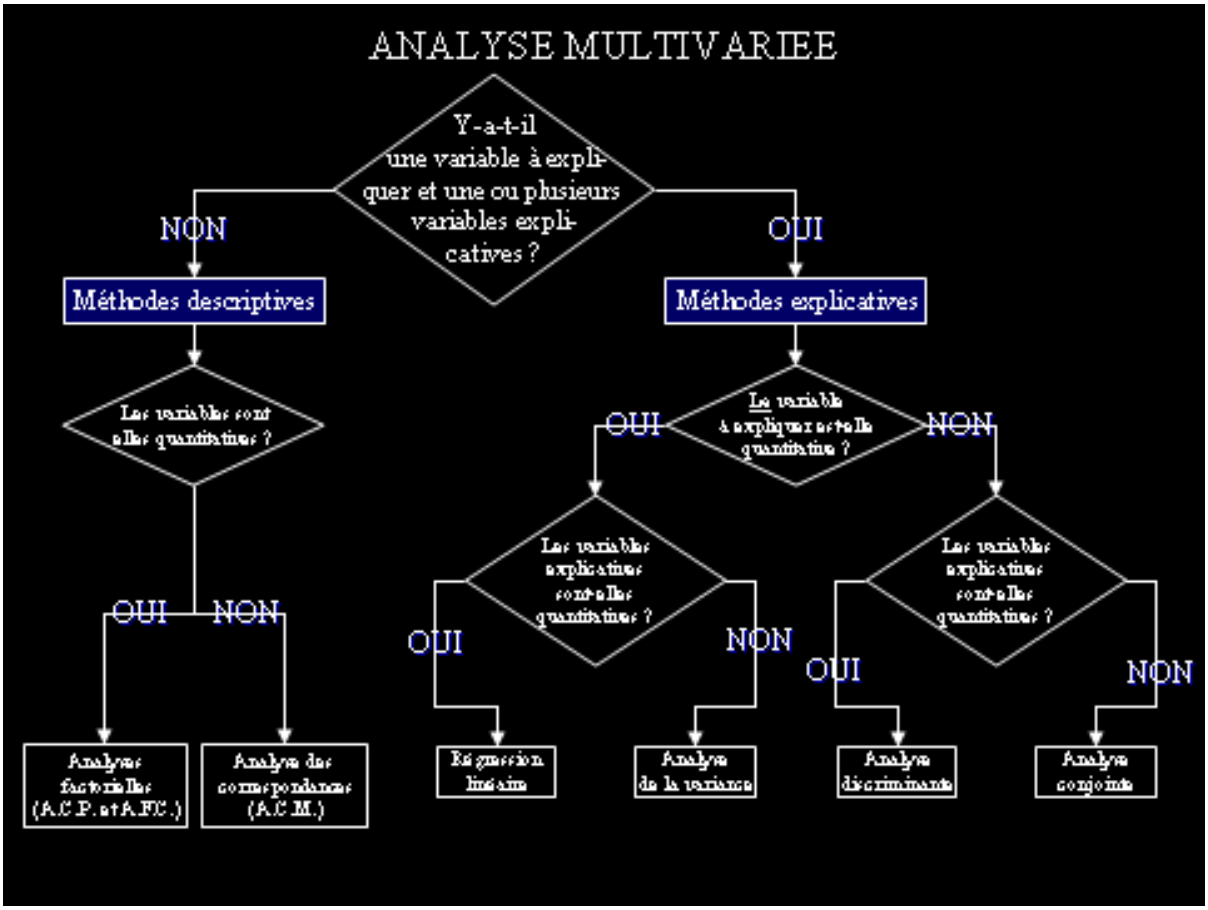


IV.4. Analyse multivariée : ACP

L'analyse multivariée vise l'étude de plusieurs variables en même temps, elle recouvre un ensemble de méthodes destinées à synthétiser l'information issue de plusieurs variables. Ces méthodes visent à structurer et simplifier les données issues de plusieurs variables, sans privilégier l'une d'entre elles en particulier. Les méthodes les plus utilisées dans le traitement des enquêtes sont :

- l'analyse en composantes principales (ACP),
- l'analyse factorielle des correspondances (AFC),
- l'analyse des correspondances multiples (ACM),
- la typologie et les méthodes de classification.

Le choix de l'une ou de l'autre de ces méthodes dépend des objectifs poursuivis et du type de données à analyser comme le montre la figure ci-dessous :



Dans ce cours on se contente de présenter l'analyse en composantes principales (ACP). C'est la méthode d'analyse la plus connue et la plus utilisée (dans 80% des cas) son objectif est :

1. de réduire des tableaux de grandes tailles en un petit nombre de variables (2 ou 3 généralement) tout en conservant un maximum d'information.
2. de représenter sous forme graphique l'essentiel de l'information contenue dans un tableau de données quantitatif.

IV.4.1 Postulats d'utilisation de l'ACP

Les postulats d'utilisation d'une ACP sont :

1. Taille d'observation (minimum 5 observations par variable, l'idéal est 10);
2. Echelle de mesure = variable métrique (donc pas catégorielle);
3. Variables doivent être indépendantes les unes des autres;

Pour exécuter une ACP il faut :

❶ s'assurer qu'il existe des corrélations minimales entre les variables qui feront l'objet de l'analyse. Dans le cas où les corrélations sont très faibles ou inexistantes, l'ACP n'est probablement pas l'analyse à conseiller. À cet égard, on peut créer une matrice de corrélation avec toutes les variables de l'analyse. Cette matrice est une option disponible dans le menu SPSS de l'analyse factorielle.

❷ Vérifier l'Indice KMO (Kaiser-Meyer-Olkin) : varie entre 0 et 1 et donne un aperçu global de la qualité des corrélations à l'examen de la matrice de corrélation. Son interprétation se fait comme suit:

- ✓ • 0,50 et moins est misérable
- ✓ • entre 0,60 et 0,70, c'est médiocre
- ✓ • entre 0,70 et 0,80 c'est moyen
- ✓ • entre 0,80 et 0,90 c'est méritoire
- ✓ • et plus 0,9 c'est merveilleux.

❸ Test Bartlett de la sphéricité : Cette mesure indique si la matrice de corrélation est une matrice identité à l'intérieur de laquelle toutes les corrélations sont égales à zéro. Nous espérons que le test soit significatif ($p < 0,05$) pour que nous puissions rejeter l'hypothèse nulle voulant qu'il s'agisse d'une matrice identité qui signifie que toutes les variables sont parfaitement indépendantes les unes des autres.

IV.4.2 Procédure de l'ACP de SPSS

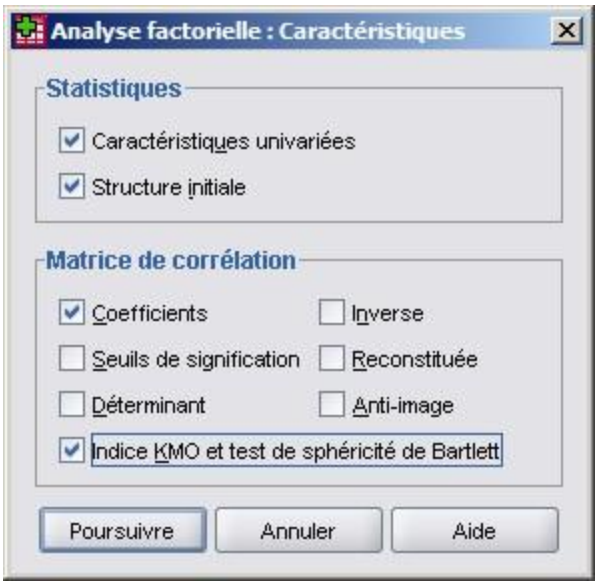
La mise en œuvre d'une Analyse en Composantes principales (ACP) peut être effectuée au moyen de la procédure d'Analyse Factorielle de SPSS : vous sélectionnez **Analyse, Réduction des dimensions**, puis **Analyse factorielle**. La boite de dialogue suivante s'ouvre.

A) Sélection des variables

Sélectionner les variables numériques choisies pour l'ACP (minimum : 2 variables) parmi celles figurant dans la liste source en les transférant dans la liste des Variables à l'aide du bouton. Il suffit alors de cliquer sur le bouton OK pour effectuer une analyse factorielle avec les paramètres prévus par défaut. On obtient alors le listage de la structure initiale, la matrice des corrélations variables-facteurs et les statistiques concernant la structure finale.

B) Statistiques descriptives

Afin de choisir les statistiques optionnelles de la procédure **Descriptives**, pour ouvrir la boîte de dialogue secondaire **Caractéristiques** permettant d'effectuer ces choix.



La boite Caractéristiques vous offre différentes options :

Statistiques

Vous pouvez choisir l'une ou plusieurs des statistiques suivantes :

- ✓ Caractéristiques univariées. Affichage du nombre d'observations valide, de la moyenne et de l'écart-type pour chaque variable.
- ✓ Structure initiale. Communautés de la solution initiale, valeurs propres et pourcentage d'inertie expliquée. Il s'agit de la statistique descriptive choisie par défaut.

Matrice des corrélations

Vous pouvez choisir l'un ou plusieurs des indicateurs statistiques suivants :

- ✓ Coefficients. Matrice des coefficients de corrélation pour les variables actives.
- ✓ Seuils de signification. Seuils unilatères de signification des coefficients de corrélations.
- ✓ Déterminant. Déterminant de la matrice des corrélations.
- ✓ Indice KMO et test de Bartlett. Indice de Kaiser-Meyer-Olkin pour la mesure de la qualité d'échantillonnage et test de sphéricité de Bartlett.

- ✓ Inverse. Inverse de la matrice des corrélations.
- ✓ Reconstituée. Matrice des coefficients de corrélations reconstitués et leurs résidus. Les coefficients de corrélation sont affichés en dessous de la diagonale tandis que les résidus sont situés au-dessus.
- ✓ Anti-image. Anti-images des matrices de corrélation et de variance-covariance. La mesure de la qualité de l'échantillonnage pour chaque variable est affichée sur la diagonale de l'anti-image de la matrice des corrélations.

C) Extraction des facteurs

Afin de choisir une méthode d'extraction des facteurs, obtenir un histogramme des valeurs propres ou contrôler le nombre de facteurs à extraire, cliquez sur le bouton Extraction ... pour ouvrir la boîte de dialogue secondaire permettant d'effectuer ces choix.

Méthode

Vous pouvez choisir une ou plusieurs des méthodes d'extraction suivantes :

- ✓ Composantes principales. Analyse en composantes principales. Il s'agit de la méthode d'extraction par défaut.
- ✓ Moindres carrés non-pondérés. Méthode des moindres carrés ordinaires (MCO).
- ✓ Moindres carrés généralisés. Méthode des moindres carrés généralisés (MCG).
- ✓ Maximum de vraisemblance. Méthode du maximum de vraisemblance (EMV).
- ✓ Factorisation en axes principaux. Méthode de la factorisation en axes principaux.
- ✓ Alpha-maximisation. Méthode d'alpha-maximisation.
- ✓ Factorisation en projections. Méthode de la factorisation en projections.

Extraire

Vous pouvez choisir l'un des critères d'extraction suivants :

- ✓ **Valeurs propres supérieures à**. Dans l'option par défaut, les facteurs correspondant aux valeurs propres supérieures à 1 sont extraits. Pour obtenir un nombre de facteurs différents, modifiez cette valeur par un nombre compris entre 0 et le nombre total de variables actives.

- ✓ **Nombre de facteurs**. Permet d'extraire un nombre de facteurs spécifié. Entrez un entier positif.

Afficher

Vous pouvez choisir une ou plusieurs options d'affichage :

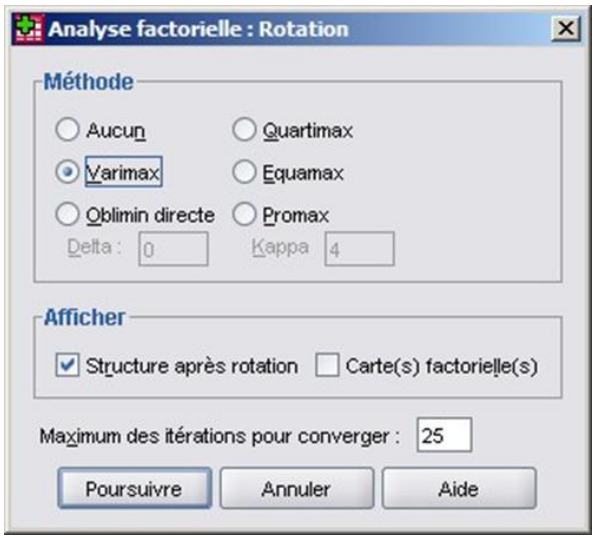
- ✓ **Structure factorielle sans rotation**. Coordonnées factorielles (matrice des corrélations variables-facteurs), communautés et valeurs propres pour la structure factorielle. Option par défaut.
- ✓ **Graphique des valeurs propres**. Histogramme des valeurs propres triées par ordre décroissant. Le graphique affiche les facteurs après rotation si une rotation a été demandé (cf. section « Rotation des facteurs » ci-après).

L'option suivante est également offerte :

Maximum des itérations pour converger. Le nombre maximum d'itérations pour que la procédure d'extraction converge est fixé par défaut à 25. Pour fixer une limite de convergence différente, entrez un entier positif. La valeur du critère de convergence pour l'extraction est égale à 0.001.

D) Rotation des facteurs

Afin de sélectionner une procédure de rotation des facteurs, cliquez sur le bouton Rotation ... pour ouvrir la boîte de dialogue secondaire permettant d'effectuer ces choix. La rotation facilite l'interprétation de la matrice en maximisant le poids de chaque variable sur un facteur et en le diminuant sur les autres.



Méthode. Aucune rotation n'est possible s'il n'y a qu'un seul facteur extrait. La standardisation de Kaiser est utilisée avec n'importe laquelle des méthodes de rotation. Vous pouvez choisir l'une des méthodes de rotation suivantes :

- ✓ **Aucune**. Aucune rotation n'est effectuée. C'est l'option par défaut.
- ✓ **Varimax**. Rotation orthogonale selon la méthode Varimax.
- ✓ **Equamax**. Rotation orthogonale selon la méthode Equamax.
- ✓ **Quartimax**. Rotation orthogonale selon la méthode Quartimax.
- ✓ **Oblimin** directe. Rotation oblique selon la méthode Oblimin.

- ✓ **Delta**. Pour modifier la valeur par défaut 0 du delta, entrez un nombre inférieur ou égal à 0,8 (0.8).

Afficher

Vous pouvez choisir une ou plusieurs options d'affichage :

- ✓ **Structure après rotation**. Affichage par défaut dès qu'une rotation est demandée. Pour les rotations orthogonales, les matrices de la configuration après rotation et de passage dans la nouvelle base sont affichées. Pour les rotations obliques, les matrices de configuration, de structure et de corrélation des facteurs sont affichées. Pour supprimer l'affichage de la structure après rotation, désélectionner cet item.
- ✓ **Carte(s) factorielle(s)**. Graphique à 3 dimensions pour les trois premiers facteurs. Pour une structure à deux facteurs, un graphique-plan est édité. Le graphique ne s'affiche pas s'il n'y a qu'un seul facteur extrait. Les graphiques affichent les structures après rotation, dès qu'une rotation a été demandée.

L'option suivante est disponible si vous avez demandé une rotation :

Maximum des itérations pour converger. Par défaut, un maximum de 25 itérations est prévu pour effectuer une rotation des facteurs. Pour spécifier un maximum différent, entrez un entier positif.

E) Coordonnées factorielles

Afin de sauvegarder les coordonnées factorielles pour les réutiliser dans d'autres analyses, cliquez sur le bouton Facteurs ... pour ouvrir la boîte de dialogue secondaire permettant d'effectuer ce choix.

Pour sauvegarder les coordonnées factorielles des individus, sélectionnez le choix Enregistrer dans des variables, puis choisissez une méthode de calcul de ces coordonnées factorielles :

Enregistrer dans des variables. Sauvegarde les coordonnées factorielles comme variables du fichier SPSS courant. Le listage de l'exécution résume dans un tableau le nom de chacune de ces nouvelles variables ainsi créées et l'étiquette de variable indiquant la méthode de calcul utilisée.

Méthode

Ces options permettent de contrôler la méthode de calcul des coordonnées factorielles :

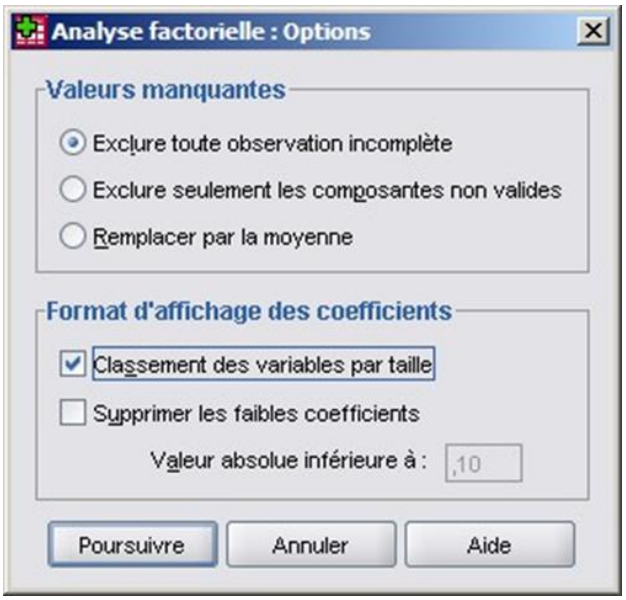
- ✓ • Régression. Méthode de la régression. C'est l'option retenue par défaut, utilisée si les corrélations des facteurs sont acceptables.
- ✓ • Bartlett. Méthode de Bartlett.
- ✓ • Anderson-Rubin. Méthode d'Anderson-Rubin. Si vous désirez que les facteurs ne soient pas corrélés

L'option suivante est également disponible :

Afficher la matrice des coefficients factoriels. Cette option permet d'afficher la matrice des coefficients ainsi que la matrice de variance-covariance des coordonnées factorielles.

F) Options d'analyse

Pour changer le traitement des valeurs manquantes ou permuter l'affichage des matrices de facteurs selon leur ordre d'importance, cliquez sur le bouton Options ... pour ouvrir la boîte de dialogue secondaire permettant d'effectuer ces choix.



Valeurs manquantes

Vous pouvez choisir l'une des options suivantes :

- ✓ Exclure toute observation incomplète. Seules les observations ayant des valeurs valides pour l'ensemble des variables actives sont retenues pour l'analyse.
- ✓ Exclure seulement les composantes non valides. Les observations sont exclues selon une approche bivariée : dans le calcul d'une corrélation, SPSS utilise toute observation ayant des valeurs valides pour les deux variables (même si certaines de ces observations possèdent des valeurs manquantes pour d'autres variables actives).
- ✓ Remplacer par la moyenne. Remplace les valeurs manquantes par la moyenne de la variable puis retient toutes les observations pour l'analyse factorielle.

Affichage des projections

Vous pouvez choisir un ou plusieurs des formats d'affichage suivants pour les projections :

- ✓ **Classement des variables**. Tri des tableaux de structure et de coordonnées factorielles pour que les variables soient triées selon la qualité de leurs projections sur chacun des facteurs.
- ✓ **Supprimer les valeurs absolues inférieures à.** Supprime les coefficients dont la valeur absolue est inférieure à la valeur spécifiée. La valeur par défaut est 0.1. Pour modifier cette valeur par défaut, entrez un nombre compris entre 0 et 1.

Vous pouvez combiner ces deux options en classant les variables et en omettant les coefficients à faible valeur absolue.

G) Analyser les résultats d'une ACP

Analyser les résultats d'une ACP (TP5 Annexe), c'est répondre à trois questions :

1. Les données sont-elles factorisables ?
2. Combien de facteurs retenir ?
3. Comment interpréter les résultats ?

Pour répondre à la première question il faut :

1) Dans un premier temps, il convient d'observer la matrice des corrélations. Si plusieurs variables sont corrélées (> 0.5), la factorisation est possible. Si non, la factorisation n'a pas de sens et n'est donc pas conseillée.

2) Dans un deuxième temps, il faut observer l'indice de KMO (Kaiser-Meyer-Olkin) qui doit tendre vers 1. si ce n'est pas le cas, la factorisation n'est pas conseillée.

3) Enfin, on utilise le test de sphéricité de Bartlett. : si la signification (Sig.) tend vers 0.000, c'est très significatif, inférieur à 0.05 significatif, entre 0.05 et 0.10 acceptable et au-dessus de 0.10, on rejette.

Remarque : Si l'ACP satisfait au moins à deux de ces trois conditions, on peut continuer l'analyse.

Pour savoir le nombre de facteurs à retenir, trois règles sont applicables :

- ➢ **1ere règle** : la règle de Kaiser qui veut qu'on ne retienne que les facteurs aux valeurs propres supérieures à 1.
- ➢ **2eme règle** : on choisit le nombre d'axe en fonction de la restitution minimale d'information que l'on souhaite. Par exemple, on veut que le modèle restitue au moins 80% de l'information.
- ➢ **3eme règle** : le test du coude. On observe le graphique des valeurs propres et on ne retient que les valeurs qui se trouvent à gauche du point d'inflexion. Graphiquement, on part des composants qui apportent le moins d'information (qui se trouvent à droite), on relie par une droite les points presque alignés et on ne retient que les axes qui sont au-dessus de cette ligne.

Remarque : Pour les deux premières règles, on examine le tableau de la variance totale expliquée.

Comment interpréter les résultats ?

C'est la phase la plus délicate de l'analyse. On doit donner un sens à un axe grâce à une recherche lexicale (ou recherche de mots) à partir des coordonnées des variables. Pour cela on doit déterminer la combinaison de variables qui est la plus associée à chacun des facteurs significatifs. Nous allons procéder en trois étapes.

1. Nous examinons la matrice des composantes (sans rotation)
   - ✓ Nous observons les variables qui saturent plus fortement sur un facteur et permettent par conséquent de le définir.
   - ✓ Nous observons également que certaines corrélations se ressemblent d'un facteur à l'autre. Il est donc difficile d'établir quelles variables vont réellement avec quel facteur.
2. Nous examinons la matrice des composantes après rotation VARIMAX
   - ✓ Nous observons que l'écart entre les corrélations est plus élevé, ce qui permet de trancher quelles variables vont réellement avec quel facteur.
   - ✓ Nous remarquons aussi que certaines variables saturent de façon importante sur plus d'un facteur. Ceci signifie qu'il faudrait probablement retirer ces variables qui ne se positionnent pas de façon adéquate sur un seul facteur et recommencer l'analyse.
3. Étiqueter les facteurs : nommer les facteurs et tenter d'identifier le contenu caché qu'ils permettent de mesurer
   - ✓ Si nous prenons, par exemple, les variables suivantes qui saturent sur un facteur F:

     Var 1 : J'ai assez de temps pour mes activités de loisir.

     Var 4 : J'ai assez de temps pour faire les choses que j'ai envie de faire.

     Var 5 : J'ai suffisamment de temps pour le travail domestique.

     Var 7 : J'ai suffisamment de temps pour aider les autres.

     Var 9 : J'ai assez d'énergie pour mes loisirs.

   - ❖ Nous voyons que la notion de temps revient dans les quatre premières variables. Nous pourrions nommer le facteur F « temps disponible ».
   - ❖ Probablement, la dernière question va un peu moins avec ce facteur, on vérifier sa corrélation avec les facteurs et on décide si nous pouvons penser à l'éliminer de l'analyse.

Annexe : Travaux pratiques

TP 0 SPSS : Présentation générale de SPSS

Le logiciel SPSS (Statistical Package for Social Sciences) sous Windows est un outil d'analyse statistique particulièrement conçu pour les étudiants en méthodologie d'enquêtes ou ceux des disciplines reliées aux sciences humaines. Il permet principalement de réaliser la codification, la saisie, la manipulation et l'utilisation des données à l'aide des fonctions informatiques et statistiques de base disponibles dans ce programme. Son utilisation exige une certaine familiarité avec l'environnement Windows et la connaissance de logiciels tels que Word, Excel etc.

MENU DE LA FEUILLE DE DONNÉES

1. **Fichier** : permet d'ouvrir des fichiers de données existants en format SPSS; d'ouvrir une feuille de données vierge; d'importer des données en format texte, Excel, Access notamment; permet de sauvegarder les fichiers créés ou modifiés; permet de quitter l'application.
2. **Édition**: copier – coller – couper – effacer les données d'une colonne et d'autres options.
3. **Affichage**: permet de déterminer l'affichage des données.
4. **Données**: édition des données, permet notamment d'insérer des lignes ou des colonnes, de créer des gabarits pour le format des variables, de transposer les lignes en colonnes et inversement, d'agréger des données, de diviser un échantillon en plusieurs groupes, de sélectionner des objets dans l'échantillon.
5. **Transformer** : permet de transformer les scores des variables, ceci par différentes formules, permet de compter des valeurs, de recoder des variables (changer de niveau de mesure), d'attribuer le rang de chaque score, de créer des séries de nombres, de remplacer les valeurs manquantes.
6. **Analyse** : dans ce menu vous trouvez tous les outils d'analyse statistique que ce logiciel propose; description, analyse de dépendance et inférence.
7. **Graphes** : les fonctions réunies sous ce menu permettent de créer les graphiques issus de l'analyse statistique; différents types de graphiques sont proposés : ligne – barres – points – en secteurs – histogrammes – etc.
8. **Outils** : permet d'obtenir des informations sur les données, les fichiers, de créer et d'utiliser des groupes de variables (afin d'éviter de manipuler à chaque fois l'ensemble des données), d'exécuter des scripts, de personnaliser le menu de SPSS.
9. **Fenêtre** : permet de minimiser la taille de la fenêtre et d'activer celles disponibles.
10. **Aide** : procure l'aide en ligne, ainsi que différentes sources d'information relatives à SPSS.

MENU DE LA FEUILLE DES RÉSULTATS (OUTPUTS

1. **Fichier** : mêmes fonctions que celles de la feuille des données, avec quelques spécificités.
2. **Édition**: idem que la feuille des données, avec en plus le choix pour l'éditeur de texte ou de graphique (selon l'élément sélectionné sur la feuille des résultats).
3. **Affichage** : mêmes finalités que celles de la feuille des données.
4. **Insérer** : permet l'insertion de nouveaux graphiques, de titres, d'objets créés à partir d'autres logiciels dans le but de compléter les résultats obtenus.
5. **Format** : permet d'aligner les éléments de la feuille, de modifier la fonte du texte,ainsi que la mise en page des résultats.
6. **Analyse**: idem que pour la feuille des données ; ainsi il n'est pas nécessaire de revenir à la feuille des données pour compléter une analyse ou en commencer une nouvelle.
7. **Graphes** : idem que pour la feuille des données.
8. **Outils** : idem que pour la feuille des données.
9. **Fenêtre**: idem que pour la feuille des données.
10. **Aide**: idem que pour la feuille des données

TP1 SPSS : Création d'un Fichier de données sur SPSS

Dans le but de bien comprendre le processus de création d'une variable, nous allons tenter de créer un fichier de données à partir de l'exemple qui suit.

Le but de l'exercice est de recréer exactement la même matrice. Le questionnaire qui supporte l'exercice se trouve ci-dessous. Il s'agit d'un court questionnaire portant sur la satisfaction des clients d'un CJE (Carrefour Jeunesse Emploi) quelconque au Québec.

Questionnaire d'évaluation des services offerts au Carrefour-Jeunesse Emploi Machin-Chouette

Bonjour, dans le but de vérifier l'adéquacité des services offerts dans notre établissement, nous vous demandons en tant que client(e) recevant de tel service, d'évaluer la qualité des interventions dont vous avez bénéficié.

Numéro de sujet : _____ Date : ____ / ____ / ____

Date de naissance : ____ / ____ / ____ Age : _____

Sexe: _____

Quel est votre statut professionnel ?

1) En formation
2) Au chômage (recherche d'emploi)
3) Assistance-emploi
4) Autre

Question 1) Quel est votre niveau de satisfaction au niveau de la communication avec votre conseiller en emploi ?

1) très insatisf 2) insatisfait 3) ni satisf/ni insatisf 4) satisf 5) très satisf

Question 2) Quel est votre niveau de satisfaction au niveau du plan d'intervention construit par votre conseiller en emploi ?

1) très insatisf 2) insatisfait 3) ni satis/ni insatis 4) satisf 5) très satisf

Question 3) Quel est votre niveau de satisfaction au niveau des progrès réalisés en termes de votre situation professionnelle suite à l'intervention du CJE ?

1) très insatisf 2) insatisfait 3) ni satis/ni insatis 4) satisf 5) très satisf

Voici les étapes à suivre pour créer la première variable IDENTIF « Numéro de sujet » : **NOM**: Changer le nom pour **identif , TYPE** : Mettre **Numérique**, **LARGEUR** : Laisser à 8, **DÉCIMALES** : Descendre à 0, **ÉTIQUETTE** : **Numéro de sujet, VALEURS** : ignorer, seulement pour variables nominales ou ordinales, **MANQUANT** : ignorer, car aucune donnée manquante n'est prévue, **COLONNES** : Laisser à 8, **MESURE** : **Échelle** (plutôt que nominales ou ordinales)

La démarche est la même pour les autres variables.

Attention, quelques variables sont de types différents et possèdent les particularités suivantes : **DATETEST** est la **date** (TYPE = date, format *jj-mm- aaaa*) de l'entrevue. Vous pouvez choisir le type de variable en plaçant votre curseur dans la colonne

Type, puis en appuyant sur le bouton …. Vous n'avez qu'à cocher le type de variable désiré dans la liste de la boite de dialogue qui apparaît. Vous devez cliquez dessus et la boite de dialogues suivante s'ouvrira. Vous indiquez la valeur de la variable dans la première boite, son étiquette dans la deuxième, puis cliquez sur Ajouter. Vous répétez l'opération jusqu'à ce que toutes vos valeurs soient définies. Cliquez ensuite sur OK.

SEXE est une variable numérique, le niveau de mesure est nominal.

STATUT est le « Statut professionnel » (à indiquer dans la colonne Étiquette) (Mesure = nominal) et les étiquettes des valeurs sont les suivants

1 = « En formation » **2** = « Au chômage » **3** = « Assistance-emploi » **4** =«Autre»

Q01, Q02 et Q03 sont trois questions (variables) de satisfaction (MESURE = Ordinales) qui correspondent aux valeurs suivantes :

1 = « Très insatisfait » **2** = « Insatisfait » **3** = « Ni sat, ni insat »**4** = « Satisfait »

5 = « Très satisfait » **8** = « Ne sait pas » (à mettre MANQUANT)

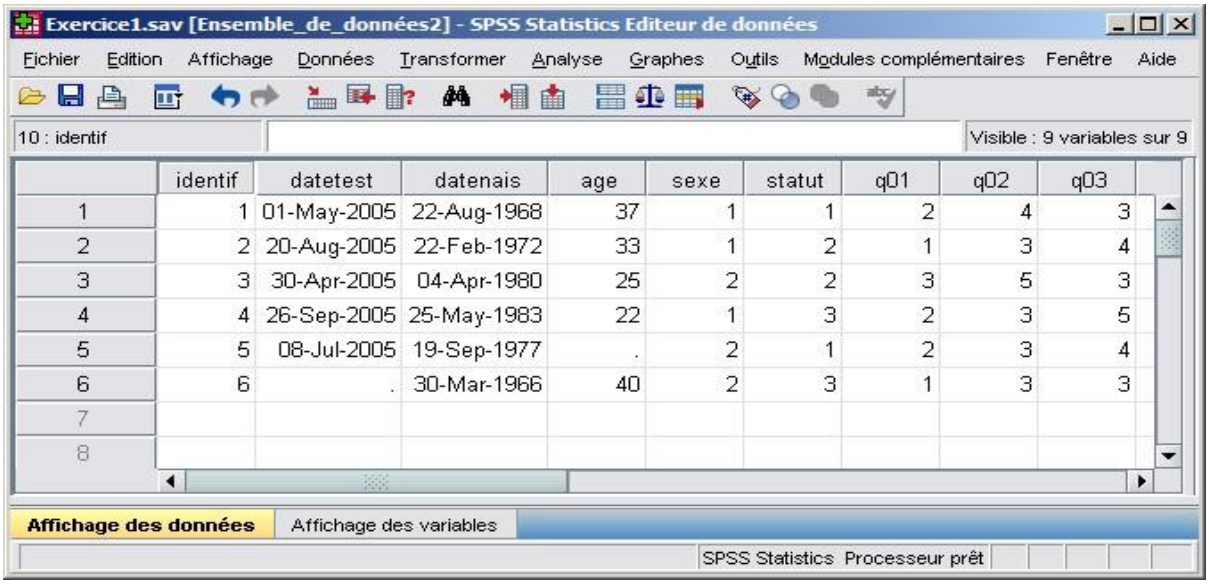
9 = « Ne s'applique pas » (à mettre MANQUANT)

Si la personne n'a pas répondu on laisse la cellule vide (,)

Pour indiquer que les valeurs 8 et 9 sont des données manquantes, placez votre curseur dans la colonne **Manquant**, puis cliquez sur …. Dans la boite de dialogue, cochez **Valeurs manquantes discrètes**. Les trois petites boites deviendront blanches et vous pourrez ajouter les deux valeurs. Cliquez ensuite sur OK.

Après avoir défini les variables, faites la saisie des données du tableau de la page suivante. Il vous faut cliquer sur l'onglet **Affichage des données** pour avoir accès au chiffrier !

Vous entrez les données, **par ligne**, telles qu'elles apparaissent sur l'image ci-dessous :



Lorsque l'entrée de données est complétée, vous pouvez sauvegarder la base de la façon suivante : Dans le menu **FICHIER**, cliquez sur **ENREGISTRER SOUS (lors d'une 1ère sauvegarde), choisissez votre répertoire de travail et** Inscrivez **Exercice1** comme nom de fichier.

TP2 SPSS : Ajout de cas/variables dans un Fichier SPSS existant

Il est possible d'ajouter de **nouvelles observations (cas)** ou **variables** en ajoutant des lignes supplémentaires (observations) ou des colonnes supplémentaires (variables) à un fichier de données existant.

Cependant, il est aussi possible d'ajouter de nouvelles observations ou variables contenues dans d'autres fichiers SPSS à l'aide des commandes **Ajouter des variables** et **Ajouter des observations** à partir de la fonction **Fusionner des fichiers** du menu **Données.** Dans ce cas, les nouvelles observations ou variables sont contenues dans des fichiers SPSS distincts créés avant de procéder à la fusion.

1. Création d'un fichier contenant de nouvelles OBSERVATIONS

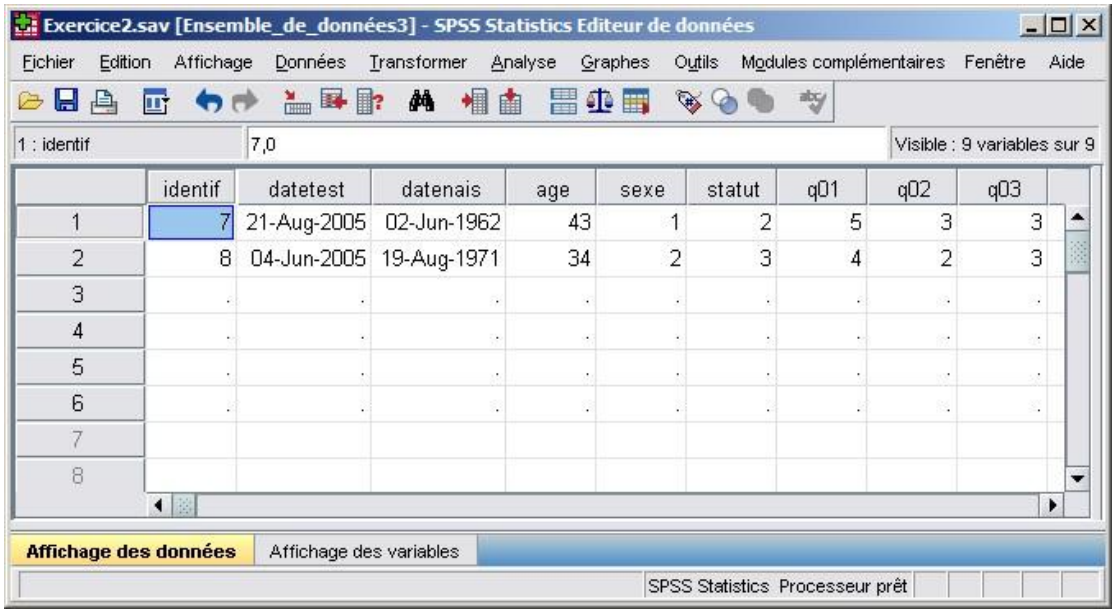
Dans un premier temps, nous allons créer un nouveau fichier contenant les nouvelles observations à ajouter ultérieurement à « Exercice1.sav ».

Pour créer un fichier avec de nouvelles observations, sélectionner les LIGNES de la matrice et tapez sur « EFFACER » (ou cliquez sur **Effacer** dans le menu **Édition**). Cette procédure purge la matrice de toutes ses données.

Vous repartez avec une matrice vide. Remarquez que les variables demeurent définies.

Vous êtes donc prêts à entrer les données se trouvant sur la prochaine image. Lorsque vous aurez terminé, sauvegardez ce fichier sous le nom de **« Exercice2.sav ».**

Attention !! Dans la saisie d'écran ci-dessous, l'utilisateur a créé par inadvertance 4 observations valides sans données. Remarquez que les cas 3 à 6 sont activés : il n'y a pas de données, mais des points apparaissent dans la matrice. Avant de sauvegarder, il faut surligner ces 4 lignes et les supprimer. La présence d'observations fantômes peut faire obstacle à la fusion de fichiers.



2. Procédure « Ajouter des observations »

Pour ajouter de nouveaux cas à **Exercice1.sav**, il faut suivre les étapes suivantes :

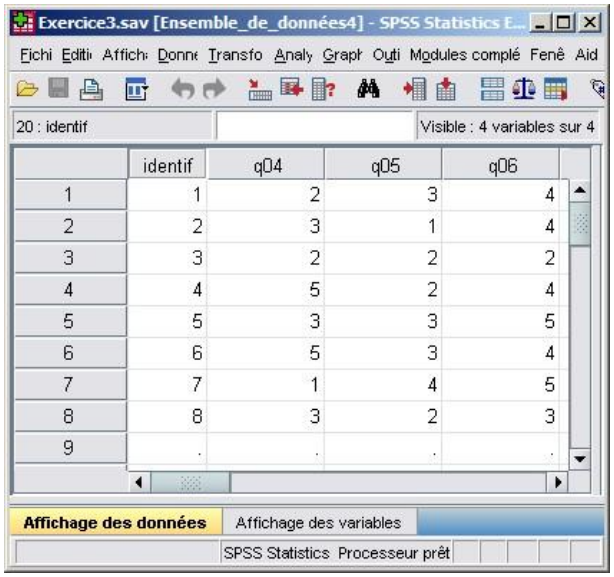
1) Vous devez ouvrir le fichier **Exercice1.sav**.
2) À l'aide du menu **Données,** cliquez sur **Fusionner les fichiers** et choisissez l'option **Ajouter des observations**.
3) Pour fusionner le fichier **Exercice2.sav** (que vous venez à peine de créer) au fichier actif à l'écran **(Exercice1.sav)**, vous devez choisir dans la boite de dialogue le fichier dans lequel les nouveaux cas sont enregistrés.

4) Vous cliquez ensuite sur **Poursuivre**.
5) Dans la deuxième boite de dialogue, vous voyez à droite les variables communes aux deux fichiers (Variables dans un nouvel ensemble de données actif). Si certaines variables n'avaient pas été communes aux deux fichiers, elles seraient apparues dans l'encadré de gauche (Variables non communes). Comme ce n'est pas le cas ici, on peut cliquer sur **OK** pour procéder à la fusion.
6) Une fois l'opération accomplie, vérifiez que les nouvelles OBSERVATIONS (7 et 8) apparaissent dans les deux dernières lignes de la matrice de données.
7) Par conséquent, sauvegardez le fichier manipulé en sélectionnant **Enregistrer sous** dans le menu **Fichier** sous le nom de « **Exercice final** ».

3. Procédure « Ajouter des variables »

Pour cette dernière partie de l'exercice sur la fusion de fichiers de données, vous devez créer un nouveau fichier.

1) Pour ce faire, allez dans le menu **Fichier**, sélectionnez **Nouveau**, puis **Données**.
2) Enregistrez le nouveau fichier sous le nom « **Exercice3** », puis entrez les données ci-dessous.



	identif	q04	q05	q06
1	1	2	3	4
2	2	3	1	4
3	3	2	2	2
4	4	5	2	4
5	5	3	3	5
6	6	5	3	4
7	7	1	4	5
8	8	3	2	3
9	.	.	.	.

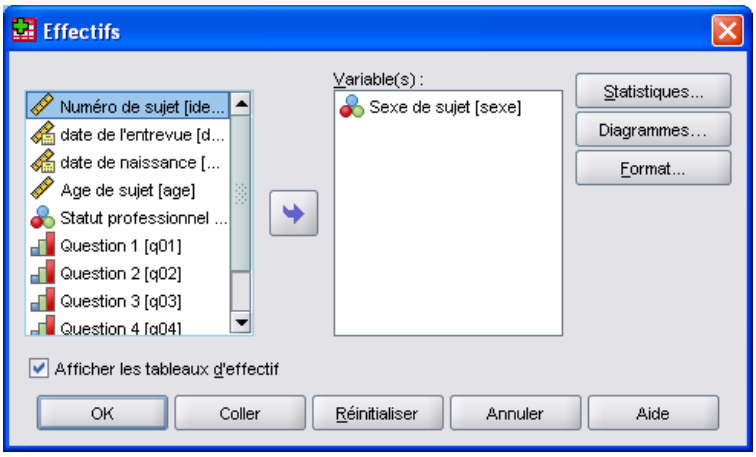
3) Vous devez bien entendu définir ces nouvelles variables sous l'onglet **Affichage des variables** de sorte que :
4) Maintenant, pour effectuer la fusion des nouvelles variables, nous devons (comme dans le cas d'ajouts de nouvelles observations) ouvrir notre fichier référence **Exercice final**.
5) Nous allons procéder comme précédemment à l'aide du menu **Données**, **Fusionner les fichiers**, mais nous choisirons **Ajouter des variables** comme commande de fusion**.**
6) À l'ouverture de la boite, choisissez le fichier qui vous intéresse, soit **Exercice3.** Si ce dernier n'est pas ouvert, comme dans le cas présent, vous n'avez qu'à choisir la deuxième option, à cliquez sur le bouton **Parcourir** et à choisir le fichier en question. Cliquez ensuite sur **Poursuivre**.
7) Dans cet exercice, la seule variable commune « IDENTIF » est la clé de fusion. En effet, il est essentiel dans l'ajout de nouvelles variables à partir d'un autre fichier d'avoir un point de repère nous permettant d'assigner les bonnes valeurs aux bons cas! Assurez-vous de cocher **« Apparier les observations sur les clés des fichiers triés »** après avoir surligné la variable « IDENTIF » dans la partie gauche de la boite.
8) Ensuite, cliquez sur  pour insérer IDENTIF dans la petite boite **variables-clés**. Vous venez de dire au logiciel que vous voulez qu'il se serve de IDENTIF pour pairer les deux fichiers…
9) Cliquez maintenant sur **OK**

TP 3 SPSS : LES ANALYSES UNIVARIÉES

L'analyse univariée permet de décrire et de synthétiser les résultats de l'étude en analysant les variables une par une.

1. On veut savoir combien d'hommes et de femmes se retrouvent dans la base de données du fichier **Exercice final 2.sav**, on exécuterait la procédure de fréquences sur la variable nominale SEXE (ayant comme étiquettes 1=Homme, 2=Femme).

Pour ce faire, il faut sélectionner :**Analyse statistiques descriptives Effectifs**; puis vous sélectionnez la variable Sexe pour laquelle nous souhaitons connaître les fréquences et la fait passer dans le cadre "variable(s)" en cliquant sur la flèche.



Pour avoir la distribution des fréquences, vous cliquez sur OK et vous aurez les tableaux suivants

Statistiques

Sexe de sujet

N	Valide	8
	Manquante	0

8 participants ont répondu au questionnaire et qu'il n'y pas de valeur manquante

Sexe de sujet

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	Homme	4	50,0	50,0	50,0
	Femme	4	50,0	50,0	100,0
	Total	8	100,0	100,0	

- La première colonne décrit les valeurs valides et manquantes qui ont été assignées à la variable sexe
- La deuxième colonne indique la fréquence associée à chaque valeur valide spécifique de la variable Sexe
- Le pourcentage donne la proportion de personnes pour chaque valeur possible
- La colonne « Pourcentage valide » affiche les pourcentages relatifs uniquement pour les sujets ayant donné une réponse valide.
- La dernière colonne de la table de fréquence permet de calculer le cumul des pourcentages des catégories précédentes à partir de la première

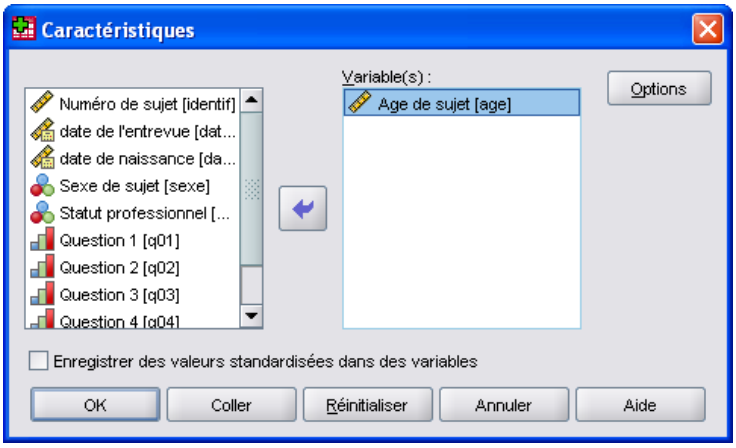
2. Représenter la variable **sexe** sur un digramme en bâtons

Remarque

- ✓ Si vous désirez calculer certaines statistiques, vous cliquez sur statistiques et vous sélectionnez les éléments désirés :
- ✓ Pour une variable nominale : mode, distribution de fréquences, minimum, maximum …

- ✓ Pour une variable ordinale : mode, distribution de fréquences, minimum, maximum, médiane
- ✓ Pour une variable métrique : écart-type, moyenne, minimum, maximum …

3. Pour réaliser une analyse descriptive de la variable âge, cliquez sur **Analyse** Choisissez **Statistiques descriptives** dans le menu déroulant, puis choisissez **Descriptives** dans le second menu déroulant.



Choisissez les options qui vous permettent d'avoir comme résultat le tableau suivant :

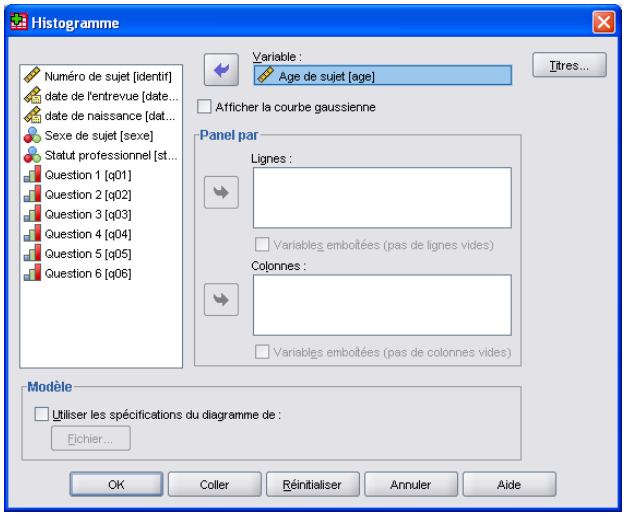
Statistiques descriptives

	N	Intervalle	Minimum	Maximum	Moyenne	Ecart type	Variance
Age de sujet	7	21	22	43	33,43	7,635	58,286
N valide (listwise)	7						

Le tableau montre le nombre d'observations valides pour chaque variable choisie. La ligne « N valide » représente le nombre d'observations pour lesquelles il y a une valeur valide pour toutes les variables étudiées dans la procédure.

4. Interprétez ces résultats

5. Créer un histogramme représentant la variable âge, pour cela cliquez sur **Graphes**, puis sur **Boites de dialogue ancienne version** et sur **Histogramme** dans les menus déroulants.



TP4 SPSS : TABLEAUX CROISÉS, TEST D'INDÉPENDANCE DU KHI-DEUX ET CORRELATION

1. Créer un tableau croisé contient les fréquences correspondant au croisement des caractéristiques des deux variables **STATUT** et **Q03** dans la base de données du fichier **Exercice final 2.sav**, et réaliser le test Khi-deux relatif aux deux variables.

LE RÉSULTAT ET LA LECTURE D'UN TABLEAU CROISÉ

Tableau croisé Question 3 * Statut professionnel

			Statut professionnel			
			En formation	Au chômage	Assistance emploi	Total
Question 3	Ni sat, ni insat	Effectif	1	2	2	5
		% compris dans Question 3	20,0%	40,0%	40,0%	100,0%
		% compris dans Statut professionnel	50,0%	66,7%	66,7%	62,5%
		% du total	12,5%	25,0%	25,0%	62,5%
	Satisfait	Effectif	1	1	0	2
		% compris dans Question 3	50,0%	50,0%	,0%	100,0%
		% compris dans Statut professionnel	50,0%	33,3%	,0%	25,0%
		% du total	12,5%	12,5%	,0%	25,0%
	Trés satisfait	Effectif	0	0	1	1
		% compris dans Question 3	,0%	,0%	100,0%	100,0%
		% compris dans Statut professionnel	,0%	,0%	33,3%	12,5%
		% du total	,0%	,0%	12,5%	12,5%
Total		Effectif	2	3	3	8
		% compris dans Question 3	25,0%	37,5%	37,5%	100,0%
		% compris dans Statut professionnel	100,0%	100,0%	100,0%	100,0%
		% du total	25,0%	37,5%	37,5%	100,0%

L'INTERPRÉTATION DES RÉSULTATS DU TEST DU Khi-deux

Ce test permet de vérifier si une relation entre deux variables (non métriques) existe dans la population.

Le test permet donc d'accepter ou de rejeter l'hypothèse H0 "il n'y pas de relation entre les deux variables ". Dans un test **Khi deux**, il y a trois données importantes :

1. Le résultat du test ou Value (3,067[a] dans le tableau ci-bas).
2. Le dl ou degré de liberté, ici df =4.
3. Le Sig. ou valeur de Signification asymptotique dans ce cas-ci 0,547.

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	3,067[a]	4	,547
Rapport de vraisemblance	3,993	4	,407
Association linéaire par linéaire	,090	1	,765
Nombre d'observations valides	8		

a. 9 cellules (100,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de ,25.

La valeur de Signification asymptotique permet de **confirmer ou d'infirmer** votre hypothèse statistique (H0). Rappelons qu'en sciences humaines, le seuil de signification est de 0,05 (fixer par convention). Ainsi nous avons une signification de 0,547 ce qui nous permet de confirmer qu'il n'y a pas de relation entre les deux variables étudiées. On peut donc conclure que le statut n'influe pas sur les réponses relatives à la question 3

Dans la lecture du tableau Khi-deux, il est préférable de se référer au seuil de signification statistique qui est toujours le même (0,05) plutôt qu'à la valeur du Khi-deux qui varie selon le nombre de degré de liberté.

2. Ouvrir le fichier "cars.sav" disponible dans SPSS. On veut vérifier s'il y a corrélation entre toutes les variables métriques de ce fichier

La procédure SPSS est la suivante :Analyse corrélation bivariée. Nous aurons les résultats suivants qui peuvent être interprétés comme suit ;

Corrélations

		miles par gallon	cylindrée (cu. inches)	puissance	poids du véhicule (lbs.)	temps d'accélération de 0 à 60 mph (sec)
miles par gallon	Corrélation de Pears	1,000	-,789**	-,771**	-,807**	,434**
	Sig. (bilatérale)	,	,000	,000	,000	,000
	N	398	398	392	398	398
cylindrée (cu. inches	Corrélation de Pears	-,789**	1,000	,897**	,933**	-,545**
	Sig. (bilatérale)	,000	,	,000	,000	,000
	N	398	406	400	406	406
puissance	Corrélation de Pears	-,771**	,897**	1,000	,859**	-,701**
	Sig. (bilatérale)	,000	,000	,	,000	,000
	N	392	400	400	400	400
poids du véhicule (lb	Corrélation de Pears	-,807**	,933**	,859**	1,000	-,415**
	Sig. (bilatérale)	,000	,000	,000	,	,000
	N	398	406	400	406	406
temps d'accélération	Corrélation de Pears	,434**	-,545**	-,701**	-,415**	1,000
de 0 à 60 mph (sec)	Sig. (bilatérale)	,000	,000	,000	,000	,
	N	398	406	400	406	406

**.La corrélation est significative au niveau 0.01 (bilatéral).

Pour chaque couple de variable (Xi, Xj), les résultats indiquent le coefficient (de Pearson) estimé, et le risque d'erreur de première espèce ou signification (sig.) -soit le risque de se tromper sur le sens de la corrélation-. Si sig. < 0,05, on peut conclure à l'existence d'une corrélation, au seuil 0,05 entre les deux variables (au seuil de signification indiqué par la statistique sig.). Le symbole ** indique tous les sig. inférieurs à 0.01. Ceci permet une lecture rapide du tableau.

TP 5 SPSS : ANALYSE EN COMPOSANTES PRINCIPALES (ACP)

PROCÉDURE SPSS Pour réaliser l'ACP, il faut sélectionner : **Analyse factorisation analyse factorielle** (nous employons le terme de factorise car il s'agit bien de réduire en une combinaison linéaire plusieurs variables ensemble). Ses étapes consistent à :

1. Sélectionner les variables à factoriser (nous pouvons sélectionner toutes les variables et cliquer sur la flèche vers la droite) ;
2. Sélectionner le type de méthode de facturation ; par défaut, conserver : analyse en composantes principales ;
3. Sélectionner des options de présentation pour classer les variables les plus importantes et cacher celles qui n'expliquent pas les dimensions ;
4. Sélectionner dans facteur l'affichage des scores factoriels.

La procédure SPSS propose en sortie des tableaux et un graphique à analyser :

1-La variance expliquée : Ce tableau présente les dimensions qui permettent de résumer l'information. Par exemple, si la première dimension extraite permet d'expliquer 45% de la variance du phénomène, cela veut dire que les variables qui composent la première dimension synthétise 45% du phénomène. La variance cumulée permet d'évaluer si la réduction des différentes variables à quelques composantes permet de conserver l'essentiel du phénomène mesuré par les variables de départ.

2-La qualité de représentation : Elle permet de répondre à la question suivante : dans quelle mesure mes variables de départ sont elles prises en compte par les composantes extraites ? Ainsi, si la qualité de représentation d'une variable X est de 0.930 cela veut dire que 93% de la variance de la variable est prise en compte par l'une des dimensions extraites. Par contre, si la qualité de représentation d'une variable Y est de 0.510 cela veut dire que 51% seulement de la variance de la variable est prise en compte par l'une des dimensions extraites ce qui signifie que cette variable sera mal représentée par les composantes retenues.

3-La matrice des composantes: nous pouvons trouver plusieurs manières d'étudier les coefficients qui sont présentés dans la matrice des composantes. D'une part, les colonnes correspondent à chacune des dimensions extraites, elles contiennent des coefficients de saturation qui s'interprètent comme des coefficients de corrélation. D'autre part, tous les coefficients forment des coefficients a, b, c… d'une droite de régression qui est la composition linéaire de la composante. Par exemple, si nous prenons une composante 1, elle peut être définie par l'équation suivante : a* X + b*Y + c*Z…Enfin, la matrice des composantes nous permet de nommer les dimensions extraites en étudiant les coefficients de saturation de chacune des variables par rapport aux dimensions ; par exemple nous pouvons nommer une composante « potentiel de développement » car les variables fortement corrélées avec elle sont des indicateurs du niveau de développement.

4-La représentation graphique : elle peut être réalisée pour la matrice des composante set surtout des scores factoriels en créant un graphique de dispersion, et ce suivant les étapes suivantes :

Sélectionner dans le menu graphes interactif diagramme de dispersion, sur la boite de dialogue **« créer un diagramme de dispersion »** Faire glisser la variable REGR factor score 1 vers l'axe X et REGR factor 2 vers l'axe Y; afin d'afficher sur le graphique les observations, faire glisser observations vers "étiqueter les observations par " ; ensuite cliquer sur options; Cliquer sur "afficher le diagramme avec les étiquettes " pour les faire apparaître ;

Le graphique ainsi obtenu nécessite quelques modifications pour en améliorer la lisibilité. Il faut double cliquer sur le graphique pour ouvrir la fenêtre d'édition des graphiques (sélectionner Options lignes de références pour l'axe X, cliquer ensuite sur Appliquer; répéter la même opération pour échelle Y.

Exercice ACP

Le fichier « **cherche_partenaire.sav** » contient les données d'une étude fictive en psychologie du couple qui porte sur les motivations des gens à la recherche d'un partenaire. Le chercheur a mis au point un questionnaire mesurant différents aspects pouvant être privilégiés dans la recherche d'un partenaire. Pour chacune des motivations suivantes les participant(e)s devaient donc indiquer si elles avaient été plus ou moins importantes dans leurs critères de recherche du partenaire idéal: a) l'honnêteté du partenaire, b) son salaire, c) son sens de l'humour, d) sa simplicité, e) son empathie, f) son apparence physique et g) sa sociabilité. En dehors du questionnaire de motivation une variable de groupes a également été mesurée, soit le contexte dans lequel la rencontre des partenaires s'est produite (1 = dans un cruising bar, 0 = autres endroits).

1. Charger le fichier de données « **cherche_partenaire.sav** »
2. Les données sont-elles factorisables ?
3. Combien de facteurs à retenir dans cette analyse?
4. Comment interpréter les résultats ?

Correction

2. Le tableau ci-bas présente la matrice d'intercorrélation entre les sept variables. Comme toutes les matrices de corrélation, il s'agit d'une matrice de forme carrée, c'est-à-dire comportant un nombre égal de rangées et de colonnes correspondant au nombre de variables. On observe évidemment la présence de la diagonale principale où s'alignent les valeurs 1.00 correspondant à la corrélation parfaite de chaque variable avec elle-même. La diagonale principale divise la matrice en deux portions triangulaires symétriques où l'on retrouve les mêmes coefficients de corrélation, puisque, par exemple, la corrélation entre l'importance du salaire et la sociabilité du partenaire (.825) est égale à la corrélation entre la sociabilité et le salaire (.825). Il est donc suffisant d'avoir accès à la portion triangulaire inférieure de la matrice de corrélation pour pouvoir procéder à une ACP.

Matrice de corrélation

		Honnêteté	Salaire	Humour	Simplicité	Empathie	Apparence_physique	Sociabilité
Corrélation	Honnêteté	1,000	,131	-,245	,780	,766	-,049	,059
	Salaire	,131	1,000	-,353	,047	-,050	,714	,825
	Humour	-,245	-,353	1,000	-,308	-,327	-,347	-,407
	Simplicité	,780	,047	-,308	1,000	,754	-,114	,011
	Empathie	,766	-,050	-,327	,754	1,000	-,149	-,026
	Apparence_physique	-,049	,714	-,347	-,114	-,149	1,000	,710
	Sociabilité	,059	,825	-,407	,011	-,026	,710	1,000

Que pouvons-nous dire de la taille des coefficients de corrélation apparaissant dans cette matrice? Certains de ces coefficients sont particulièrement petits, par exemple entre la sociabilité et la simplicité (.011) ou entre l'empathie et la sociabilité (-.026). Vous comprendrez que si tous les coefficients de corrélation étaient aussi faibles que ceux-là, il n'y aurait absolument aucun intérêt à procéder à une analyse en composantes principales de ces données. En effet, pour pouvoir extraire une composante correspondant à une fonction linéaire des variables initiales, il faut nécessairement que ces variables soient intercorrélées. Heureusement la matrice de corrélation comporte aussi un certain nombre de coefficients de tailles intéressantes (.766, .780, .754, etc.).

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,777
Test de sphéricité de Bartlett	Khi-deux approximé	1552,286
	ddl	28
	Signification de Bartlett	,000

L'indice **KMO** de 0,777 peut être qualifié d'excellent. Il nous indique que les corrélations entre les variables sont de bonne qualité. Ensuite, le résultat du test de sphéricité de Bartlett est significatif (p < 0,0005). Nous pouvons donc conclure que nos données sont Factorisables et nous pouvons donc poursuivre l'analyse.

3.Combien de facteurs à retenir dans cette analyse?

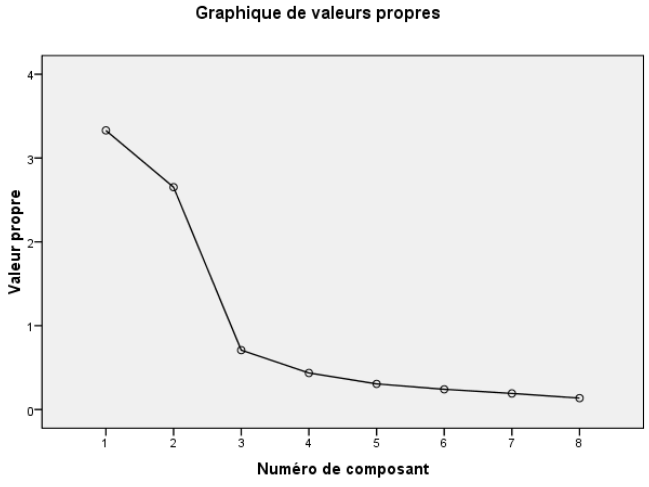
Variance totale expliquée

Composante	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus			Somme des carrés des facteurs retenus pour la rotation		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	3,330	41,626	41,626	3,330	41,626	41,626	3,110	38,875	38,875
2	2,652	33,151	74,777	2,652	33,151	74,777	2,579	32,243	71,118
3	,708	8,848	83,625	,708	8,848	83,625	1,001	12,507	83,625
4	,436	5,445							
5	,306	3,821							
6	,241	3,012							
7	,192	2,395							
8	,136	1,701							

Méthode d'extraction : Analyse en composantes principales.

Pour choisir le nombre de facteurs à extraire, nous analysons le tableau de la variance totale expliquée ci dessus. En regardant la deuxième colonne, nous constatons que deux facteurs (ou composantes) ont une valeur propre plus élevée que 1. Nous les conservons donc pour l'analyse. Le premier facteur explique à lui seul 41,626% de la variance totale des 7 variables de l'analyse. Mis en communs, les deux facteurs permettent d'expliquer 71,118% de la variance. Comme les facteurs 3 à 8 n'expliquent pas suffisamment de variance, ils ne sont pas retenus.

Nous désirons toutefois être certains de bien choisir le bon nombre de facteurs à extraire. Nous regardons donc le graphique des valeurs propres ci-dessous et examinons où se situe la rupture du coude de Cattell. Nous voyons un changement après le deuxième facteur. Nous ne retenons donc que deux facteurs pour l'analyse, puisque ce critère est plus rigoureux que celui des valeurs propres.



4.Interprétation des résultats

L'examen de la matrice des **composantes** après rotation permet de constater facilement que la première composante est définie par les motivations reliées à la sociabilité du partenaire, à son salaire et à son apparence physique. La deuxième composante, quant à elle, se définit en termes de d'empathie, de simplicité et d'honnêteté du partenaire idéal. Ces conclusions sont bien illustrés sur la carte factorielle ci après qui correspond au diagramme de composantes dans l'espace après rotation Je vous laisse le soin de déterminer si ces deux composantes correspondent à votre perception des motivations reliées à la recherche d'un partenaire.

Matrice des composantes après rotation

	Composante	
	1	2
Sociabilité	**,919**	,006
Salaire	**,910**	,027
Apparence_physique	**,873**	-,140
Humour	-,541	-,412
Empathie	-,057	**,917**
Simplicité	,001	**,915**
Honnêteté	,058	**,903**

Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.
a. La rotation a convergé en 3 itérations.

Diagramme de composantes dans l'espace après rotation