RÉPUBLIQUE DU NIGER
Fraternité - Travail - Progrès
MINISTÈRE DU PLAN
INSTITUT NATIONAL DE LA STATISTIQUE
PLATEFORME NATIONALE D'INFORMATION POUR LA NUTRITION

NIGER
MANUEL
JUILLET 2020

MANUEL SUR LES MÉTHODES D'ANALYSES

OUTIL DE RENFORCEMENT DES CAPACITÉS EN ANALYSE DES DONNÉES

SOFRECO

PNiN
Plateformes nationales d'information pour la nutrition

HC3N
LES NIGERIENS NOURRISSENT LES NIGERIENS

Institut National de la Statistique
NIGER

PRÉFACE

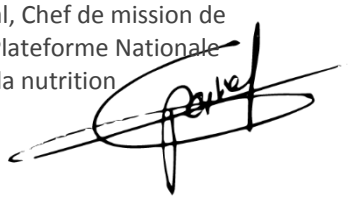
L'initiative « Plateformes Nationales d'Information pour la Nutrition (PNIN) », portée par la Commission Européenne, vise à aider les pays à renforcer leurs systèmes d'information et leurs capacités d'analyse de données pour la nutrition, de manière à mieux étayer les décisions stratégiques auxquelles ils sont confrontés pour prévenir la malnutrition et ses conséquences. L'approche développée par l'initiative PNIN consiste à renforcer les capacités des pays bénéficiaires du programme en matière d'exploitation optimale des données et informations existantes en lien avec la nutrition, de manière à ce qu'ils puissent mettre en œuvre des politiques et programmes efficaces et définir des priorités dans l'allocation des ressources avec l'appui des délégations locales de la Commission Européenne.

Au Niger, la plateforme PNIN est mise en œuvre par l'Institut National Statistique (INS) qui coordonne l'ensemble des activités avec les différentes parties prenantes et le Haut-Commissariat à l'Initiative 3N (HC3N) qui en est le leader stratégique. Le programme PNIN regroupe d'autres bénéficiaires en particulier les Directions des Statistiques et les Directions des Études et de la Programmation des ministères sectoriels : Santé, Éducation, Agriculture et Élevage, Hydraulique et Assainissement, Environnement.

Dans ce contexte, une Assistance Technique (AT) internationale (SOFRECO) intervient en appui à l'INS. Le présent toolkit de formation entre dans le cadre de l'objectif spécifique 2 de l'Assistance Technique, à savoir « créer les capacités, au sein des parties prenantes au Niger, de formuler des questions/demandes en termes d'analyse, d'analyser les données afin de répondre à celles-ci et de mesurer les progrès effectués vers l'atteinte des objectifs nationaux de réduction de la prévalence de sous-nutrition ».

Afin d'alimentation le débat public autour des questions sociodémographiques, économiques ou de compréhension de la stabilité de la malnutrition au Niger, plusieurs formations ont été mises en œuvre dans le cadre du programme PNIN qui apparaissent comme des conditionnalités à la valorisation des données. Suite à la formation effectuée par l'Assistance Technique, le présent toolkit permet de rappeler les principales méthodes d'analyses approfondies les plus utilisées. Ces méthodes d'analyses permettent l'identification et la compréhension des tendances, le croisement des informations et l'identification des liens de causalité entre variables, la mise en relief d'explications et la concrétisation d'hypothèses. Ces méthodes doivent permettre à l'Équipe PNIN Niger (INS, HC3N) d'assurer la mise en œuvre du Plan Cadre d'Analyses 2019-2020 et la production d'analyses approfondies régulières et ainsi promouvoir la valorisation des bases de données disponibles et par conséquent l'alimentation du débat public autour des questions dans le domaine de la nutrition au Niger.

Guillaume **POIREL**
Expert Technique International, Chef de mission de l'Assistance Technique de la Plateforme Nationale d'Information pour la nutrition



SIGNALÉTIQUE

OURS

Unité responsable : Plateforme Nationale d'Information pour la Nutrition

Firmin SEKA, Expert Court-Terme, consultant en méthodes d'analyses des données, Assistant Technique PNIN (AT/PNIN)

Guillaume POIREL, Chef d'Equipe, Statisticien-Analyste, Assistant Technique PNIN (AT/PNIN)

Editeur de la publication : Assistance Technique de la Plateforme Nationale d'Information pour la Nutrition

SOMMAIRE

LISTE DES TABLEAUX

LISTE DES FIGURES

1 INTRODUCTION À L'ANALYSE DES DONNÉES

1.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin de ce module de formation, les participants seront capables de distinguer les différents types d'analyse de données et les types de données. Les participants seront à même de définir une variable et catégoriser les types de variable. En combinant ces différents points, les participants seront à même de choisir le type de test statistique selon la/les question(s) de recherche de l'étude.

1.2 ASPECT THÉORIQUE

1.2.1 QU'EST-CE QUE L'ANALYSE DES DONNÉES

L'analyse des données peut se définir comme l'ensemble des méthodes permettant une étude approfondie d'informations quantitatives. Selon Jean de Lagarde : « Le propre de l'analyse des données, dans son sens moderne, est justement de raisonner sur un nombre quelconque de variables, d'où le nom d'analyse multivariée qu'on lui donne souvent[1]. » Pour certains, le rôle principal de l'analyse des données est « de mettre en relief les structures pertinentes de grands ensembles de données[2]».

L'analyse des données, telle qu'on la connaît aujourd'hui, s'inscrit dans la convergence :

- De disciplines particulières des sciences de la gestion ou des sciences sociales ;
- Des méthodes de la statistique appliquée ;
- De l'existence de logiciels très performants de traitement des données.

1.2.2 QUELLES SONT LES DIFFÉRENTS TYPES D'ANALYSE DE DONNÉES

Dans l'analyse des données, on distingue habituellement :

- L'analyse univariée, qui porte sur l'étude des variables prises une à une dans la présentation et l'interprétation ;
- L'analyse bivariée, qui a pour objectif d'examiner les relations de deux (2) variables en même temps ;
- L'analyse multivariée qui vise l'étude de plusieurs variables en même temps.

1.2.3 QUEL EST LA PLACE DE L'ANALYSE DES DONNÉES DANS LE PROCESSUS DE RECHERCHE

L'analyse des données correspond à certaines étapes bien spéciales du processus de recherche. Voici par exemple les principales étapes d'une recherche faite à l'aide d'un questionnaire fermé :

1. La définition du sujet de recherche ;
2. La délimitation du terrain d'étude ;
3. La définition de l'objet de recherche ;

1 J. de Lagarde (1995), *Initiation à l'analyse des données*, Paris, Dunod, p. 2.
2 J.-P. Crauser, Y. Harvatopoulos et P. Sarnin (1989), *Guide pratique de l'analyse des données*, Paris, Éditions d'Organisation, p. 9.

4. La revue de la littérature ;
5. La question de recherche ;
6. La problématique ;
7. L'échantillonnage ;
8. La construction du dispositif du recueil des données ;
9. Les investigations sur le terrain (observation, questionnaire, entrevue) ;
10. L'analyse des données recueillies et leur interprétation en vérifiant s'il y a une possible relation entre les variables de l'hypothèse de départ au moyen de tests statistique ;
11. Les commentaires, les enseignements/leçons apprises ;
12. Conclusion ;
13. Recommandations.

Le premier intérêt ici dans toutes ces étapes est l'analyse des données.

1.2.4 QUELS SONT LES TYPES DE DONNÉES

Tableau 1 : Types de données

Données primaires	les données primaires sont construites par le chercheur dans un but bien précis. Les enquêtes qualitatives ou quantitatives réalisées à l'aide de sondages aléatoires ou non produisent des données primaires	le chercheur ou l'équipe de recherche qui décide de la forme que prendront les variables
Données secondaires	Les données secondaires sont des données recueillies par des gouvernements ou des organismes officiels internationaux ou nationaux	Elles découlent de décisions politiques et administratives prises à un haut niveau.

Source : Auteurs, PNIN

On distingue deux types de données : (1) les données primaires ; (2) des données secondaires. Très souvent, les données primaires et secondaires sont complémentaires : la sécurité alimentaire peut s'expliquer, en partie par la situation économique, mais aussi par des dimensions socio-culturelle, sociales, psychologiques et politiques. Une véritable étude de marché doit tenir compte non seulement de la demande d'un bien révélée par un sondage, mais aussi de la situation économique générale, du revenu disponible et aussi des habitudes de consommation etc.

1.2.5 QU'EST-CE QU'UNE VARIABLE

Au plan strictement sémantique le terme « variable » suppose qu'une réponse à une question donnée peut varier (dans un certain écart) d'un individu à un autre. Si la caractéristique mesurée peut prendre différentes valeurs, on dit alors que cette caractéristique est une variable.

Au plan mathématique, une variable est perçue comme un ensemble de règles qui permettent de ranger les éléments d'un ensemble donné dans des catégories définies au départ. À partir de cette définition, toute variable est comprise dans le sens d'une norme de classification. Une variable est donc un critère par lequel on classe des individus dans des catégories.

1.2.6 QUELS SONT LES TYPES DE VARIABLES

On classifie les variables selon leur degré d'abstraction et leur pouvoir explicatif.

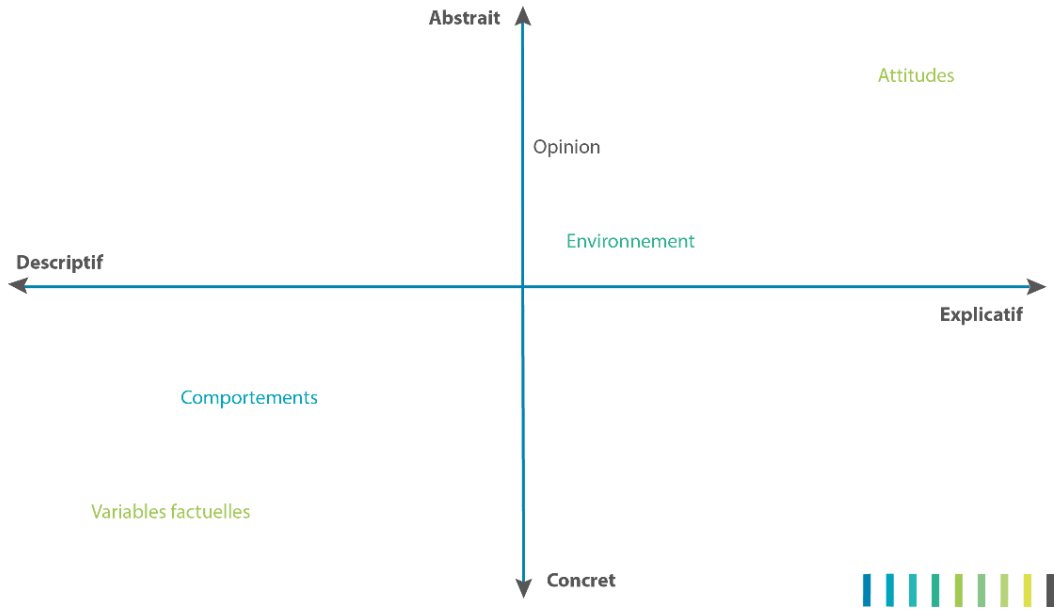
Tableau 2 Types de variables

Types de variables	Degré d'abstraction	Pouvoir explicatif	Exemple
Variables factuelles concernant • **la personne**	Faible	Faible	Variable de personne (qui ?) Age Sexe Niveau d'étude Revenu
• **son environnement**			Variable environnementale Situation de famille Emploi Lieu de résidence Type d'habitat
• **Ses comportements**			Variable de comportement Comportement économique Consommation (Alimentation, vêtement, transport...) Comportement politique Comportement sociologique et psychologique
Les variables reliées aux opinions de la personne	Moyen	Moyen	
Les variables reliées aux attitudes de la personne	Elevé	Elevé	Les attitudes sont plus abstraites et complexes que les opinions

Une attitude… : la personne aime les aliments sucrés - **qui s'exprime dans une opinion :** la personne aime le chocolat - **qui s'actualise dans un comportement** : la personne achète/ consomme le chocolat au lait

Source : Auteurs, PNIN

Figure 1 : Dimensions des types de variables



Source : Auteurs, PNIN

Si les attitudes sont relativement stables, les liens avec les opinions exprimées et les comportements effectifs peuvent bien ne pas l'être. La personne peut refuser d'exprimer son opinion ou différer ses comportements pour des raisons économiques, sociologiques, psychologiques ou même politiques.

1.2.7 ECHELLE DE MESURE DES VARIABLES

Les niveaux de mesure sont le parachèvement de la problématique et de la conceptualisation du problème de la recherche. Il s'agit de faire correspondre un concept à une mesure ; c'est dans cette opération que la démarche de recherche devient empirique. Mesurer, c'est relier des nombres à des entités plus ou moins abstraites : l'âge, le sexe, la satisfaction, l'intérêt…

L'analyse des données est basée en grande partie sur les principes des mathématiques et plus particulièrement de la statistique appliquée. Les nombres possèdent certaines propriétés mathématiques dont il faut tenir compte. Ces propriétés sont les suivantes :

1. La propriété de classer des individus dans des catégories ;
2. La propriété d'établir un ordre de préséance, un ordre hiérarchique entre ces catégories ;
3. la propriété de fixer des intervalles égaux dans cet ordre hiérarchique construit en fonction de la deuxième propriété ;
4. la propriété de fixer une origine 0 à cet ordre hiérarchique (en plus d'avoir des intervalles égaux).

Tableau 3 : Echelle des variables

Echelle	Classement	Ordre	Distance	Zéro absolu	Exemple
Nominale	Oui	Non	Non	Non	Sexe Les aliments
Ordinale	Oui	Oui	Non	Non	Statut nutritionnel Consommation des protéines Niveau d'étude
Intervalles	Oui	Oui	Oui	Non	Revenu du ménage Indice BMI
Rapport	Oui	Oui	Oui	Oui	Dépenses alimentaires du ménage au cours de la dernière semaine

Source : Auteurs, PNIN

1.3 SYNTAXE ET TESTS STATISTIQUES

Le choix d'un test statistique le plus approprié pour votre recherche dépend de la question de recherche, de l'hypothèse, du type de mesure des données et de l'échantillonnage.

Le choix du test approprié n'est pas une chose évidente étant donné l'existence d'un nombre important de tests. Afin de faciliter le choix adéquat du test, il importe de se concentrer uniquement sur certains les tests paramétriques (les données analysées ont une distribution normale) et non-paramétriques.

Les tests statistiques paramétriques les plus couramment utilisés sont :

- Test de Khi2 ;
- T-test apparié ;
- T-test à échantillon indépendant ;
- ANOVA à un facteur ;

- Corrélation / régression simple ;
- Régression multiple / régression logistique

Les tests statistiques non-paramétriques les plus couramment utilisés sont :

- Test de KRUSKAL-WALLIS ;
- Test WILCOXON ;
- Test MANN-WHITNEY ;
- Test de MCNEMAR.

1.3.1 QUESTION DE RECHERCHE ET CHOIX DE MÉTHODE D'ANALYSE DE DONNÉES

Toute recherche doit commencer par une question. Cette question de recherche fait partie du processus de conceptualisation. En d'autres termes, qu'est- ce qu'on veut chercher au juste ? Pour chaque type de question correspond des tests.

- Est-ce qu'on veut faire ***une description ?***
- Est-ce qu'on cherche à ***comparer*** les échantillons ou les variables ?
- Est-ce qu'on cherche ***une relation ou association*** entre variables ?
- Est-ce qu'on veut faire ***une prédiction ?***

Tableau 4 : Types d'hypothèses et méthodes d'analyse

Type d'hypothèse	Ce qu'on cherche à faire	Méthodes d'analyse et Statistiques
Hypothèse monovariée (à une seule variable	**Description**	Effectifs Moyen Ecart-type Pourcentage Médiane Pourcentage
Hypothèse bivariée (à 2 variables) ou trivariée (3 variables)	**Comparaison**	Test de Khi2 T-test apparié T-test à échantillon indépendant ANOVA à un facteur
	Association	Corrélation de Person
	Prédiction	Régression simple
Hypothèse multivariée	**Association**	ACP (corrélation multiple)
	Prédiction	Régression multiple Régression logistique

Source : Auteurs, PNIN

➥ Test à utiliser si l'hypothèse est monovariée

Si l'hypothèse est monovariée, c'est-à-dire avec une seule variable, on utilise l'analyse descriptive (effectifs ou descriptives). En effet, il s'agit de faire uniquement une description avec la descritpion de certaines caractéristqiues : moyenne, écart-type, mode, médiane, pourcentage etc. C'est ce qui est appelé **le tri à plat**. On peut aussi faire appel au ***T test à échantillon apparié*** : 2 tests, un avant et un autre après, tous les deux appliqués au même échantillon ou (pour la même variable).

✓ **Exemple**

Soit la 1ère l'hypothèse de recherche suivante : « Les aliments contenant les protéines animales sont disponibles au Niger ».

Ici, il y a une seule variable, quelle est cette variable ?

On peut faire appel à la moyenne, l'écart-type, la médiane, le graphique, etc.

➥ Test à utiliser si l'hypothèse est bivariée ou trivariée

Si l'hypothèse est bivariée ou trivariée, deux (2) types d'analyses se présentent à savoir : la *comparaison* des données ou la *recherche de relation ou association entres variables* (ou échantillons).

Pour la comparaison

Il s'agit de comparer les moyennes ou les écart-types des variables (ou de 2 échantillons) et de voir s'ils sont les mêmes ou différents. Ici, on ne soucie pas de la relation qui pourrait exister entre elles.

✓ **Exemple**

1. Comparaison entre le taux de malnutrition des enfants de Niamey et ceux de l'intérieur du pays.

2. Comparaison entre le taux de malnutrition des enfants du Niger avant et après la fortification des aliments par de base en vitamine A.

On peut utiliser le T test à échantillons indépendants (différence entre les 2 échantillons) ou le test ANOVA (différence entre plus de 2 échantillons).

Pour la recherche d'une relation ou d'association entre variables (ou échantillon)

La question à laquelle on veut répondre est : « est-ce que l'une dépend de l'autre » ou « est-ce-que l'une prédit l'autre ». Il s'agit de déterminer la variable indépendante (appelée aussi variable prédicteur, variable des données) et la variable dépendante (appelée aussi variable à prédire, variable test, variable à observer). En d'autres termes, la variable indépendante agit sur la variable dépendante.

Ici, nous pouvons faire appel à:

- **L'analyse de corrélation (ou de Pearson)** (voir s'il y a association entre variables) ;
- L'analyse de régression linéaire et analyse de régression logistique (association et prédiction) ;
- **Le test de Khi2** qui permet de se prononcer sur une relation éventuelle entre variables qualitatives uniquement ;
- **L'ACP (Analyse en Composantes Principales)** consiste à analyser les corrélations entre plusieurs variables (quantitatives).
- **L'ACM (Analyse de Correspondance Multiples)** consiste à analyse les liens entre plusieurs variables qualitatives

✓ **Exemple**

1. Il y a une association entre malnutrition chronique des enfants de 24 mois à 59 mois et l'ethnie au Niger.

2. Le taux de malnutrition des enfants de 6 à 59 mois augmente avec l'âge des enfants.

3. La pratique de l'allaitement maternel exclusif permet de réduire les malnutritions des enfants

de moins de 6 mois.

4. Comment évoluent le niveau de pauvreté et le pouvoir d'achat des ménages en comparaison de l'évolution des indicateurs de malnutrition chronique chez les enfants de 6 mois à 59 mois au Niger ?

1.3.2 TESTS EN FONCTION DU TYPE DE DONNÉES

Le choix des tests dépend du type de données utilisé. Ainsi pour, les données qualitatives, le test de Khi2 qui permet de rechercher la relation entre variables qualitatives uniquement. Enfin pour des données quantitatives, le test à échantillons appariés, le test à échantillons indépendants, la corrélation linéaire (corrélation de Pearson) ou de Spearman, l'analyse de régression, l'ANOVA à un facteur, l'ACP (Analyse en Composantes Principales) peuvent être utilisés.

Tableau 5 : Tests utilisables selon les différents types de données

	Les calculs statistiques utilisables	*Les tests des relations entre les variables*
Nominale	Fréquence absolue et relative Mode	Khi carré – Coefficient de contingence – Coefficient phi – Lambda – Régression logistique
Ordinale	Ceux de l'échelle nominale plus : – Médiane – Mesures de positions	Ceux de l'échelle nominale plus : – Corrélation de rang – Autres tests non paramétriques – Régression logistique ordinale
Intervalles	Ceux des deux premières échelles plus : Mesures de tendance centrale et de dispersion (moyenne, écart-type…)	Ceux des deux premières échelles plus : – Analyse de variance – Corrélation de Pearson Régression simple et multiple
Rapport	Tous	Tous

Source : Auteurs, PNIN

Tableau 6 : Types d'études et d'analyses selon l'approche et les différents types de données

	Approche « descriptive »	*Approche « explicative »*
Nominale	Étude des fréquences Analyse factorielle	Tableaux croisés Analyse discriminante Régression logistique
Ordinale	Étude des fréquences Mesures de position Analyse factorielle	Tableaux croisés Analyse discriminante Régression logistique
Intervalles	Étude des fréquences Mesures de position Mesures de tendance centrale et de dispersion Analyse factorielle	Tableaux croisés Analyse discriminante Analyse de variance Régression logistique Régression simple et multiple
Rapport	Étude des fréquences Mesures de position Mesures de tendance centrale et de dispersion Analyse factorielle	(Tableaux croisés) Analyse discriminante Analyse de variance (Régression logistique) Régression simple et multiple

Source : Auteurs, PNIN

1.3.3 MÉTHODE D'ANALYSE ET NOMBRE D'ÉCHANTILLON

Certaines recherches reposent soit sur un, deux ou un jeu d'échantillons de la même population d'étude.

Un seul échantillon

Deux cas peuvent se présenter pour les tests paramétriques à savoir une ***simple mesure*** et une ***mesure avant et après***.

- Une simple mesure (moyenne, écart-type, médiane, etc.) ; pas de comparaison, il s'agit d'une **analyse descriptive**.
- Une mesure avant et après : c'est-à-dire avec le même échantillon réaliser 2 mesures différentes (**comparaison** de la différence des moyennes des 2 échantillons).

✓ **Exemple**

Le poids de quelqu'un avant et après un régime alimentaire sévère (à comparer la différence des résultats) (T test apparié).

Deux cas peuvent se présenter pour les tests non-paramétriques

- Si les données ne suivent pas la loi normal ou ordinale, on fait appel au **Test Wilcoxon**. Ainsi, l'hypothèse alternative est vérifiée si la **différence des médianes** de 2 échantillons est différente de zéro ;
- Le Test de McNemar est utilisé pourcomparer deux échantillons appariés lorsque les données sont nominales et dichotomiques (Oui/non, homme/femme, Noir/Blanc).

Deux échantillons

Les mesures des 2 échantillons sont à comparer. C'est-à-dire le **même test** est appliqué à **2 échantillons différents** et on utilise le **T-test à échantillons indépendants**.

✓ **Exemple**

On compare les moyennes des notes des élèves de 2 classes différentes du même niveau.

Si les données sont des intervalles qui ne satisfont pas les conditions de normalité et sont ordinales, on utilisera le **test Mann-Whitney** au lieu de **T test à échantillons indépendants** pour comparer deux échantillons indépendants. C'est **un test non-paramétrique**. Ici on compare les médianes des 2 échantillons.

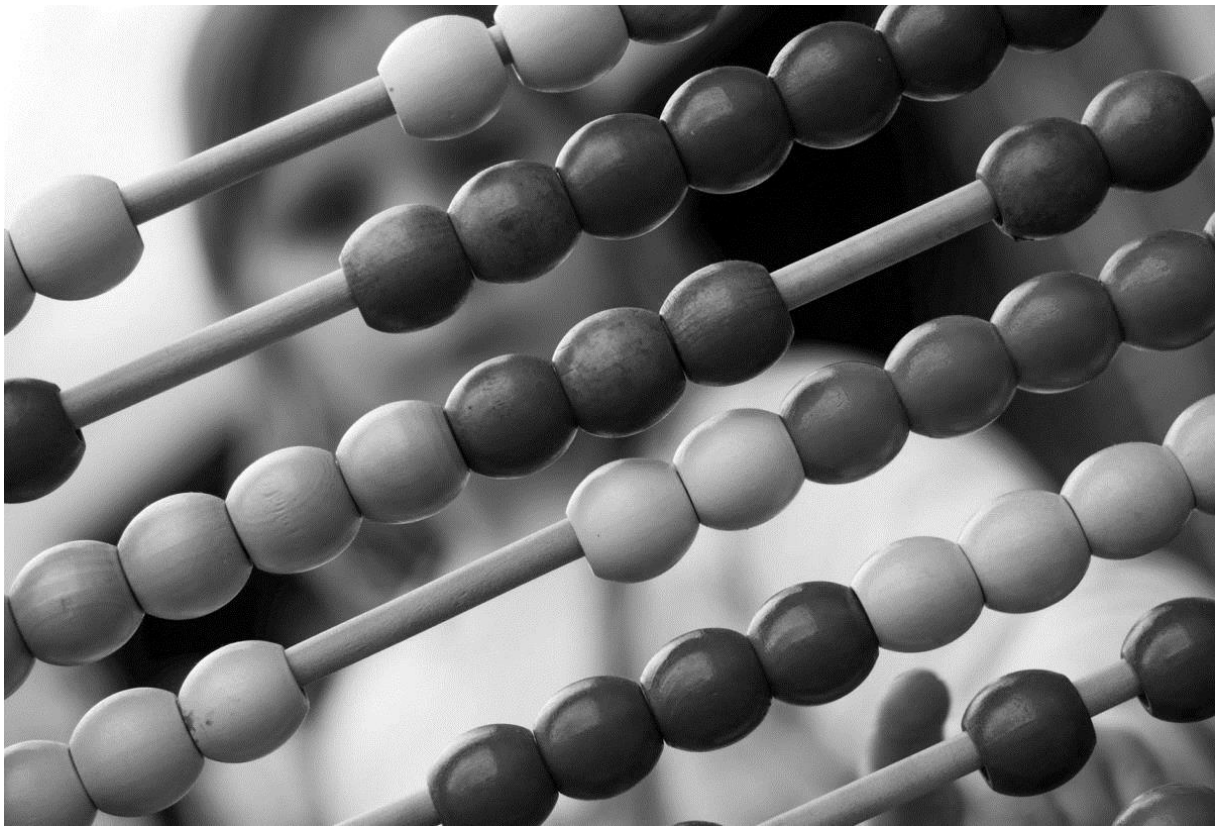
Plus de 2 échantillons : comparaison de leurs mesures

Deux tests se présentent, à savoir :

- **ANOVA** à 1 facteur (test paramétrique) ;
- Si les données ne suivent pas la loi normale ou sont ordinales, on utilisera **le Test Kruskal-Wallis** à la place du ANOVA à 1 facteur.

1.3.4 DIFFÉRENCE ET LE LIEN ENTRE ANALYSE DE DONNÉES, INTERPRÉTATION DES DONNÉES ET RÉDACTION DU RAPPORT DE RECHERCHE

L'analyse des données regroupe surtout un ensemble de méthodes statistiques susceptibles de « faire parler » les données. L'interprétation des données nous ramène au problème du « sens », du caractère explicatif des informations présentées. L'interprétation des données est aussi liée aux aspects « utiles » de ces informations d'où la notion de la pertinence et de la signification. Ainsi, faire de l'analyse de données, c'est accepter de perdre du détail des informations pour gagner en signification. Le rapport de recherche est du domaine de la rédaction en utilisant le style littéraire qui est le seul capable d'expliquer les choix ou de présenter une pensée en évolution, dont la communication peut avoir un intérêt même si elle n'est pas formalisée.



1.4 EXERCICES

Parmi les propositions suivantes, choisir celle(s) qui est (sont) exactes(s).

QCM1. Dans un échantillon, on s'intéresse à la taille et à la masse des individus par rapport à leur pays de naissance.

A. La variable « taille » est une variable quantitative continue
B. La variable « masse » est une variable quantitative discrète
C. La variable « pays de naissance » est une variable quantitative
D. Aucune des propositions précédentes n'est exacte

QCM2. L'objectif de l'analyse de données est de :

A. Perdre de l'information
B. Gagner de l'information
C. Perdre de la significativité
D. Gagner de la significativité

QCM3. Parmi les tests paramétriques, il y a :

A. Le test de Khi2
B. Le T-test apparié
C. Le test de WILCOXON
D. La régression multiple

QCM4. Le test du Khi2 est utilisé :

A. Dans le cadre d'une hypothèse bivariée
B. D'une hypothèse monovariée
C. Pour rechercher la relation entre variable quantitative uniquement
D. Pour rechercher la relation entre variable qualitative uniquement

QCM1 (A, B) - QCM2 (A, D) - QCM3 (A, B, D) - QCM4 (B, D).

2 INTRODUCTION AU TRAITEMENT DES DONNÉES SOUS SPSS

2.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin de ce module, les participants sauront ce que c'est que SPSS et les éléments qui le composent. Ils pourront ainsi, saisir des données ou lire les données dans SPSS à partir de fichier externe. Aussi, les participants pourront transformer les données et faire les première analyses statistiques (représentations graphiques, mesures descriptives) et avoir des base pour entamer les analyses approfondies (régression).

2.2 ASPECT THÉORIQUE

2.2.1 LOGICIELS DE TRAITEMENT DE DONNÉES

Les logiciels de traitement des données sont nombreux ; non allons citer quelques-uns :

- **Excel**, produit par Microsoft ; la version la plus récente contient une partie des procédures statistiques utilisées dans les analyses des données ;
- **StatBox et Question**, mis au point par la firme Grimmer Logiciels qui sont des logiciels conçus spécialement pour l'analyse des données d'enquête ;
- **Sphinx** qui est un logiciel utilisé surtout pour la recherche marketing ;
- **Minitab**, logiciel statistique puissant qui propose un grand nombre de procédures statistiques ;
- **SAS** (Système d'Analyse Statistique), a été conçu au départ pour le calcul économique et les modèles de régression ; par la suite, a été adapté de façon à y inclure les méthodes les plus connue de l'analyse des données ;
- **SPSS** (Statistical Package for the Social Science), a été créé au tout début pour les besoins des psychologues. Par la suite, un grand nombre de procédures statistiques y a été intégré.

2.2.2 DÉCOUVERTE DE SPSS

2.2.2.1 Qu'est-ce-que SPSS

SPSS signifie Statistical Package for the Social Science, est un logiciel spécialement conçu au début pour les analyses statistiques en sciences sociales. Avec le temps, un grand nombre de procédures statistiques ont été intégrées en vue de faciliter le travail de manipulation des données. C'est un logiciel spécialisé de le traitement statistique des données dont l'objectif qui permet de réaliser la totalité des tests statistiques. SPSS comprend plusieurs modules :

- Système de base ;
- Modèles de régression (regression models) ;
- Modèles avancés (advanced models) ;
- Tableaux (tables) ;
- Tests exacts (exact tests) ;
- Catégories (categories) ;
- Tendances (trends) ;
- Autres modules spécialisés.

2.2.2.2 Comment démarrer SPSS

SPSS installe dans le menu "Démarrer : Programme" une ligne pour démarrer le logiciel, toutefois, la localisation exacte peut varier d'un ordinateur à l'autre. Cette présentation est basée sur la version IBM SPSS Statistics 21 disponible pour les ordinateurs de type PC.

Pour lancer SPSS, deux méthodes peuvent être utilisée :

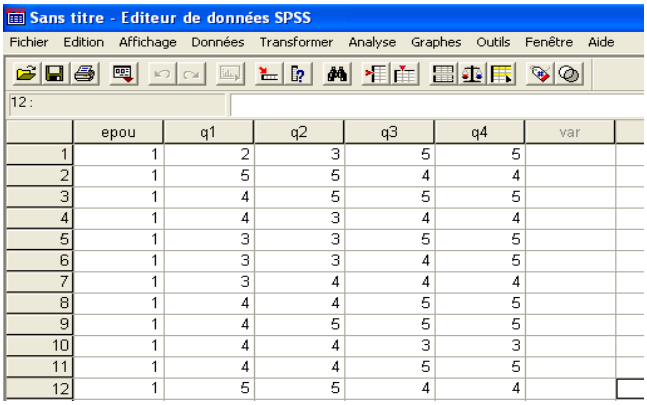
1. Faites un double clic sur l'icône SPSS apparaissant sur le bureau ; ou,
2. Cliquez sur Démarrer, puis Programmes et *IBM SPSS Statistics.*

2.2.2.3 Types de fenêtre dans SPSS

Il existe différents types de fenêtres dans le logiciel SPSS, toutefois, une session typique SPSS a toujours 3 fenêtres :

L'éditeur de données/ Data Editor

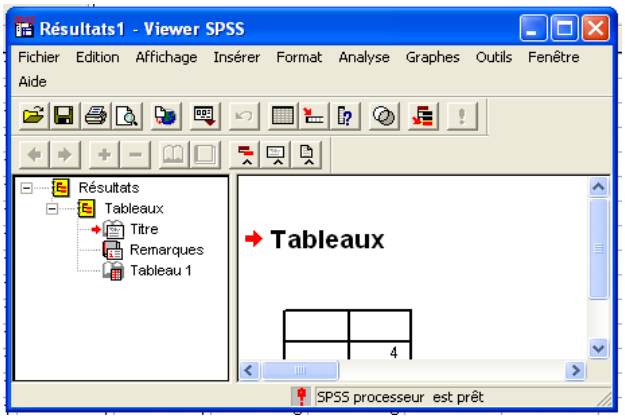
Cette fenêtre permet créer de nouveaux fichiers de données ou modifier des fichiers de données existants. Elle affiche le contenu du fichier de données actif (*fichier d'extension « .sav. »)*. Si vous avez plus d'un fichier de données ouvert, alors il y a une fenêtre Data Editor distincte pour chaque fichier de données.



	epou	q1	q2	q3	q4	var
1	1	2	3	5	5	
2	1	5	5	4	4	
3	1	4	5	5	5	
4	1	4	3	4	4	
5	1	3	3	5	5	
6	1	3	3	4	5	
7	1	3	4	4	4	
8	1	4	4	5	5	
9	1	4	5	5	5	
10	1	4	4	3	3	
11	1	4	4	5	5	
12	1	5	5	4	4	

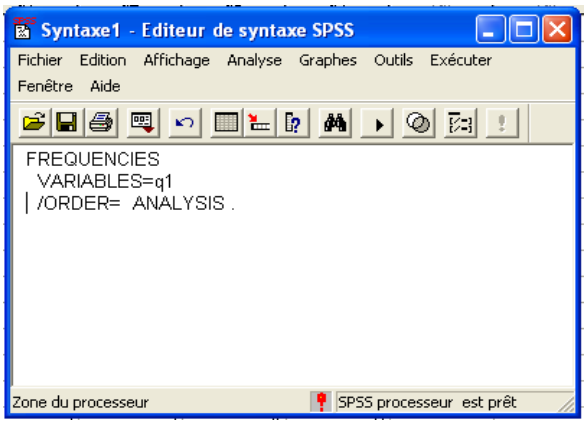
La fenêtre des résultats/ Viewer

Cette fenêtre s'ouvre automatiquement la première fois que vous exécutez une procédure qui génère des résultats (tableaux et diagrammes) : *c'est un fichier d'extension « .spo ».* Les résultats proprement dit apparaissent à droite. Sur la gauche, une table des matières des résultats générée par SPSS est affichée. Les résultats peuvent être imprimés tels quels, ou encore, on peut faire copier-coller vers un autre logiciel (tel que vers un logiciel de traitement de texte).



La fenêtre de syntaxe/ Syntax Editor

Cette fenêtre permet d'écrire les commandes d'analyses statistiques. Elle fonctionne comme un traitement de texte simple (fichier .sps.). Lorsqu'une commande est complète, il est possible de l'exécuter en allant dans le menu « Run », puis « Current » (ou encore en tapant « Ctrl-R »). Pour obtenir une fenêtre de syntaxe vide, aller dans le menu "Fichier » puis « Nouveau », puis « syntaxe ».



2.2.2.4 Barre des menus et Barre des boutons

C'est à partir de la barre des menus que l'utilisateur peut exécuter divers commandes permettant d'effectuer les opérations statistiques, graphiques et autres besoins. La barre des boutons est uniquement un raccourci de la barre des menus et ne contient que les commandes les plus fréquemment utilisés.

- Le menu **FICHIER FILE** : permet la gestion des fichiers (ex : ouvrir un nouveau fichier, fermer, enregistrer, etc.).
- Le menu **EDITION / EDIT** : permet d'effectuer les opérations de traitement de texte (ex : copier, couper, coller, sélectionner, etc.) ;
- Le menu **AFFICHAGE / VIEW** : permet de définir les options de l'écran (ex : barres d'outils) ;
- Le menu **DONNÉES / DATA** : traite de tout ce qui est lié à la gestion de la barre de données (ex : définir ou insérer une variable, trier les données, etc.) ;
- Le menu **TRANSFORMER / TRANSFORM** : présente les différentes opérations de transformation possibles sur les variables de la barre de données (ex : recodification, catégorisation, création d'indices, etc.) ;
- Le menu **ANALYSE /ANALYZE** : permet d'accéder à toutes les analyses statistiques que SPSS rend possibles (ex : analyses descriptives, corrélations, etc.) ;
- Le menu **GRAPHES / GRAPHS** : présente tous les types de graphiques que SPSS permet de créer (ex : histogrammes, boîtes à moustaches, courbes, etc.) ;
- Le menu **UTILITAIRES / UTILITIES** : comprend les utilitaires du programme (ex : informations sur les fichiers, informations sur les variables, etc.) ;
- Le menu **FENÊTRE / WINDOWS** : permet la gestion des fenêtres ;
- Le menu **AIDE / HELP** : propose des rubriques d'aide à l'utilisation de SPSS.

2.3 SYNTAXES SPSS

2.3.1 ENTRER LES DONNÉES À PARTIR DU QUESTIONNAIRE

2.3.1.1 Saisie des données

Les données constituent un élément de base pour le fonctionnement d'un logiciel de statistiques. Sans les données, il est impossible d'effectuer les différentes opérations mathématiques et statistiques. Avec SPSS, il est possible d'ajouter les données de deux façons différentes. La première façon consiste à les saisir directement dans l'écran **AFFICHAGE /VIEW** de SPSS. La deuxième

façon consiste à ***importer*** les données d'un autre logiciel (Excel ou Access, etc). Il est également possible d'importer les données d'un fichier ASCII. Ce type d'importation est surtout utilisé lorsque les données proviennent d'un ordinateur sur une plate-forme autre que IBM PC, les ordinateurs centraux de marque VAX et HP, par exemple.

2.3.1.2 Encoder le questionnaire

Il est recommandé de résumer les informations les plus importantes sur les variables rassemblées dans un « tableau de codage ». Ce tableau de codage à deux utilités à deux moments bien précis :

- **Pendant l'entrée des données** : comme règle de codage des variables ;
- **Après l'entrée des données** : comme description compacte du fichier des données.

Le tableau de codage de la base de données doit contenir les informations suivantes :

- ***Nom de la variable*** : les items qui appartiennent au même questionnaire devraient porter le même radical de leur nom (ex : satis1, satis2, satis3… pour un questionnaire mesurant le degré de satisfaction des individus concernant des sujets précis) ;
- Etiquette de la variable (variable label) ;
- ***Etiquettes des valeurs*** (value labels) : il s'agit d'une variable d'identification qui établit une relation entre les documents d'un cas et les données dans le fichier (ex : questionnaire et preuve de carnet de vaccination) ;
- ***Numéro d'identification (ID variable)*** : doit être noté sur les questionnaires pour que l'on puisse facilement retrouver le document d'un sujet afin de contrôler ou corriger des valeurs dans la base de données.

NB : Encoder vos données en suivant quelques règles suivantes :

- Une ligne par sujet. N'oubliez pas d'identifier vos sujets (colonne Identifiant) ;
- Une colonne par variable ;
- Tous les résultats en chiffres (pas de lettres !) ;
- Nom de la variable : maximum 8 caractères (commencer par une lettre).

2.3.1.3 Créer un nouveau fichier de données dans SPSS

Lorsque l'on démarre SPSS, une boîte de dialogue « Que voulez-vous faire ? / What would you like to do ? » apparaît par défaut. Si vous sélectionnez « Saisir des données / Type in data », vous obtenez un éditeur vide. Si l'on se trouve déjà dans un éditeur de données, il faut cliquer sur File/New Data. Une fois l'éditeur de données ouvert, il faut définir les variables dans la vue des variables (Vaiable View). Pour cela, vous vous aisdez du tableau de codage déjà créé à partir de votre questionnaire. Vous pouvez commencer avec le nom de la première variable (pour passer à la cellule suivante, puuyer sur « TAB » ou sur la flèche vers la droit (⟶). Il est possible d'attribuer une ou toutes les caractéristiques d'une variable à une ou plusieurs autres variables :

- Pour une caractéristique : copier la cellule (cliquant sur « copy », ou en cliquant sur le menue « Edit/Paste) ;
- Pour toutes les caractéristiques d'une variable : copier et coller toute la variable (en cliquant sur le numéro de la ligne, cette action la grise) ;

La seule caractéristique qui ne peut pas être copiée est le nom d'une variable, car chaque variable doit avoir un nom unique.

Pour créer plusieurs nouvelles variables avec le même radical dans leur nom (exemple : bf1, bf2, …, bfN), on procède comme suit :

- Entrer la variable bf1 avec son type, son étiquette, etc ;
- Copier cette variable ;
- Sélectionner « **Copy variables** » ;
- Dans la boîte de dialogue qui apparaît entrer le nombre de nouvelles variables à créer, leur radical (nom des nouvelles variables :bf) et le numéro de la première variable qui suivra le radical(2 car bf1 a été déjà créé).

2.3.1.4 Comment coder les réponses

Le codage dépend du type de variable qui se présente.

- **Coder les variables alphanumériques/série de caractères :** il faut entrer les caractères et définir la variable comme ***String*** (chaîne de caractères). Ex : Nom du répondant.
- **Coder les réponses à réponses courtes :** coder les réponses ouvertes avec des valeurs numériques. Si la varaible correspond à la question « *quelle est votre nationalité ? »,* vous pouvez coder les différentes réponses de la façon suivante : 1=Nigérienne ; 2=Ivoirienne ; 3=Tchadienne, etc.
- **Coder les réponses multiples :** il faut créer une variable pour chaque catégorie, de sorte que si la catégorie est choisie, la variable prend « 1 » et elle prend « 0 » dans le cas contraire. Si la varaible correspond à la question « ***quelles sont les activités que vous exercez ? »,*** vous pourver codes « A » pour éElevage », « B » pour « Commerce », « C » pour « Artisans », etc.

On va créer une variable Activ1-Activ4 en codant « 1 » (cette activité est choisie) et « 0 » (cette activité pas choisie). Pour la réponse « autre à préciser », il est possible de créer une variable alphanumérique (chaîne de caractère) qui permettra de recueillir les réponses qui seront données.

- **Coder les réponses ouver**tes : il faut regrouper les informations en catégorie grâce à l'analyse de contenu. Ce type de codification sera traité dans les prochains modules.

2.3.1.5 Comment coder les valeurs manquantes

Dès que vous crée une variable numérique, toutes les cellules de cette variable sont désignés par un point «.» (cela correspond à **Sysmis** c'est-à-dire **System Defined Missing**) qui est remplacé quand on entre une valeur. Garder le point «.» dans la cellule signifie que la valeur pour cette cellule est manquante. Ce type de données manquante n'a pas à être définie comme telle car elle est reconnue automatiquement par SPSS comme valeur manquante.

Entrer un chiffre en dehors de l'étendue de valeurs valables (ex : 99) et définir ces valeurs dans la vue des variables comme valeurs manquantes. Pour cela, entrer un espace dans « **discret missing values** ».

Concernant les variables alphanumériques, il ne suffit pas de garder la cellule vierge pour que SPSS la considère comme valeur manquante. Il faut aussi la définir comme telle dans la vue des variables, en entrant un espace dans « **discret missing values** ».

2.3.1.6 Réduire les erreurs en entrant les données

Vous devez toujours entrer les données « brutes » telles qu'elles sont. En effet, il est préférable de ne pas recoder à la main un item. Si vous avez un format de réponse bipolaire avec des valeurs positives et négatives, il est mieux d'utiliser un codage avec uniquement des valeurs positives.

Mais il est important d'avoir défini cela dans le tableau de codage du questionnaire. Exemple du tableau ci-dessous.

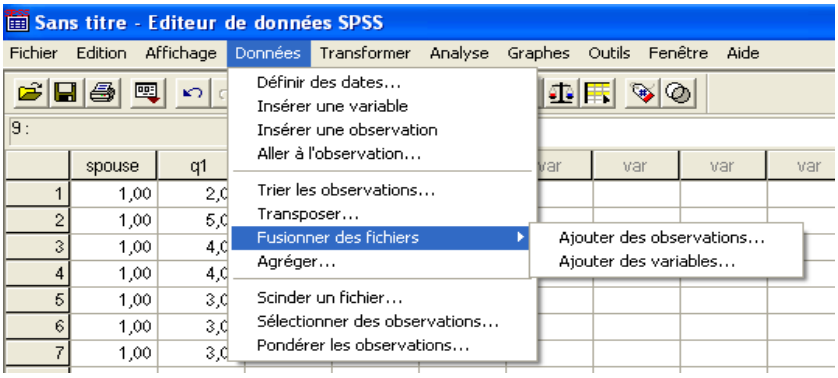
Format de réponse	--	-	0	+	++
Sujet à erreurs	-2	-1	0	1	2
Mieux pour la saisie	1	2	3	4	5

2.3.1.7 Eliminer et insérer des observations/variables dans la vue de données

- **Pour éliminer un cas** : sélectionnez la ligne et appuyer sur « **Delete** » sur le clavier.
- **Pour éliminer une variable** : sélectionnez la colonne et appuyer sur « **Delete** » sur le clavier.
- **Pour insérer un cas entre deux autres cas** : sélectionnez la ligne au-dessus de laquelle vous voulez insérer une observation et cliquez sur **Data/Insert cases** ou cliquez droit et sélectionnez **Insert cases.**
- **Pour insérer une variable entre deux autres variables** : sélectionnez la colonne avant celle où voulez insérer une variable et cliquez sur **Data/Insert Variables** ou cliquez droit et sélectionnez **Insert Variables**.

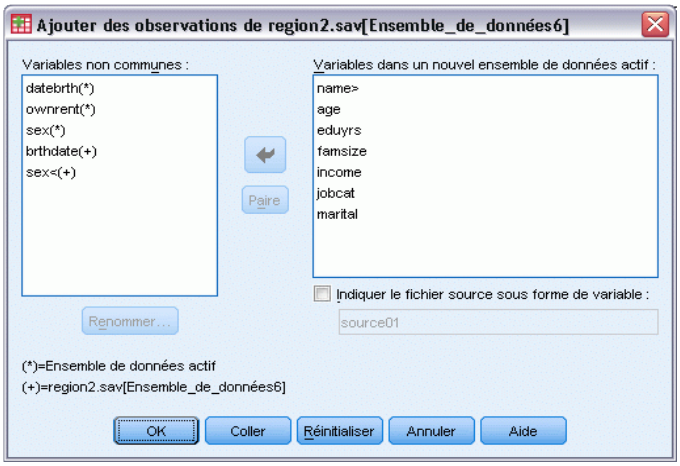
2.3.1.8 Fusionner des fichiers de données

Il existe deux type de fusion de données: soit on ajoute des observations, soit on ajoute des variables.



2.3.1.9 Ajouter des observations

Il est possible d'avoir deux fichiers contenant des variables similaires, mais des observations différentes. Par exemple, vous pouvez saisir des questionnaires sur des postes différents et chercher en fin de compte à obtenir un fichier unique. Pour cela il faut ouvrir le premier fichier, c'est-à-dire celui qui sera votre fichier de travail. A partir du menu, faites Données/Data→Fusionner des fichiers/Merge files→Ajouter des observations/Add cases : cherchez votre deuxième fichier.



La boîte de dialogue qui apparaît vérifie si les deux fichiers contiennent les mêmes variables (avec les mêmes noms). Par défaut, toutes les variables de même nom seront incluses Si ce n'est pas le cas, les variables non appariées se trouvent dans la section « **Variables non communes/Unpaired**

Variables ».

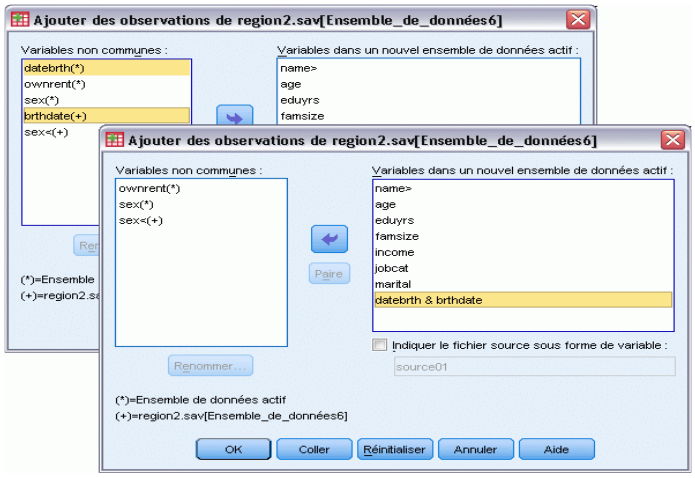
Attention

Il est important que le nom des variables, ainsi que le type, l'étiquette et les valeurs et les données manquantes soient définis exactement de la même manière dans les deux fichiers, sinon SPSS définira une variable avec même nom comme « Variables non communes/Unpaired Variables » (fenêtre de gauche) - toujours vérifier que les variables communes aux deux fichiers qui nous intéressent se trouvent bien dans « Variables dans un nouvel ensemble de données actif/Variables in new working data file ».

Les variables qui apparaissent seulement dans un fichier de données peuvent quand même être ajoutées en appuyant sur « les cas de l'autre fichier qui n'ont pas de valeurs à ces variables reçoivent des valeurs manquantes = **Sysmis »**.

S'il y a des variables dans les deux fichiers qui mesurent la même chose mais qui ne portent pas le même nom (ex : par erreur), on peut les apparier. Pour faire cela, il faut sélectionner les deux variables (on sélectionne la deuxième variable en pressant sur la touche « **CTRL »**), puis appuyer sur « **Apparier/Paired** », on obtient ainsi dans « ***Variables dans un nouvel ensemble de données actif* /Variables in new working data file** », une nouvelle variable qui se nomme datebirth & birthdate (dans le fichier fusionné, la variable portera le nom du premier fichier).

Une fois toutes les variables qui nous intéressent sélectionnées, il faut cliquer sur **Ok**, ce qui ajoute les observations du deuxième fichier au premier. On a maintenant un nouveau fichier de données. Si nous sauvons ce fichier en faisant **Fichier/File>Enregistrer/Save** , cela va écraser le premier fichier. Ce qu'il faut faire c'est **Fichier/File > Enregistrer sous/Save as** et on obtient ainsi une nouvelle base de données, différentes des deux initialement utilisées pour sa création.



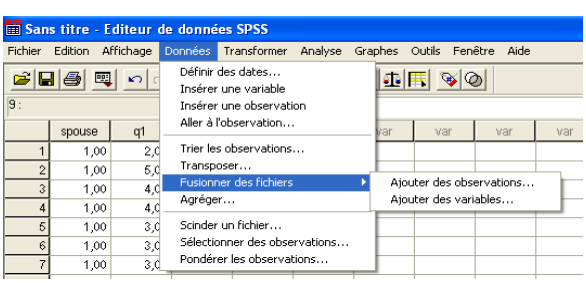
2.3.1.10 Ajouter des variables : mêmes observations

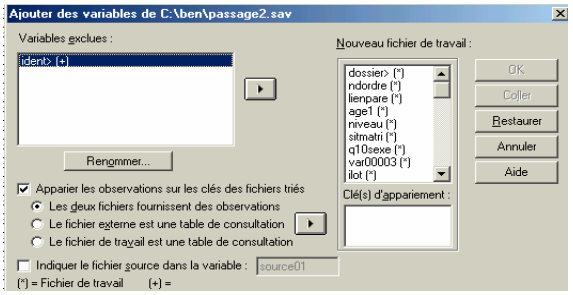
Ajouter des variables permet de fusionner le fichier de données de travail avec un fichier de données SPSS externe qui contient les mêmes observations mais pas les mêmes variables. Par exemple, vous souhaitez ***fusionner*** un fichier de données contenant les résultats d'une ***enquête préalable*** (fichier passage1.sav) avec un autre fichier contenant les résultats d'une ***autre enquête*** (fichier passage2.sav) sur les ***mêmes individus***.

Pour fusionner les deux fichiers :

- Ouvrir un des fichiers à fusionner (passage2.sav), le trier par ordre croissant selon la ou les clés d'appariement puis l'enregistrer **(Données/Data > Trier/Sort cases)** ;
- Ouvrir l'autre fichier (passage1.sav), le trier également par ordre croissant selon la ou les clés d'appariement puis l'enregistrer **(Données/Data > Trier/Sort cases)** ;
- A partir des menus, sélectionnez : Données > Fusionner des fichiers > Ajouter des variables ;

- Sélectionner l'ensemble de données ou le fichier de données au format SPSS Statistics à fusionner avec l'ensemble de données actif (passage2.sav), la boîte de dialogue ci-dessous apparaît :





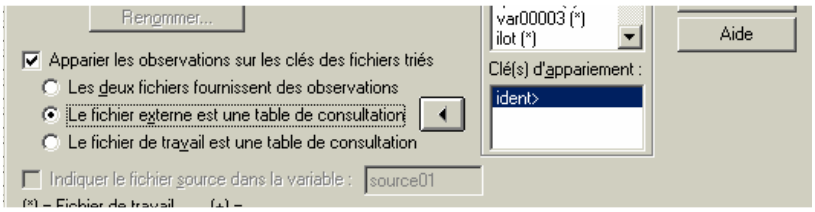
- Dans la zone « variables exclues », les noms de variable du second fichier de données de travail sont exclus par défaut ;
- Les variables du nouveau fichier de données de travail apparaissent dans la zone « variables à inclure » dans le nouveau fichier actif. Par défaut, tous les noms de variables uniques dans les deux fichiers sont inclus dans la liste ;
- Sélectionner la variable « clé d'appariement » pour la basculer dans la zone clé(s) d'appariement ;
- Cocher la ligne les deux fichiers fournissent des observations ;
- Terminer en cliquant sur « OK ».

2.3.1.11 Ajouter des variables : fichiers hiérarchisés

Il s'agit ici d'une fusion où l'un des fichiers est une table codée. **Une table codée** ou un fichier de consultation de table est un fichier dans lequel les données de chaque observation peuvent s'appliquer à plusieurs observations de l'autre fichier. Par exemple, si l'un des fichiers contient des informations sur les différents membres d'une famille (sexe, âge, niveau scolaire) et l'autre des informations générales sur la famille (revenu global, taille, habitat), vous pouvez utiliser le fichier sur la famille comme fichier de consultation et appliquer les informations générales à chaque membre de la famille dans le fichier fusionné.

La démarche à suivre pour la fusion de tels fichiers reste semblable à l'ajout de variables avec les mêmes observations. La différence est que :

- Les clés d'appariement doivent avoir le même nom dans les deux fichiers de données ;
- Il faut préciser lequel des deux fichiers est le fichier consultation de table (voir le schéma ci-dessous) :



- Les deux fichiers de données doivent présenter les variables par ordre croissant ou décroissant et l'ordre des variables de la liste clés d'appariement doit être le même que l'ordre de tri. Les valeurs de cette variable clé doivent être identiques (par exemple, une variable alphanumérique doit être entrée de la même façon - par rapport aux majuscules/minuscules et nombre de caractères - attention aux espaces après les derniers caractères !) ;

- Les observations qui n'ont pas de correspondance dans les clés d'appariement sont incluses dans le fichier fusionné mais elles sont fusionnées avec les observations de l'autre fichier. Les observations sans correspondance contiennent des valeurs uniquement pour les variables du fichier duquel elles sont issues. Les variables de l'autre fichier contiennent la valeur manquante par défaut.

2.3.1.12 Utiliser Excel pour entrer les données dans SPSS

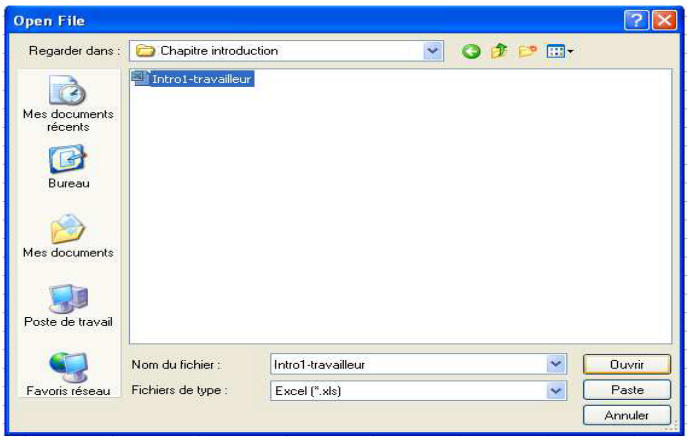
Pour utiliser Excel pour entrer les données dans SPSS, deux étapes s'appliquent :

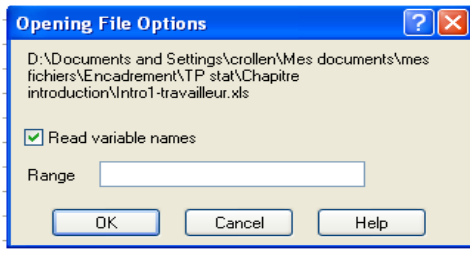
Créer une grille de données avec Excel

- Saisir les noms des variables (pas les étiquettes) dans la première ligne ;
- Les lignes au-dessous des noms représentent les observations (une observation par ligne) ;
- Toutes les données à traiter doivent se trouver sur la même feuille (ne pas séparer par exemple en créant une feuille pour les filles et une autre pour les garçons) ;
- Sauvegarder les données comme classeur Excel (extension **.xls**).
- Fermer la feuille Excel (sinon, vous ne pourrez pas l'ouvrir dans SPSS).

Importer les données dans SPSS

- Ouvrir SPSS, ce logiciel permet d'importer directement des tableaux de format Excel, si on précise dans « Ouvrir » l'extension « .xls » ou « .xlsx » ;
- Cliquer sur **Fichier/File > Ouvrir > Données** (ou cliquer sur l'icône correspondante) ;
- Sélectionner « Fichier Excel », (voir figure ci-dessous).





Cochez cette case si la première ligne dans Excel contient le nom des variables

Les données apparaissent alors dans la fenêtre SPSS.

- Terminer et sauvegarder les données comme fichier SPSS (extension **.sav**).

2.3.1.13 Utiliser un fichier texte pour entrer les données dans SPSS

Deux étapes peuvent être suivies :

Créer une grille de données avec un fichier texte

- Ouvrir un fichier texte (ex : notepad) ;
- Saisir les noms des variables dans la première ligne que vous séparez soit par un point-virgule, soit par un séparateur comme « tab ») ;

- Les lignes au-dessous des noms représentent les observations (une observation par ligne) et chaque valeur de la ligne doit être séparé pat un séparateur (toujours utiliser le même séparateur) ;
- Sauvegarder les données comme classeur texte (extension **.txt**).

Importer les données dans SPSS

- Ouvrir SPSS ;
- Cliquer sur Fichier/File > Lire données texte/Read text data ;
- Spécifier l'arrangement de vos variables : types de séparateur, est-ce que chaque ligne correspond à un cas, est-ce que le fichier contient le nom des variables sur la première ligne, etc ;
- Terminer et sauvegarder les données comme fichier SPSS (extension **.sav**).

2.1.3.14 Eliminer les erreurs de l'entrée de données

Deux types d'erreurs sont possibles :

- Erreurs de type (a) : la valeur fausse est une valeur dans l'étendue des valeurs valables ;
- Erreurs de type (b) : la valeur fausse est une valeur en dehors de l'étendue des valeurs valables.

On peut réparer les erreurs de type (a) qu'en comparant les données vraies avec les données entrées. Mais cela est très coûteux, d'où l'importance d'être très minutieux en entrant les données.

Donc seules les erreurs de type (b) peuvent être repérées et éliminées après l'entrée des données.

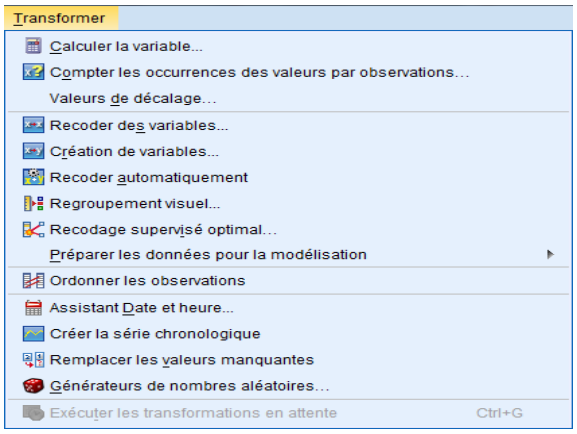
Pour repérer et éliminer les erreurs de type (b), il faut inspecter les valeurs minimales et maximales de toutes les variables en passant par : **Analyse/Analyze > Statistiques descriptives/Desriptive Statistics > Effectifs/Frequencies**. Il faut sélectionner toutes les variables numériques et ajouter les dans la section droite, puis cliquer dans « **Statistics** », choisir « **Minimum** » et « **Maximum** ».

2.3.2 PRÉPARATION DES DONNÉES : TRANSFORMER LES DONNÉES

2.3.2.1 Transformer les données

Transformer les variables dans une base de données fait partie des tâches à réaliser avant l'analyse de données proprement dite (statistiques descriptives, régression). Les données collectées se présentent généralement sous la forme brute et elles sont inadaptées à une analyse statistique poussée. Par exemple, la date de naissance est moins utile à l'analyse statistique que l'âge. La transformation des variables permet à partir de certains paramètres bruts de créer de nouveaux paramètres plus pertinents à considérer.

Le logiciel SPSS permet certaines procédures de transformation des variables, dont les plus utilisées sont :



- Créer une nouvelle variable à partir d'une formule de calcul, faisant intervenir un ou plusieurs paramètres. Ces formules mathématiques peuvent être simples ou complexe (construire des indicateurs comme calculer des scores d'échelle, des sous échelle, centrer et réduire une variable, etc.) ;
- Changer la présentation des données d'une variable, en regroupant certaines valeurs d'une ou des variables cela s'appelle **« Recodage »** (regrouper des catégories, inverser les sens de l'échelle, etc.). Le recodage change les valeurs d'une variable, soit en écrasant les anciennes valeurs ou en copiant les nouvelles valeurs dans une variable distincte. Cette dernière procédure est plus sécuritaire, puisqu'elle permet de garder une copie de la variable d'origine ;
- Compter l'occurrence d'une ou plusieurs valeurs dans plusieurs variables distincte ;
- D'autres procédures de transformation sont disponibles également sous SPSS : fusionner les valeurs de deux ou plusieurs variables ou au contraire séparer les valeurs d'une variable ou en extraire une partie, etc.

Quelques conseils au moment d'effectuer des transformations de données :

- Après avoir éliminé les erreurs, gardez toujours un fichier des données brutes ;
- Avant d'effectuer des modifications, sauvegardez les données sous un autre nom pour conserver le fichier original ;
- Toutes les variables initiales restent dans le fichier et les variables recodées obtiennent un nouveau nom.

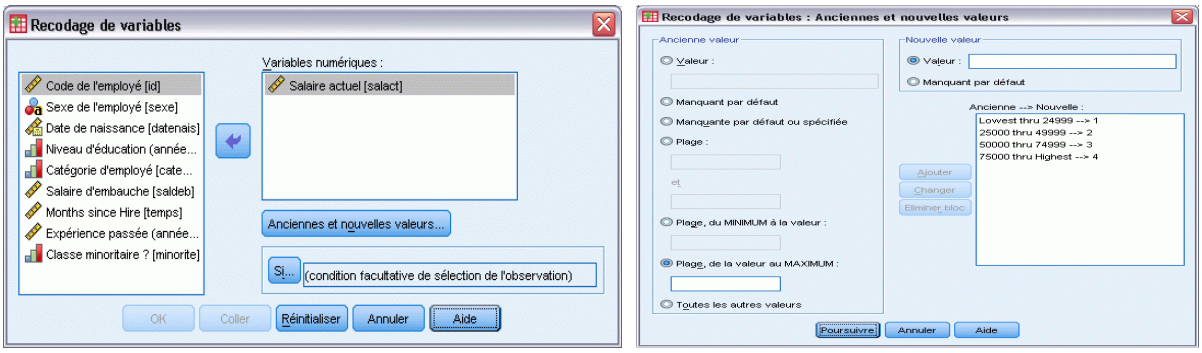
2.3.2.2 Recoder les variables

L'opération de recodage permet de modifier directement les modalités d'une variable ou de créer une nouvelle variable en recodant les modalités de l'ancienne. Le deuxième cas offre l'avantage de conserver l'ancienne variable. Si vous **sélectionnez plusieurs variables, elles doivent toutes être du même type. Vous ne pouvez pas recoder** ensemble des variables numériques et chaîne (Par exemple, fusionner des salaires dans des modalités d'intervalles de salaires).

Pour recoder les valeurs d'une variable il faut :

- A partir des menus, sélectionner : **Transformer** > **Recoder des variables** ;
- Sélectionner les variables sue vous désirez recoder. Si vous sélectionnez plusieurs variables, elles doivent être du même type (numérique ou alphanumérique) ;
- Cliquer sur « Anciennes » et « nouvelles valeurs » et spécifier comment recoder les valeurs.

Figure 2 : Recoage de variable



- Le bouton « **Valeur** » du cadre nommé « **ancienne valeur** » permet de choisir une valeur précise ancienne et de lui attribuer un code ;
- Le bouton « **manquante par défaut** » ou « **manquante par défaut ou spécifiée** » permet de choisir la valeur manquante par défaut ou spécifier ancienne pour lui attribuer un code.
- Le bouton « **plage** » permet de choisir un intervalle de valeur dans les anciennes valeurs et de lui donner un code. Les intervalles sont parfois fermés parfois ouverts. Il faut suivre le sens des intervalles.
- Le bouton « toutes les autres valeurs » englobe toute autre valeur restante non incluse dans l'une des spécifications de la liste **Ancienne-Nouvelle**. Ces valeurs apparaissent sous l'intitulé **ELSE** dans la liste Ancienne-Nouvelle.
-

Attention

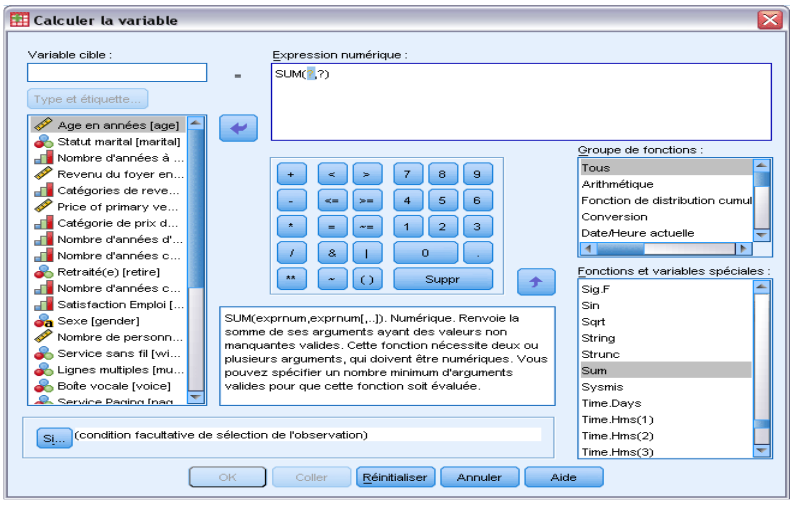
Il est préférable de faire le processus : **Transformer** > **Recoder des variables** > **Dans une variable différente**. Il faut éviter d'utiliser **Transformer** > **Recoder des variables** > **Dans une même variable** car cette option écraserait votre variable initiale.

2.3.2.3 Construire les indicateurs

Pour construire une nouvelle variable à partir de plusieurs variables de départ, il faut passer par :

- Transformer > Calculer la variable.

On obtient la boîte de dialogue ci-dessus.



On définit le nom de la nouvelle variable (« **Variable cible** »), ainsi que « **type et l'étiquette** » et les **valeurs** en cliquant en haut à gauche sous le nom de la variable.

En haut à droite sous « **Expression numérique** » on définit la transformation des variables initiales qu'on souhaite effectuer.

Les transformations qui peuvent être effectuées sont définies en bas à droite dans « **Fonctions et variables spéciales** ». En cliquant sur une des fonctions, on obtient au centre de la boîte de dialogue une explication de ce que fait l'opération choisie et comment rentrer les variables initiales (qui se trouvent à gauche dans la boîte). Par exemple, pour construire la sous-échelle d'expression de joie, nous allons utiliser la moyenne des deux éléments qui composent cette sous échelle : exp02 (joie) et exp08 (bonheur).

L'une des manières usuelles pour agréger différente variables est l'utilisation de la moyenne ou la somme des différentes variables initiales.

Dans notre exemple, au lieu de mettre dans « **Expression numérique** » : MEAN (exp02, exp08), il est possible de calculer la moyenne des items seulement pour les personnes qui ont une valeur valable (non manquante) sur au moins « g » de ces items (sinon, la nouvelle valeur sera manquante). Pour cela il faut spécifier le nombre de valeurs valables après l'expression MEAN : MEAN.g(exp02,exp08). Par exemple, MEAN.2(exp02,exp08) ne fait la moyenne que s'il n'y a pas de valeur manquante.

N.B : Il est recommandé d'avoir au moins 80 % de valeurs valables pour calculer le score (la moyenne) d'une échelle. Si on choisit MEAN (exp02,exp08), cela correspond à MEAN.1(exp02,exp08), c'est-à-dire qu'on fait la moyenne au moins une variable (mais pas forcément plus).

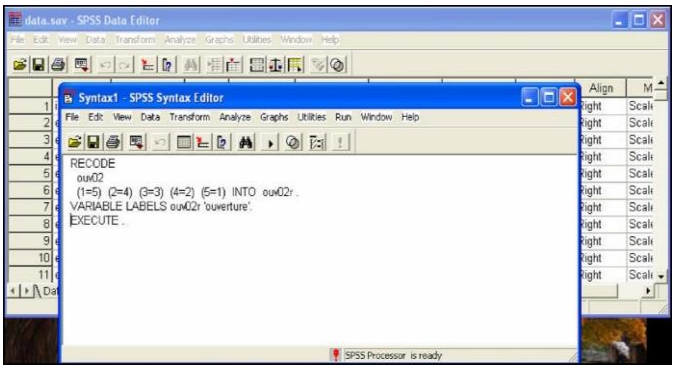
Attention

On pourrait aussi calculer la somme de tous les items d'une échelle. Cependant, il n'est pas évident comment gérer les valeurs manquantes en créant la somme des items - si on enlève tous les cas avec les valeurs manquantes on risque de fortement diminuer l'échantillon. Le remplacement des valeurs manquantes par la moyenne de l'échantillon est aussi problématique! C'est pour cela que calculer la moyenne est recommandé.

2.3.2.4 Transformer les données en utilisant la page syntaxe

Pour transformer les données en utilisant la page syntaxe, il faut utiliser l'option « **coller »** dans la boîte de dialogue qui s'affiche lorsque vous faites la procédure de transformation (**Transformer > …** et appuyer sur le bouton « **Coller** » au lieu de « **OK** »). La commande exécutée s'inscrira dans la page de syntaxe.

La syntaxe est un fichier de texte qui peut être sauvegardé (extension **.sps**) et imprimé. Les commandes peuvent être copiées, collées et changées. Aussi, on peut ajouter du texte commentaire qui doit être précédé d'un astérisque et suivi par un point. Pour exécuter les commandes, on les sélectionne et on envoie la syntaxe en appuyant sur le bouton « ▶ ».



Les commandes enregistrées peuvent être réutilisées dans les séances de travail ultérieures.

Les règles les plus importantes à suivre :

- Une commande se compose de son nom (ex : recode) et de ses spécifications (ex : noms des variables, règle de recodage) ;

- Chaque commande doit débuter sur une nouvelle ligne et se terminer par un point ;
- Il n'y a pas de lignes vides à l'intérieur d'une commande.

2.3.3 ANALYSE DES DONNÉES : REPRÉSENTATIONS GRAPHIQUES

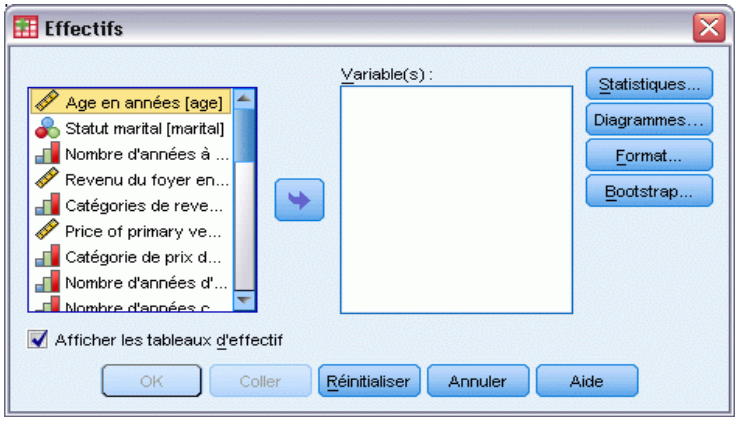
2.3.3.1 Fréquences

Les tableaux de fréquences indiquent pour une variable donnée, toutes les valeurs prises par cette variable, le nombre de fois que chaque valeur apparaît et la proportion qu'elle représente par rapport à l'ensemble des autres valeurs de la variable.

A partir du menu, sélectionner :

Analyse > Statistiques descriptives > Effectifs

La boîte de dialogue s'affiche.



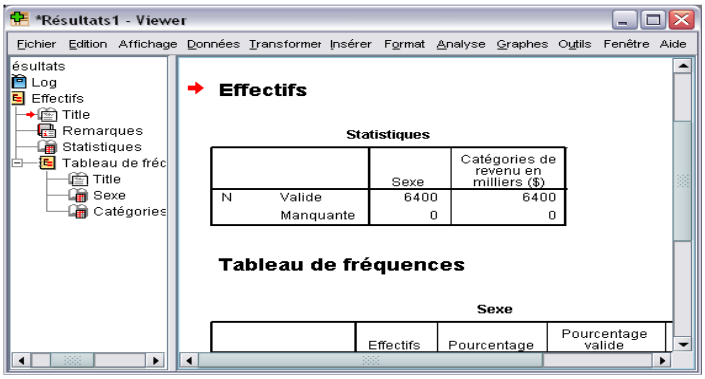
Une icône à côté de chaque variable fournit des informations sur le type de données et le niveau de mesure.

	Numérique	Chaîne	Date	Heure
Echelle (continue)		n/a		
Ordinal				
Nominal				

- Dans la boîte de dialogue, choisissez les variables à analyser dans la liste source à gauche et faites-les glisser dans la liste **Variable(s)** à droite ;
- Le bouton « **OK** », qui exécute l'analyse, est désactivé jusqu'à ce qu'une variable soit placée dans la liste **Variable(s).**

Vous pouvez obtenir des informations supplémentaires en cliquant avec le bouton droit sur tout nom de variable dans la liste.

Exemple d'affichage de résultats



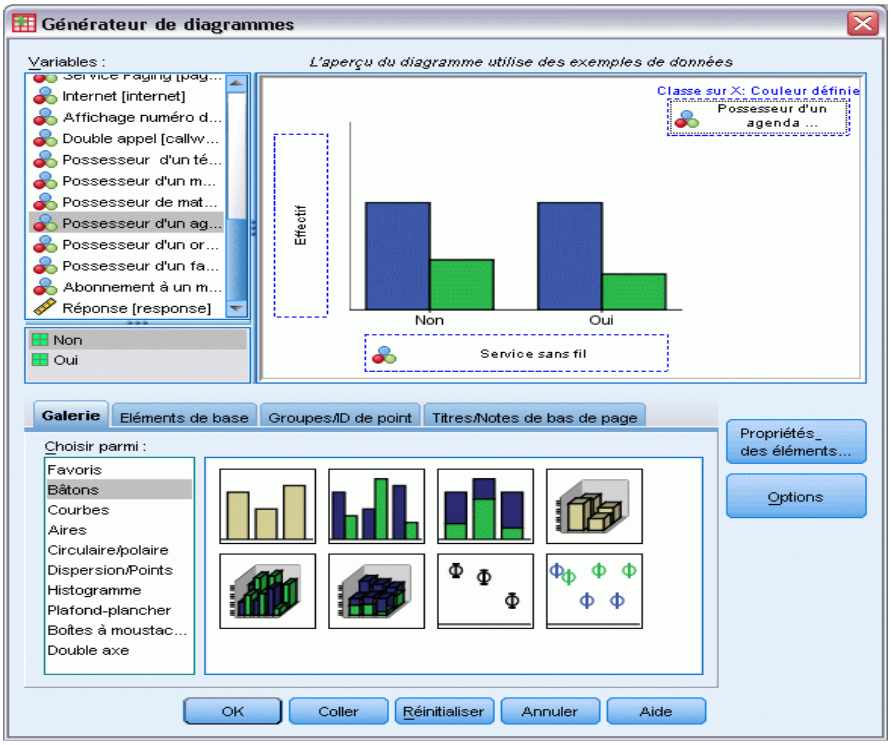
2.3.3.2 Graphiques pour les variables nominales et ordinales (fréquences)

La procédure **GRAPHES** génère des graphiques en élaborant des statistiques à partir des données du fichier actif. Plusieurs catégories de graphiques peuvent être obtenus : bâtons, bâtons 3D, courbes, aires, etc.

A partir du menu, sélectionner :

- Graphes > Générateur de diagrammes > Galerie (s'il n'est pas sélectionné) ;
- Cliquer sur le diagramme dont vous avez besoin et faites le glisser dans l'espace réservé qui est la zone étendue au-dessus de la galerie ;
- Renseigner les axes à partir des variables qui sont à gauche ; sélectionner et faire glisser dans le cadre de l'axe réservé.

Boîte de dialogue « **Générateur de diagrammes** » avec comme exemple le diagramme en bâtons.



2.3.3.3 Graphiques pour les variables métriques

Pour présenter la distribution des fréquences d'une variable métrique dans tout l'échantillon, on peut faire un histogramme, pour cela, à partir du menu, sélectionner :

- Graphes > Générateur de diagrammes > Galerie (s'il n'est pas sélectionné) ;
- Cliquer sur le diagramme dont vous avez besoin (histogramme) et faites le glisser dans l'espace réservé qui est la zone étendue au-dessus de la galerie ;
- Renseigner les axes à partir des variables qui sont à gauche, sélectionner et faire glisser dans le cadre de l'axe réservé.

Vous pouvez également faire :

- Graphes > Boîtes de dialogue ancienne version > Histogramme
- Entrer la variable pour laquelle vous voulez obtenir les fréquences dans « **Variable** ».

2.3.3.4 Modifier les caractéristiques d'un graphique

Pour pouvoir modifier un graphique, il faut commencer par double-cliquer dessus. Dans **l'éditeur de diagrammes / éditeur de graphiques (chart editor)** qui s'ouvre on peut modifier différents paramètres :

- **Taille du diagramme** : redimensionner la hauteur et la largeur du diagramme ;
- **Remplissage et bordures** : changer les couleurs et le style des traits ;
- **Variables** : changer de variable, de position des axes.

2.3.3.5 Exporter des tableaux et des graphiques dans Word

Pour exporter un tableau ou un graphique, sélectionner dans la fenêtre **résultats / Viewer,** le tableau et cliquez sur le bouton droit de la souris. Sélectionnez « **Copier** » et les résultats seront collés dans un tableau Word qui peut être modifié. Dans Word, cliquez sur le bouton droit de la souris et sélectionnez « **coller** ».

La taille du graphique peut être modifiée en traînant un coin du cadre du graphique. Parfois, le graphique n'est pas bien reproduit dans Word. Dans ce cas, il faut sauvegarder le graphique comme un fichier et l'insérer après dans Word. Dans la fenêtre **résultats / Viewer** :

- cliquez droit sur le graphique, sélectionnez « **Exporter…** » ;
- Dans la boîte-dialogue, option « **Document > Type**», choisir « **Aucun (diagrammes uniquement)** » ;
- Pour « **Graphiques > Type** » choisir « **Enhanced Metafile *.emf** ». En cliquant sur « **Browse** », choisir l'endroit où vous voulez sauvegarder le fichier et son nom.
- Dans Word : **menu Insertion > Image > a partir du fichier**. Si le graphique est de nouveau mal reproduit, faire la même opération, mais pour « **Graphiques > Type** » choisir « **Fichier JPEG (*jpg)** ». Ce type de format est très universelle, cependant la définition (en points par pouce) du graphique ne sera pas très bonne et typiquement pas acceptable pour une publication.

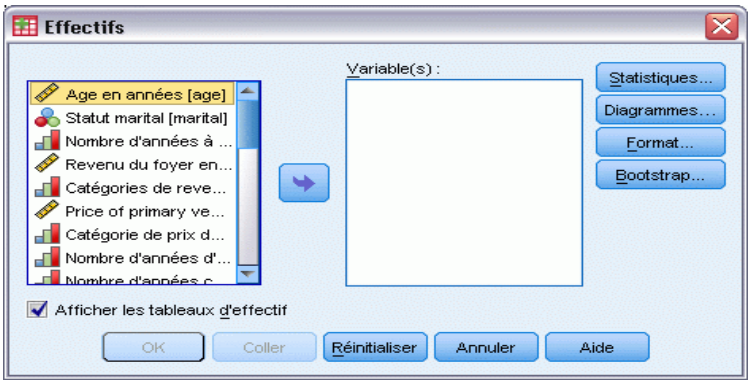
2.3.4 ANALYSE DES DONNÉES : MESURES DESCRIPTIVES

2.3.4.1 Mesures descriptives

Pour obtenir des informations (en forme de tableaux ou graphiques) sur la distribution d'une variable (fréquences, tendance (mode, médiane, moyenne, …), dispersion (variance, écart type, intervalle interquartile, …), on utilise le menu:

- Analyse > Statistiques descriptives > Effectifs.

La boîte de dialogue suivante s'affiche



- Choisir une plusieurs variables ;
- **Afficher les tableaux d'effectifs *:*** tableaux de distribution de fréquences (coché par défaut) (si vous ne voulez que des statistiques et/ou graphique, vous pouvez décocher).
- *Le bouton* **Statistiques** : permet d'ajouter des statistiques de Fractiles (quartiles, centile, …), de tendances centrale (moyenne, médiane, mode, …), de dispersion (écart type, variance, minimum-maximum, …) et de distribution (skewness = coefficients d'asymétrie, kurtosis = coefficient d'aplatissement).

Attention

Le choix des statistiques dépend de l'échelle de mesure, mais SPSS calcule tous les coefficients pour toutes les variables choisies - même si ça n'a pas de sens!

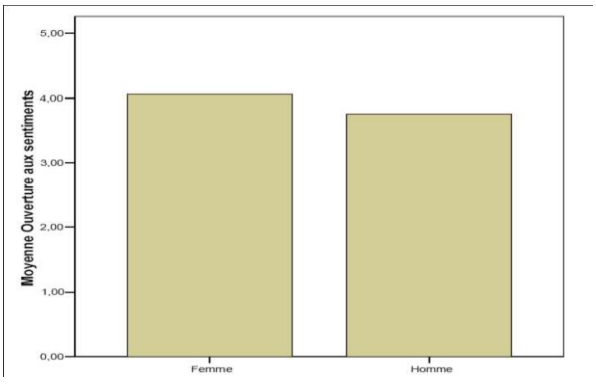
- *Le bouton* ***Diagrammes*** : permet d'ajouter un diagramme au tableau de fréquences. On a le choix entre diagramme en bâtons, diagramme, diagramme circulaire/camembert/pie chart, pour les variables nominales ou ordinales et histogramme pour les variables métriques.
- *Le bouton* ***Format*** : permet de choisir des options d'affichage des tableaux de fréquences. On choisir dans quel ordre les catégories de la variable seront affichées (*Ordre d'affichage* : valeurs dans l'ordre croissant, …) et la forme de présentation des coefficients de plusieurs variables (*variables multiples* : comparer variables, séparer résultats par variables).
- *Le bouton* ***Bootstrap*** : permet d'estimer la distribution d'échantillonnage d'un estimateur en procédant à un rééchantillonnage avec des remplacements par rapport à l'échantillon initial. On peut y définir les intervalles de confiance et le type d'échantillonnage.

2.3.4.2 Graphiques de mesures descriptives

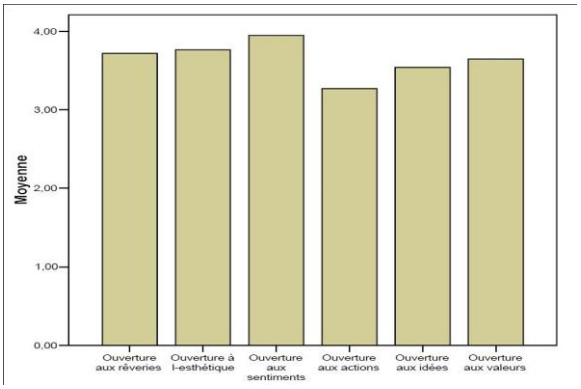
Les graphiques permettant de matérialiser la distribution d'une variable métrique sont : les diagrammes en bâtons (Bar), ou les barres d'erreurs (Error bar).

Diagramme en bâtons (Bar)

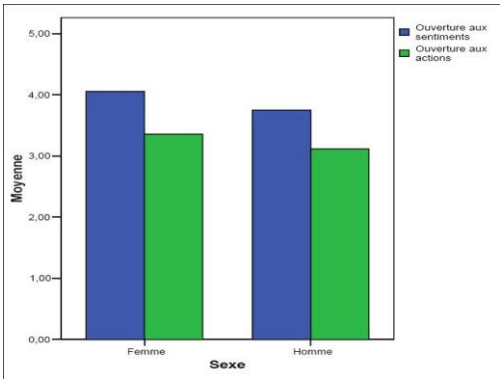
Pour présenter graphiquement la moyenne d'une variable métrique dans différents sous-groupes, il faut cliquer sur : **Graphes > Boîtes de dialogue ancienne version > choisir « Simple » >** choisir **« Récapitulatifs pour groupes d'observations »**. Dans la boîte de dialogue « *Définir le diagramme en bâtons simples…* », choisir **« Autre statistique (moyenne, par exemple)** » > dans « l'*Axe des modalités »*, choisir la variable qui intéresse > « **ok »**.



Pour représenter la moyenne de plusieurs variables métriques dans tout l'échantillon, il faut cliquer sur : **Graphes > Boîtes de dialogue ancienne version > choisir « Simple » >** choisir **« Récapitulatifs pour variables distinctes »**. Dans la boîte de dialogue « *Définir le diagramme en bâtons simples…* », dans *les bâtons représentent*, faire glisser toutes les variables qui intéressent > « **ok** ».



Pour représenter la moyenne de plusieurs variables métriques dans différents sous-groupes, il faut cliquer sur : **Graphes > Boîtes de dialogue ancienne version >** choisir **« Juxtaposé » >** choisir **« Récapitulatifs pour variables distinctes »**. Dans la boîte de dialogue « *Définir le diagramme en bâtons juxtaposés…* », dans « *les bâtons représentent* », faire glisser toutes les variables qui intéressent > dans « *Axe des modalités* » entrez la variable de groupement > « **ok** ».



Exemple : la moyenne de l'ouverture pour deux dimensions en fonction du sexe.

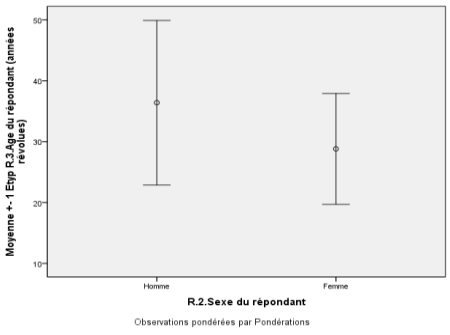
Remarque :

La moyenne devrait être représentée par un point et non par une hauteur ou une surface comme c'est le cas dans les diagrammes en bâtons, c'est pourquoi nous vous conseillons d'utiliser les graphes error bar/barres d'erreurs (ci-après) qui sont corrects d'un point de vue strictement du sens de la mesure de la moyenne.

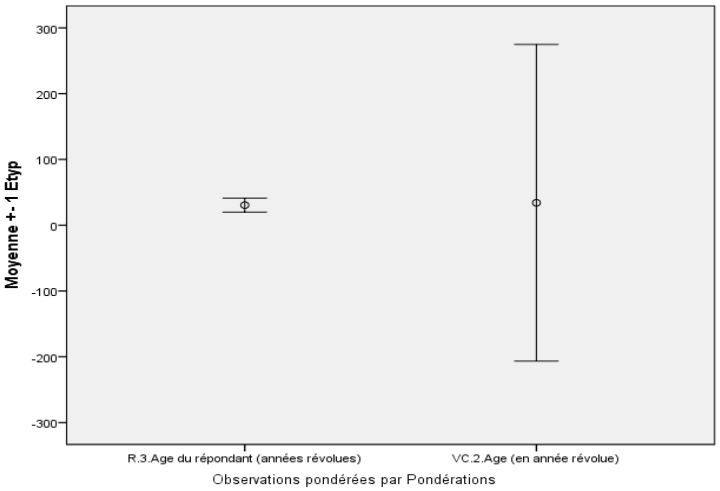
Barres d'erreur (Error bar)

La barre d'erreur permet de représenter la moyenne et la variabilité de variables métriques (moyenne plus ou moins l'écart-type).

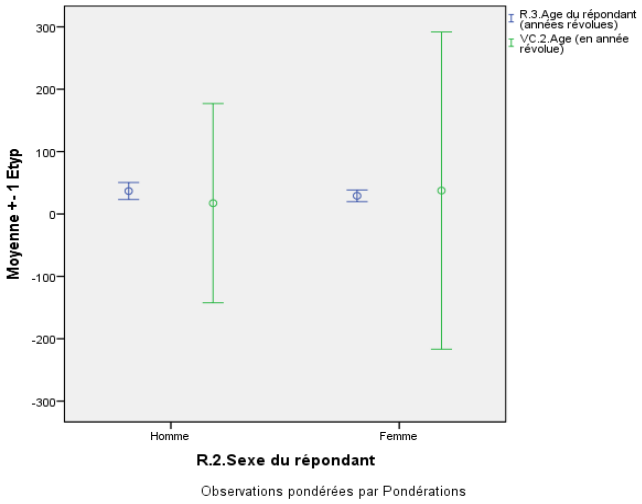
Pour réaliser un graphique de la moyenne et la variabilité d'une variable métrique dans différents sous-groupes, cliquer sur : **Graphes** > **Barre d'erreur** > choisir **« Simple »** et **« Récapitulatifs pour groupes d'observations »**. Dans l'onglet « *Variable* » choisir la variable pour laquelle l'on veut obtenir les moyennes et leur variabilité. Dans l'onglet « *Axe des modalités* », choisir la variable de groupement. Dans « *Les bâtons représenten* »', choisir « **Ecart type** » et « *Multiplicateur* », écrire « **1** » > **« ok »**.



Pour représenter la moyenne et la variabilité de plusieurs variables métriques dans tout l'échantillon, choisir **Graphes** > **Barre d'erreur** > choisir « **Simple** » et « **Récapitulatifs pour variables distinctes** ». Dans l'onglet « *Bâtons de variation* » entrer les variables pour lesquelles l'on veut obtenir les moyennes et leur variabilité. Dans « *Les bâtons représentent* », choisir « **Ecart type** » et « *Multiplicateur* », écrire « **1** » > **« ok »**.



Pour obtenir un graphique de la moyenne et la variabilité de plusieurs variables métriques dans différents sous-groupes, sélectionner : **Graphes** > **Barre d'erreur** > choisir **« Juxtaposé »** et **« Récapitulatifs pour variables distinctes »**. Dans l'onglet « *Variables* » entrer les variables pour lesquelles l'on veut obtenir les moyennes et leur variabilité. Dans « *Axe des modalités* », entrer la variable de groupement, par exemple « sexe du répondant ». Dans « *Les bâtons représentent* », choisir « **Ecart type** » et « *Multiplicateur* », écrire « **1** » > **« ok »**.



2.3.4.3 Boxplots / Boîte à moustaches

Le Boxplot est un moyen de représenter graphiquement d'aspects de la distribution d'une variable, comme la médiane ou la dispersion. La variable doit être au minimum ordinale.

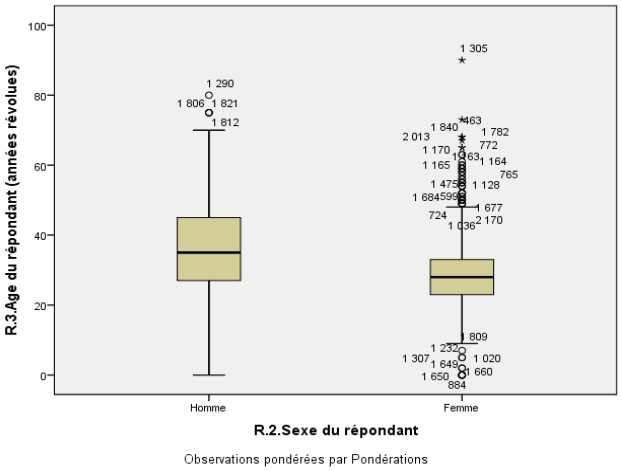
Boxplots pour une variable dans différents sous-groupes : sélectionner **Graphes** > **Boîtes de dialogue ancienne version** > B**oîte à moustaches** > choisir « **Simple** » et « Récapitulatifs pour groupes d'observations ». Dans l'onglet « *Variable* » glisser la variable pour laquelle l'on veut obtenir la médiane et la dispersion. Dans « *Axe des modalités* », glisser la variable de groupement. Dans « *Etiqueter les observations par* », si on a une variable d'identification des individus, on peut l'y entrer pour identifier les valeurs extrêmes (quand on laisse cette option vide, SPSS utilise le numéro de ligne).

Une alternative est la suivante : Analyse > Statistiques descriptives > Explorer > dans « *Liste Variable dépendantes* », entrer la variable pour laquelle l'on veut obtenir le boxplot. Dans « *Liste des facteurs* », entrer la variable de groupement.

Lecture du graphique

La hauteur de la boîte correspond à l'intervalle interquartile (IQ), le bord inférieur de la boîte représente le 1er quartile et le bord supérieur de la boîte représente le 3ème quartile.

Le trait traversant la boîte représente la médiane, le trait en dessous de la boîte relie le 1er quartile à l'extrême inférieur (c'est-à-dire l'observation égale ou juste supérieure à la valeur du 1erquartile moins 1.5 fois l'IQ).

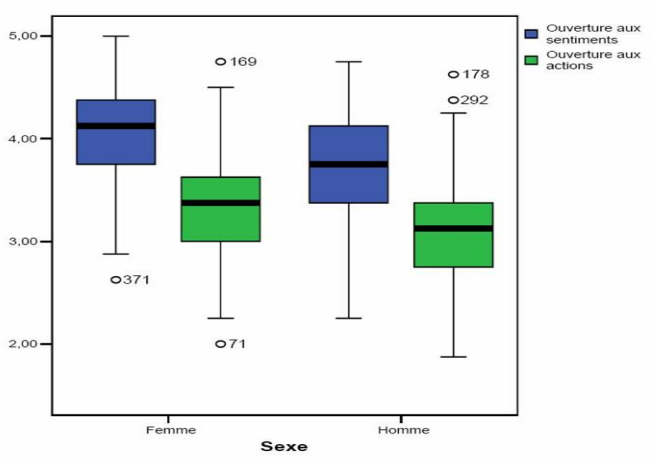


Le trait au-dessus de la boîte relie le 3èmequartile à l'extrême supérieur (c'est-à-dire l'observation égale ou juste inférieure à la valeur 3ème quartile plus 1.5 fois l'IQ).

Les points sont des valeurs extrêmes qui se trouvent entre 1,5 et 3 fois l'IQ en dessous ou au-dessus de la boîte. Les astérisques sont des valeurs extrêmes qui se trouvent plus de 3 fois l'IQ au-dessous ou au-dessus la boîte. Les chiffres à côté d'un point ou d'un astérisque sont le numéro des individus.

Pour les *Boxplots pour plusieurs variables dans tout l'échantillon,* cliquer sur **Graphes** > **Boîtes de dialogue ancienne version** > **Boîte à moustache** > choisir « **Simple** » et « **Récapitulatifs pour variables distinctes** ». Dans « *Les zones représentent »*, entrer les variables qui intéressent.

Pour les Boxplots pour plusieurs variables dans différents sous-groupes, cliquer sur **Graphes** > **Boîte à moustaches** > choisir « **Juxtaposé** » et « **Récapitulatifs pour variables distinctes** ». Dans *« Les zones représentent »*, glisser les variables qui intéressent. Dans « *Axe des modalités »*, glisser la variable de groupement.



2.3.4.4 Modifier la présentation des résultats (Output Labels)

Pour modifier la présentation des résultats, cliquer sur **Edition** > **Options**. La boîte de dialogue qui s'affiche permet de modifier la présentation des résultats. Dans la partie « Tableaux pivotants », on peut choisir entre : (1) les étiquette des variables, (2) les noms des variables ; (3) les noms des variables et leurs étiquette.

Sélectionner des cas

Pour sélections des cas en fonction des valeurs d'une variable (ou de plusieurs variables), il faut cliquer sur : **Données > Sélectionner des observations >** choisir **« Selon une condition logique »** et cliquer sur **« Si »** et définir la condition à l'aide de la variable, une valeur, ainsi que des

opérateurs relationnels et des opérateurs logiques. Cliquer sur **Poursuivre > « ok ».**

- Dans *la vue des données*, les cas non sélectionnés sont barrés dans la première colonne.
- Dans le coin en bas à droite, SPSS informe que la fonction est active (**Filtre actif**).
- Dans *l'affichage des variables*, une nouvelle variable est nommée « **filter_$** ». Les valeurs de cette variable filtre sont ***1=le cas est sélectionné*** ou ***0=le cas n'est pas sélectionné***.
- Une fois que cette fonction est en marche, seuls les cas sélectionnés (avec une valeur de 1 sur la variable filtre) seront utilisés dans les analyses qui suivront.
- Si on change la condition sous laquelle les cas sont sélectionnés, SPSS remplace la variable « *filter_$* ». Pour garder une variable filtre, on peut renommer cette variable « filter_$ ».

Attention

La fonction « **Sélectionner des observations** » reste active jusqu'à ce qu'on la désactive. Pour la désactiver, il faut passer par **Données** > **Sélectionner des observations** > **Cocher** « **Toutes les observations** » > « **ok** ».

2.3.4.5 Comparer les groupes

La fonction ***« Scinder un fichier »*** dans le menu ***« Données »*** permet de fragmenter un fichier et de comparer des cas en fonction des valeurs d'une variable (ou de plusieurs variables), c'est-à-dire que l'on peut analyser séparément des sous-groupes de l'échantillon afin de les comparer (comme par exemple, les femmes et les hommes). Pour le faire, il faut suivre la procédure suivante :

- Données > Scinder un fichier > Comparer les groupes.

L'option « Comparer les groupes » donne un tableau commun pour les sous-groupes alors que l'otion « Séparer les résultats par groupe » donne des tableaux séparés pour les sous-groupes.

Dans le coin en bas à droite, spss informe que la fonction est active (Diviser par…).

Attention

La fonction « Scinder un fichier » reste active jusqu'à ce qu'on la désactive. Pour la désactiver, il faut passer par **Données** > **Scinder un fichier** > **Analyser toutes les observations**, ne pas créer de groupes.

2.3.5 ANALYSE DES DONNÉES : CORRÉLATION ET RÉGRESSION

2.3.5.1 Corrélation de rangs

Pour calculer une corrélation de rang, cliquer sur : **Analyse** > **Corrélation** > **Bivariée**.

Dans l'onglet ***« Coefficients de corrélation »***, choisir le type résultat dont vous avez besoin en plus de **« Person » (**soit **« Tau-b de Kendall »** ou **« Spearman »**). Les variables qui sont entrées seront corrélées et on obtient une matrice complète, c'est-à-dire un tableau avec toutes les corrélations des variables deux à deux.

2.3.5.2 Scatterplot / Nuage de point

Le Scatterplot permet de présenter la relation entre deux variables métriques. Chaque point dans le graphique représente un cas, c'est-à-dire un couple de valeurs issue de deux variables. Pour obtenir un scatterplot, cliquer sur : **Graphes** > **Boîtes de dialogue ancienne version** > **Dispersion/Points…** > **Dispersion simple**.

Dans la boîte de dialogue *« Diagramme de dispersion simple »* qui s'affiche :

- Choisir les variables : **Axe des Y** pour l'axe des ordonnées et **Axe des X** pour l'axe des abscisses.
- ***Définir les marques par*** : on peut entrer un variable de groupement ; les différents sous-groupes seront affichés par des couleurs différentes.

Attention

Un point dans le Scatterplot peut représenter plusieurs cas. La couleur affichée correspond dans cette situation au sous-groupe du premier cas. L'utilisation de cette option n'est donc pas recommandée.

Etiqueter les observations par : on peut entrer la variable ID (numéro identifiant des observations) pour obtenir le numéro d'un cas individuel dans l'éditeur de graphique. Quand on n'entre pas de variable dans cette section, SPSS prend la numérotation des lignes (qui ne correspond pas forcément à la variable ID).

Pour obtenir le numéro d'un cas individuel dans l'éditeur de graphique :

- Double-cliquer sur le graphique pour activer l'éditeur ;
- Cliquer sur le symbole ⊞ ;
- Cliquer sur le point dont l'on veut obtenir le numéro ;
- Cliquer encore une fois au-dessus pour effacer le numéro.

2.3.5.3 Corrélation de Bravais-Pearson

Pour calculer une corrélation pour des variables numériques, cliquer sur : **Analyse** > **Corrélation** > **Bivariée**, le *coefficient de corrélation de Pearson* est le coefficient par défaut. Les variables qui sont entrées seront corrélées et on obtient une matrice complète.

- Pour obtenir la moyenne et l'écart-type pour chaque variable, ainsi que les produite croisés et les covariances, cliquer sur « **Options** », cocher les options dont vous avez besoin dans l'onglet ***« Statistiques »***.
- Si vous choisissez **« Exclure toutes les observations incomplète »**, pour chaque coefficient de corrélation, les cas qui n'ont pas de valeurs valables sur le couple de variables seront exclus.
- Si vous choisissez « **Exclure seulement les composantes non valides** », pour chaque coefficient de corrélation, seulement les cas qui ont des valeurs sur toutes les variables seront analysés (même N pour chaque coefficient).

Tout ce qui concerne la significativité n'a pas été traité dans ce module. Mais en général, une très bonne corrélation correspond à au moins 0,80 ; une bonne corrélation à 0,50 et une faible corrélation à 0,20.

Une alternative est de passer par la syntaxe. Parfois, on veut obtenir des corrélations entre deux groupes de variables. Dans ce cas, le tableau standard (obtenu par le menu « **Corrélation** ») devient très grand et il y aura des corrélations dont on n'a pas besoin. Pour obtenir seulement les corrélations entre les deux groupes de variables, il est préférable d'utiliser la syntaxe :

- Dans la boîte de dialogue, entrer toutes les variables et cliquez sur « **Coller** » (au lieu de OK) ;
- Dans la fenêtre de syntaxe, il faut ajouter le mot '**WITH**' entre les deux groupes de variables. Par exemple :

```
CORRELATIONS
  /VARIABLES=Q09Tailleduménage Q10Nombretotaldefemmesenâgedepro
Q12Nombretotaldesenfantsde1223mo
  WITH Q11Nombredefemmesquiontaccouchéa Q16Ageduchefdeménage
Acombiendetempsremontevotrederni
  /PRINT=TWOTAIL NOSIG
/STATISTICS XPROD
  /MISSING=PAIRWISE.
```

Corrélations

		Q11. Nombre de femmes qui ont accouché au cours des 12 derniers mois	Q16. Quel est l'âge du chef de ménage?	Q9_1. A combien de temps remontre votre dernière grossesse?
Q09. Quel est la taille du ménage?	Corrélation de Pearson	,338	,362	,002
	Sig. (bilatérale)	0,000	0,000	,833
	Somme des carrés et produits croisés	4417,421	228106,615	519,353
	Covariance	,340	17,544	,040
	N	13003	13003	13003
Q10. Nombre de femmes en âge de procréer dans le ménage	Corrélation de Pearson	,387	,299	-,007
	Sig. (bilatérale)	0,000	,000	,408
	Somme des carrés et produits croisés	983,782	36522,706	-397,064
	Covariance	,076	2,809	-,031
	N	13003	13003	13003
Q12. Nombre total d'enfants âgés de 12 à 23 mois	Corrélation de Pearson	,150	,060	,057
	Sig. (bilatérale)	,000	,000	,000
	Somme des carrés et produits croisés	138,079	2662,758	1136,843
	Covariance	,011	,205	,087
	N	13003	13003	13003

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

2.3.5.4 Régression linéaire simple

Pour effectuer une régression linéaire simple, cliquer sur **Analyse** > **Régression** > **Linéaire**.

- « Dépendant » : entrer la variable Y (variable expliquée) ;

- « Variables indépendantes » : entrer les variables explicatives ;
- « Statistiques » : on peut cocher les paramètres descriptifs que l'on veut obtenir (variation de R-deux, intervalles de confiance, …) ;
- Pour la régression simple, ne pas cocher d'option.

Syntaxe

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /DEPENDENT D17a
  /NOORIGIN
  /CRITERIA=PIN(.05) POUT(.10)
  /METHOD=ENTER D1 A9Nombretotaldepersonnesdansleménage.
```

SPSS-Viewer (Résultats)

« Model summary » :

- **R** : correspond à la corrélation xy dans le cas de la régression simple ;
- **R-deux (R2)** : coefficient de détermination ou pourcentage de ka variation de Y expliqué par la régression.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,367[a]	,135	,117	79840,1621

a. Valeurs prédites : (constantes), A9. Nombre total de personnes dans le ménage, D1_Avant_Le ménage pratiquaie membre dans une AVEC ?

Tableau « ANOVA »: sommes des carrés de la régression

ANOVAa

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	95370511555,851	2	47685255777,926	7,481	,001[b]
	Résidu	611947341979,503	96	6374451478,953		
	Total	707317853535,354	98			

a. Variable dépendante : D1.7a.la production de_Banane plantain

b. Valeurs prédites : (constantes), A9. Nombre total de personnes dans le ménage, D1_Avant_Le ménage pratiquaient-il la culture du cacao avant l'entrée d'un membre dans une AVEC ?

Tableau « Coefficients »

- ✓ Ligne « (Constant) » et colonne « A » : l'ordonnée à l'origine de la droite de régression ;
- ✓ Ligne de la variable et colonne « A » (valeur non-standardisée) : pente de la droite de régression ;
- ✓ Colonne « Bêta » (valeur standardisée) : correspond à la corrélation xy dans le cas de la régression simple ;
- ✓ « Erreur standard » et les valeurs t et sig. concernent la significativité des coefficients (ne sera pas traité ici). En général, au niveau de significativité, on accepte la valeur supérieure ou égale à **0.05**.

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	174677,957	31067,562		5,623	,000
	D1_Avant_Le ménage pratiquaient-il la culture du cacao avant l'entrée d'un membre dans une AVEC ?	-107886,904	29696,452	-,348	-3,633	,000
	A9. Nombre total de personnes dans le ménage	-1986,901	2359,346	-,081	-,842	,402

a. Variable dépendante : D1.7a.la production de_Banane plantain

2.4 EXERCICES

Parmi les propositions suivantes, choisir celle(s) qui est (sont) exactes(s).

QCM1. Quelle est l'extension que porte le fichier de données ?

A. .sav
B. .spo
C. .sps
D. .txt

QCM2. Où les données sont-elles affichées ?

A. Variable view
B. File
C. Output
D. Data view

QCM3. Dans variable view, à quoi chaque ligne correspond-elle ?

A. Un item
B. Un chiffre
C. Une variable
D. Un menu

QCM3. Comment coder les variables alphanumériques/série de caractère ?

A. Créer une variable pour chaque catégorie
B. Regrouper l'information en catégorie grâce à l'analyse de contenu
C. Entrer les caractères et définir la variable comme chaîne de caractère (string)
D. Ne rien faire

QCM4. Comment coder les réponses (ouvertes) courtes ?

A. Ne rien faire
B. Créer une variable pour chaque catégorie
C. Créer une variable alphanumérique
D. Coder les réponses avec des valeurs numériques

QCM5. Comment réduire les erreurs en entrant les données ?

A. Ne recoder jamais les items à la main
B. Utiliser des valeurs positives et négatives
C. Recoder les items à la main
D. Pas besoin de définir le format de réponse au préalable

QCM6. Comment peut-on fusionner des fichiers en ajoutant des variables ?

A. Avec Données > Scinder un fichier …
B. Avec Données > Sélectionner des observations …
C. Avec Données > Fusionner des fichiers > Ajouter des variables
D. Avec Données > Fusionner des fichiers > Ajouter des observations

QCM7. Comment peut-on transformer des variables ?

A. Par le menu Transformer > Recoder des variables > Dans une même variable
B. Par le menu Transformer > Recoder des variables > Dans une variable différente
C. Par le menu Transformer > Calculer
D. Par la syntaxe

QCM8. Comment peut-on obtenir un tableau de fréquences ?

A. Par le menu Analyse > Statistiques descriptives > Description
B. Par le menu Analyse > Tables > Table d'effectifs …
C. Par le menu Analyse > Statistiques descriptives > Effectifs
D. Par le menu Analyse > Tables > Tables basic

QCM9. Pourquoi fait-on un histogramme ?

A. Pour présenter la moyenne de l'échantillon
B. Pour présenter l'écart-type de la distribution de l'échantillon
C. Pour présenter la distribution des fréquences d'une variable métrique dans tout l'échantillon
D. Pour présenter la distribution des fréquences d'une variable nominale dans tout l'échantillon

QCM10. Qu'est-ce que la barre d'erreur ?

A. La barre d'erreur permet de présenter la fréquence de la distribution
B. La barre d'erreur permet de présenter la taille de la distribution
C. La barre d'erreur permet de présenter la moyenne et la variabilité des variables métriques
D. La barre d'erreur permet de présenter la médiane et l'écart-type

QCM11. Comment appel-t-on le graphique qui représente les différents aspects de la distribution d'une variable (tendance centrale et dispersion) ?

A. Nuage de point/Scatterplot
B. Boxplot
C. Histogramme
D. Camembert

QCM1 (A) - QCM2 (D) - QCM3 (C) - QCM4 (D) - QCM5 (A) - QCM6 (C) - QCM7 (B) - QCM8 (C) - QCM9 (C) - QCM10 (C) - QCM11 (B).

3 INFÉRENCE STATISTIQUE ET THÉORIE DES TESTS D'HYPOTHÈSES

3.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin de ce chapitre, les participants seront capables de faire la description des variables sous SPSS, de décrire la population et l'échantillon d'une étude. Ils pourront également réaliser tous les tests possibles selon le type d'étude. Par ailleurs, les participants sauront faire la différence entre les différents types de risque d'erreur.

3.2 ASPECT THÉORIQUE

3.2.1 RAPPEL SUR LA DESCRIPTION D'UNE VARIABLE

3.2.1.1 Tableau statistique simple

Un tableau statistique est un ensemble de cases rangées en lignes et en colonnes, contenant des données chiffres. Pour transformer ces données en arguments, il faut d'abord les traiter, puis analyser le tableau en sélectionnant les informations les plus significatives. Par ailleurs, le tableau de fréquence donne le nombre de fois qu'une modalité est observée. Les modalités peuvent être des caractères, des nombres, des intervalles.

Tableau 7 : Exemple de tableau statistique simple

Modalité ou classe	Effectif (fréquence absolue)
X1	N1
X2	N2
…	…
Xi	Ni
…	…
Xk	Nk
Total	N

3.2.1.2 Description d'une variable

Une variable est considérée comme quantitative ou métrique lorsque ses modalités peuvent être mesurées (par exemple : l'âge, la valeur d'une action, etc.).

- La description d'une variable qualitative consiste à présenter les effectifs, c'est-à-dire le nombre d'individus de l'échantillon pour chaque modalité de la variable et les fréquences, c'est-à-dire la proportion des réponses associées à chaque modalité de la variable étudiée (fréquence absolue ou fréquence relative). *Dans le langage des études de marchés, on parle tri à plat.*

Il existe plusieurs possibilités dans SPSS pour décrire les données collectées. Il est possible par exemple dans un premier temps, de générer un rapport sur les observations pour s'assurer qu'elles ne comportent pas d'erreurs de saisie, de valeurs aberrantes (**Analyse** > **Rapport** > **Récapitulatif des observations)** ou pour simplement prendre connaissance des variables dans un tableau synthétique (utile en début d'analyse) (**Utilitaires** > **Variables** > …). La procédure **Fréquence** permet d'obtenir les affichages statistiques et graphiques qui servent à décrire des variables quantitatives et qualitatives.

Plusieurs indicateurs permettent de décrire une variable quantitative : les indicateurs de tendance centrale (moyenne, médiane, mode) ; les indicateurs de dispersion (étendue,variance, écart-type, coefficient de variation) ; les indicateurs de forme de la distribution (asymétrie, aplatissement) ; les représentations graphiques (histogramme ou boîtes à moustaches….

3.2.1.3 Fréquence

En statistique, on appelle ***fréquence absolue***, l'effectif des observations d'une classe et ***fréquence relative ou simplement fréquence***, le quotient de cet effectif par celui de la population. Si la valeur est un nombre compris entre 0 et 1 ; ou un pourcentage, il s'agit de la fréquence relative.

3.2.1.4 Mode

En statistique, le mode ou valeur dominante est la valeur la plus représentée d'une variable quelconque dans une population donnée. Si plusieurs valeurs à la fois présentent la plus grande fréquence d'apparition, chacune d'entre elles est un mode ; on dit que la distribution est ***plurimodale***.

3.2.1.5 Moyenne

La moyenne d'une série statistique est égale au quotient de la somme de toutes les valeurs de cette série par l'effectif total. La moyenne révèle la tendance centrale, dans le sens où les réponses se trouvent de part et d'autre de la moyenne. La moyenne est la statistique utilisée en premier dès qu'on a des variables quantitatives. Mais elle est sensible aux valeurs extrêmes ou atypiques, et ce, d'autant plus que le nombre d'observations est petit.

3.2.1.6 Variance

En statistique et en théorie des probabilités, la variance est une mesure de la dispersion des valeurs d'un échantillon ou d'une distribution de probabilité. La variance est la mesure de la dispersion autour de la moyenne. Lorsque les données se concentrent autour de la moyenne, la variance est faible. Si les données sont dispersées autour de la moyenne, la variance est élevée. Il s'agit d'une mesure plus fine de la dispersion car toutes les données sont prises en compte. En revanche, elle est comme la moyenne, sensible aux valeurs extrêmes.

3.2.1.7 Ecart-type

En mathématiques, l'écart-type est une mesure de la dispersion des valeurs d'un échantillon statistique ou d'une distribution de probabilité. Il est défini comme la racine carrée de la variance. L'écart-type est la mesure de la dispersion autour de la moyenne, exprimée dans la même unité que la variable.

3.3 CONCEPT DE POPULATION ET ÉCHANTILLON

Le but principal de la statistique est de déterminer les caractéristiques d'une ***population*** données à partir de l'étude d'une partie de cette population, appelée ***échantillon***. Lorsque l'on travaille à partir d'un échantillon, la statistique descriptive rend simplement compte des observations faites à partir de cet échantillon. Aussi, on projette les résultats sur l'ensemble de la population lorsque les données d'un échantillon permettent de généraliser sur l'ensemble de la population. Il faut donc prendre les moyens pour s'assurer que l'échantillon est représentatif avant de faire cette opération de généralisation. Cependant, les statistiques n'auront de sens que si les données utilisées sont crédibles. Pour cela, il faut que le nombre d'observations soit suffisant et fiable.

3.3.1 POPULATION

La population est un ensemble d'objets ou d'individus ayant des caractéristiques qui leurs sont propres. Les caractéristiques de la population sont inconnues. Pour l'étude d'une caractéristique d'une population, chaque échantillon représentatif donnera une valeur différente, comme par exemple, la population masculine Nigérienne.

3.3.2 ECHANTILLON

Un échantillon est un sous-ensemble d'une population. Les caractéristiques de l'échantillon sont connues. On peut avoir plusieurs échantillons d'une même population et pour chaque échantillon des observations spécifiques. Le bon échantillon est celui qui est représentatif de la population et le meilleur moyen d'y parvenir est l'échantillonnage.

La représentativité d'un échantillon n'est toujours que partiellement vérifiable. Un échantillon peut être représentatif suivant une, deux, trois variables ou plus, mais jamais totalement identique à la population totale.

L'analyse des observations de chaque échantillon pour une même caractéristique permet d'avoir la marge d'erreur, on parle dans ce cas de **fluctuation d'échantillonnage**.

3.3.3 INFÉRENCE STATISTIQUE

La statistique descriptive se distingue de l'inférence statistique qui vise, elle, à extrapoler sur la population entière les résultats d'une enquête portant sur un échantillon. Par exemple, lorsqu'il devient impossible d'interroger toutes les personnes qui composent une classe, nous utiliserons une fraction des personnes présentes. Le résultat obtenu sera par la suite utilisé pour représenter tous les membres. L'échantillon est la fraction et c'est elle qui sert donc à décrire l'ensemble, la population (toute la classe).

Pour un gain de temps et d'argent, il est recommandé de conduire les études à partir d'un échantillon au lieu de prendre en compte toute la population. Toutefois, il faut tirer des conclusions comme si on avait étudié toute la population : c'est ce qu'on appelle l'inférence statistique.

La question principale : dans quelle mesure ce qu'on a obtenu dans l'échantillon est différent de ce qu'on aurait eu si on étudiait toute la population ? Par exemple, à partir d'un échantillon d'enfants, on constate que le taux de malnutrition des enfants est plus élevé que le taux de malnutrition des enfants de Maradi.

3.4 TESTS STATISTIQUES

Il existe de très nombreux tests qui permettent d'évaluer des aspects différents de significativité. Les objectifs principaux auxquels peuvent répondre les tests statistiques sont :

- L'évaluation de la représentativité des répartitions observées par rapport aux valeurs connues pour l'ensemble de la population ;
- La mesure de la significativité de la différence constatée sur les observations de deux groupes d'individus ou d'un même groupe pour deux variables observées ;
- L'existence et l'intensité d'une liaison entre deux variables.

3.4.1 FONCTIONNEMENT DES TESTS STATISTIQUES

Les tests statistiques fonctionnent tous sur le même principe qui consiste à énoncer une hypothèse sur la « population mère », puis à vérifier, sur les observations constatées, si celles-ci

sont vraisemblables dans le cadre de cette hypothèse.

Autrement dit, on cherche à estimer la probabilité de tirage au sort dans la « population mère », d'un échantillon ayant les caractéristiques observées. Si cette probabilité est minime, on rejette l'hypothèse énoncé. Dans le cas contraire, celle-ci peut être adoptée, au moins provisoirement, dans l'attente de validations complémentaires.

L'hypothèse à tester est appelée H0 ou hypothèse nulle. Cette hypothèse s'accompagne impérativement de son hypothèse alternative appelée H1. Ainsi, le test s'attachera à valider ou à rejeter H0.

Un test d'hypothèse est un procédé d'inférence permettant de contrôler (accepter ou rejeter) à partir de l'étude d'un ou plusieurs échantillons aléatoires, la validation d'hypothèses relatives à une ou plusieurs populations.

Les méthodes de l'inférence statistique nous permettent donc de déterminer, avec une probabilité donnée, si les différences constatées au niveau des échantillons peuvent être imputables au hasard ou si elles sont suffisamment importantes pour signifier que les échantillons proviennent de populations vraisemblablement différentes.

3.4.2 PRINCIPE DES TESTS STATISTIQUES

On distingue deux classes de tests, à savoir :

- **Les tests paramétriques** requièrent un modèle à fortes contraintes (normalité des distributions ou approximation normale pour des grands échantillons). Ces hypothèses sont d'autant plus difficiles à vérifier que les effectifs étudiées sont plus réduits.
- **Les tests non paramétriques** sont des tests dont le modèle ne précise pas les conditions que doivent remplir les paramètres de la population dont a été extrait l'échantillon. Il n'y a pas d'hypothèse de normalité au préalable.

Les tests paramétriques, quand leurs conditions sont remplies, sont plus puissants que les tests non paramétriques. Les tests non paramétriques s'emploient lorsque les conditions d'applications des autres méthodes ne sont pas satisfaites, même après d'éventuelles transformations de variables. Ils peuvent s'utiliser même pour des échantillons de taille très faible.

3.4.3 DIFFÉRENTS TESTS STATISTIQUES

3.4.3.1 Test de conformité

Le test de conformité consiste à confronter un paramètre calculé sur l'échantillon à une valeur préétablie. Les plus connus sont certainement les tests portant sur la moyenne, la variance ou sur les proportions. On connaît la loi théorique en général la loi normale.

3.4.3.2 Test d'ajustement ou d'adéquation

Le test d'ajustement ou d'adéquation consiste à vérifier la compatibilité des données avec une distribution choisie a priori. Le test le plus utilisé dans cette optique est le test d'ajustement à la loi normale, qui permet ensuite d'appliquer un test paramétrique.

3.4.3.3 Test d'homogénéité ou de comparaison

Le test d'homogénéité ou de comparaison revient à la question « Y a-t-il une différence entre la prévalence de la diarrhée des enfants malnutris pris en charge dans un centre de réhabilitation nutritionnelle et ceux qui sont pris en à domicile ? ».

3.4.3.4 Test d'indépendance ou d'association

Le test d'indépendance ou d'association consiste à éprouver l'existence d'une liaison entre 2 variables : « Est-ce que le taux de malnutrition est indépendant de la région au Niger ? ».

3.4.4 PRINCIPE DES TESTS D'HYPOTHÈSE

Le principe des tests d'hypothèse est de poser une hypothèse de travail et de prédire les conséquences de cette hypothèse pour la population ou l'échantillon.

- On compare ces prédictions avec les observations et l'on conclut en acceptant ou en rejetant l'hypothèse de travail à partir de règles de décisions objectives.
- Définir les hypothèses de travail, constitue un élément essentiel des tests d'hypothèses de même que vérifier les conditions d'application de ces dernières.

3.4.5 ETAPES DES TESTS D'HYPOTHÈSE

- Définir l'hypothèse nulle, notée H0, à contrôler ;
- Choisir une statistique pour contrôler H0 ;
- Définir la distribution de la statistique sous l'hypothèse « H0 est réalisée » ;
- Définir le niveau de signification du test α et la région critique associée ;
- Calculer à partir des données fournies par l'échantillon, la valeur de la statistique ;
- Prendre une décision concernant l'hypothèse posée.

3.4.6 TYPES D'HYPOTHÈSES

Pour réaliser un test statistique, deux hypothèses sont posées à savoir H0 et H1.

- L'hypothèse nulle notée H0 est l'hypothèse que l'on désir contrôler : elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative ; et est due aux fluctuations d'échantillonnage.
- Cette hypothèse H0 est formulée dans le but d'être rejetée ;
- L'hypothèse alternative notée H1 est la « négation » de H0, elle est équivalente à dire « H0 est fausse ».
- Selon la formulation de H1, on peut distinguer si le test est bilatéral (les deux sens) ou unilatéral (un sens).

3.4.7 RISQUES D'ERREUR

Dans la réalisation des tests d'hypothèse, il existe deux types de risque d'erreurs ; à savoir le risque d'erreur de première espèce et le risque d'erreur de seconde espèce.

- **Risque d'erreur de première espèce** : on appelle risque d'erreur de première espèce, la probabilité de rejeter H0 et d'accepter H1 alors que H0 est vraie.
- La valeur du risque **α** doit être fixée à priori par l'expérimentateur et jamais en fonction des données. C'est un compromis entre le risque de conclure à tort et la faculté de conclure.
- **Risque d'erreur de seconde espèce :** on appelle risque d'erreur de seconde espèce, notée **β**, la probabilité de rejeter H1 et d'accepter H0 alors que H1 est vraie. Pour quantifier le risque **β**, il faut connaitre la loi de probabilité de la statistique sous l'hypothèse H1.

3.4.8 PUISSANCE DU TEST

- Rappelons que les tests ne sont pas faits pour « démonter » H0, mais pour « rejeter » H0.
- L'aptitude d'un test à rejeter H0 alors qu'elle est fausse constitue la puissance du test.
- On appelle ***puissance d'un test***, la probabilité de rejeter H0 et d'accepter H1 alors que H1 est vraie ; sa valeur est **1–β.**
- La puissance d'un test est fonction de la nature de H1. Un test unilatéral est plus puissant qu'un test bilatéral. Elle augmente avec la taille de l'échantillon N étudié et diminue lorsque **α** diminue.
- La robustesse d'une technique statistique représente sa sensibilité à des écarts aux hypothèses faites.

Tableau 8 : Test et hypothèses

		En réalité	
		H0 est vraie	**H1 est vraie**
Décision	**H0 accepté**	Correct 1-α	Manque de puissance risque de seconde espèce β
	H1 accepté	Rejet à tort au risque α	Puissance du test 1- β

3.5 EXERCICES

Dans une commune de Niamey, (noté A2), on tire au sort un échantillon de 200 enfants. On observe que 40 d'entre eux sont atteints malnutrition. On se demande si le taux de malnutrition infantile dans cette commune diffère de 10 %, le taux dans la commune étudié en première partie (noté A1).

QCM1. Pour répondre à cette question, on envisage de faire un test statistique.

A. Cela n'est pas vraiment nécessaire, car 0,2≠0,1
B. On choisit un test de comparaison de 2 proportions observées
C. On choisit un test de comparaison d'une proportion observée à une proportion théorique
D. Pour pouvoir effectuer le test en utilisant la loi normale, il faut que le nombre d'individus soit au moins de 30.

QCM2. L'hypothèse nulle est :

A. 0,2=0,1.
B. Les probabilités de malnutrition sont les mêmes dans les 2 communes.
C. Les proportions de malnutrition ne sont pas les mêmes dans les 2 communes.
D. Les proportions observées de malnutrition sont les mêmes dans les 2 communes.

QCM3. Le risque de première espèce du test :

A. C'est la probabilité de rejeter l'hypothèse nulle.
B. C'est la probabilité de rejeter à tort l'hypothèse nulle.
C. C'est la probabilité de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse.
D. Le test serait plus puissant si on diminuait ce risque.

QCM4. Lors d'un test d'hypothèse, combien y-a-t-il de sortes d'erreurs ?

A. 1
B. 4
C. 0
D. 2

QCM5. Parmi les tests suivants, lequel est un test de conformité ?

A. Comparaison d'une moyenne observée et d'une moyenne théorique.
B. Comparaison de 2 moyennes observées.
C. Comparaison de 2 variances observées.
D. Comparaison d'une moyenne et d'une variance.

QCM6. L'hypothèse nulle H0 est appelée ainsi car :

A. C'est une hypothèse irréalisable.
B. Elle est toujours vérifiée.
C. On part du fait que les différences entre les valeurs observées et les valeurs théoriques sont uniquement dues aux fluctuations d'échantillonnage.
D. L'hypothèse est égale à 0.

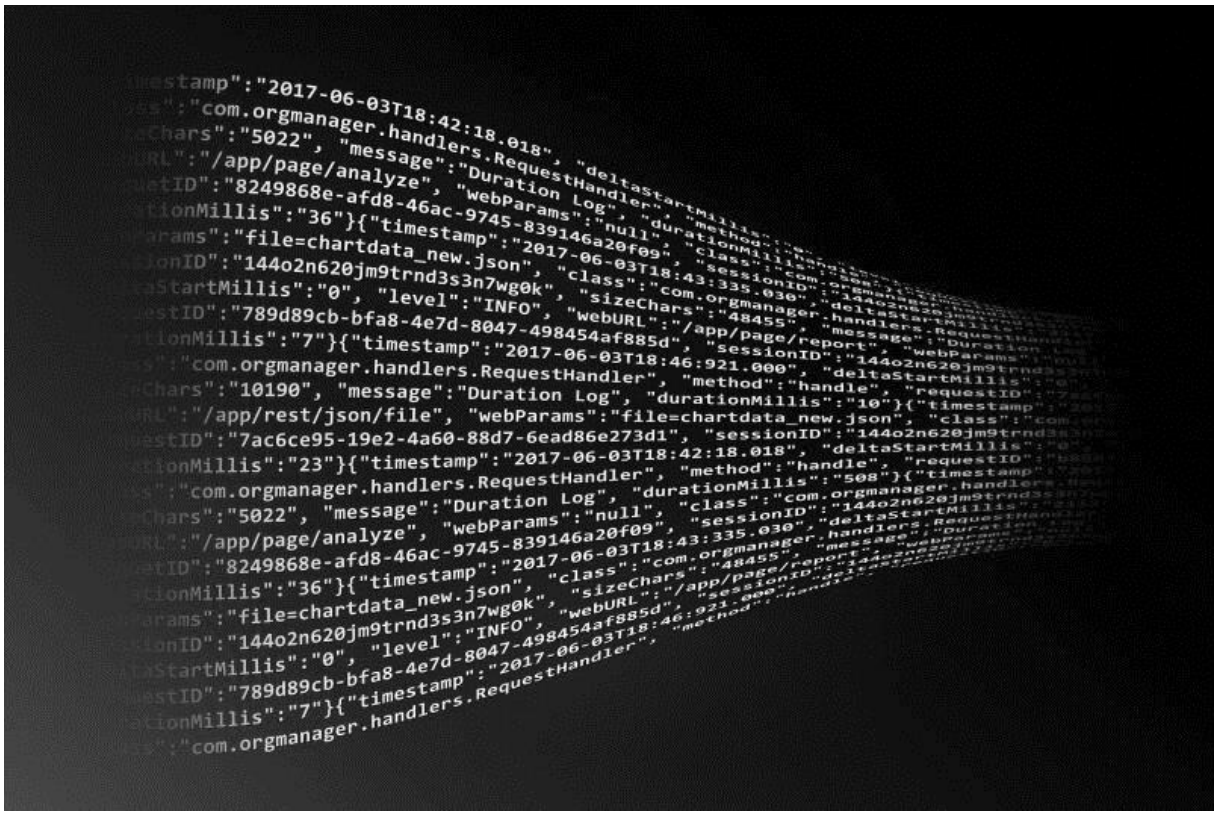
QCM7. Si l'on pose les 2 hypothèses H0 : « μ=10 » et H1 : « μ<20 » alors le test est :

A. Bilatéral.
B. Unilatéral à gauche.
C. Unilatéral à droit.
D. Mal posé

QCM8. Si l'on pose les 2 hypothèses H0 : « μ=50 » et H1 : « μ>50 » alors le test est :

A. Bilatéral.
B. Unilatéral à gauche.
C. Unilatéral à droit.
D. Mal posé

QCM1 (C) - QCM2 (B) - QCM3 (B) - QCM4 (D) - QCM5 (A) - QCM6 (C) - QCM7 (D) - QCM8 (C)



4 ANALYSE BIVARIÉE / ANOVA

4.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin de ce chapitre, les participants seront capables de formuler la ou les question(s) de recherche pour un test ANOVA. Ils seront également à même de mettre en œuvre de façon pratique le test ANOVA sous SPSS et d'interpréter les résultats donnés par le logiciel.

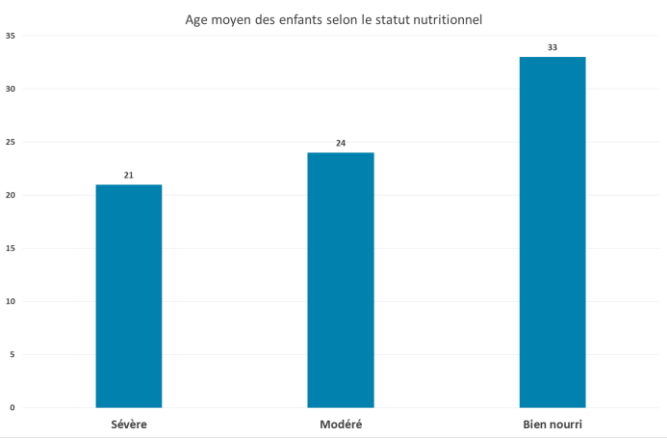
4.2 ASPECT THÉORIQUE

4.2.1 PRINCIPE

Lorsqu'on confronte une variable quantitative à une variable qualitative (nominale ou ordinale), on a recourt très généralement à la comparaison de moyenne ou à l'analyse de variance (ANOVA). Les logiciels d'analyse statique offrent diverses alternatives pour étudier la relation entre une variable quantitative et une variable qualitative. Ces méthodes ont toutes en commun le fait qu'elles font une comparaison des moyennes résultant de la répartition des données de la variable quantitative selon les catégories que comporte la variable qualitative.

L'analyse de variance permet de confronter les données d'une variable aux données d'une variable qualitative comportant deux catégories ou plus. On se demande par exemple, dans quelle mesure l'âge des enfants (variable quantitative) est associé aux différents niveaux (observés) du statut nutritionnel (variable qualitative ordinale).

Dans certaines conditions, on peut généraliser la conclusion relative à une relation entre l'âge et le statut nutritionnel.



4.2.2 QUESTIONS DE RECHERCHE

- Les différences de moyennes qui permettent de visualiser le graphique sont-elles dues aux particularités aléatoires de l'échantillon ou reflètent-elles aussi des différences réelles dans les trois populations correspondant aux trois niveaux de statut nutritionnel ?
- Y-a-t-il un lien entre l'âge et le statut nutritionnel des enfants ?

4.2.3 MISE EN ŒUVRE DU TEST ANOVA

4.2.3.1 Raisonnement intuitif

Le principe du raisonnement est le suivant :

- Plus les différences entres les moyenne dans l'échantillon sont importante, plus il est difficile d'admettre que ces différences résultent simplement du hasard ; et plus on est porté à admettre qu'il existe des différences entre les moyennes de populations (correspondant aux différents niveaux du statut nutritionnel par exemple).
- Par ailleurs, on sera plus confiant sur ce type de conclusion si la variation autour des moyennes observées est petite.

4.2.3.2 Prémisses du test d'analyse de variance

Avant de procéder à l'analyse proprement dite, il faut s'assurer de respecter certaines prémisses.

- **Les groupes sont indépendants et tirés au hasard de leur population respective.** Ceci signifie qu'il n'y a ni relation entre les observations à l'intérieur d'un groupe, ni relation entre les observations entre les groupes. Par exemple, si on propose 4 traitements aux individus, il existe forcément une relation entre les observations et on ne pourra pas utiliser l'ANOVA dans ce contexte.
- **Les valeurs des populations sont normalement distribuées.** L'ANOVA n'est pas très sensible aux écarts de la normalité. Il est donc possible de procéder sans avoir une normalité parfaite. Par contre, avec un petit échantillon, il faut faire attention à l'impact des valeurs extrêmes (on peut faire le test avec et sans les valeurs extrêmes).
- **Les variances des populations sont égales.** Cette prémisse peut être vérifiée par l'examen visuel du graphique « ***boîte à moustaches »*** ou encore par le ***test de Levene,*** qui est disponible dans les options de l'ANOVA.
    - Si les groupes sont de tailles identiques, on peut passer outre cette prémisse ;
    - Si la taille des groupes est très inégale, la prémisse d'égalité des variances doit être vérifiée systématiquement. Si le test est significatif, il est possible d'utiliser d'autres procédures disponibles dans le menu ANOVA : ***test Brown-Forsythe ou le Welch Robust F***.

4.2.3.3 Formulation des hypothèses d'une ANOVA

- **Hypothèse nulle** : Les groupes proviennent de la même population ; leurs moyennes sont semblables H_0: $\bar{X}_1 = \bar{X}_2 = \bar{X}_3 = \cdots \bar{X}_n$
- **Hypothèse alternative** : il y a une différence entre les moyennes, c'est-à-dire qu'au moins une des moyennes est différentes des autres. H_1: $\bar{X}_1 \neq \bar{X}_2 \neq \bar{X}_3 \neq \cdots \bar{X}_n$

4.2.3.4 Statistique Fisher (F) ou distribution F

L'ANOVA utilise le mécanisme du **F de Fisher**, non pas pour comparer deux variances d'échantillons, mais bien les deux composantes d'une même variance. Le test ne « fonctionnera » que si les moyennes sont les mêmes dans tous les échantillons.

La statistique **F** produite par l'ANOVA est le rapport entre la variabilité inter et intra-groupes. Elle permet de déterminer s'il existe une différence significative entre les groupes. Comme la variabilité inter-groupe est le numérateur de ce rapport, plus les moyennes sont éloignées les unes des autres, plus la valeur **F** est élevée. Par la suite, il faut comparer la valeur **F** obtenue à la

distribution **F**. le degré de signification va dépendre de trois facteurs : la valeur **F** et les 2 degrés de liberté (inter et intra-groupes).

4.2.3.5 Démarche générale

- Identifier le facteur (la variable indépendante) et on sélectionne les données ;
- Faire le test de normalité ;
- Faire le test d'homogénéité des variances (homoscédasticité) ;
- Poser les hypothèses. H0 : les moyennes sont égales ; H1 : au moins une moyenne est différente des autres ;
- Calculer la statistique F de Fisher / ou utiliser la procédure de son logiciel favori (nous utilisons dans ce module le logiciel SPSS) ;
- Examiner ce que restitue le logiciel, c'est-à-dire le tableau de l'analyse de la variance puis, interpréter. **H0 est rejetée si F est supérieur à sa valeur critique**.

4.2.4 COMPARAISONS MULTIPLES

Le test d'analyse de variance ne nous dit qu'une chose : l'hypothèse nulle est rejetée ou non. Il ne nous dit pas où se situe la ou les différences. Il faut donc effectuer d'autres tests pour savoir entre quels groupes se trouvent cette ou ces différences.

Ces tests sont appelés **post-hoc** ou tests a posteriori. Ils indiquent quels groupes se distinguent

4.3 SYNTAXES SPSS

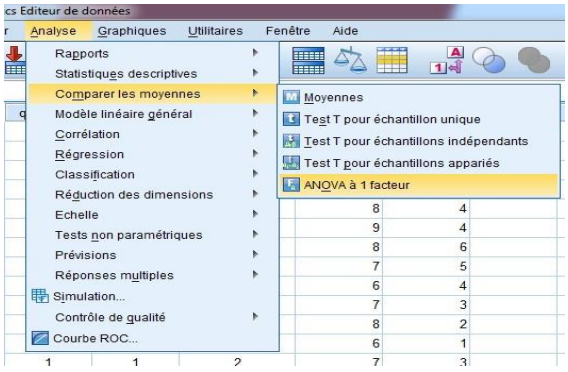
4.3.1 FORMALISATION DU PROCESSUS SOUS SPSS

Pour les fins de cet exemple, notez qu'à la droite de la matrice une nouvelle variable nominales a été créée : Formation (1=science ; 2=technique ; 3=autres). Chacune des trois niveaux de cette variable compte 15 participants pour un total n = 45.

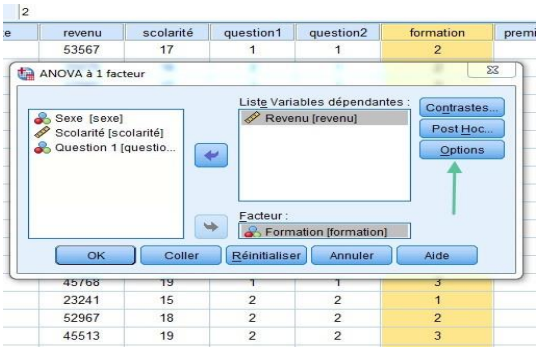
Etape 1 : Ouvrir votre matrice de données sous SPSS

sexe	revenu	scolarite	question1	formation
1	56784,00	17	1	2
2	34342,00	16	2	1
1	67564,00	19	2	3
2	23456,00	17	1	2
1	56453,00	18	2	3
2	45634,00	17	2	2

Etape 2 : Choisir le menu Analyse > Comparer les moyennes > ANOVA à 1 facteur



Etape 3 : Une fenêtre s'ouvre …



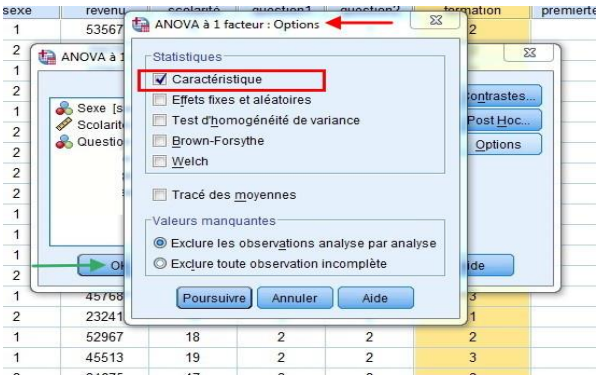
Etape 4 : Au moyen des flèches , choisir la variable dépendante que vous souhaitez analyser (Revenu ici).

Etape 5 : choisir en fonction de quelle variable indépendante, ou facteur vous désirez analyser le revenu (Formation ici).

Etape 6 : Le but de ce test est de comparer la variance des revenus de trois types de formation (science, technique, autres). Alors, la question de recherche est la suivante : « ***la formation d'une personne influence-t-elle son revenu ?*** ».

Etape 7 : cliquer sur ***« options »*** pour choisir vos options.

Etape 8 : Une fenêtre s'ouvre …



Etape 9 : Cocher « ***Caractéristiques »*** pour obtenir les moyennes et les variances de vos trois groupes ; et cliquer sur « ***Poursuivre » > « ok ».*** Voici le résultat final :

A 1 facteur

Descriptives

revenu annuel

1	N	Moyenne	Ecart-type	Erreur standard	Intervalle de confiance à 95% pour la moyenne		Minimum	Maximum
					Borne inférieure	Borne supérieure		
science	15	50030,2667	1,73846E4	4488,67342	40403,0197	59657,5137	23456,00	89098,00
technique	15	47197,0667	1,65014E4	4260,63092	38058,9222	56335,2111	23443,00	76876,00
autres	15	53205,4000	1,06726E4	2755,65726	47295,1030	59115,6970	34444,00	67897,00
Total	45	50144,2444	1,50068E4	2237,07887	45635,7082	54652,7807	23443,00	89098,00

ANOVA

revenu annuel

2	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	2,710E8	2	1,355E8	2 ,591	3 ,559
Intra-groupes	9,638E9	42	2,295E8		
Total	9,909E9	1 44			

Le premier tableau est une analyse descriptive (moyenne) de l'échantillon. Le second tableau est une analyse comparative (ou inférentielle).

4.3.2 INTERPRÉTATION DES RÉSULTATS

Dans l'analyse de la variance, il y a 2 tableaux importantsn à savoir : le ***tableau des moyennes*** qui décrit les trois groupes (science, technique, autres) et le ***tableau du test F*** qui permet de comparer les trois groupes.

Dans le ***tableau du test F***, il y a **3 données importantes** :

- Le ddl ou degré de liberté qui est ici 44 (soit n-1) ;
- Le résultat du F (0,591) ;
- La valeur de p ou signification (ici 0,559).

A 1 facteur

Descriptives

revenu annuel

	N	Moyenne	Ecart-type	Erreur standard	Intervalle de confiance à 95% pour la moyenne		Minimum	Maximum
					Borne inférieure	Borne supérieure		
science	15	50030,2667	1,73846E4	4488,67342	40403,0197	59657,5137	23456,00	89098,00
technique	15	47197,0667	1,65014E4	4260,63092	38058,9222	56335,2111	23443,00	76876,00
autres	15	53205,4000	1,06726E4	2755,65726	47295,1030	59115,6970	34444,00	67897,00
Total	45	50144,2444	1,50068E4	2237,07887	45635,7082	54652,7807	23443,00	89098,00

ANOVA

revenu annuel

2	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	2,710E8	2	1,355E8	2 ,591	3 ,559
Intra-groupes	9,638E9	42	2,295E8		
Total	9,909E9	1 44			

Le **F** et le **ddl** permettent de calculer la valeur de **p ou signification**.

La valeur de **p** permet de confirmer ou d'infirmer l'hypothèse nulle (H0).

En choisissant un seuil de signification α=0,05, la règle de décision est la suivante :

- Si la valeur de **p** est supérieure à α=0,05 ; alors on accepte H0 ;
- Si la valeur de **p** est inférieure à α=0,05 ; alors on rejette H0.

Dans cet exemple, p=0,559 > α=0,05 et donc on accepte H0. Conclusion : **le type de formation n'influence pas le revenu des individus**.

4.3.3 DÉMARCHE LORSQUE L'HYPOTHÈSE NULLE (H0) EST REJETÉE

Si la signification ou la valeur de **p** est inférieure à **α**, alors il faut faire deux choses :

- Rejeter l'hypothèse nulle (H0) et conclure qu'il y a une différence significative entre les groupes ;
- Procéder à un **test post-hoc** afin de savoir laquelle des comparaisons de groupes, pris deux à deux, est significativement différente.

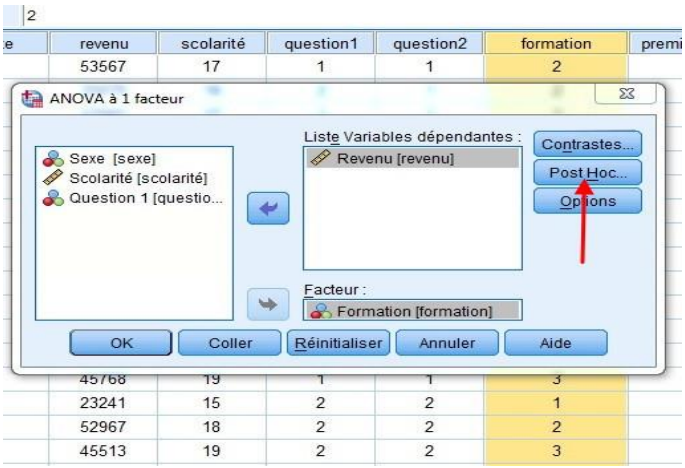
Un test post-hoc significatif indique quelle paire de groupes est différente.

4.3.3.1 Analyse des résultats d'un test post-hoc

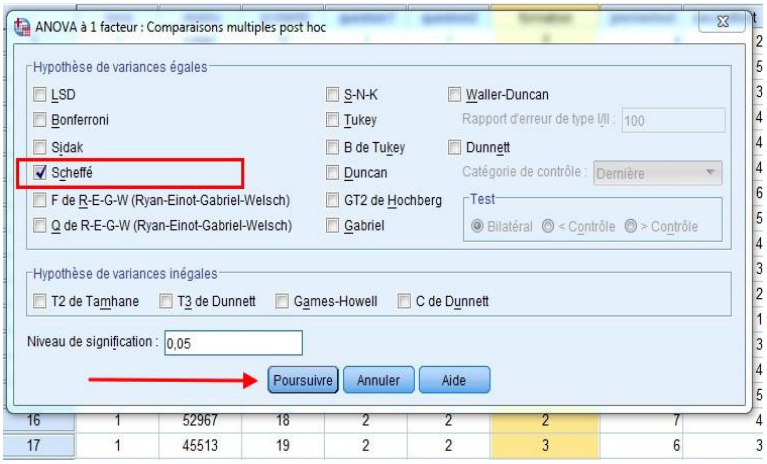
L'analyse de variance compare généralement trois groupes ou plus. Le résultat significatif de ce test indique que certains groupes ou échantillons pris deux à deux sont significativement différents. Mais lorsque le test est effectué avec trous groupes ou plus, il ne précise pas quelle paire de groupes présente un écart significatif.

Pour résoudre ce problème, il faut faire un test supplémentaire ou test post-hoc, qui indiquera laquelle des trois paires de groupes sont différents ***(exemple : science et technique ou autres et sciences ou technique et autres)*** : ce test compare les trois groupes ou échantillons deux à deux.

Pour choisir ce test, cliquer sur le bouton **« post-hoc »**.



Une nouvelle fenêtre apparaît ; et il existe un vaste de choix de test post-hoc



Dans le cas ce cet exemple, choisir le test **Scheffé** ; puis cliquez sur **Poursuivre**.

Une fenêtre de résultats apparaît :

Comparaisons multiples

revenu annuel
Scheffe

	(I) Type de formation	(J) Type de formation	Différence de moyennes (I-J)	Erreur standard	Signification	Intervalle de confiance à 95% Borne inférieure	Borne supérieure
1	science	technique	2833,20000	5531,41403	,877	-1,1204E4	16870,2294
		autres	-3,17513E3	5531,41403	,849	-1,7212E4	10861,8961
2	technique	science	-2,83320E3	5531,41403	,877	-1,6870E4	11203,8294
		autres	-6,00833E3	5531,41403	,559	-2,0045E4	8028,6961
3	autres	science	3175,13333	5531,41403	,849	-1,0862E4	17212,1627
		technique	6008,33333	5531,41403	,559	-8028,6961	20045,3627

Selon les résultats de cette fenêtre, la colonne **Signification** indique qu'il n'y a pas de différence significative entre les trois groupes comparés deux à deux (Sig.< 5 %).

- La première ligne compare la formation scientifique à la formation technique et autres ;
- La seconde ligne compare la formation technique à la formation scientifique et autres ;
- Finalement, les autres formations sont comparées à la formation scientifique et technique.

4.4 EXERCICES

Parmi les propositions suivantes, choisir celle(s) qui est (sont) exactes(s).

QCM1 : Le test d'ANOVA est appliqué pour détecter s'il y a un lien entre :

A. Deux variables quantitatives
B. Une variable quantitative et une variable qualitative
C. Deux variables qualitatives
D. Plus de deux variables quantitatives

QCM2. Le test d'ANOVA compare :

A. Les variances de deux échantillons
B. Les fréquences de deux échantillons
C. Les proportions de deux échantillons
D. Les moyennes de deux échantillons

QCM3. Il y-t-il des tests à effectuer avant une ANOVA ? Si oui lesquels ?

A. Oui, seulement le test de normalité
B. Oui, le test de normalité et le test d'homogénéité des variances
C. Oui, le test de normalité, d'homogénéité des variances et des représentations graphiques
D. Non, pas besoin de test

QCM4. Combien d'hypothèses faut-il pour faire une ANOVA

A. 1
B. 4
C. 2
D. 0

QCM5. L'hypothèse alternative (H1) se libelle comme suit :

A. Au moins une des moyennes est différente des autres
B. Toutes les moyennes sont les mêmes
C. Les moyennes sont les mêmes deux par deux
D. Toutes les moyennes sont différentes

QCM6. Le résultat d'un test ANOVA permet de :

A. Rejeter ou accepter l'hypothèse H0
B. Rejeter ou accepter l'hypothèse H1
C. Accepter l'hypothèse H0 et rejeter H1 seulement
D. Aucune de ces propositions

QCM7. Le Degré De Liberté (ddl) se calcul comme suit (n étant le taille de l'échantillon) :

A. ddl = n-1
B. ddl = n-3
C. ddl = n+1
D. ddl = n+3

QCM8. La règle de décision d'une ANOVA

A. Si H0 est inférieur à α, on accepte H0
B. Si H0 est supérieur à α, on accepte H0
C. Si H0 est inférieur à α, on rejette H0
D. Si H0 est supérieur à α, on rejette H0

QCM9. La statistique utilisée par l'ANOVA est :

A. La statistique de Fisher
B. La statistique du Khi2
C. La statistique de Student
D. La statistique de Durbin-Watson

QCM10. Si l'hypothèse H0 est rejetée, pour trouver où se trouve la différence, il faut :

A. Refaire l'ANOVA en prenant les variables deux à deux
B. Faire le test post-hoc
C. Pas possible de trouver où se trouve la différence

QCM1 (B) - QCM2 (D) - QCM3 (C) - QCM4 (C) - QCM5 (A) - QCM6 (A) - QCM7 (A) - QCM8 (B C) - QCM9 (A) - QCM10 (B)

5 TABLEAU DE CONTINGENCE / TEST DU KHI2

5.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin de ce module, les participants auront une meilleure connaissance des principes du test de Khi2, des conditions d'utilisation de ce test et la démarche de son exécution. Ils seront également capables de réaliser toutes les étapes d'exécution d'un test de Khi2 à l'aide de SPSS et d'interpréter les résultats fournis par le logiciel.

5.2 ASPECT THÉORIQUE

5.2.1 PRINCIPE DU TEST DE KHI2

Le test de Khi2 est une analyse bivariée qui consiste à déterminer s'il existe une association entre deux variables qualitatives. C'est-à-dire déceler une éventuelle relation d'indépendance ou d'influence d'une variable sur une autre. Le Khi2 est une analyse dite non-paramétrique, pas de prémisses des paramètres de la distribution de la variable (moyenne, écart-type et normalité).

En d'autres termes, le Khi2 est un test statistique conçu pour déterminer si la différence entre deux distributions de fréquences est attribuable à l'erreur d'échantillonnage (le hasard) ou est suffisamment grande pour être statistiquement significative.

Questions de recherche

- Les variables X et Y sont-elles indépendantes ?
- Existe-t-il un lien entre les variables X et Y ?

X et Y étant des variables qualitatives.

5.2.2 TABLEAU DE CONTINGENCE OU TABLEAU CROISÉ

Le test de Khi2 utilise le tableau croisé (aussi appelé tableau de contingence), pour examiner la relation entre deux variables catégorielle. C'est un arrangement dans lequel les données sont classées selon deux variables catégorielles. Les catégories d'une variable apparaissent dans les lignes et les celles de l'autre variable apparaissent dans les colonnes.

Tableau 9 : Exemple de tableau de contingence

Y / X	Y1	Y2	Yj		Yp	Total
X1	N11	N12	N1j		N1p	N1.
X2	N21	N22	N2J		N1p	N2.
Xi	Ni1	Ni2	Nij		Nip	Ni.
Xk	Nk1	Nk2	Nkj		Nkp	Nk.
Total	N.1	N.2	N.j		N.p	N

5.2.3 MISE EN ŒUVRE DU TEST DE KHI2

5.2.3.1 Conditions d'utilisation du test

Avant de faire un test de Khi2, il est nécessaire de vérifier les conditions suivantes :

- Les deux variables sont qualitatives ;
- Chaque variable comporte deux ou plusieurs modalités ;
- Les observations sont indépendantes ;
- La taille de l'échantillon es relativement grande et aléatoire ;
- Les modalités des variables sont mutuellement exclusives (par exemple : oui/non ou homme/femme ; pas de réponse intermédiaire) ;
- Les occurrences attendues (effectifs théoriques) doivent être supérieures ou égales à 5 et aucune occurrence attendue ne doit être inférieure à 1.

5.2.3.2 Formulation des hypothèses

Hypothèse nulle (H0) : il n'y a pas de relation entre les 2 variables catégorielles. Nous pouvons dire aussi que les 2 variables sont indépendantes. L'indépendance signifie que la valeur d'une des 2 variables ne nous donne aucune information sur la valeur possible de l'autre variable. ***H0 : les variables X et Y sont indépendantes.***

Hypothèse alternative (H1) : il existe une relation entre les variables ou les deux variables sont dépendantes. ***H1 : les variables X et Y ne sont pas indépendantes.***

5.2.3.3 Calcul de la statistique Khi2

La statistique se calcul en utilisant les effectifs observés et les effectifs théoriques : $X^2 \; ou \; Q = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$

*Le « o » est l'effectif observé et le « e » est l'effectif de l'hypothèse nulle (effectif théorique).

5.2.3.4 Détermination du risque d'erreur et du degré de liberté

Le risque d'erreur est le pourcentage de chances de se tromper, c'est-à-dire de rejeter à tort l'hypothèse nulle le plus souvent, ***α=5 % (0,05),*** mais d'autres seuils de probabilité peuvent être choisis.

La distribution Khi2 demande (tout comme l'analyse de variance, le calcul du degré de liberté. En effet, cette distribution varie de forme en fonction du degré de liberté du tableau croisé. Cependant, le calcul du degré de liberté ne dépend pas du nombre de sujets, mais plutôt du nombre de ligne et de colonne dans le tableau croisé.

Degré de liberté (ddl) = (nombre de lignes -1) * (nombre de colonnes -1)

5.2.3.5 Démarche générale

- Identifier si le test de Khi2 est applicable (vérifier les conditions d'utilisation du test) ;
- Si oui, formuler les hypothèses. H0 : les variables sont indépendantes ; H1 : les variables ne sont pas indépendantes ;
- Calculer l'indicateur de Khi2, en calculant utiliser les effectifs théoriques et observés ou utiliser un logiciel statistique (nous utilisons dans ce module le logiciel SPSS) ;
- Déterminer le risque d'erreur et le degré de liberté ;

- Confronter l'indicateur de Khi2 à la table de la loi du Khi2. On compare X^2 à une valeur $X^2_{k-1.\infty}$ issue de la table de la loi du Khi2 ;
- Interpréter les résultats. Si $X^2 < X^2_{k-1.\infty}$, on accepte H0 au seuil ***α*** ; si $X^2 > X^2_{k-1.\infty}$, on rejette HO au seuil ***α***.

5.2.4 FORCE DE LA RELATION

Le Khi2 donne la signification, c'est-à-dire l'existence d'un lien entre les deux variables. Cependant, il ne donne pas d'information sur la force de la relation.

Il est possible d'apprécier la force de l'association entre les variables à partir de tests complémentaires sur les mesures symétriques. Ces mesures sont basées sur la statistique Kki2 qui a été modifiée pour tenir compte de la taille de l'échantillon et des degrés de libertés. Le résultat de ces tests se situe entre « 0 » et « 1 ». Les plus fréquemment utilisés sont le **Phi** et le **V de Cramer**.

- **Coefficient Phi**. Cette mesure d'association est pertinente pour les tableaux 2*2 seulement. La valeur s'interprète directement selon les balises de taille d'effet de la corrélation de Pearson.
- **V de Cramer.** Cette mesure d'association est valable pour tous les tableaux plus grands que 2*2. Cependant, pour l'interpréter simplement, il faut transformer le coefficient pour tenir compte de l'inflation de la valeur du Khi2 en fonction de la taille du tableau. Pour ce faire, Cohen (1988) propose de calculer la valeur Oméga **(Ω)** où **k** représente le plus petit nombre de catégories du croisement.

La valeur de **Ω** résultante s'interprète ensuite selon les balises de Cohen pour la corrélation de Pearson.

Valeur	Force du lien statistique
0	Absence de relation
Entre 0,05 et 0,10	Très faible
Entre 0,10 et 0,20	Faible
Entre 0,20 et 0,40	Modérée
Entre 0,40 et 0,80	Forte
Entre 0,80 et 1	Louche (Colinéarité)

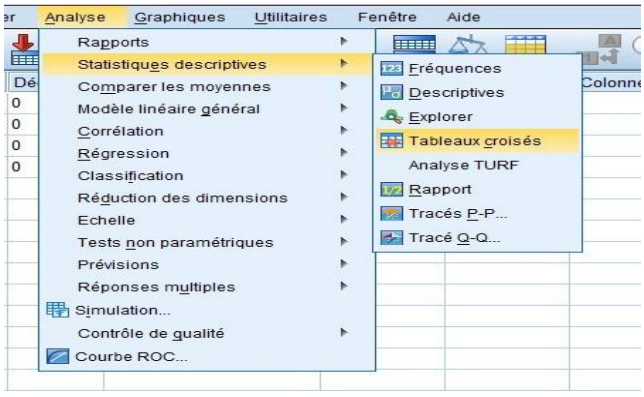
5.3 SYNTAXES SPSS

5.3.1 FORMALISATION DU PROCESSUS SOUS SPSS

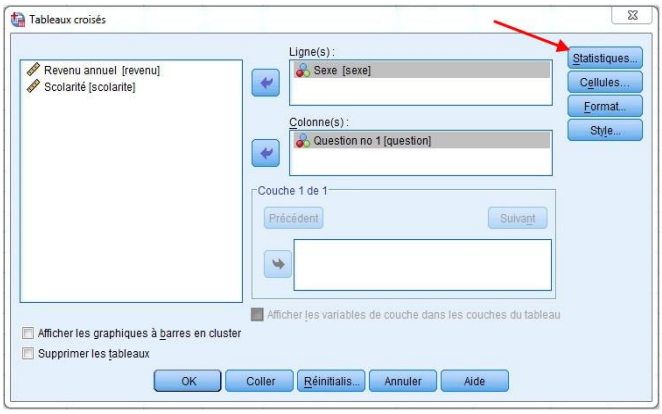
Etape 1 : ouvrir votre matrice de données sous SPSS.

Etape 2 :

Choisir le menu **Analyse** > **Statistiques Descriptives** > **Tableaux Croisés**



Une fenêtre s'ouvre ...

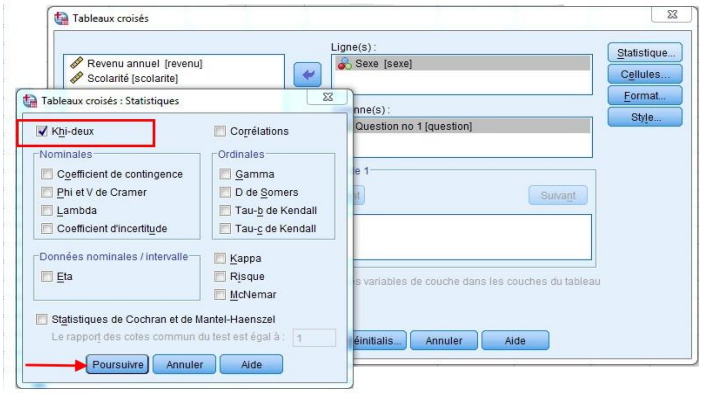


Etape 3 :

Au moyen des flèches choisir votre Ligne (la variable indépendante de votre recherche). Dans cet exemple = [Sexe]. Choisir ensuite votre Colonne (la variable dépendante de votre recherche). Dans cet exemple = [Question no 1].

Etape 4 :

Cliquez ensuite sur **Statistiques** pour choisir le test khi-deux. Une fenêtre s'ouvre...



Etape 5 : cocher **Khi-Deux**

Le but de ce test est de comparer les effectifs de OUI et de NON des deux groupes - HOMMES et FEMMES - afin de vérifier l'hypothèse selon laquelle, au sein de la population, la fréquence des hommes est différente de la fréquence des femmes.

- Ensuite cliquez sur **Poursuivre.**
- Puis sur **OK** dans la fenêtre principale.

Voici le résultat final :

Tableau croisé Sexe * Question 1

Effectif

		Question 1		
		oui	non	Total
Sexe	homme	5	10	15
	femme	4	11	15
Total		9	21	30

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,159[a]	1	,690		
Correction pour la continuité[b]	,000	1	1,000		
Rapport de vraisemblance	,159	1	,690		
Test exact de Fisher				1,000	,500
Association linéaire par linéaire	,153	1	,695		
Nombre d'observations valides	30				

a. 2 cellules (50,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 4,50.
b. Calculé uniquement pour un tableau 2x2

Le premier tableau est l'analyse descriptive de votre échantillon (fréquence ou proportion)

Le second tableau est l'analyse comparative (ou inférentielle).

5.3.2 INTERPRÉTATION DES RÉSULTATS

Dans l'analyse d'un Khi2, il y a 2 tableaux importantsn à savoir : le **tableau des effectifs ou des fréquences** qui décrit les 2 groupes et le **tableau du Khi2** qui permet de comparer ces groupes.

Tableau croisé Sexe * Question 1

Effectif

		Question 1		
		oui	non	Total
Sexe	homme	5	10	15
	femme	4	11	15
Total		9	21	30

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,159[a]	1	,690		
Correction pour la continuité[b]	,000	1	1,000		
Rapport de vraisemblance	,159	1	,690		
Test exact de Fisher				1,000	,500
Association linéaire par linéaire	,153	1	,695		
Nombre d'observations valides	30				

Dans le **tableau du Khi2**, il y a 3 résultats importants :

- Le résultat du test ou **valeur** (dans cet exemple il est de 0,159) ;
- Le **ddl** ou **degré de liberté** : *ici ddl= (2-1)*(2-1) ;*
- La signification asymptotique (bilatérale) ou valeur de p (ici 0,690).

La valeur du test et le ddl permettent à SPSS de calculer la Signification asymptotique (bilatérale) ou **p**. Cette valeur de **p** permet de confirmer ou d'infirmer l'hypothèse nulle (H0) et partant l'objectif de la recherche.

En choisissant un seuil de signification α=0,05, la règle de décision est la suivante :

- Si la valeur de **p** ou **Signification asymptotique** est supérieure à α=0,05 ; alors on accepte H0 ;
- Si la valeur de **p** ou **Signification asymptotique** est inférieure à α=0,05 ; alors on rejette H0.

5.4 EXERCICES

Parmi les propositions suivantes, choisir celle(s) qui est (sont) exactes(s).

QCM1. Le test de Khi2 permet de comparer :

A. Deux variables qualitatives
B. Deux variables quantitatives
C. Un variable quantitative et une variable qualitative
D. Plusieurs types de variables quantitatives et/ou qualitatives

QCM2. Les conditions de l'application du test de Khi2 sont :

A. Les effectifs théoriques doivent être supérieurs ou égales à 30
B. Les observations doivent être dépendantes
C. Les observations doivent être indépendantes
D. Les effectifs théoriques doivent être supérieurs ou égales à 5

QCM3. L'hypothèse nulle (H0) est formulée comme suit :

A. Les deux variables sont dépendantes
B. Il n'existe pas de lien entre les deux variables
C. Les deux variables sont les mêmes
D. Les deux variables sont indépendantes

QCM4. Le risque d'erreur noté α est :

A. Le risque de rejeter H0, alors que H0 est vrai
B. Le risque d'accepter H0, alors que H0 est vrai
C. Le risque de ne pas trouver la bonne statistique

QCM5. La décision du test de Khi2 se fait par rapport :

A. L'hypothèse nulle (H0)
B. L'hypothèse alternative (H1)
C. Le degré de liberté
D. La probabilité p

QCM6. La décision du test de Khi2 est le suivant :

A. Si la valeur de la probabilité (p) est inférieure à α, on accepte H0
B. Si la valeur de la probabilité (p) est inférieure à α, on accepte H1
C. Si la valeur de la probabilité (p) est égale à α, on accepte H0
D. Si la valeur de la probabilité (p) est supérieure à α, on accepte H0

QCM7. Le calcul du Degré De Liberté (ddl) dans l'application du test de Khi2 dépend de :

A. La taille de l'échantillon
B. Le nombre de lignes
C. Le nombre de colonnes
D. Le nombre de lignes et de colonnes

QCM8. La formule de calcul du ddl est :

A. ddl = n-1
B. ddl = (nombre de lignes – 1) * (nombre de colonnes – 1)
C. ddl = nombres de ligne – 1
D. ddl = nombre de colonne - 1

QCM9. Pour déterminer la force de la relation si elle existe, il faut :

A. Utiliser la probabilité fournie par le test de Khi2
B. Utiliser un t-test
C. Utiliser le coefficient Phi ou le coefficient de Cramer
D. Impossible de déterminer la force de la relation

QCM10. Vous voulez savoir s'il existe un lien entre les motivations d'achat des consommateurs et leur statut matrimonial. Quelle procédure statistique préconisez-vous ?

A. Régression simple
B. Test de Khi2 d'ajustement
C. Test du Khi2 de contingence
D. ANOVA

QCM1 (A) - QCM2 (C D) - QCM3 (B D) - QCM4 (A) - QCM5 (A D) - QCM6 (D) - QCM7 (D) - QCM8 (B) - QCM9 (C) - QCM10 (C).

6 CORRÉLATION ET RÉGRESSION SIMPLE

6.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin de ce module, les participants connaitront l'utilité de l'application d'un test de corrélation ainsi que d'un test de régression linéaire simple. Ils pourront faire la différence entre le coefficient de Pearson et celui de Spearman et savoir pourquoi calculer le coefficient de Pearson. Par ailleurs, les participants seront capables de réaliser les tests de corrélation et de régression simple avec SPSS et interpréter tous les résultats importants restitués par le logiciel.

6.2 ASPECT THÉORIQUE

Les instruments d'exploration de données empiriques se diversifient et s'affinent lorsque les deux variables sont quantitatives. La représentation graphique, la corrélation et la régression simple sont ici, généralement privilégiées.

Les trois approches sont en pratique indissociables, la corrélation explorant la relation entre deux variables en supposant qu'elles sont liées par une relation « linéaire ».

6.2.1 CORRÉLATION

La corrélation est une quantification de la relation linéaire entre des variables continues. Un coefficient de corrélation calcule dans quelle mesure deux variables tendent à changer ensemble. Le coefficient décrit l'importance et le sens de la relation.

6.2.1.1 Corrélation bivariée simple : Pearson ou Spearman

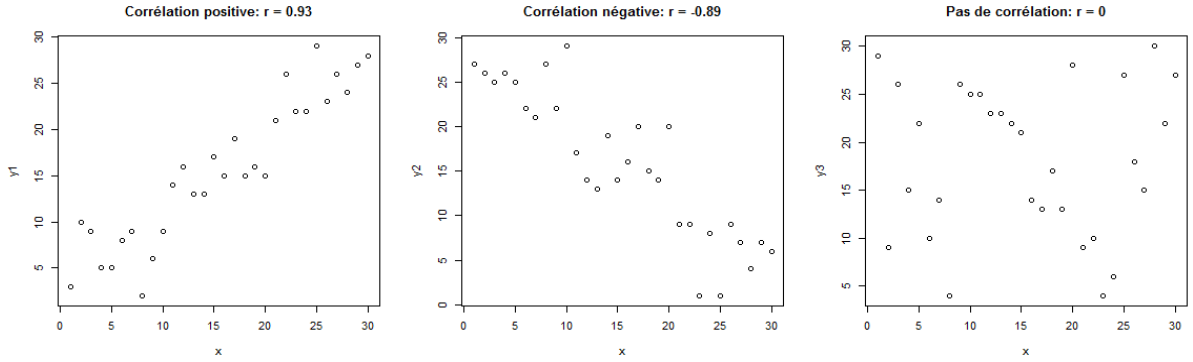
La corrélation de Pearson évalue la relation linéaire entre deux variables continues. Une relation est dite linéaire lorsqu'une modification de l'une des variables est associée à une modification proportionnelle de l'autre variable. Par exemple, on peut utiliser une corrélation de Pearson afin d'évaluer si les augmentations de température sur le site de production sont associées à la diminution de l'épaisseur de l'enrobage de chocolat.

Coefficient de corrélation de Pearson

Le coefficient de corrélation le plus connu est le coefficient **r** de Pearson, également appelé coefficient de corrélation linéaire. Ce coefficient implique que les deux variables soient au moins mesurées sur des échelles d'intervalles et détermine dans quelle mesure les valeurs des deux variables sont proportionnelles les unes aux autres. Le coefficient de corrélation ne dépend pas des unités de mesure utilisées.

Le caractère proportionnel signifie une liaison linéaire, c'est-à-dire que la corrélation sera forte si les points s'alignent sur une droite (de pente positive ou négative).

- Le coefficient de corrélation est compris entre « -1 » et « 1 » ;
- Plus le coefficient est proche de 1, plus la relation linéaire positive entre les variables est forte.



6.2.1.2 Coefficient de corrélation de Spearman

La corrélation de Spearman évalue la relation monotone entre deux variables continues ou ordinales. Dans une relation monotone, les variables ont tendance à changer ensemble, mais pas forcément à une vitesse constante. Le coefficient de corrélation de Spearman est fondé sur les valeurs classées pour chaque variable plutôt que sur les données brutes.

La corrélation de Spearman est souvent utilisée dans le but d'évaluer les relations comprenant des variables ordinales. Par exemple, une corrélation de Spearman peut être utilisée afin d'évaluer si l'ordre dans lequel des employés effectuent un exercice d'un test est lié au nombre de mois d'ancienneté.

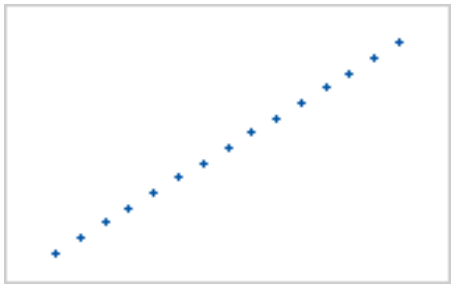
Il est toujours judicieux d'examiner la relation entre les variables à l'aide d'un nuage de points. ***Les coefficients de corrélation ne mesurent que des relations linéaires (Pearson) ou monotones (Spearman).***

6.2.1.3 Comparaison des coefficients de Pearson et de Spearman

La valeur des coefficients de corrélation de Pearson et de Spearman est comprise entre « - 1 » et « + 1 ». Pour que le coefficient de corrélation de **Pearson soit « +1 »**, ***l'une des variables doit augmenter de façon constante lorsque l'autre augmente.***

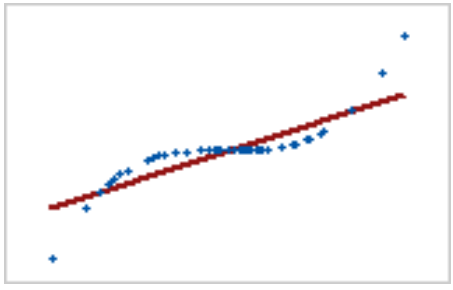
Cette relation forme une droite parfaite. Le coefficient de corrélation de **Spearman est également de « +1 »** dans ce cas.

Pearson = +1, Spearman= +1

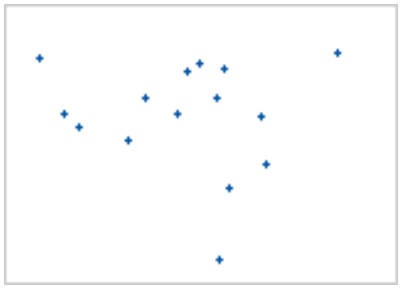


Si dans la relation, une variable augmente lorsque l'autre augmente, mais que cette **augmentation n'est pas constante**, le coefficient de corrélation de Pearson est positif mais inférieur à « +1 ».

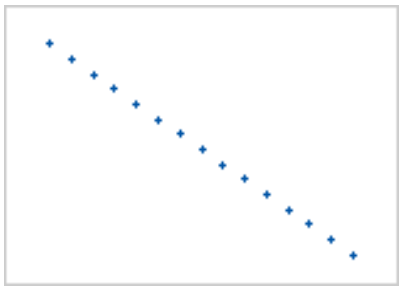
Dans ce cas, le coefficient de Spearman est lui toujours égal à +1.



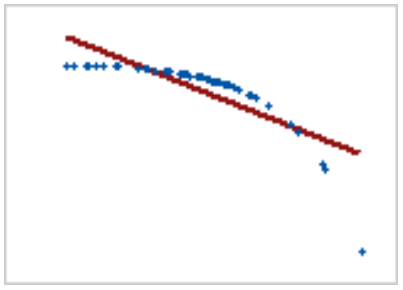
Lorsqu'une relation est aléatoire ou inexistante, les deux coefficients de corrélation sont proches de 0.



Si la relation est une droite parfaite décrivant une relation décroissante, les deux coefficients de corrélation sont de « -1 ».



Si dans la relation ***une variable diminue lorsque l'autre augmente***, mais que cette ***diminution n'est pas constante***, le coefficient de corrélation de Pearson est négatif mais supérieur à « -1 ». Dans ce cas, le coefficient de Spearman est toujours égal à -1.



Les valeurs de corrélation « - 1 » ou « + 1 » impliquent une relation linéaire exacte. Toutefois, le véritable intérêt des valeurs de corrélation est de permettre la quantification de relations non parfaites. L'existence d'une corrélation entre deux variables est souvent une information importante pour une analyse de régression qui essaie d'affiner la description de ce type de relation. Dans la suite de ce chapitre, nous utiliserons le coefficient de corrélation de Pearson.

6.2.1.4 Quand et pourquoi calculer le coefficient de corrélation

Quand. Si votre recherche comporte une variable indépendante quantitative (X) et une variable dépendante quantitative (Y).

Pourquoi ?

- Pour établir l'existence d'un lien entre X et Y ;
- Pour mesurer la force ou l'intensité de ce lien ;
- Pour inférer l'existence d'une corrélation au sein de la population (r + test de signification de la pente).

6.2.1.5 Formulation des hypothèses du test de corrélation

Hypothèse nulle (H0) : les deux variables ne sont pas associées ou il n'y a pas de relation entre les deux variables. **H0 : Pas de corrélation entre les deux variables ρ = 0.**

Hypothèse alternative (H1) : il existe une relation entre les deux variables. **H1 : Corrélation entre les deux variables ρ ≠ 0**

6.2.1.6 Conditions d'utilisation du test de corrélation

- Les variables X et Y doivent être aléatoires : les observations de chaque variable doivent être indépendantes les unes des autres. ***Cette condition n'est pas remplie lorsque l'on compare des données Y en fonction du temps X. Dans cas, il y a autocorrélation et il faut faire appel à des techniques d'analyse de séries chronologiques.***
- Chaque paire de variables bivarie normalement : dans la population d'où est issu l'échantillon, les distributions conditionnelles de Y liées à chaque valeur de X doivent être normales et de variance égales ; et symétriquement. Cette condition est impossible à vérifier en pratique mais est souvent vraie.

6.2.2 REGRESSION SIMPLE

Cette section est une introduction à la modélisation par le modèle le plus élémentaire, la régression linéaire simple. Le but de cette régression est d'établir un lien entre une variable dépendante ***Y*** et une variable indépendante ***X*** pour pouvoir ensuite faire des prévisions sur ***Y*** lorsque ***X*** est mesuré. Avant tout travail de modélisation, une approche descriptive ou exploratoire est nécessaire pour dépister au plus tôt des difficultés dans les données : dissymétrie des distributions, valeurs atypiques, liaison non linéaire entre les variables peut s'avérer nécessaire.

6.2.2.1 Définir le modèle

On note ***Y*** la variable aléatoire réelle à expliquer (variable endogène, dépendante ou réponse) et ***X*** la variable explicative ou effet fixe (exogène). Le modèle revient à supposer, qu'en moyenne, E (Y) est une fonction affiné de ***X***. L'écriture du modèle suppose implicitement une notion préalable de causalité dans le sens où ***Y*** dépend de ***X*** car le modèle n'est pas systématique. Le modèle s'écrit comme suit : $E(Y) = f(X == \beta o + \beta 1X \; ou \; Y = \beta 0 + \beta 1X + \varepsilon$

Où **Y** est la variable dépendante ; **β0 et β1** sont les coefficients (ordonnée à l'origine et pente) ; **X** est la variable indépendante ; **ε** est une erreur aléatoire.

Les hypothèses relatives à ce modèle sont les suivants :

- La distribution de l'erreur ***ε*** est indépendante de ***X*** ;
- L'erreur est centrée et la variance constante (homoscédasticité) : pour tout i=1, …., n E(**ε**i)=0 ; V(**ε**i)= σ^2 ;
- β0 et β1 sont constants, pas de rupture du modèle ;
- Hypothèse complémentaire pour les inférences : **ε** ~ N (0, σ^2).

On cherche à estimer les paramètres β0, β1 et ε et à vérifier si le modèle est adéquat.

6.2.2.2 Estimation des paramètres

L'estimation des paramètres **β0, β1** et **ε** est obtenue en maximisant la vraisemblance sous l'hypothèse que les erreurs sont gaussiennes ou encore par minimisation de la somme des carrés des écarts entre observations et modèles (moindres carrés).

6.2.2.3 Formulation des hypothèses de la régression linéaire simple

Hypothèse nulle (H0). Il n'y a pas de relation entre la variable dépendante et la variable indépendante. La variable indépendante ne permet pas de prédire la variable dépendante.

Hypothèse alternative (H1). Il est possible de prédire la variable dépendante à partir de la variable indépendante.

6.2.2.4 Condition d'utilisation de la régression linéaire simple

Avant d'utiliser la régression linéaire simple, certaines conditions doivent être vérifiées :

- **Distribution normale** : les valeurs de la variable dépendante sont normalement distribuées ;
- **Homogénéité des variances** : la variance dans la distribution de la variable dépendante doit être constante pour toutes les valeurs de la variable indépendante ;
- **Le prédicteur (la variable indépendante)** : doit présenter une certaine variance dans les données (pas de variance nulle) ;
- **Le prédicteur** n'est pas corrélé à des variables externes (qui n'ont pas été intégrées au modèle) qui influencent la variable dépendante ;
- **Homoscédasticité :** pour toutes les valeurs du prédicteur, la variance des résiduels (erreur de mesure) est homogène. Cette condition peut être vérifiée par l'examen du nuage de points du croisement entre les valeurs prédites standardisées et les résiduels standardisés ;
- **Distribution normale et aléatoire des résiduels** : cette condition signifie que la différence entre le modèle et les valeurs observées sont près de « 0 ». Elle peut être vérifiée par l'examen du nuage de points qui a servi à vérifier la condition d'homoscédasticité ;
- **Les valeurs de la variable dépendante** sont **indépendantes :** chaque valeur de la variable dépendante vient d'une observation distincte. Les observations ne sont pas reliées entre elles ;
- **Relation linéaire entre la variable indépendante et la variable dépendante** : la relation modélisée est linéaire. Cette condition peut être vérifiée par le nuage de points du croisement entre ces deux variables.

6.2.3 CORRELATION VS REGRESSION SIMPLE

Tableau 10 : Corrélation VS Régression simple

	CORRELATION	REGRESSION SIMPLE
Variables	X = quantitative Y = quantitative	X = quantitative Y = quantitative
Symétrie de la relation	Oui/Non Y lié à X X lié à Y	Non Y dépend de X -
Prédiction	Non	Oui (équation)
Test	Coefficient de corrélation $-1 \leq r \leq 1$	Pente de la droite de régression
Conditions	Indépendance des observations Liaison linéaire Distribution normale et de variance constante	

6.3 SYNTAXE SPSS

6.3.1 SYNTAXE SPSS POUR LE TEST DE CORRÉLATION

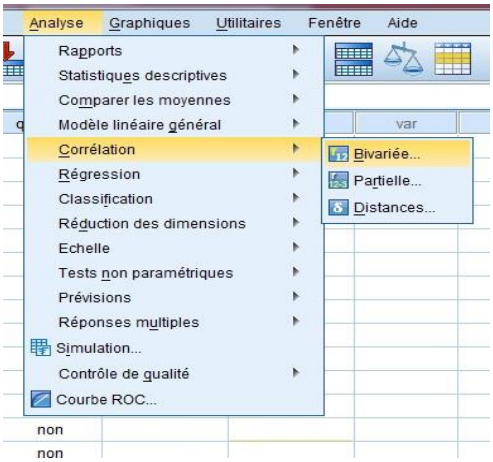
6.3.1.1 Formalisation du processus

Etape 1.

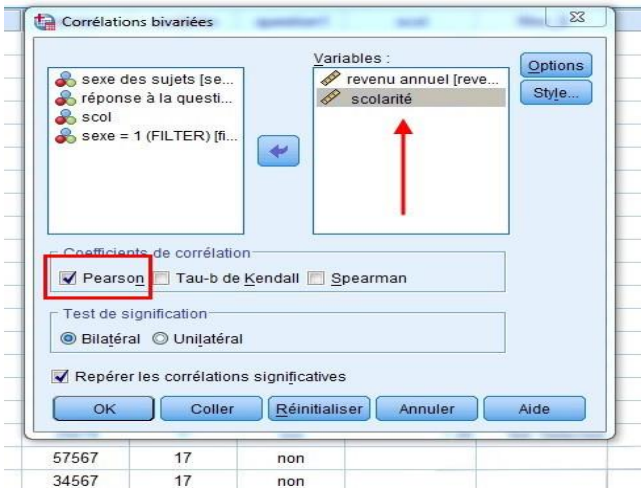
Ouvrir votre matrice de données sous SPSS

Etape 2.

Choisir le menu Analyse > Corrélation > Bivariée



Une fenêtre s'ouvre …



Etape 3. Au moyen des flèches  choisir les deux variables que vous souhaitez analyser. Notez que le coefficient de corrélation de **Pearson** est déjà coché (par défaut). Cliquez sur « **OK ».**

Voici le résultat final.

Corrélations

		revenu annuel	année de scolarité
revenu annuel	Corrélation de Pearson	1	,179
	Sig. (bilatérale)		,345
	N	30	30
année de scolarité	Corrélation de Pearson	r= ,179	1
	Sig. (bilatérale)	Sig. = ,345	
	N	30	30

6.3.1.2 Interprétation des résultats

Dans l'analyse d'une corrélation, il y a deux résultats importants :

- Le résultat du test de corrélation ou corrélation de Pearson **(r)**, dans notre exemple 0,179 ;
- La valeur de **p** du test de la pente ou **Sig. (bilatérale)**, ici 0,345.

Corrélations

		revenu annuel	année de scolarité
revenu annuel	Corrélation de Pearson	1	,179
	Sig. (bilatérale)		,345
	N	30	30
année de scolarité	Corrélation de Pearson	r= ,179	1
	Sig. (bilatérale)	Sig. = ,345	
	N	30	30

Le premier résultat (**r**) mesure le degré de liaison linéaire entre les variables dépendante (**Y**) et indépendante (**X**) de l'échantillon.

- La valeur **r = 0** équivaut à une absence de lien, alors que **r = 1** constitue un lien parfait entre ***X*** et ***Y*** ;
- Le signe « + » signifie que la relation entre ***X*** et ***Y*** est proportionnelle ; quand ***X*** augmente (ou diminue), ***Y*** diminue (ou augmente).

Par convention, on dira que la relation entre ***X*** et ***Y*** est :

- Parfaite si **r** =1 ;
- Très forte si **r** > 0,8 ;
- Forte si **r** se situe entre 0,5 et 0,8 ;
- D'intensité moyenne si **r** se situe entre 0,2 et 0,5 ;
- Faible si **r** se situe entre 0 et 0,2 ;
- Nulle si **r** = 0.

Le second résultat (Sig.) est obtenu au moyen d'un test d'hypothèse. Ce test, celui de **signification de la pente du r,** permet de décider si ce lien donné par **r** est significatif, autrement dit, si la corrélation observée entre ***X*** et ***Y*** (= votre échantillon) existe bel et bien au sein de la population de l'étude.

En choisissant un seuil de signification **α**=0,05, la règle de décision est la suivante :

- **Si la valeur de p** ou **Sig. est supérieure à α=0,05, alors on accepte H0** et on conclut que la corrélation observée entre ***X*** et ***Y*** est due au hasard ;
- **Si la valeur de p** ou **Sig. est inférieure à α=0,05, alors on rejette H0** et on conclut qu'une corrélation entre ***X*** et ***Y*** existe bel et bien au sein de la population.

6.3.2 SYNTAXE SPSS POUR LE TEST DE RÉGRESSION LINÉAIRE

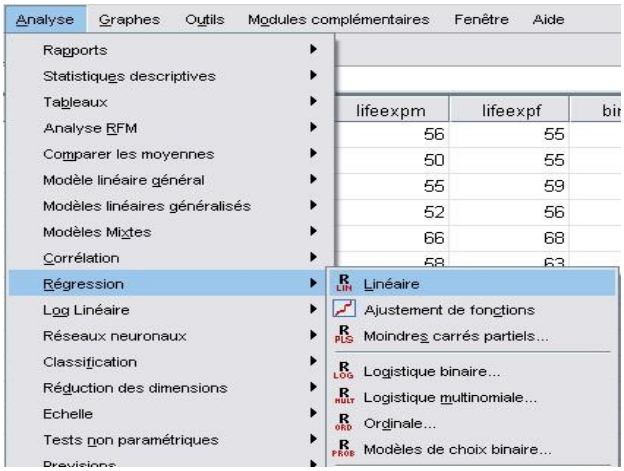
6.3.2.1 Formalisation du processus

Etape 1.

Ouvrir votre matrice de données sous SPSS

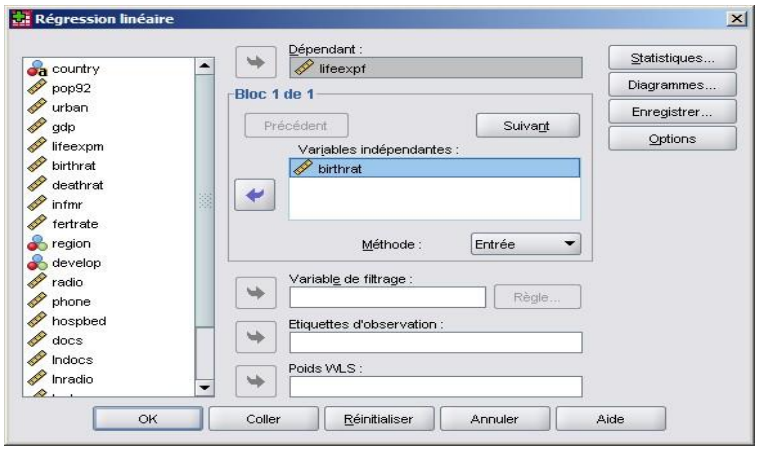
Etape 2.

Choisir Analyse > Régression > Linéaire

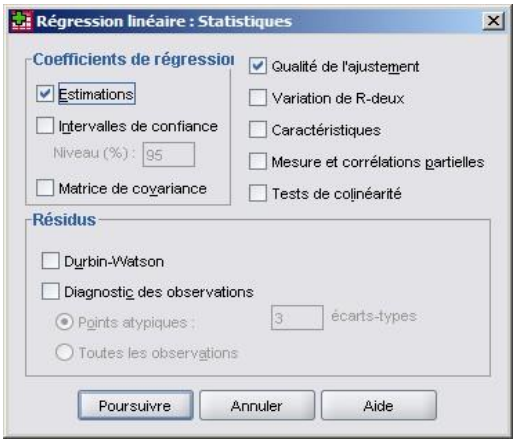


Etape 3. Cliquer sur  pour insérer la variable dépendante dans la boîte « ***Dépendant*** » et la ou les variables indépendantes. Puisqu'il s'agit ici d'une régression simple, une seule variable indépendante suffit.

Etape 4. Laisser la méthode d'analyse par défaut, c'est-à-dire le modèle « Entrée » qui utilise toutes les variables choisies pour prédire la variable dépendante.



Etape 5. Dans la régression linéaire simple, vous pouvez conserver les statistiques par défaut fournies par SPSS  . Vous obtiendrez d'une part les estimations des coefficients de régression qui permettent de reconstituer l'équation de la droite de régression. Vous obtiendrez d'autre part un tableau basé sur la distribution F vous informant de la qualité de l'ajustement du modèle.



Cliquer sur « ***Poursuivre*** ».

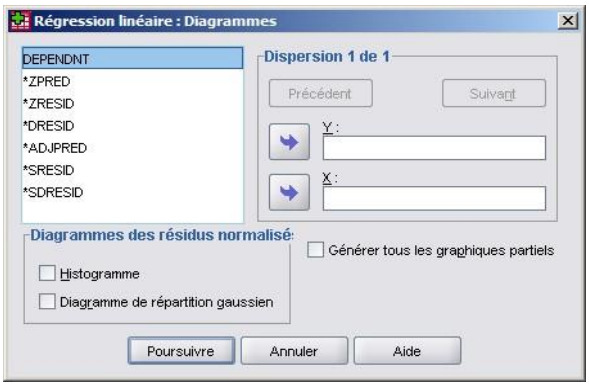
Les autres options fournis par le bouton Statistiques... sont les suivantes :

- **Intervalle de confiance :** indique les intervalles de confiance pour les coefficients de régression ;
- **Matrice de covariance :** affiche une matrice de covariance, les coefficients de corrélation et les variances entre les coefficients de régression et les variables du modèle ;
- **Variation de R-deux :** indique les changements du R^2 lorsque l'on ajoute un (ou un ensemble de) prédicteurs. Cette mesure est très utile dans la régression multiple pour voir la contribution des nouveaux prédicteurs à la variance expliquée ;
- **Caractéristiques :** affiche non seulement un tableau qui inclut le nombre d'observations, la moyenne et l'écart-type de chaque variable, mais aussi une matrice de corrélation entre les variables inclues dans le modèle ;
- **Mesure et corrélation partielles :** effectue une corrélation de Pearson entre la variable dépendante et la variable indépendante. Elle effectue une deuxième corrélation en contrôlant l'effet des autres variables indépendantes (dans la régression multiple).

Etape 6.

Le bouton Diagrammes... permet de réaliser plusieurs graphiques qui peuvent aider à vérifier certaines conditions de la régression.

Insérer les variable pour lesquelles vous voulez produire un graphique dans les boîtes **X** et **Y**. Cliquer sur « ***Poursuivre*** » pour revenir à la boîte de dialogue principale



Etape 7. Le bouton Enregistrer... permet de sauvegarder les valeurs calculées par le modèle de régression et d'en faire de nouvelles variables dans la base de données.

Cliquer sur « ***Poursuivre*** » pour revenir à la boîte de dialogue principale.

6.3.2.2 Interprétation des résultats

Pour la régression, SPSS ne fournit pas de statistiques descriptives à moins que vous ne les ayez demandées en cochant « **Caractéristiques »** dans la boîte de dialogue des statistiques. Ainsi, le premier tableau indique les variables qui ont été introduites dans le modèle. Puisque nous avons effectué une régression simple et choisi la méthode « **Entrée »**, les deux variables choisies ont été inclues dans le modèles.

Etape 1. Evaluation de la qualité du modèle de régression

La première chose à faire lors de l'examen des résultats est de vérifier si le modèle avec prédicteur explique significativement plus de variabilité de la variable dépendante qu'un modèle sans prédicteur. Ceci peut se faire en interprétant le tableau ANOVA.

Analyse de variance

Pour qu'un modèle soit pertinent, l'amélioration obtenue avec la variable indépendante doit être grande et les résiduels entre les valeurs observées et la droite de régression doivent être faibles. Pour tester cela, SPSS procède au test de la valeur **F**. Le calcul de la valeur de **F** se fait automatiquement et le degré de signification associé se trouve dans la dernière colonne.

Etape 2. Evaluation de l'ajustement des données au modèle de régression

Lorsque le modèle apporte une amélioration significative, on doit rapporter dans quelle mesure les données sont ajustées à ce modèle. En quelque sorte, il est possible de quantifier dans quelle mesure le modèle représente bien la dispersion des points dans le graphique.

Cette information se trouve dans le tableau « **Récapitulatif du modèle** » avec l'indice **R** qui présente la valeur de la corrélation multiple du modèle. La corrélation **R** s'interprète de la même manière que la corrélation simple **r**. Elle représente la corrélation combinée de toutes les variables indépendantes d'un modèle avec la variable dépendante.

Comme dans cette section nous n'avons qu'une seule variable indépendante, ce coefficient est identique (en valeur absolue) au coefficient de corrélation **r**.

La valeur de **R^2 ajusté** est un estimé de la robustesse du modèle si on prenait un échantillon différent provenant de la même population.

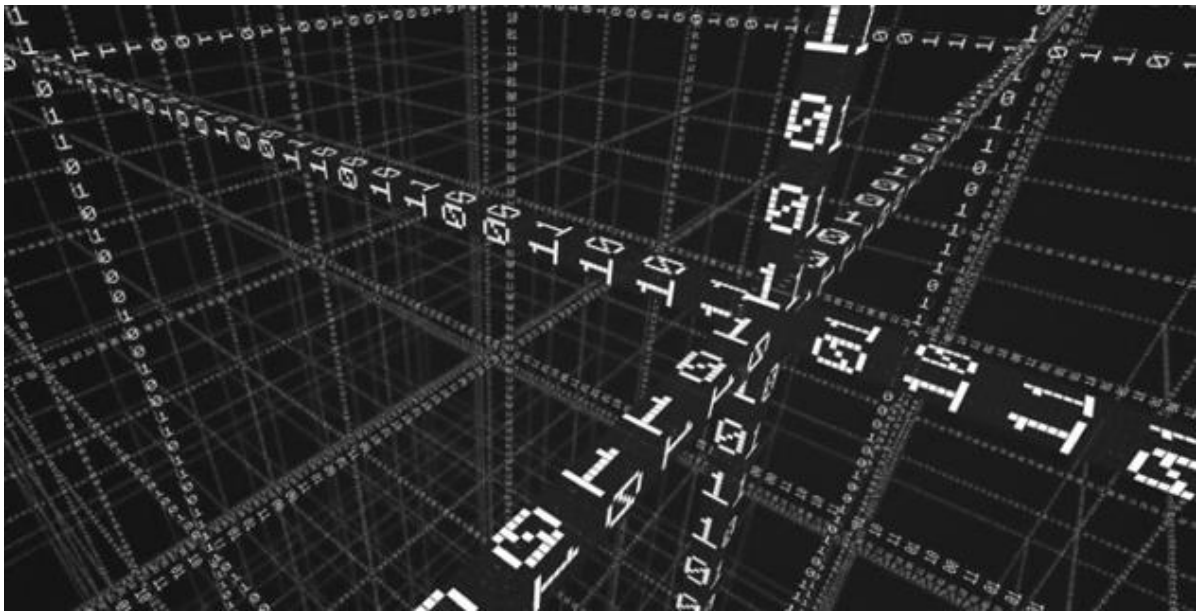
Etape 3. Evaluation de la variabilité expliquée par le modèle de régression

Enfin, on doit rapporter la proportion de la variance totale qui est expliquées par le modèle. Cette information se situe dans le même tableau sous la colonne **R-deux**.

Etape 4. Les paramètres du modèle

Le dernier tableau donne les paramètres de l'équation du modèle de régression. Il est alors possible de construire la droite de régression à l'aide des coefficients **β** non standardisés. Ce tableau est très utile dans les cas de régression multiple, car il permet de déterminer laquelle ou lesquelles des variables indépendantes contribue(nt) significativement au modèle. En effet, chaque coefficient **β** est testé en fonction de l'hypothèse nulle voulant que **β** = 0.

Les coefficients standardisés permettent de connaître de sens de la relation entre chaque prédicteur et la variable dépendante (relation positive ou négative) et la valeur absolue des coefficients standardisés significatifs permet de déterminer le poids relatif des variables dans le modèle. Les coefficients standardisés représentent la variation moyenne de **Y** pour une variation d'un écart-type de **X**.



6.4 EXERCICES

Parmi les propositions suivantes, choisir celle(s) qui est (sont) exactes(s).

QCM1. La corrélation est utilisée pour :

A. Etablir l'existence d'un lien entre une variable quantitative et une variable qualitative
B. Etablir l'existence d'un lien entre plusieurs variables quantitatives
C. Déterminer la mesure du lien qui existe entre deux variables quantitatives
D. Etablir l'existence d'un lien entre deux variables quantitatives

QCM2. Un coefficient de corrélation linéaire r :

A. Toujours positif
B. Prend des valeurs toujours comprise entre -1 et +1
C. Prend des valeurs toujours comprise entre 0 et +1
D. Prend des valeurs toujours comprise entre -1 et +∞

QCM3. Notons deux variables U et V. L'hypothèse nulle dans le cadre du test de corrélation se formule ainsi :

A. La variable U est liée à la variable V
B. La variable V est liée à la variable V
C. La variable U dépend de la variable V
D. La variable V dépend de la variable U

QCM4. Notons deux variables U et V. Si la valeur du coefficient r est égale à 0, cela correspond à :

A. U n'est pas lié à V
B. V n'est pas lié à U
C. U ne dépend pas de V
D. V ne dépend pas de U

QCM5. Les conditions d'application du test de corrélation sont :

A. Autocorrélation des variables
B. Les variances des variables doivent être égales
C. Normalité de la distribution des variables
D. Indépendance des observations

QCM6. La régression linéaire simple fait intervenir :

A. Plusieurs variables qualitatives
B. Deux variables qualitatives
C. Plusieurs variables quantitatives
D. Deux variables quantitatives

QCM7. La régression linéaire simple permet de :

A. Déterminer si l'une des deux variables dépend de l'autre (les variables quantitatives)
B. Déterminer si une variable explique les autres variables (les variables quantitatives)
C. Déterminer s'il existe un lien entre les deux variables quantitatives
D. Déterminer s'il existe un lien entre plusieurs variables qualitatives

QCM8. Dans le modèle linéaire simple, la variable dépendante est aussi appelée :

A. Variable endogène
B. Variable exogène
C. Variable réponse
D. Le prédicteur

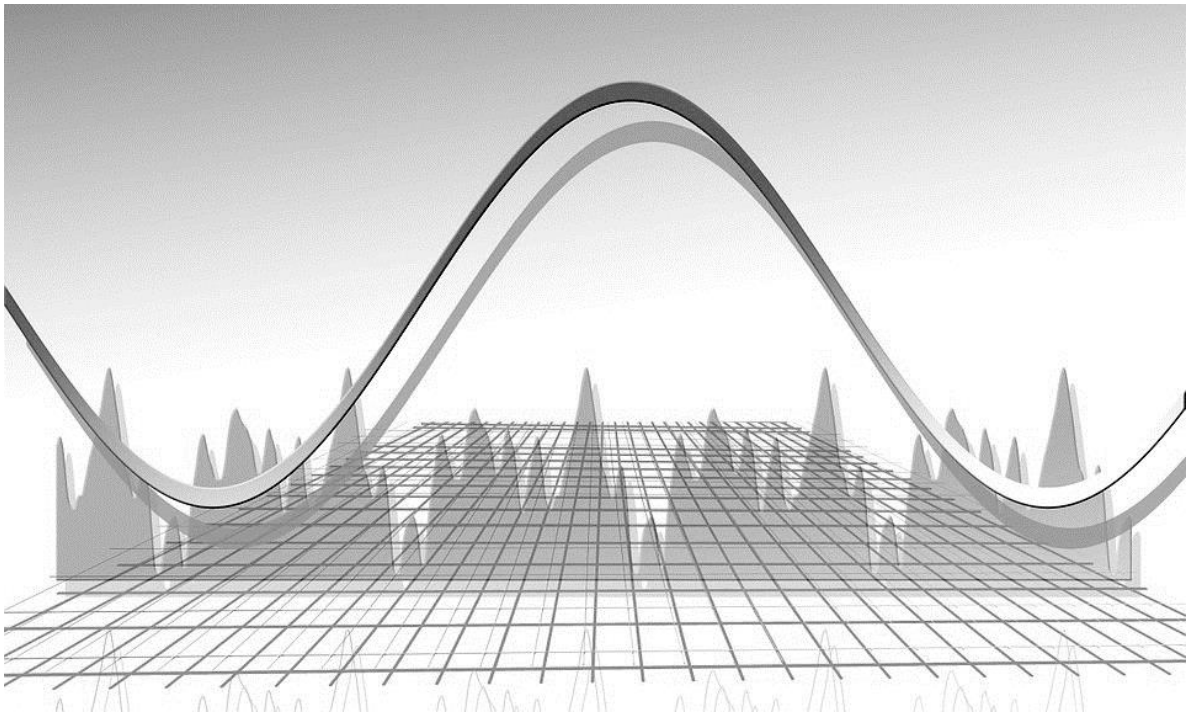
QCM9. On pose un modèle de régression linéaire simple : **$Y=\beta0+\beta1X+\varepsilon$.** Les paramètres sont définis comme suit :

A. Y est la variable indépendante, X est la variable dépendante, β0 est l'ordonné à l'origine, β1 est la pente et ε est une erreur aléatoire
B. Y est la variable indépendante, X est la variable dépendante, β0 est l'ordonné à l'origine, β1 est la pente et ε est une erreur type
C. Y est la variable dépendante, X est la variable indépendante, β0 est l'ordonné à l'origine, β1 est la pente et ε est une erreur aléatoire
D. Y est la variable indépendante, X est la variable dépendante, β0 et β1 sont les coefficients et ε est un écart-type

QCM10. La sortie d'une régression simple donne les résultats suivants : $R^2 = 0,77$ (Pr>F) = 0,50 (Pr>|t|) = 0,045 pour le coefficient β0 et (Pr>|t|) =0,005 pour le coefficient β1. Que pouvez-vous conclure ?

A. Le modèle n'est ni explicatif, ni significatif
B. Le modèle est significatif mais non explicatif
C. Le modèle est explicatif mais non significatif
D. Le modèle est explicatif et non significatif

QCM1 (C D) - QCM2 (B) - QCM3 (A B) - QCM4 (A B) - QCM5 (B C D) - QCM6 (D) - QCM7 (A) - QCM8 (A C) - QCM9 (C) - QCM10 (C).



7 RÉGRESSION MULTIPLE – RÉGRESSION LOGISTIQUE

7.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin de ce module, les participants seront capables de définir le modèle pour réaliser une régression linéaire multiple et une régression logistique, de formuler les différentes hypothèses pour ces deux régressions. Ils pourront réaliser assez facilement les régressions au niveau du logiciel SPSS grâce aux étapes formulées dans les processus.

7.2 ASPECT THÉORIQUE

En général, les modèles de régression sont construits dans le but d'expliquer (ou prédire, selon la perspective de l'analyse) la variance d'un phénomène (variable dépendante) à l'aide d'une combinaison de facteurs explicatifs (variables indépendantes).

7.2.1 RÉGRESSION MULTIPLE

Dans la régression linéaire multiple, la variable dépendante est toujours une variable continue tandis que les variables indépendantes peuvent être continues ou catégorielles. La régression linéaire est appelée multiple lorsque le modèle est composé d'au moins deux variables indépendantes. A l'inverse, un modèle de régression linéaire simple ne contient qu'une seule variable indépendante.

La régression linéaire multiple généralise l'approche adoptée dans la régression linéaire simple. Dans la régression multiple, le nombre de variables indépendantes est supérieur ou égal à 2, mais inférieur au nombre de situations (observations) considérées. Les variables indépendantes sont de types variées. Elles peuvent être quantitatives (par exemple l'âge ou le nombre d'années de scolarité) ou qualitatives. Les variables indépendantes ne doivent pas être corrélées entre elles.

Les questions auxquelles la régression linéaire multiple permet de répondre sont nombreuses. Par exemple : « est-ce que la satisfaction au travail varie en fonction de l'augmentation des défis à relever et de l'esprit d'équipe ? ».

7.2.1.1 Définir le modèle

De manière générale, les modèles statistiques se présentent globalement ainsi : **Observation$_i$: (Modèle$_i$) + erreur$_i$.** Chaque valeur de la variable dépendante (Observation) peut être expliquée en partie par un modèle statistique. La partie que le modèle ne peut pas expliquer est **l'erreur spécifique** associée à cette valeur.

L'équation de la régression linéaire multiple est en fait la généralisation du modèle de régression simple : $Y_i : (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon_{i/}$

On observe que chaque variable indépendante (**X**) est multipliée par son propre coefficient (**b**) qui sous sa forme standardisée correspond à sa contribution relative dans le modèle. La constante ($\mathbf{b_0}$) correspond à la valeur de la variable dépendante lorsque toutes les variables indépendantes égalent « 0 ». On appelle aussi $\mathbf{b_0}$ l'ordonnée à l'origine.

Exemple : *Revenu (Y) = b_0 + b_1 (expérience en année) + b_2 (niveau scolaire en année)*

7.2.1.2 Formulation des hypothèses de la régression linéaire multiple

Hypothèse nulle (H0). Il n'y a pas de relation linéaire entre la combinaison des variables indépendantes ($X_1, X_2, X_3, ..., X_n$) et la variable dépendante (Y).

Hypothèse alternative (H1). La combinaison des variables indépendantes est associée significativement à la variable dépendante.

7.2.1.3 Condition d'utilisation de la régression linéaire multiple

Avant d'utiliser la régression linéaire multiple, certaines conditions doivent être vérifiées :

- **Type de variables à utiliser** : les variables indépendantes doivent être continue ou catégorielle et la variable dépendante doit être continue ;
- **Pas de variance égale à 0** : la distribution des prédicteurs (variables indépendantes) doit comprendre une certaine variance, donc ne pas être constante ;
- **Aucune multicolinéarité parfaite** : il ne doit pas y avoir de relation linéaire parfaite entre deux ou plusieurs variables indépendantes. Par conséquent, les corrélations ne doivent pas être trop fortes entre celles-ci ;
- **Pas de corrélation entre les variables indépendantes et les variables externes** : les variables d'influence doivent être toutes incluses dans le modèle ;
- **Homoscédasticité** (homogénéité des variances des résiduels) : la variance des valeurs résiduelles doit être similaire à tous les niveaux de la variable indépendantes ;
- **Indépendance des erreurs** : les valeurs résiduelles ne doivent pas être corrélées entre les individus. Cette condition peut être vérifiée avec la statistique de **Durbin-Watson**. La règle arbitraire est que la valeur de Durbin-Watson ne doit pas être plus petite que « 1 » ou plus grande que « 3 » ;
- **Distribution normale des résiduels** : bien que les variables indépendantes ne doivent pas nécessairement suivre une loi normale, il importe que les résiduels en suivent une. Ils doivent donc avoir une moyenne de « 0 » ou la majorité des valeurs doivent s'en rapprocher ;
- **Indépendance de la variable prédite** : toute les observations formant la distribution des valeurs de la variable dépendante sont indépendantes et viennent d'un individu différent ;
- **Relation linéaire entre les variables indépendantes et la variable dépendante** : la variation de la variable dépendante pour chaque augmentation d'une unité d'une variable indépendante suit une ligne droite.

7.2.1.4 Conception d'un modèle de régression linéaire multiple

La conception d'un modèle de régression doit faire l'objet d'une réflexion préalable portant sur : **le choix des variables indépendantes et le choix de la méthode de régression**.

Choix des variables indépendantes

Le choix des variables indépendantes dit être guidé par le principe de parcimonie qui veut qu'un modèle comprenne un nombre optimal de variables et par la présence d'un lien théorique connu ou présumé avec la variable dépendante. Les quelques éléments importants à considérer lors du choix des variables indépendantes sont :

- **La nature des objectifs ou hypothèses de recherche :** les variables mises en cause dans l'énoncé d'une hypothèse ou d'un objectif doivent forcément se retrouver dans le modèle. L'énoncé peut également avoir un impact sur le choix de la méthode de régression ;

- **La présence de variables confondantes :** il est possible que certaines variables n'apparaissant pas dans l'énoncé de l'objectif ou de l'hypothèse soient importantes dans un modèle dans la mesure où elles peuvent influencer les résultats ;
- **La présence de corrélation avec la variable dépendante :** dans certains contextes, il est possible de choisir les variables indépendantes en fonction de leur degré d'association avec la variable dépendante. Des variables n'ayant pas de lien assez forts avec celle-ci pourrait être exclues du modèle ;
- **La puissance statistique du devis :** le nombre d'observations détermine la quantité maximale de variables qu'un modèle peut supporter. Plus on a d'observations, plus on peut inclure de variables dans le modèle.

Choix de la méthode de régression

La méthode choisie n'est pas la même lorsque l'on désire tester un modèle théorique précis, contrôler l'effet de variables confondantes ou tout simplement explorer une combinaison particulière de variables indépendantes. De même, la façon d'introduire les variables ou blocs de variables indépendantes dans ce modèle doit faire également l'objet d'une justification rationnelle.

Dans un premier temps, on doit choisir une des deux stratégies suivantes :

- **La modélisation globale** : la combinaison de toutes les variables est évaluée globalement.
- **La modélisation par blocs** : les variables sont regroupées en bloc et les résultats évaluent le modèle global ainsi que la contribution de chaque bloc.

Dans un second temps, on doit déterminer la manière dont les variables indépendantes seront insérées dans le modèle global ou dans les blocs : **par entrée forcée** ou **par entrée progressive**.

Type de méthode de régression

7.2.1.5 La régression hiérarchique

Cette méthode permet au chercheur de déterminer l'ordre d'entrée des variables dans le modèle à l'aide de la création des blocs de variables qui seront entrées de manière hiérarchique dans le modèle. Cela permet d'observer plus en détail comment se comporte le modèle.

Les résultats indiquent l'apport de chaque bloc en termes de pourcentage de variance expliquée ($\mathbf{R^2}$). Pour les blocs constitués de plus d'une variable, il est possible de faire entrer celle-ci en un seul temps (entrée forcée) ou progressivement.

7.2.1.6 La régression avec entrée forcée

Toutes les variables indépendantes sont introduites simultanément et un test **F** évalue l'ensemble du modèle. Cette méthode doit être utilisée si l'on veut déterminer l'équation de la droite de régression avec toutes les variables indépendantes. Elle est exclusivement utilisée si l'on pense qu'une des variables est plus importante que les autres.

Le choix des variables à inclure repose sur la théorie, cependant, le chercheur n'influence pas l'ordre d'entrée des variables.

7.2.1.7 La régression avec entrée progressive

Contrairement aux deux autres méthodes, la sélection des variables à inclure est basée sur un critère mathématique. Une fois les variables indépendantes choisies, leur inclusion dans le modèle dépendra de leur contribution mathématique à son amélioration. Il existe 3 méthodes

progressives.

- **Méthode ascendante (forward).** Dans ce cas, le modèle initial ne contient que la constante (b_0). Celui-ci servira de base de comparaison pour déterminer si l'ajout d'une variable contribue significativement à l'amélioration du modèle. SPSS choisit parmi les variables indépendantes soumises, celle qui a la plus forte corrélation avec la variable dépendante et évalue si cet ajout est significatif :
  - ✓ Si l'ajout est significatif, SPSS intègre une deuxième variable. Cette dernière a la plus forte corrélation partielle avec la variable dépendante. Le logiciel évalue encore si l'ajout de cette variable est significatif. Si c'est le cas, il la retient et détermine s'il peut ajouter un 3^{ième} prédicteur.
  - ✓ SPSS cesse d'inclure des nouvelles variables lorsque l'augmentation de la valeur de R^2 n'est plus significative
- **Méthode pas-à-pas (stepwise).** Elle ressemble beaucoup à la méthode ascendante puisque le choix de la première variable est basé sur la corrélation la plus élevée et celui des variables suivantes sur la corrélation partielle. Toutefois, lorsque SPSS ajoute une variable au modèle, il évalue si elle apporte une contribution significative, mais également si celle qui contribuait le moins au modèle demeure significative. Si c'est le cas, il la retire. De cette manière, il est possible d'éliminer les variables redondantes.
- **Méthode descendante (backward).** Le modèle initial comprend toutes les variables, comme pour la régression forcée. SPSS va cette fois retirer la variable ayant la plus faible contribution au modèle si la variation du **R^2** n'est pas significative en l'éliminant. La procédure va être répétée jusqu'à ce que toutes les variables conservées contribuent significativement à l'amélioration du **R^2**. Si le retrait de la variable affaiblit significativement le modèle, elle est réintroduite. On répète la procédure jusqu'à ce que l'on n'ait que les variables utiles.

7.2.1.8 Type de méthode de régression à privilégier

De manière générale, on suggère qu'un modèle bien balisé par la théorie devrait utiliser une stratégie globale avec une ***méthode d'entrée forcée, hiérarchisée ou non***. Pour les travaux de nature davantage exploratoire, les méthodes progressives sont adaptées. Parmi les 3 méthodes présentées, ***on privilégiera la méthode descendante***, car il y a plus de risques de commettre des erreurs de type II avec la méthode ascendante.

La régression hiérarchisée est intéressante lorsque le modèle comporte plusieurs variables qui peuvent être théoriquement regroupées ou lorsque certaines variables doivent être contrôlées statistiquement (exemple : variables socioéconomiques). SPSS permet de regrouper ces variables en « blocs » dont l'ordre d'inclusion devrait représenter leur position relative par rapport à la variable dépendante. SPSS donne les résultats pour le modèle global (toutes les variables) ainsi que l'apport spécifique de chaque bloc une fois l'effet du bloc précédent considéré.

7.2.2 REGRESSION LOGISTIQUE

La régression logistique est une méthode d'analyse statistique qui consiste à prédire une valeur de données d'après les observations réelles d'un jeu de données. Un modèle de régression logistique prédit une variable de données dépendante en analysant la relation entre une ou plusieurs variables indépendantes. Par exemple, la régression logistique pourrait répondre à la question suivante : « quelles sont les chances d'admission d'un bachelier à une grande école particulière ? ».

Le modèle analytique qui en résulte peut tenir compte de plusieurs critères en entrée. Dans cet exemple, il pourrait s'agir de la moyenne générale de l'élève, de ses notes aux épreuves de baccalauréat et du nombre d'activités parascolaires.

Lorsque la variable dépendante est une variable qualitative (nominale ou ordinale), l'approche par régression « classique » telle que présentée dans la section précédente est inadéquate.

La question qui se pose est la suivante :

- A quelle catégorie de la variable qualitative renvoient les valeurs prises par les variables explicatives ?
- Selon quelles probabilités les valeurs prises par les variables explicatives renvoient-elles aux différentes catégories de la variable qualitative dépendante ?

Les fonctions qui associent les variables explicatives aux probabilités d'occurrence des catégories d'une variable qualitative dépendante doivent au minimum respecter certains critères :

- Elles doivent faire en sorte que la variable dépendante prenne ses valeurs entre « 0 » et « 1 », puisqu'il s'agit de probabilités ;
- Elles doivent tenir compte du fait qu'elles renvoient à des catégories exhaustives et exclusives ;
- Les probabilités relatives à l'occurrence de chaque catégorie sont soumises à la contrainte que la somme des probabilités est égale à « 1 ».

La fonction dite « logistique » est d'usage courant en raison de ses caractéristiques et parce qu'elle se plie relativement bien à l'estimation de ses paramètres.

Trois cas seront pris en considération :

- Le cas où la variable dépendante qualitative comporte deux catégories exhaustives et exclusives (exemple : la situation nutritionnelle : malnutri, bien nourri) ;
- Le cas où la variable dépendante qualitative nominale comporte plus de deux catégories (exemple : résider à Niamey, dans une ville de l'intérieur ou en milieu rural) ;
- Le cas où la variable qualitative dépendante est ordinale (exemple : malnutrition sévère, malnutrition modérée, bien nourri).

7.2.2.1 Définir le modèle

La régression logistique propose de tester un modèle de régression dont la variable dépendante est dichotomique (codée « 0 » / « 1 ») et dont les variables indépendantes peuvent être continues ou catégorielles. Le modèle de régression logistique ressemble à celui de la régression linéaire, mais on y ajoute la transformation logarithmique :

$$Ln\left(\frac{p}{1-p}\right) = logit(p) = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \ldots + \beta_n * X_n + \epsilon$$

Le modèle de régression logistique multiple : le logit de la probabilité (**p**) de la réalisation de la variable à expliquer (***Y***) est exprimé en fonction d'un intercept (ou ordonnée à l'origine) β_0, des variables explicatives (**Xi**) rattachées à leurs coefficients β_i et à un terme de bruit **ε**.

7.2.2.2 Formulation des hypothèses de la régression logistique

Hypothèse nulle (H0). La combinaison des variables indépendantes (le modèle) ne parvient pas à mieux expliquer la présence / absence de la variable dépendante qu'un modèle sans prédicteur.

Hypothèse alternative (H1). Au moins un prédicteur du modèle est associé significativement à la variable dépendante.

7.2.2.3 Condition d'utilisation de la régression logistique

Avant d'utiliser la régression logistique, certaines conditions doivent être vérifiées :

- **Types de variables à utiliser :** les variables indépendantes (prédicteurs) doivent être continues ou catégorielles dichotomiques. La variable dépendante (prédite) doit être catégorielle dichotomique. Cette dernière doit être une vrai variable dichotomique et non une variable continue recodée en 2 groupes, ce qui serait associé à une importante perte d'information.
- **Inclure les variables pertinentes :** toutes les variables pertinentes doivent être comprises dans le modèle et celles qui ne le sont pas doivent être éliminées.
- **Indépendance des observations et des résiduels :** un individu ne peut pas faire partie de 2 groupes de variables dépendantes.
- **Relation linéaire entre les variables indépendantes et la transformation logistique de la variable dépendante.**
- **Aucune multicolinéarité parfaite ou élevée :** il ne doit pas y avoir de relation linéaire parfaite, ni très élevée entre deux ou plusieurs prédicteurs. Par conséquent, les corrélations ne doivent pas être trop fortes entre ceux-ci.
- **Pas de valeurs extrêmes des résiduels :** comme dans la régression multiple, des valeurs résiduelles standardisées plus élevées que 2,58 ou moins élevées que -2,58 influencent les coefficients du modèle et limitent la qualité de l'ajustement.
- **Taille de l'échantillon :** l'échantillon doit être suffisant pour que l'on puisse procéder à l'analyse.
- **Echantillon adéquat pour les prédicteurs catégoriels :** lorsqu'une variable indépendante est croisée avec la variable dépendante, aucune cellule ne doit avoir moins d'une observation et un maximum de 20 % des cellules peuvent comprendre 5 observations en moins.

7.2.2.4 Conception d'un modèle de régression logistique

L'estimation du modèle Logit s'effectue généralement par la **méthode du maximum de vraisemblance**. Il s'agit dans un premier temps de définir la fonction de vraisemblance **L**, représentant la probabilité d'observer les données d'échantillon sous l'hypothèse que le modèle est vrai. La procédure revient à choisir à l'aide d'un processus itératif les estimations des paramètres qui permettent de maximiser la fonction **L**.

Les logiciels fournissent en même temps des informations « omnibus » permettant d'évaluer globalement la qualité des estimations obtenues, des informations plus spécifiques sur les coefficients estimés, leur erreur type et les tests associés et des informations permettant d'évaluer la distribution des résidus.

Méthodes de régression logistique

Les méthodes de régression disponibles sont les mêmes que pour la régression linéaire. Toutefois, le critère de sélection pour les méthodes progressives est différent.

- Vous pouvez opter pour la **méthode Entrée** et insérer toutes les variables prédictrices en même temps. De plus, si vous préférez sélectionner l'ordre d'entrée des variables, choisissez la **méthode hiérarchique**. Les paramètres seront calculés pour chaque bloc de variables.

Parmi les méthodes progressives, le choix entre **ascendante ou descendante** est donnée.

- Dans **la méthode ascendante**, SPSS introduit la variable ayant le score le plus élevé en premier jusqu'à ce qu'aucune variable n'ait un score statistiquement significatif (soit plus petit que 0,05).
- Dans la **méthode descendante**, le contraire se produit puisque le premier modèle évalué contient toutes les variables et SPSS retire celles qui ne contribuent pas significativement à l'amélioration de la prédiction.

La différence pour la régression multiple est que SPSS évalue à chaque étape si certaines variables devraient être retirées en se basant sur :

- **Le rapport de vraisemblance (likelihood-ratio, LR)** : SPSS conserve la variable si le changement du **LR** est significatif quand la variable est retirée, ce qui indique que cette variable contribue à la qualité de l'ajustement.
- **La statistique conditionnelle :** il s'agit d'un critère moins exigeant que le **LR**, il est donc préférable de prioriser le 1er.
- **La statistique de Wald :** SPSS retire toutes les variables pour lesquelles la statistique Wald est inférieur à 0,1. Cette méthode peut être utilisée avec un petit échantillon ; sinon, il est préférable de privilégier le **LR**.

7.2.3 REGRESSION LINEAIRE MULTIPLE VS REGRESSION LOGISTIQUE

La principale différence entre la régression logistique et la régression linéaire multiple est que la régression logistique fournit un résultat constant, tandis que la régression linéaire fourni un résultat continu. Dans la régression logistique, le résultat est tel qu'une variable dépendante n'a qu'un nombre limité de valeurs possibles. Cependant, en régression linéaire, le résultat est continu, ce qui signifie qu'il peut avoir n'importe laquelle parmi un nombre infini de valeurs possibles.

La régression logistique est utilisée lorsque la variable réponse est catégorique (oui/non ; vrai/faux ; réussite/échec). La régression linéaire est utilisée lorsque la variable réponse est continue, comme le nombre d'heures, la taille et le poids.

7.3 SYNTAXE SPSS

7.3.1 SYNTAXE SPSS POUR LA RÉGRESSION LINÉAIRE MULTIPLE

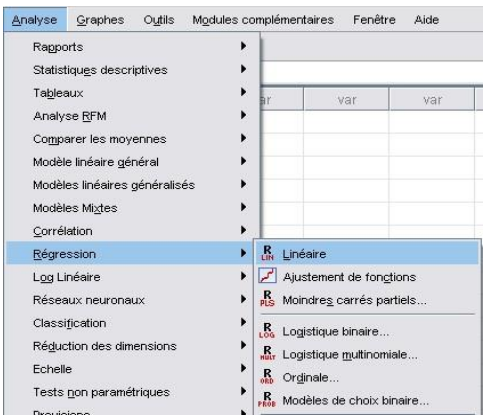
7.3.1.1 Formalisation du processus

Etape 1.

Ouvrir votre matrice de données sous SPSS

Etape 2.

Cliquer sur **Analyse** > **Régression** > **Linéaire.**

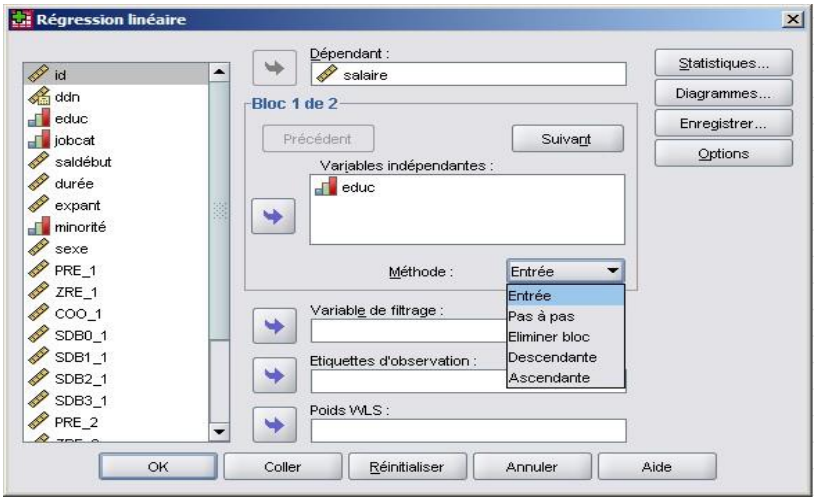


Etape 3.

Cliquer sur  pour insérer la variable dépendante et les variables indépendantes dans la boîtes appropriées.

Etape 4.

Si vous désirez absolument que la première variable indépendante soit incluse, privilégiez la méthode « **Entrée** ».

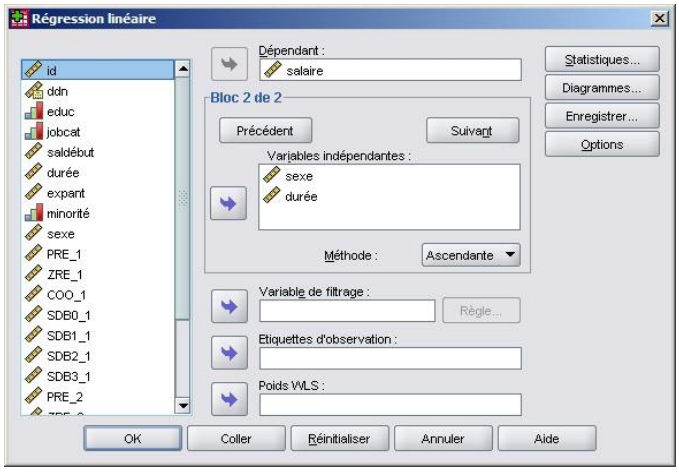


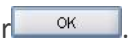
Etape 5. Vous pouvez choisir une variable de filtrage pour limiter l'analyse à un sous-échantillon formé par les participants ayant obtenu une ou des valeurs particulières à cette même variable.

Etape 6. Vous pouvez aussi spécifier une variable qui permettra d'identifier les coordonnées sur le graphique (**Etiquette d'observation**).

Etape 7.

Vous pouvez choisir une variable numérique pondérée (Poids WLS) pour effectuer l'analyse des moindres carrés. Par cette analyse, les valeurs sont pondérées en fonction de leurs variances réciproques, ce qui implique que les observations avec de larges variances ont un impact moins important sur l'analyse que les observations associées à de petites variances.



Etape 8. Pour procéder à l'analyse cliquer sur  .

7.3.1.2 Interprétation des résultats

Les différentes étapes pour l'interprétation des résultats d'un modèle de régression linéaire multiple seront ponctuées par l'introduction de résultats d'un exemple permettant de mieux s'orienter lors des interprétations.

Nous voulons savoir quelles variables influencent le salaire annuel d'un employé (SALAIRE). La théorie indique que le Nombre d'années de scolarité a une importante influence (EDUC). Aussi nous voulons savoir si le Sexe des employés (SEXE) et le Nombre de mois d'expérience dans l'entreprise (DUREE) exercent également une influence.

Etape 1. **Evaluation de la qualité du modèle de régression**

Tout comme la régression simple, l'interprétation débute en évaluant la qualité du modèle. On vérifie si la première étape du modèle explique significativement plus de variabilités qu'un modèle sans prédicteurs (variables indépendante). Ensuite, il s'agit de s'assurer que toutes les variables

introduites contribuent à améliorer significativement la variabilité expliquée par le modèle final.

Analyse de variance

Le tableau ANOVA permet de déterminer si l'hypothèse nulle (H0) est rejetée ou non.

Dans l'exemple, nous voulons savoir dans un 1er temps si le nombre d'années de scolarité prédit mieux le SALAIRE que ne le fait un modèle sans prédicteur (avec seulement la moyenne). Dans un 2ième temps, si le nombre d'années de scolarité et le sexe prédisent mieux le SALAIRE qu'un modèle sans prédicteur.

L'hypothèse nulle (H0), est donc que les 2 modèles sont équivalents à la moyenne du salaire.

ANOVA[c]

Modèle		Somme des carrés	ddl	Moyenne des carrés	F	Sig.
1	Régression	60178217760,000	1	60178217760,000	365,381	,000[a]
	Résidu	77738277676,340	472	164699740,840		
	Total	137916495436,340	473			
2	Régression	67463175600,070	2	33731587800,035	225,505	,000[b]
	Résidu	70453319836,269	471	149582420,035		
	Total	137916495436,340	473			

a. Valeurs prédites : (constantes), educ
b. Valeurs prédites : (constantes), educ, sexe
c. Variable dépendante : salaire

On constate à la lecture du tableau que selon la valeur **F** obtenue pour les deux modèles, on peut rejeter l'hypothèse nulle (H0) au seuil de 1 %.

Etape 2. Evaluation de l'ajustement du modèle de régression aux données

Maintenant que l'on sait que le modèle est significatif, le tableau récapitulatif des modèles permet de déterminer la contribution de chaque bloc de variables.

Ce tableau indique le **R^2** cumulatif à chaque étape du modèle (colonne R-deux) ainsi que l'apport spécifique de chaque bloc (colonne Variation de R-deux).

Récapitulatif des modèles[c]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Changement dans les statistiques					Durbin-Watson
					Variation de R-deux	Variation de F	ddl1	ddl2	Sig. Variation de F	
1	,661[a]	,436	,435	12833,540	,436	365,381	1	472	,000	
2	,699[b]	,489	,487	12230,389	,053	48,702	1	471	,000	1,956

a. Valeurs prédites : (constantes), educ
b. Valeurs prédites : (constantes), educ, sexe
c. Variable dépendante : salaire

- **La valeur de la corrélation multiple (R)** représente la force de la relation entre la variable dépendante et la combinaison des variables indépendantes de chaque modèle. Les valeurs de **R** suggèrent que les données sont ajustées de manières satisfaisantes au modèle (0,661 et 0,699).
- Ce modèle explique une proportion significative de la variance de la variable SALAIRE. Nous sommes passés de **R^2**=0 à **R^2**=0,436. Le deuxième modèle fait passer le **R^2** de 0,436 à 0,489. Cette variation est de 0,053 (0,489-0,436). SPSS détermine si la différence (0,053) entre le **R^2** du modèle 2 (0,489) et celui du modèle 1 (0,436) est significative : c'est le cas ($p < 0,001$). Chaque étape contribue donc significativement à l'amélioration de l'explication de la variabilité de la variable dépendante.
- La dernière colonne concerne le **test de Durbin-Watson**, il n'y a pas de seuil de signification associée, seulement la valeur de la statistique est acceptable lorsqu'elle se situe entre 1 et 3.

Etape 3. Evaluation de la variabilité expliquée par le modèle de régression

La valeur du **R^2** lorsqu'elle est multipliée par 100, indique le pourcentage de variabilité de la variable dépendante expliquée par le modèle (les prédicteurs). Les résultats suggèrent que 43,6 % du salaire est expliqué par le nombre d'année de scolarité et que 48,7 % du salaire est expliqué par la combinaison de la scolarité et du sexe de l'employé.

Étape 4. Évaluation des paramètres du modèle

Maintenant que nous savons que notre modèle est significatif et que le deuxième est celui qui explique le plus de variance, il est possible de construire l'équation de régression pour prédire une valeur de **Y**. L'équation de base était la suivante : $Y_i : (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon_i$.

Remplaçons maintenant les **b** par les coefficients fournis dans le tableau.

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	95,0% % intervalles de confiance pour B	
		B	Erreur standard	Bêta			Borne inférieure	Limite supérieure
1	(Constante)	-18331,178	2821,912		-6,496	,000	-23876,242	-12786,114
	educ	3909,907	204,547	,661	19,115	,000	3507,971	4311,842
2	(Constante)	-7500,990	3104,940		-2,416	,016	-13602,237	-1399,742
	educ	3391,683	208,599	,573	16,259	,000	2981,784	3801,582
	sexe	-8423,462	1207,028	-,246	-6,979	,000	-10795,289	-6051,636

$Y_{prédit}$ = (-7500,99 + 3391,68educ – 8423,46sexe). Pour un homme ayant complété 16 années de scolarité, nous obtiendrions un salaire prédit de : $Y_{prédit}$ = (-7500,99 3391,68*16 – 8423,46*1), soit **$Y_{prédit}$ = 37 942,43 francs CFA/Année**

- Le signe du coefficient nous indique le sens de la relation. Dans le cas de l'exemple, plus le nombre d'années de scolarité augmente, plus le salaire augmente. Aussi, quand le sexe diminue (passant de « 2 » pour les femmes à « 1 » pour les hommes), le salaire augmente.
- Le coefficient informe sur le degré auquel chaque prédicteur influence la variable dépendante si tous les autres prédicteurs sont constants. Par exemple, chaque année de scolarité de plus est associé à 3 391,68 francs CFA de plus annuellement.
- L'erreur standard renseigne sur la variabilité du coefficient dans la population. Elle permet également de calculer la valeur de **t** qui indique si le coefficient est significatif. La signification de « **t** » nous permet de répondre à la question « est-ce-que le **b** du prédicteur est différent de 0 ? », donc si chaque variable contribue significativement au modèle. Plus la valeur de « **t** »est élevée et plus celle de « **p** » est petite, plus le prédicteur contribue au modèle.

Dans notre exemple, les 2 variables sont significatives, mais la variabilité expliquée par le nombre d'années se scolarité est plus importante que celle expliquée par le sexe.

- La valeur du bêta standardisé (**β**) indique le changement de la variable dépendant pour chaque augmentation d'un écart-type de la variable indépendante quand toutes les autres valeurs sont constantes.

Ce tableau présente la valeur des corrélations et des corrélations partielles. Ce sont ces valeurs sur lesquelles se base SPSS lorsqu'il choisit d'introduire des variables quand on sélectionne une méthode d'entrée progressive.

Modèle		Corrélations			Statistiques de colinéarité	
		Corrélation simple	Partielle	Partie	Tolérance	VIF
1	(Constante)					
	educ	,661	,661	,661	1,000	1,000
2	(Constante)					
	educ	,661	,600	,535	,873	1,145
	sexe	-,450	-,306	-,230	,873	1,145

a. Variable dépendante : salaire

La 1ère variable est choisie à partir de la corrélation simple la plus forte (ici 0,661 pour EDUC). Le choix des variables suivantes est basé sur la corrélation partielle, c'est-à-dire la plus forte corrélation entre les variables toujours disponibles et la partie de variance qui reste à expliquer une fois que l'on a retiré ce qui est expliqué par le 1er prédicteur.

7.3.2 SYNTAXE SPSS POUR LA RÉGRESSION LOGISTIQUE

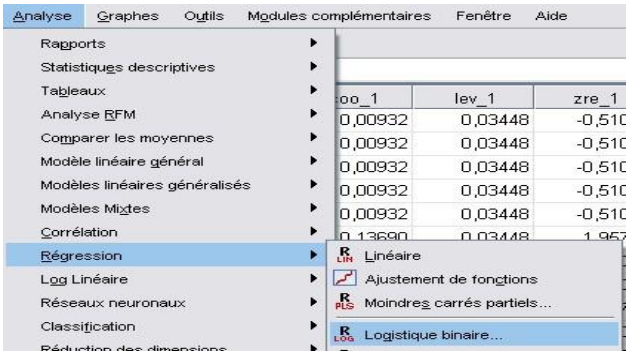
7.3.2.1 Formalisation du processus

Etape 1.

Ouvrir votre matrice de données sous SPSS

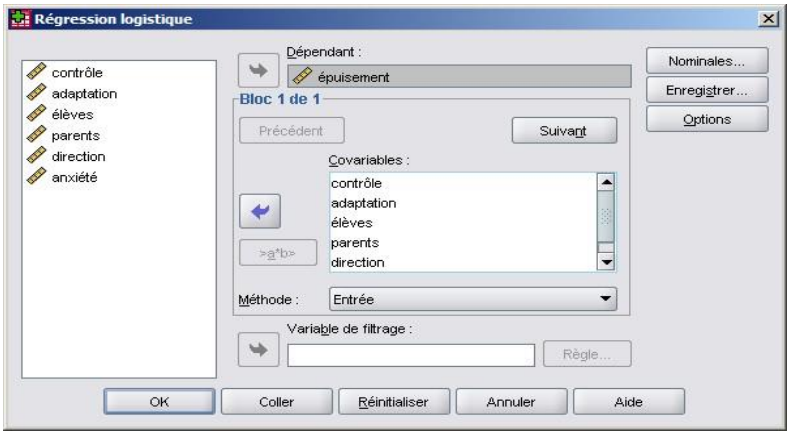
Etape 2.

Cliquer sur **Analyse** > **Régression** > **Logistique binaire**



Etape 3.

Dans la première boîte dialogue, insérer la variable dépendante dichotomique dans la boîte **Dépendant** et les variables prédictrices dans la boîte Covariables.



Etape 4. Si vous voulez insérer un terme d'interaction entre 2 prédicteurs, sélectionner les 2 variables dans la boîte de gauche et cliquer sur le bouton >a*b>.

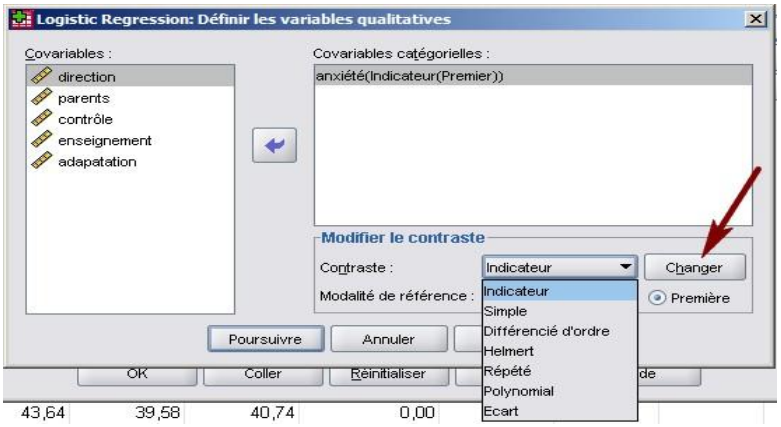
Etape 5.

Choisir ensuite la méthode de régression. Et cliquer sur OK

Le bouton Nominales...

Ce bouton permet d'indiquer à SPSS quelles variables soumises sont catégorielles dichotomiques. Il faut les insérer dans la boîte « **Covariance catégorielle**». Le « Contraste » par défaut est « **Indicateur** ». Il est à recommander. Il s'agit de la technique de base pour catégoriser une variable catégorielle à plusieurs groupes, en plusieurs variables n'ayant que les valeurs « 0 » et « 1 ».

Il est toutefois possible d'opter pour les autres contrastes en les sélectionnant dans le menu déroulant.

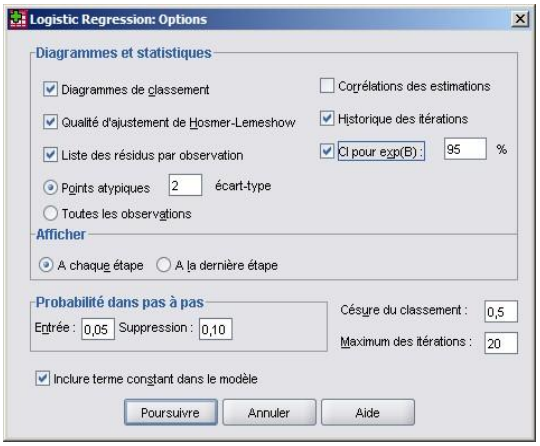


Le bouton Enregistrer... permet d'enregistrer en tant que nouvelles variables certaines informations d'intérêt, telles que les résiduels standardisés, les prévisions probabilités (valeurs **P(Y)** prédites à partir de l'équation) et les prévisions groupe d'affectation (prédiction du groupe dans lequel les individus seront inclus en fonction du modèle). Le résidu « **logit** » est calculé à partir des coefficients logit. Ces nouvelles variables serviront essentiellement à examiner la qualité d'ajustement du modèle.

Le bouton Options comprend les options pour réaliser l'analyse ainsi que les tableaux qui peuvent s'afficher dans les résultats. Il faut conserver les options par défaut.

D'autres options disponibles peuvent être aussi pertinentes :

- **Diagramme de classement** : histogramme illustrant les valeurs prédites et observées ; il permet de juger de l'ajustement du modèle.
- **Qualité de l'ajustement de Hosmer-Lemeshow** : test évaluant s'il y a une différence significative entre les valeurs observées et les valeurs prédites.
- **Liste des résidus par observation** : soit pour les valeurs extrêmes, soit pour toutes les observations



7.3.2.2 Interprétation des résultats

Les différentes étapes pour l'interprétation des résultats d'un modèle de régression logistique seront ponctuées par l'introduction de résultats d'un exemple permettant de mieux s'orienter lors des interprétations.

Nous cherchons à identifier les variables qui permettent de prédire le plus efficacement la probabilité de vivre un épuisement professionnel chez les enseignants (EPUISEMENT). Nous vérifions donc l'effet du stress généré par les élèves (ELEVES), par les parents (PARENTS) et par la Direction (DIRECTION), le sentiment d'autocontrôle (CONTROLE), les stratégies d'adaptation (ADAPTATION) et la présence d'un trouble anxieux (ANXIETE) sur la présence ou non d'un épuisement professionnel. Toutes les variables prédictrices évaluées sont continues, mises part la variable « ANXIETE » qui est catégorielle.

Etape 1 : Le modèle de base

Ce 1er tableau indique que SPSS a conservé les mêmes valeurs que celles utilisées pour coder les variables.

Codage de variables dépendantes

Valeur d'origine	Valeur interne
0 Non épuisé	0
1 Épuisé	1

Ce second tableau illustre les valeurs utilisées pour la variable prédictrice catégorielle. Puisque nous avons choisi le contraste « indicateur », nous conservons également les mêmes valeurs que pour coder la variable.

Codages des variables nominales

		Fréquence	Codage des paramètres (1)
anxiété	,00	406	,000
	1,00	61	1,000

Ce troisième tableau présente l'historique des itérations pour le modèle de base. Nous retenons particulièrement la probabilité **log** (-2LL) initiale. Dans l'exemple, **-2LL** = 530,107 et c'est cette probabilité que nous cherchons à améliorer en ajoutant des variables prédictrices.

Historique des itérations[a,b,c]

Itération		-2log-vraisemblance	Coefficients Constant
Etape 0	1	530,875	-,981
	2	530,108	-1,071
	3	530,107	-1,073
	4	530,107	-1,073

a. La constante est incluse dans le modèle.

b. -2log-vraisemblance initiale : 530,107

c. L'estimation a été interrompue au numéro d'itération 4 parce que les estimations de paramètres ont changé de moins de ,001.

Le tableau des variables dans l'équation nous indique la valeur du coefficient **b0**. Dans le cas de notre exemple, b0 = 1,073.

Variables dans l'équation

		B	E.S.	Wald	ddl	Sig.	Exp(B)
Etape 0	Constante	-1,073	,106	102,111	1	,000	,342

Etape 2. Evaluation de la signification du modèle de régression

Le tableau récapitulatif des modèles fournit les valeurs **-2LL** pour chaque étape du modèle. Nous pouvons déterminer si la probabilité -2LL de chaque étape du modèle est inférieur à la probabilité -2LL de base (530,111) et si cette différence est significative. Ceci nous indiquera si les termes de l'équation logistique finale prédisent mieux la probabilité de vivre en « EPUISEMENT » professionnel que ne le fait la probabilité initiale observée.

Pour l'étape 1, nous pouvons calculer 530,11 – 399,033, ce qui donne 131,74. Cette valeur est évaluée dans une distribution χ^2 et sa signification est présentée dans le tableau tests de spécification du modèle.

Récapitulatif des modèles

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	399,033[a]	,245	,361
2	364,179[a]	,299	,441
3	336,770[a]	,339	,500
4	324,710[b]	,356	,524

a. L'estimation a été interrompue au numéro d'itération 5 parce que les estimations de paramètres ont changé de moins de ,001.

b. L'estimation a été interrompue au numéro d'itération 6 parce que les estimations de paramètres ont changé de moins de ,001.

Dans les étapes suivantes du présent tableau, la ligne « étape » et la ligne « modèle » n'indiquent pas les mêmes valeurs. La ligne « étape » montre en effet la différence entre la probabilité -2LL de l'étape précédente et celle obtenue par l'ajout du nouveau prédicteur. Nous cherchons à ce qu'à chaque étape, le modèle présente une diminution significative du -2LL.

Tests de spécification du modèle

		Khi-Chi-deux	ddl	Sig.
Etape 1	Etape	131,074	1	,000
	Bloc	131,074	1	,000
	Modèle	131,074	1	,000
Etape 2	Etape	34,853	1	,000
	Bloc	165,928	2	,000
	Modèle	165,928	2	,000
Etape 3	Etape	27,409	1	,000
	Bloc	193,337	3	,000
	Modèle	193,337	3	,000
Etape 4	Etape	12,060	1	,001
	Bloc	205,397	4	,000
	Modèle	205,397	4	,000

A la lumière de ces 2 tableaux, nous pouvons dire que le modèle final permet de prédire significativement mieux la probabilité de vivre un épuisement professionnel que le fait le modèle incluant seulement la constante.

Examinons le test de Hosmer-Lemeshow qui indique s'il existe un écart important entre les valeurs prédites et observées. Nous constatons à la lecture du tableau qu'il existe une différence significative entre les valeurs prédites et observée pour les étapes 1 à 3. Mais lorsque la 4ème variable est introduite, les valeurs prédites et observées sont cohérentes

Test de Hosmer-Lemeshow

Etape	Khi-Chi-deux	ddl	Sig.
1	50,018	8	,000
2	34,438	8	,000
3	15,972	8	,043
4	11,489	8	,176

Etape 3. Evaluation de l'ajustement des données au modèle de régression

Il faut évaluer la signification statistique des coefficients estimés des variables indépendantes conservées afin de s'assurer que chacune contribue à mieux prédire **P(y)** qu'un modèle qui ne l'inclurait pas. Pour ce faire, nous nous basons sur la statistique **Wald**. Cette dernière illustre la différence dans le modèle avant et après l'ajout de la dernière variable. On observe qu'à l'étape finale, tous les coefficients sont significatifs, même si plusieurs variables ont été introduites. On rejette donc pour chaque variable que le coefficient est égal à « 0 ». Par conséquent, chacune contribue à l'amélioration du modèle.

Le sens des coefficients **β** et de **Exp(β)** indiquent le sens de la relation. On constate que la relation est positive pour les variables « contrôle », « élèves » et « direction ». Ainsi, le faible sentiment de contrôle et le stress engendré par les élèves et la direction prédisent l'épuisement professionnel.

Variables dans l'équation

		B	E.S.	Wald	ddl	Sig.	Exp(B)	IC pour Exp(B) 95% Inférieur	IC pour Exp(B) 95% Supérieur
Etape 1[a]	élèves	,086	,009	90,613	1	,000	1,090	1,071	1,110
	Constante	-3,384	,283	142,632	1	,000	,034		
Etape 2[b]	contrôle	,061	,011	31,316	1	,000	1,063	1,040	1,086
	élèves	,083	,009	77,950	1	,000	1,086	1,066	1,106
	Constante	-4,484	,379	139,668	1	,000	,011		
Etape 3[c]	contrôle	,092	,014	46,340	1	,000	1,097	1,068	1,126
	adaptation	-,083	,017	23,962	1	,000	,921	,890	,952
	élèves	,131	,015	76,877	1	,000	1,139	1,107	1,173
	Constante	-1,707	,619	7,599	1	,006	,181		
Etape 4[d]	contrôle	,107	,015	52,576	1	,000	1,113	1,081	1,145
	adaptation	-,110	,020	31,660	1	,000	,896	,862	,931
	élèves	,135	,016	75,054	1	,000	1,145	1,110	1,181
	direction	,044	,013	11,517	1	,001	1,045	1,019	1,071
	Constante	-3,023	,747	16,379	1	,000	,049		

a. Variable(s) entrées à l'étape 1 : élèves.
b. Variable(s) entrées à l'étape 2 : contrôle.
c. Variable(s) entrées à l'étape 3 : adaptation.
d. Variable(s) entrées à l'étape 4 : direction.

Par contre la relation est négative pour la variable « adaptation », c'est donc dire meilleures sont les stratégies d'adaptation de l'enseignant face au stress, moins il est probable qu'il vive un épuisement professionnel.

Le présent tableau permet d'évaluer à chaque étape la présence d'un changement significatif de la probabilité -2LL lorsqu'une variable est retirée du modèle (la valeur doit être significative pour que la variable soit conservée).

Modèle si terme supprimé

Variable		Modèle log-vraisemblance	Modification dans -2log-vraisemblance	ddl	Signification de la modification
Etape 1	élèves	-265,054	131,074	1	,000
Etape 2	contrôle	-199,516	34,853	1	,000
	élèves	-236,237	108,294	1	,000
Etape 3	contrôle	-196,546	56,322	1	,000
	adaptation	-182,090	27,409	1	,000
	élèves	-231,036	125,301	1	,000
Etape 4	contrôle	-195,158	65,606	1	,000
	adaptation	-181,548	38,386	1	,000
	élèves	-224,912	125,113	1	,000
	direction	-168,385	12,060	1	,001

Etape 4 : Evaluation de l'ajustement du modèle final

Nous savons maintenant que le modèle final est significatif et que chacune des variables indépendantes contribue à mieux prédire **P(y)** qu'un modèle qui ne les inclut pas.

Nous nous intéressons maintenant à savoir si le modèle est bien ajusté aux données. Pour ce faire, revenons au tableau récapitulatif du modèle pour voir les valeurs des **R^2**. Comme le **R^2** de la régression multiple, ***plus la valeur est élevée, mieux le modèle est ajusté aux données***. Nous observons que la valeur augmente pour chaque étape et pouvons conclure que le modèle final est mieux ajusté.

Il est possible de calculer la valeur du **Pseudo-R^2** pour obtenir un estimé de la variabilité expliquée :$3{,}39 = \frac{530{,}107-324{,}710}{530{,}107}$

Le modèle final prédit donc 39 % de la variance de la probabilité de vivre un épuisement professionnel.

Récapitulatif des modèles

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	399,033[a]	,245	,361
2	364,179[a]	,299	,441
3	336,770[a]	,339	,500
4	324,710[b]	,356	,524

a. L'estimation a été interrompue au numéro d'itération 5 parce que les estimations de paramètres ont changé de moins de ,001.
b. L'estimation a été interrompue au numéro d'itération 6 parce que les estimations de paramètres ont changé de moins de ,001.

Etape 5. Evaluation de la justesse de l'ajustement du modèle

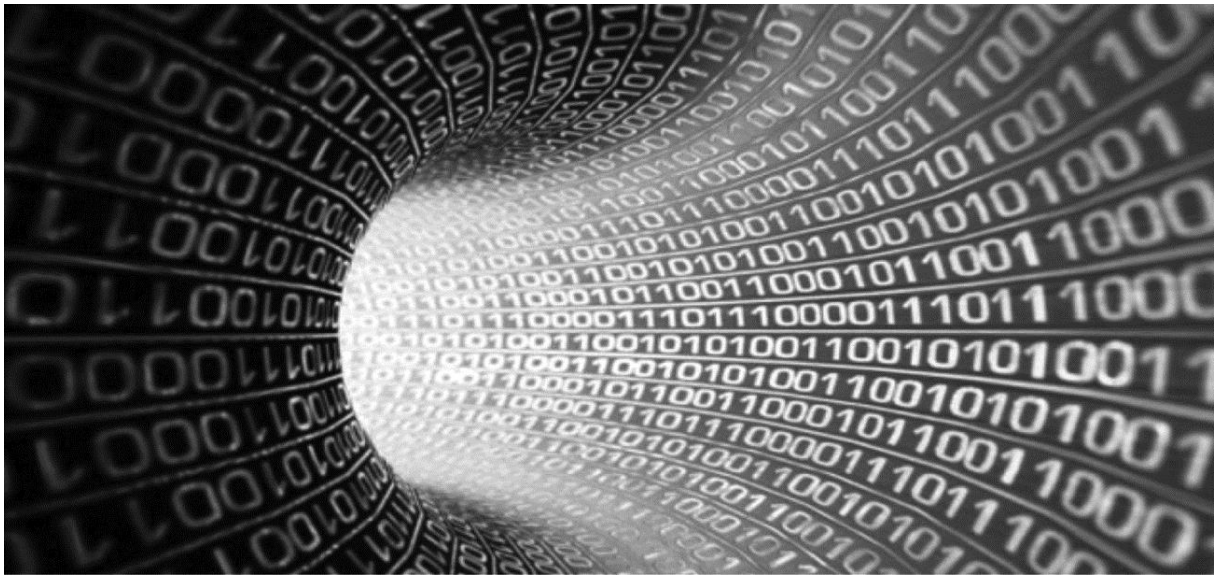
Il est possible maintenant d'examiner si le modèle permet de bien classer les sujets dans leur groupe d'appartenance à partir de l'équation finale. Le pourcentage correct de classification passe de 78,8 % avec une seule variable indépendante et monte à 83,1 % pour l'étape 3.

Tableau de classement[a]

			Prévisions		
			Épuisement professionnel		
	Observations		,00	1,00	Pourcentage correct
Etape 1	Épuisement professionnel	,00	321	27	92,2
		1,00	72	47	39,5
	Pourcentage global				78,8
Etape 2	Épuisement professionnel	,00	318	30	91,4
		1,00	61	58	48,7
	Pourcentage global				80,5
Etape 3	Épuisement professionnel	,00	322	26	92,5
		1,00	53	66	55,5
	Pourcentage global				83,1
Etape 4	Épuisement professionnel	,00	321	27	92,2
		1,00	53	66	55,5
	Pourcentage global				82,9

a. La valeur de césure est ,500

Il redescend minimalement à 82,9 % pour l'étape 4 où 92,2 % des enseignants non épuisés sont classés correctement, mais que seulement 55,5 % des épuisés le sont. Cette amélioration est significative.



7.4 EXERCICES

Parmi les propositions suivantes, choisir celle(s) qui est (sont) exactes(s).

QCM1. Dans la régression multiple,

A. Toutes les variables doivent être quantitatives
B. Tous les variables doivent être qualitatives
C. Les variables peuvent être quantitatives ou qualitatives
D. Un seul type de variable doit intervenir

QCM2. Dans la régression multiple, le nombre de variables indépendantes doit être :

A. Supérieur ou égale à 2, mais inférieur au nombre d'observations
B. Supérieur ou égal à 2, mais égal au nombre d'observations
C. Inférieur ou égal à 2, et inférieur au nombre d'observations
D. Inférieur ou égal à 2, et supérieur au nombre d'observations

QCM3. On pose un modèle de régression multiple : $\mathbf{Y_i : (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon_i}$ Les paramètres sont définis comme suit :

A. Y est la variable indépendante, X1, X2, …, Xn sont les variables dépendantes, β0 est l'ordonné à l'origine, β1, β2,…, βn sont les coefficients et ε est une erreur aléatoire
B. Y est la variable dépendante, X1, X2, …, Xn sont les variables indépendantes, β0 est l'ordonné à l'origine, β1, β2,…, βn sont les coefficients et ε est une erreur aléatoire
C. Y est la variable dépendante, X1, X2, …, Xn sont les variables indépendantes, β0, β1, β2,…, βn sont les coordonnées à l'origine et ε est une erreur aléatoire
D. Y est la variable dépendante, X1, X2, …, Xn sont les variables indépendantes, β0, β1, β2,…, βn sont les coordonnées à l'origine et ε est un écart-type

QCM4. Concernant les types de variables :

A. La variable dépendante doit être continue et les variables indépendantes doivent être continues ou catégorielles
B. La variable dépendante doit être discrète et les variables indépendantes doivent être continues
C. La variable dépendante et les variables indépendantes doivent être continues ou catégorielles
D. La variable dépendante et les variables indépendantes doivent être discrètes

QCM5. Nous voulons prédire les valeurs d'une variable qualitative en fonction d'une série de données. Le type de modèle à utiliser est :

A. La régression linéaire simple
B. La régression multiple
C. L'ANOVA
D. La régression logistique

QCM6. Les méthodes suivantes font partie de la régression multiple avec entrée progressive :

A. Méthode d'entrée forcée
B. Méthode descendante
C. Méthode ascendante
D. Méthode pas-à-pas

QCM7. Plusieurs cas peuvent être considérés dans une régression logistique :

A. Le cas où la variable dépendante qualitative comporte deux catégories exhaustives et exclusives
B. Le cas où la variable qualitative dépendant comporte moins de deux catégories
C. Le cas où la variable dépendante qualitative nominale comporte plus de deux catégories
D. Le cas où la variable qualitative dépendante est ordinale

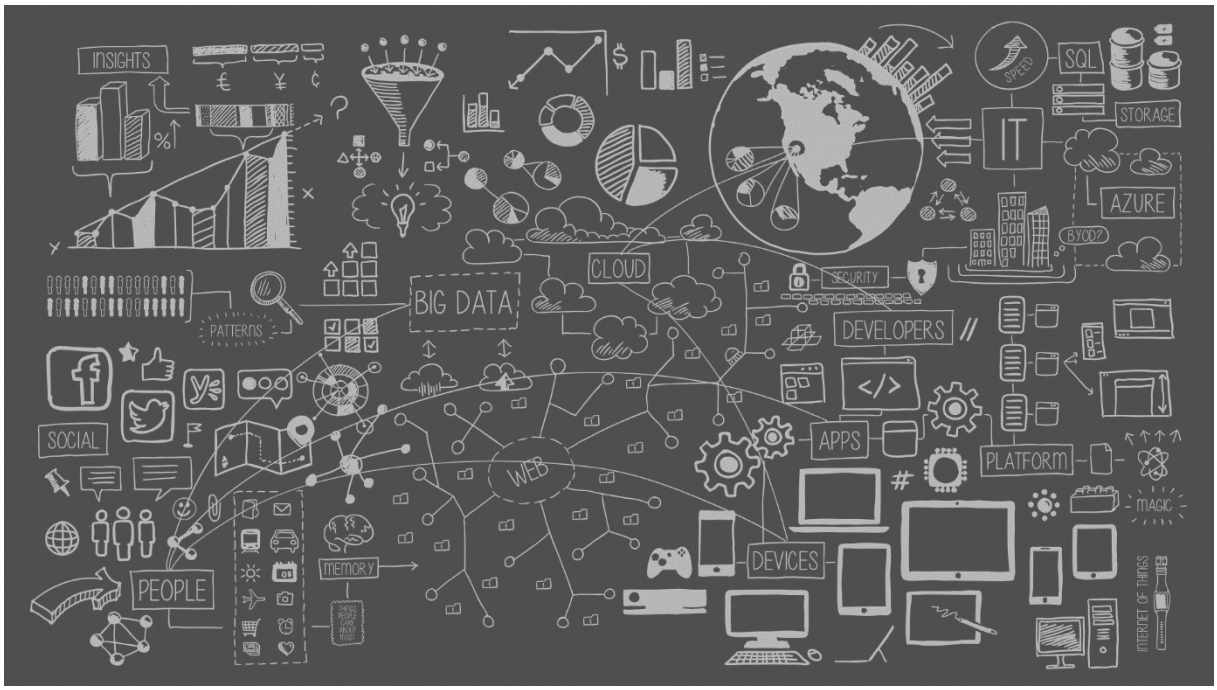
QCM8. Dans la régression logistique, il ne faut pas utiliser une variable continue recodée en variable dichotomique comme variable dépendante car :

A. Elle n'est pas considérée comme une variable qualitative
B. Elle engendre une grande perte d'informations
C. Ses modalités ne s'excluent pas mutuellement
D. Ses modalités s'excluent mutuellement

QCM9. La/les différence(s) entre une régression multiple et une régression logistique est :

A. La régression multiple donne un résultat constant, alors que la régression logistique donne un résultat continu
B. La régression multiple donne un résultat continu, alors que la régression logistique donne un résultat constant
C. La régression multiple peut avoir n'importe quelle valeur parmi un nombre infini de valeur possibles, contrairement à la régression logistique qui a un nombre limité de valeurs possibles
D. La régression multiple à un nombre limité de valeurs possibles, contrairement à la régression logistique qui peut avoir n'importe quelle valeur parmi un nombre infini de valeurs possibles

QCM1 (B) - QCM2 (A) - QCM3 (B) - QCM4 (A) - QCM5 (D) - QCM6 (B C D) - QCM7 (A C D) - QCM8 (B) - QCM9 (B C)



8 ANALYSE EN COMPOSANTES PRINCIPALES - ANALYSE DES CORRESPONDANCES MULTIPLES

8.1 OBJECTIFS À ATTEINDRE À LA FIN DU CHAPITRE

A la fin ce chapitre, les participants seront capables d'identifier les types de recherche nécessitant l'utilisation d'une Analyse en Composantes Principales (ACP) ou d'une Analyse des Correspondances Multiples (ACM). Ils pourront également appliquer ces deux méthode d'analyse, faire sortir les résultats à partir du logiciel SPSS et les analyser.

8.2 ASPECT THÉORIQUE

Les méthodes les plus classiques de la statistique descriptive multidimensionnelle sont les méthodes factorielles. Elles consistent à rechercher des facteurs en nombre restreint et résumant le mieux possible les données considérées. Elles aboutissent à des représentations graphiques des données (des individus comme des variables) par rapport à ces facteurs représentés comme des axes. Ces représentations graphiques sont du type nuage de points (ou diagramme de dispersion).

Dans chaque méthode que nous allons développer, les variables considérées seront de même nature : toutes quantitatives (Analyse en Composantes Principales) ou toutes qualitatives (Analyse des Correspondances).

8.2.1 ANALYSE EN COMPOSANTES PRINCIPALES

L'analyse en Composantes Principales (ACP) est une méthode fondamentale en statistique descriptive multidimensionnelle. Cette méthode permet de traiter simultanément un nombre quelconque de variables toutes quantitatives. Le but de cette analyse est de résumer le maximum d'informations possibles en n'en perdant le moins possible pour :

- Faciliter l'interprétation d'un grand nombre de données initiales ;
- Donner plus de sens aux données réduites.

L'ACP permet donc de réduire des tableaux de grandes tailles en petite nombre de variables (2 ou 3 varaibles généralement) tout en conservant un maximum d'information. Les variables de départ sont dites « métriques ».

L'Analyse en Composantes Principales (ACP) est une technique multivariée dite « d'interdépendance » car il n'y a pas variable dépendante ou indépendance identifiée au préalable une autre caractéristique importante de l'ACP est qu'il n'y a pas d'hypothèse nulle à tester ou à vérifier.

8.2.1.1 Principaux objectifs de l'ACP

L'Analyse en Composantes Principales (ACP) vise 3 principaux objectifs :

- Comprendre la structure d'un ensemble de variables (dans un questionnaire, voir quelles variables sont associées) ;
- Concevoir et raffiner des instruments de mesure comme les questionnaires basés sur des échelles de type **Likert** permettant de mesurer des construits latents qu'il est impossible de mesurer directement comme le degré de bonheur d'une personne ;
- Condenser l'information contenue à l'intérieur d'un grand nombre de variables en un

ensemble restreint de nouvelles dimensions composites tout en assurant une perte minimale d'informations.

De manière brève, l'ACP est utile pour identifier (avec la condition de garder le maximum d'information) :

- Les dimensions ou les facteurs fondamentaux qui expliquent les corrélations entre plusieurs variables ;
- Un nouvel et petit ensemble de variable non corrélées ;
- Un ensemble plus petit des variables les plus déterminantes.

L'ACP considère 4 types de relations :

- Les relations des variables entre elles ;
- Les relations des variables aux facteurs ;
- Les relations entre les variables d'un même facteur ;
- Les relations entre les différents facteurs.

8.2.1.2 Conditions d'application de l'ACP

L'Analyse en Composantes Principales doit respecter certaines contraintes, à savoir :

- Le nombre de variables doit être suffisant (cinq variables en plus) ;
- La forme des réponses aux questions (les items) doit être la même (par exemple : 5 choix de réponse pour tous). Dans le cas contraire, les variables doivent être réduites et normalisées ;
- On doit avoir 10 fois plus de cas qu'il y a de variables impliquées. Par exemple, si on a 10 variables, alors on doit avoir une taille ***n*** égale à ***100.***

8.2.1.3 Procédure d'utilisation de l'ACP

Pour mener une ACP, il faut :

- Formuler le problème ;
- Calculer la matrices des corrélations ;
- Extraire les facteurs ;
- Interpréter les facteurs ;
- Calculer les scores factoriels ;
- Estimer l'adéquation du modèle.

8.2.2 ANALYSE DES CORRESPONDANCES MULTIPLES

L'Analyse des Correspondant Multiples (ACM) est une méthode qui permet d'étudier l'association entre au moins 2 variables qualitatives. L'**ACM** est aux **variables qualitatives** ce que l'**ACP** est aux **variables quantitatives**. Elle permet en effet d'aboutir à des cartes de représentation sur lesquelles on peut visuellement observer les proximités entre les catégories des variables qualitatives et observations.

Le but de l'ACM est donc est le même que celui de l'ACP. Il s'agit de représenter graphiquement les individus par des points dans un sous-espace de manière à ce que le nuage de points ressemble le plus au nuage de points de l'espace original. L'ACM est adaptée aux tableaux dans lesquels un ensemble d'individus (en ligne) est décrit par un ensemble de variables qualitatives (en colonne).

Un exemple typique de ces données est celui des enquêtes d'opinion.

8.2.2.1 Principaux objectifs de l'ACM

L'Analyse des Correspondances Multiples (ACM) vise à mettre en évidence :

- Les relations entre les modalités des différentes variables ;
- Les relations entre les individus statistiques ;
- Les relations entre les variables telles qu'elles apparaissent à partir des relations entre modalités.

8.2.2.2 Principe de l'ACM

La construction du tableau disjonctif complet est l'une des étapes préalables au calcul de l'ACM. Les **p** variables qualitatives sont éclatées en **p** tableaux disjonctifs Z_1, Z_2,…., Z_n composés d'autant de colonnes qu'il y a de modalités pour chacune des variables. A chaque fois qu'une modalité **m** de la **ième** variable correspond à un individu **i**, on affecte **1** à $\mathbf{Z_{j(i,m)}}$. Les autres valeurs de $\mathbf{Z_i}$ sont nulles.

Les **p** tableaux disjonctifs sont alors concaténés en un tableau disjonctif complet.

A partir du tableau disjonctif complet, sont calculés les coordonnées des modalités des variables qualitatives, ainsi que les coordonnées des observations dans un espace de représentation optimal pour le critère d'inertie.

Dans le cas de l'ACM, on montre que l'inertie est égale au nombre moyen de modalités moins un. Elle ne dépend donc pas uniquement de l'association entre les variables.

Remarque

L'ACM est sensible aux effectifs faibles, aussi il est préférable de regrouper les classes peu représentées le cas échéant.

8.2.2.3 Procédure d'utilisation de l'ACM

La démarche de l'ACM est la suivante :

- Partir des données qualitatives ;
- Les données quantitatives peuvent être utilisées à condition de les transformer en données qualitatives avant de réaliser l'ACM avec elles ;
- Construire un tableau disjonctif complet ou ;
- Construire un tableau de BURT.

Les données doivent être des variables « chaîne » qui sont toujours converties en nombres entiers positifs par ordre croissant alphanumérique. Les valeurs manquantes par défaut et les valeurs inférieures à « 1 » sont considérées comme manquantes. Vous pouvez donc recoder ou ajouter une constante aux variables contenant des valeurs inférieurs à « 1 » pour les définir comme non manquantes.

Le tableau disjonctif complet comporte une colonne pour chaque modalité des variables étudiées et une ligne pour chaque individu statistique. Les cellules du tableau contiennent « 1 » ou « 0 » selon que l'individu considéré représente la modalité correspondante ou non.

	Q1-1	Q1-2	Q1-3	Q2-1	Q2-2	Q3-1	Q3-2	Q4-1	Q4-2	Q4-3
A	1	0	0	1	0	1	0	1	0	0
B	1	0	0	0	1	0	1	1	0	0
C	1	0	0	1	0	1	0	0	1	0
D	1	0	0	1	0	0	1	0	1	0
E	1	0	0	1	0	1	0	1	0	0
F	1	0	0	0	1	0	1	0	0	1
G	1	0	0	0	1	0	1	1	0	0
H	1	0	0	0	1	0	1	1	0	0
I	0	1	0	1	0	1	0	1	0	0
J	0	1	0	1	0	1	0	1	0	0
K	0	1	0	1	0	0	1	1	0	0
L	0	1	0	1	0	1	0	0	1	0
M	0	1	0	1	0	1	0	0	1	0
N	0	1	0	1	0	1	0	0	1	0
O	0	1	0	0	1	1	0	0	1	0
P	0	1	0	0	1	1	0	0	1	0
Q	0	0	1	1	0	1	0	1	0	0
R	0	0	1	1	0	1	0	0	0	1
S	0	0	1	1	0	0	1	0	1	0
T	0	0	1	0	1	0	1	0	1	0
U	0	0	1	0	1	0	1	0	1	0
V	0	0	1	0	1	0	1	0	1	0
W	0	0	1	0	1	0	1	0	0	1
X	0	0	1	0	1	0	1	0	0	1
Y	0	0	1	0	1	0	1	0	0	1
Z	0	0	1	1	0	0	1	0	0	1

Le tableau de Burt est issu du tableau disjonctif complet et comporte tous les tableaux de contingence des variables prises deux à deux qui sont autant de sous-matrices et qui constituent une matrice carré symétrique. L'obtention de la matrice **B** du tableau de Burt s'obtient en multipliant la matrice du disjonctif complet **Z** par sa transposée : **B = Z'Z.**

Le tableau de Burt issu du tableau disjonctif complet précédant :

Les sous-matrices et leurs transposées (de même couleur) croisent les modalités d'une question avec les modalités d'une autre question. Comme il se doit, la somme des éléments de chaque sous-matrice est égale au nombre d'interviewés.

Tableau de Burt :

	Q1-1	Q1-2	Q1-3	Q2-1	Q2-2	Q3-1	Q3-2	Q4-1	Q4-2	Q4-3
Q1-1	8	0	0	4	4	3	5	5	2	1
Q1-2	0	8	0	6	2	7	1	3	5	0
Q1-3	0	0	10	4	6	2	8	1	4	5
Q2-1	4	6	4	14	0	10	4	6	6	2
Q2-2	4	2	6	0	12	2	10	3	5	4
Q3-1	3	7	2	10	2	12	0	5	6	1
Q3-2	5	1	8	4	10	0	14	4	5	5
Q4-1	5	3	1	6	3	5	4	9	0	0
Q4-2	2	5	4	6	5	6	5	0	11	0
Q4-3	1	0	5	2	4	1	5	0	0	6

A titre d'exemple, les individus qui ont choisi la 2ième modalité à la question 1 optent majoritairement pour la modalité 1 de la question 2.

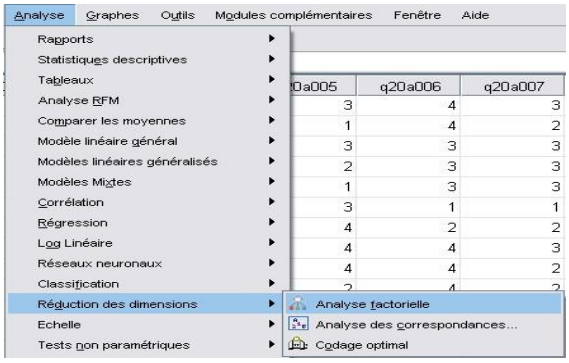
8.3 SYNTAXE SPSS

8.3.1 SYNTAXE SPSS POUR RÉALISER L'ACP

8.3.1.1 Formalisation du processus

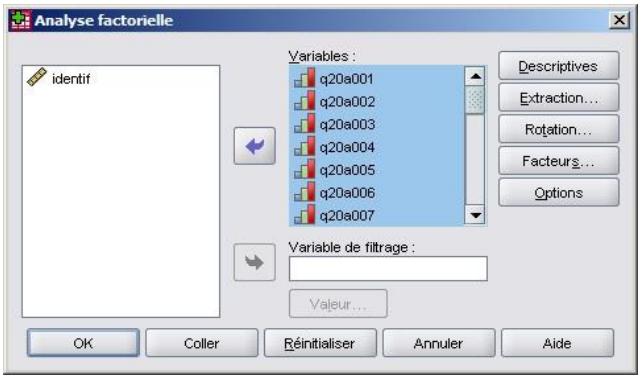
Etape 1. Ouvrir votre matrice de données sous SPSS.

Etape 2. Cliquer sur Analyse > Réduction des dimensions > Analyse factorielle.



Etape 3. Dans la boîte de dialogue, transférer toutes les variables de la boîte de gauche que vous désirez inclure dans l'analyse dans la boîte de droite en cliquant sur .

Vous devez ensuite préciser certains éléments à l'aides des boutons situés à droite (Descriptives, Extraction, …).



Etape 4 : le bouton Descriptives

L'option « Caractéristiques » offre différents types de statistiques.

Le 1er encadré permet d'obtenir les statistiques descriptives (moyenne, écart-type et nombre d'observations) pour chacune des variables : ***Caractéristiques univariées***. Vous pouvez aussi afficher un tableau indiquant, à côté de la part de variance extraite pour chaque variable, la variance que vous aviez à extraire au départ : ***Structure initiale***.

L'encadré de la ***matrice de corrélation*** permet de vérifier certains postulats de base de l'ACP :

- **Coefficients** : réalise la matrice de corrélation entre les variables ;
- **Indice KMO et test de sphéricité de Bartlett** : mesure l'adéquation de l'échantillonnage et si la matrice est une matrice identité

Cliquer sur Poursuivre pour revenir à la boîte de dialogue principale.

Le test de Kaiser-Meyer-Olkin (**KMO**) est une mesure généralisée de la corrélation partielle entre les variables de l'étude. Cette mesure est basée sur la moyenne des coefficients de corrélation qui sont situés dans la diagonale de la matrice anti-image. La lecture du test KMO se fait de la façon suivante :

- 0,90 et plus = très grande validité ;
- 0,89 à 0,80 = grande validité ;
- 0,79 à 0,70 = validité moyenne ;
- 0,69 à 0,60 = validité faible ;
- 0,59 à 0,50 = validité au seuil limite ;
- 0,49 et moins = invalide.

Le test de **sphéricité de Bartlett** permet de juger de l'inégalité des racines latentes, c'est-à-dire de l'absence significative de sphéricité du modèle mentionné. Si le modèle s'avère sphérique, on peut présumer que les corrélations ente les variables sont voisines de « 0 » et donc qu'il n'y a pas intérêt à remplacer les variables par des composantes.

Le test de Bartlett est un test d'hypothèse, une forme approchée du Khi carré.

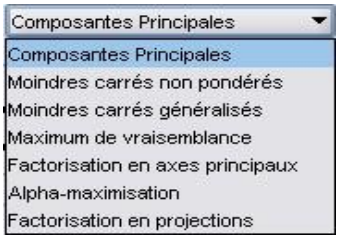
H0 : la matrice de corrélation est une matrice identité.

H1 : la matrice de corrélation est différente d'une matrice identité. Il est donc justifié de rechercher des composantes (facteurs).

Dans ce test, il est inutile de faire appel à une table de décision. On doit considérer seulement la signification du test. La valeur observée doit être inférieure ou égale à **α**.

Etape 5.

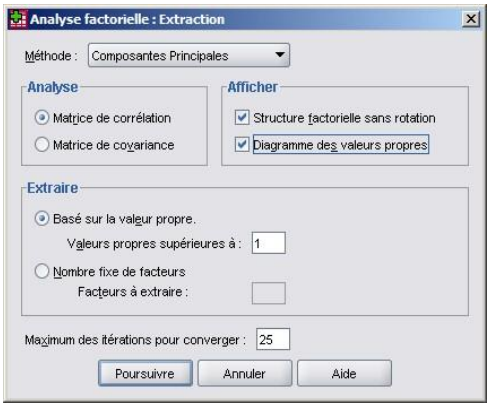
Le bouton Extraction... permet de déterminer quelle méthode vous désirez privilégier. Par défaut, SPSS réalise une **analyse en composantes principales** qui est la méthode la plus utilisée.



Etape 6.

Pour le choix de la matrice, conservez la ***matrice de corrélation.*** La matrice de covariance est en fait la matrice de corrélation non standardisée. La matrice de corrélation est donc plus pertinente. Cliquer sur « ***structure factorielle sans rotation*** » pour pouvoir comparer avant et après rotation.

Cliquer sur « ***diagramme des valeurs propres*** » pour vérifier où se situe la rupture du coude qui vous indiquera le nombre de facteurs à extraire.



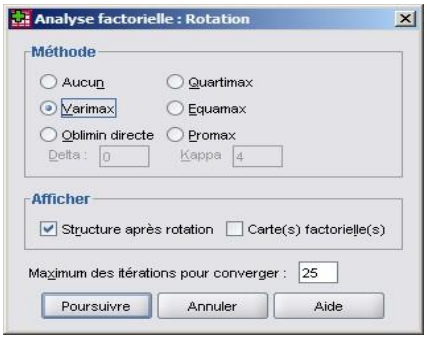
Par convention, la valeur retenue est supérieur à « 1 », mais vous pouvez la modifier. Idéalement, vous réalisez une 1ère analyse en choisissant la valeur propre initiale et vous analysez ensuite le diagramme des valeurs propres.

- Si la rupture se situe à l'endroit où les valeurs propres sont inférieures à « 1 », vous pouvez poursuivre l'interprétation de l'analyse.
- Si la rupture est avant, vous refaites l'analyse en indiquant cette fois le nombre de facteurs que vous voulez extraire (nombre de facteurs situés avant la rupture du coude).

Cliquer sur Poursuivre.

Etape 7. Le bouton Rotation... facilite l'interprétation de la matrice en maximisant le poids de chaque variable sur un facteur et en diminuant sur les autres.

- Lorsque vous croyez que les facteurs sont indépendants, choisir une rotation orthogonale comme la rotation **Varimax** ;
- Si au contraire, à partir de vos lectures, vous croyez que les facteurs sont reliés, choisir une rotation oblique comme **Oblimin directe** ou **Promax**.



Cliquer sur Poursuivre.

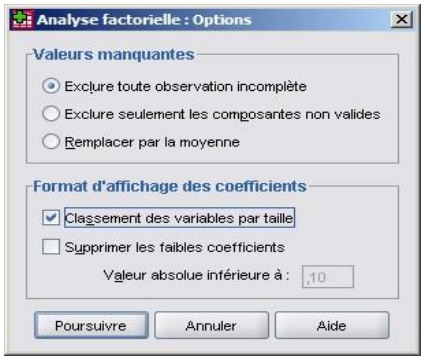
Etape 8. Le bouton Facteurs... permet de conserver les résultats de l'ACP en créant une nouvelle colonne dans la base de données pour chaque facteur, en indiquant le résultat de chaque observation pour le facteur dans cette colonne. Vous pourrez utiliser ces résultats pour une analyse ultérieure ou pour comparer les groupes de participants.

- Cocher l'option « Enregistrer » dans des variables ;
- Cocher la méthode « Anderson-Rubin » : si vous désirez que les facteurs ne soient pas corrélés ;
- Cocher la méthode « Régression » : si les corrélations sont acceptables.

Vous pouvez demander d'afficher la matrice des coefficients factoriels. Cette matrice est excellente pour épater la galerie, mais n'est pas du tout nécessaire pour atteindre l'objectif de l'ACP, c'est-à-dire émerger des construits sous-jacents à partir d'un ensemble de variables.

Etape 9.

Le bouton Options comme pour l'ensemble des autres analyses, vous pouvez choisir **d'exclure toute observation incomplète** de l'analyse. Puisque l'ACP est basée sur l'examen des corrélations.



- L'option « **Exclure seulement les composantes non valides** » permet d'exclure les variables pour lesquelles il y a des valeurs manquantes pour la paire de variables mises en relation.
- L'option **« Remplacer par la moyenne »** permet de remplacer les valeurs manquantes par la moyenne.

Nous vous conseillons de juste garder l'option par défaut (exclure toute observation incomplète) afin de conserver le même nombre d'observation pour l'ensemble de l'analyse.

L'encadré « *Format d'affichage des coefficients* » permet de placer les variables dans la matrice des composantes et la matrice des composantes après rotation à l'aide du « Classement des variables par taille » (en fonction de la taille du coefficient). Cette option facilite l'identification des poids les plus élevés sur chaque facteur. Cliquer sur Poursuivre et « **ok** ».

8.3.1.2 Interprétation des résultats

Les différentes étapes pour l'interprétation des résultats d'une ACP seront ponctuées par l'introduction de résultats d'un exemple permettant de mieux s'orienter lors des interprétations.

Les données reprennent la structure fonctionnelle des dépenses d'un Etat de 1872 à 1971. Ces dépenses sont exprimées en pourcentage suivant 11 postes : pouvoirs publics (PVP), Agriculture (AGR), commerce et industrie (CMI), transports (TRA), logement et aménagement du territoire (LOG), éducation et culture (EDU), action sociale (ACS), anciens combattants (ANC), défense (DEF), remboursement de la dette (DET), divers (DIV).

Tableau 11 : Interprétation d'un resultat ACP

année	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1872	18	0.5	0.1	6.7	0.5	2.1	2	0	26.4	41.5	2.1
1880	14.1	0.8	0.1	15.3	1.9	3.7	0.5	0	29.8	31.3	2.5
1890	13.6	0.7	0.7	6.8	0.6	7.1	0.7	0	33.8	34.4	1.7
1900	14.3	1.7	1.7	6.9	1.2	7.4	0.8	0	37.7	26.2	2.2
1903	10.3	1.5	1.4	9.3	0.6	8.5	0.9	0	38.4	27.2	3
1906	13.4	1.4	0.5	8.1	0.7	8.6	1.8	0	38.5	25.3	1.9
1909	13.5	1.1	0.5	9	1.6	9	3.4	0	36.8	23.5	2.6
1912	12.9	1.4	0.3	9.4	0.6	9.3	4.3	0	41.1	19.4	1.3
1920	12.3	0.3	0.1	11.9	2.4	3.7	1.7	1.9	42.4	23.1	0.2
1923	7.6	1.2	3.2	5.1	0.6	5.6	1.8	10	29	35	0.9
1926	10.5	0.3	0.4	4.5	1.8	6.6	2.1	10.1	19.9	41.6	2.3
1939	10	0.6	0.6	9	1	8.1	3.2	11.8	28	25.8	2
1932	10.6	0.8	0.3	8.9	3	10	6.4	13.4	27.4	19.2	0
1936	8.8	2.6	1.4	7.8	1.4	12.4	6.2	11.3	29.3	18.5	0.4
1938	10.1	1.1	1.2	5.9	1.4	9.5	6	5.9	40.7	18.2	0
1947	15.6	1.6	10	11.4	7.6	8.8	4.8	3.4	32.2	4.6	0
1950	11.2	1.3	16.5	12.4	15.8	8.1	4.9	3.4	20.7	4.2	1.5
1953	12.9	1.5	7	7.9	12.1	8.1	5.3	3.9	36.1	5.2	0
1956	10.9	5.3	9.7	7.6	9.6	9.4	8.5	4.6	28.2	6.2	0
1959	13.1	4.4	7.3	5.7	9.8	12.5	8	5	26.7	7.5	0
1962	12.8	4.7	7.5	6.6	6.8	15.7	9.7	5.3	24.5	6.4	0.1
1965	12.4	4.3	8.4	9.1	6	19.5	10.6	4.7	19.8	3.5	1.8
1968	11.4	6	9.5	5.9	5	21.1	10.7	4.2	20	4.4	1.9
1971	12.8	2.8	7.1	8.5	4	23.8	11.3	3.7	18.8	7.2	0

Toutes les étapes du processus de réalisation de l'ACP ont été réalisées. Il s'agit dans cette partie d'interpréter les résultats donnés par le logiciel SPSS.

Résultats 1. Statistiques descriptives

Ce tableau donne les moyennes, les écarts-type de toutes les variables

Statistiques descriptives

	Moyenne	Ecart-type	n analyse
PVP	12,213	2,2383	24
AGR	1,996	1,6812	24
CMI	3,979	4,5507	24
TRA	8,321	2,5209	24
LOG	4,000	4,2424	24
EDU	9,942	5,3356	24
ACS	4,817	3,4821	24
ANC	4,275	4,2442	24
DEF	30,258	7,4667	24
DET	19,142	12,4560	24
DIV	1,183	1,0478	24

Indice KMO et test de Bartlett[a]

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,172
Test de sphéricité de Bartlett	Khi-deux approximé	317,514
	ddl	55
	Signification de Bartlett	,000

a. Basé sur les corrélations

Résultats 2. Indice de KMO et test de Bartlett

Indice KMO et test de Bartlett(a)

Le test de Sphéricité de Bartlett permet de tester l'hypothèse nulle (H0) : la matrice de corrélation est une matrice identité. Le résultat du test de sphéricité de Bartlett est significatif (p<0,0005). On peut donc rejeter l'hypothèse nulle. La matrice de corrélation n'est pas une matrice identité.

La lecture du test *KMO* se fait de la façon suivante :

- 0,90 et plus = très grande validité ;
- 0,89 à 0,80 = grande validité ;
- 0,79 à 0,70 = validité moyenne ;
- 0,69 à 0,60 = validité faible ;
- 0,59 à 0,50 = validité au seuil limite ;
- 0,49 et moins = invalide

Résultats 3. **Valeur propre et variance totale expliquée**

La plus grande valeur propre de la matrice de corrélation est de **4,961**; elle est associée au premier axe qui explique **45,102** de la variabilité.

On **choisit les 3 premiers axes principaux** qui expliquent **75,499%** de la variance. Ce choix se voie clairement dans le graphique des valeurs propres (résultat 4).

Variance totale expliquée

	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus		
Composante	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	**4,961**	**45,102**	45,102	4,961	45,102	45,102
2	2,059	18,720	63,823	2,059	18,720	63,823
3	1,284	11,676	**75,499**	1,284	11,676	**75,499**
4	,995	9,046	84,546			
5	,702	6,382	90,927			
6	,568	5,165	96,093			
7	,205	1,867	97,960			
8	,128	1,161	99,121			
9	,063	,575	99,696			
10	,033	,303	99,999			
11	,000	,001	100,000			

Méthode d'extraction : Analyse en composantes principales.

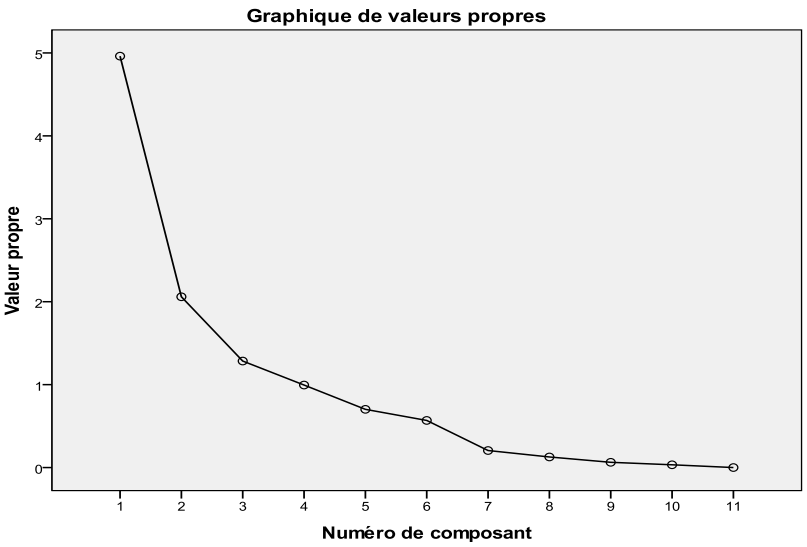
Résultats 4. Nombre d'axes principaux à retenir

Pour choisir le nombre d'axes principaux à retenir, deux règles sont applicables :

- **1ère 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 70 % de l'information.*
- **2ième règle :** on observe le graphique des valeurs propres et on 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 trouve 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.

Dans notre exemple, on retient que les trois premiers axes qui permettent de prendre en compte environ 75,499 de l'inertie totale.



Résultats 5.

Matrice des composantes

Les coefficients de corrélation entre les variables initiales et les composantes principales sont donnés dans le tableau suivant :

- La 1ère composante principale est corrélée positivement avec les variables ACS, CMI, AGR. Par contre, elle est négativement corrélée avec les variables DET et DEF. Cette opposition explique déjà près de la moitié de la variance (les autres corrélations sont relativement faibles).
- La 2ième composante est positivement corrélée avec les variables PVP et TRA, mais elle est négativement corrélée avec la variable ANC. Les autres corrélations sont plus faibles.
- La 3ième composante présente une corrélation assez importante (comparée aux autres valeurs) avec la variable DIV.

Matrice des composantes[a]

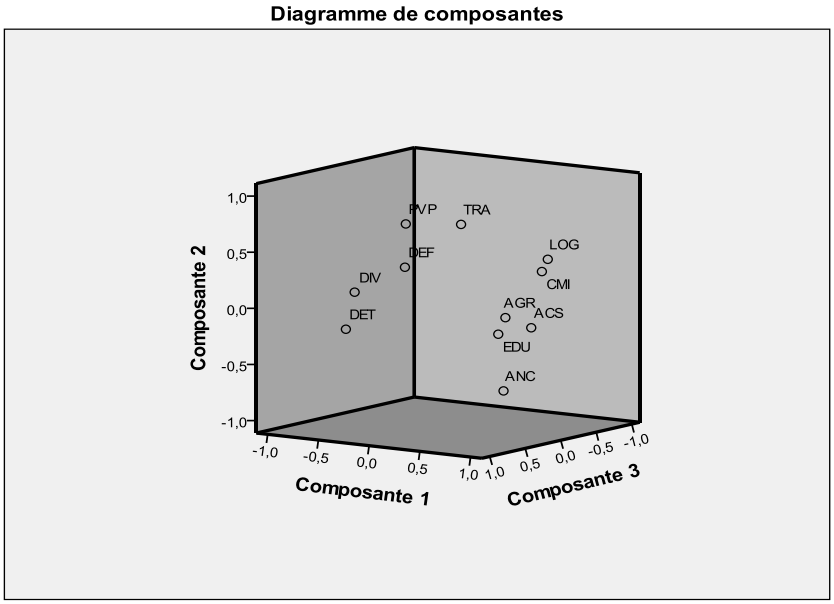
	Composante		
	1	2	3
ACS	,933	-,104	,162
DET	-,890	-,300	,167
CMI	,830	,344	-,136
AGR	,820	,004	,363
EDU	,788	-,138	,419
LOG	,718	,407	-,379
DEF	-,611	,218	-,265
DIV	-,544	,118	,538
ANC	,286	-,809	-,373
PVP	-,174	,736	,346
TRA	-,136	,631	-,378

Méthode d'extraction : Analyse en composantes principales.

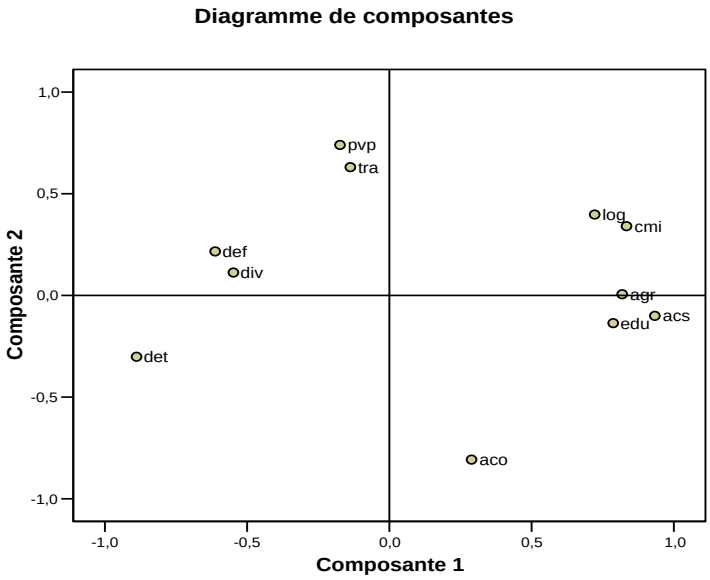
a. 3 composantes extraites.

Résultats 6. Diagramme des variables

Le diagramme des variables dans l'espace formé par les 3 axes retenus est le suivant.



Remarque : si on se limite à 2 axes principaux, on a une représentation des variables dans le plan comme dans le diagramme suivant.

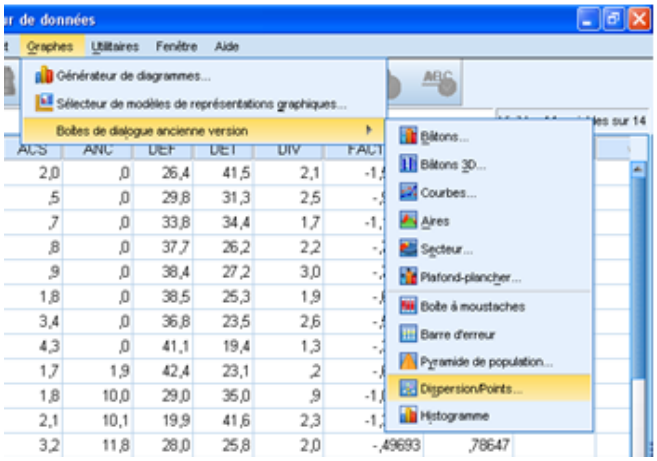


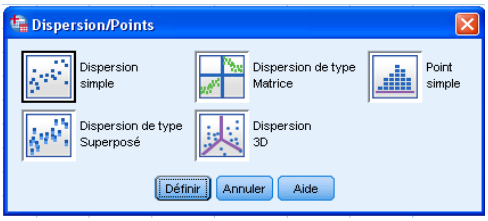
Résultats 7.

Représentation des individus dans les axes principaux.

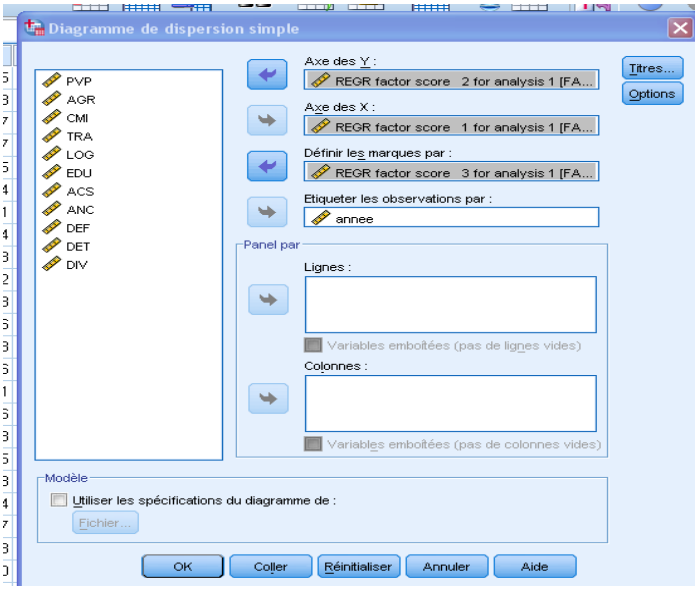
	annee	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV	FAC1_1	FAC2_1	FAC3_1
1	1872	18,0	,5	,1	6,7	,5	2,1	2,0	,0	26,4	41,5	2,1	-1,28019	,68894	1,38485
2	1880	14,1	,8	,1	15,3	1,9	3,7	,5	,0	29,8	31,3	2,5	-1,21970	1,36854	-,13178
3	1890	13,6	,7	,7	6,8	,6	7,1	,7	,0	33,8	34,4	1,7	-1,06640	,14725	,67669
4	1900	14,3	1,7	1,7	6,9	1,2	7,4	,8	,0	37,7	26,2	2,2	-,90714	,51047	,88072
5	1903	10,3	1,5	1,4	9,3	,6	8,5	,9	,0	38,4	27,2	3,0	-,99323	,15122	,51585
6	1906	13,4	1,4	,5	8,1	,7	8,6	1,8	,0	38,5	25,3	1,9	-,87579	,42088	,60277
7	1909	13,5	1,1	,5	9,0	1,6	9,0	3,4	,0	36,8	23,5	2,6	-,80687	,59507	,78695
8	1912	12,9	1,4	,3	9,4	,6	9,3	4,3	,0	41,1	19,4	1,3	-,63201	,51519	,16257
9	1920	12,3	,3	,1	11,9	2,4	3,7	1,7	1,9	42,4	23,1	,2	-,94493	,64548	-1,51000

Pour la représentation graphique, on clique sur ***Graphes*** > ***Boîte de dialogue ancienne version*** > ***Dispersion/Points*** > ***Dispersion 3D***.

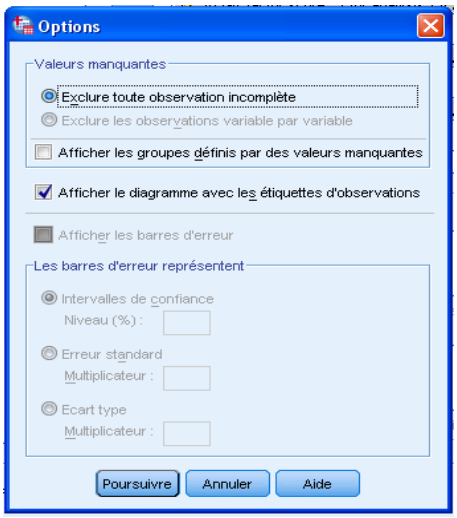




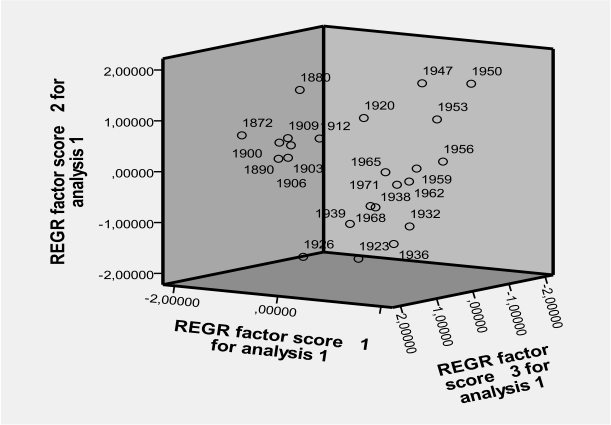
La boîte de dialogue suivante s'affiche.



Cliquer sur « **Options** » puis cocher sur « **Afficher le diagramme avec les étiquettes d'observations** ».



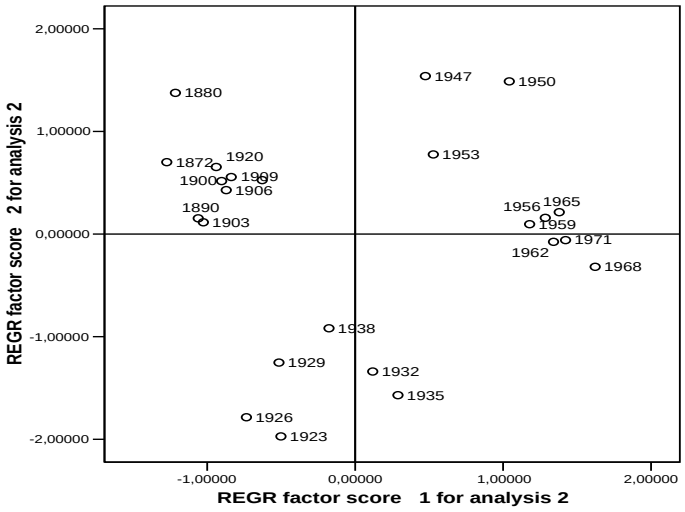
On obtient le graphique suivant.



Remarque : On peut représenter les individus dans le plan défini par les deux premiers axes principaux.

On remarque que les années se répartissent en 3 groupes (avant la 1ère guerre mondiale, entre les 2 guerres et après les 2nd guerre mondiale).

Seule l'année 1920, première année où il apparaît un poste de dépenses consacré aux anciens combattants, est placée avec le 1er groupe, alors qu'elle appartient au second groupe normalement.



8.3.2 SYNTAXE SPSS POUR RÉALISER L'ACM

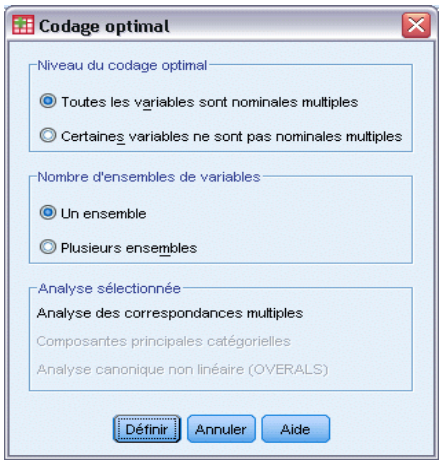
8.3.2.1 Formalisation du processus

Etape 1.

Ouvrir votre matrice de données sous SPSS.

Etape 2.

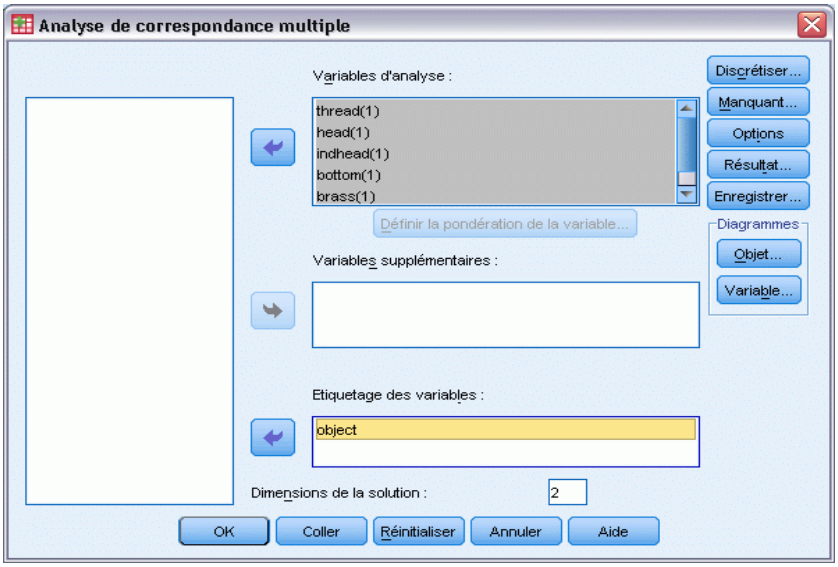
cliquer sur **Analyse** > **Réduction des dimensions** > **Codage optimal**.



Etape 3.

Sélectionner « ***Toutes les variables nominales multiples*** », puis sélectionner « ***Un ensemble*** » et cliquer sur « ***Définir*** ».

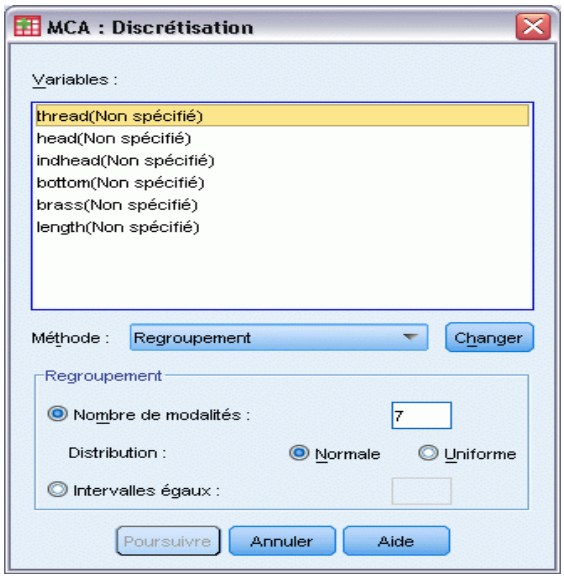
Sélectionner au moins 2 variables dans l'onglet «***Variable d'analyse*** » et spécifier le nombre de dimensions de la solution dans «***Dimensions de la solution***». Cliquer sur «**OK**».



Vous pouvez peut être spécifier des variables supplémentaires qui sont ajustées à la solution trouvée ou des variables d'étiquettes pour les diagrammes.

Etape 4.

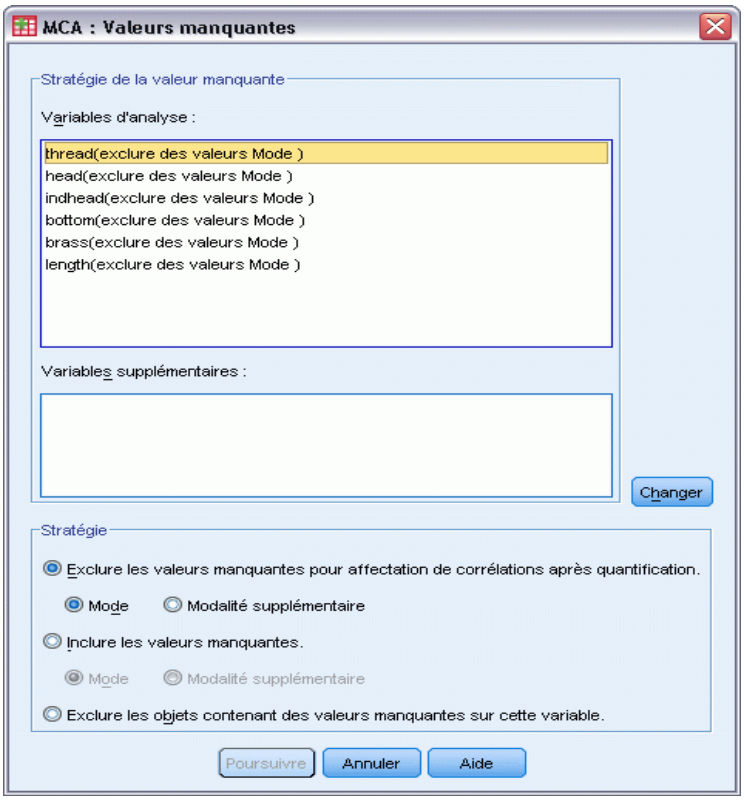
Le bouton « **Discrétisation** ». La boîte de dialogue 'Discrétisation' vous permet de choisir une méthode de recodage des variables. Les valeurs fractionnées sont regroupées en 7 modalités (ou en nombre de valeurs distinctes de variables si le nombre est inférieur à 7) avec une distribution normale approximative, à moins qu'une autre configuration ne soit spécifiée.



Dans l'onglet « Méthode », choisir entre Regroupement (recoder en un nombre spécifié de modalité ou par intervalle) ; Rang (la variable est discrétisée via le classement des observations) ; ou Multiplier (les valeurs courantes de la variable sont standardisées, multipliées par 10 et arrondies et possèdent une constante ajoutée de sorte que la valeur discrétisée la plus faible soit égale à 1.

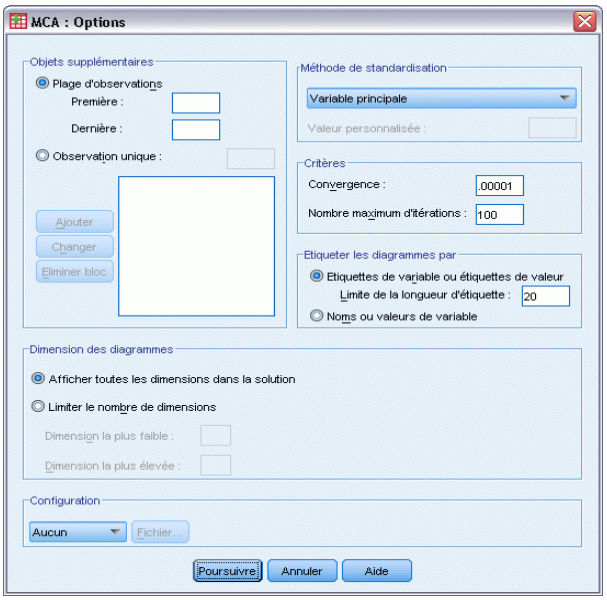
Etape 5.

Le bouton « **Manquant** » permet de choisir la stratégie de gestion des valeurs manquantes pour les variables de l'analyse et supplémentaires.



Etape 6.

Le bouton «**Options** » permet de sélectionner la configuration initiale, de spécifier les itérations et les critères de convergence, de sélectionner une méthode de standardisation, de sélectionner une méthode d'étiquetage des diagrammes et enfin de spécifier des objets supplémentaires.

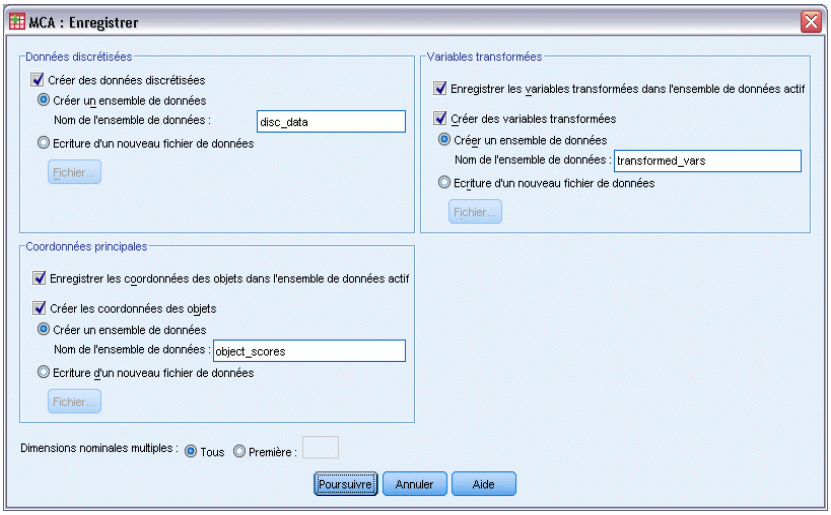


Etape 7.

Le bouton « **Résultats** » permet de créer des tableaux pour les coordonnées des objets, les mesures de discrimination, l'historique des itérations, les corrélations des variables d'origine et des variables transformées ainsi que les quantifications des modalités et statistiques descriptives des variables sélectionnées.

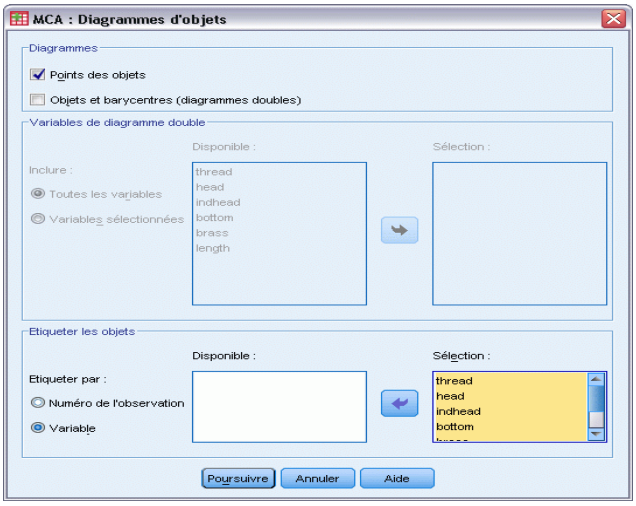
Etape 8.

Le bouton « **Enregistrer** » permet d'enregistrer les données discrétisées, les coordonnées des objets et les valeurs transformées dans un fichier de données externe ou un ensemble de données dans la session en cours.



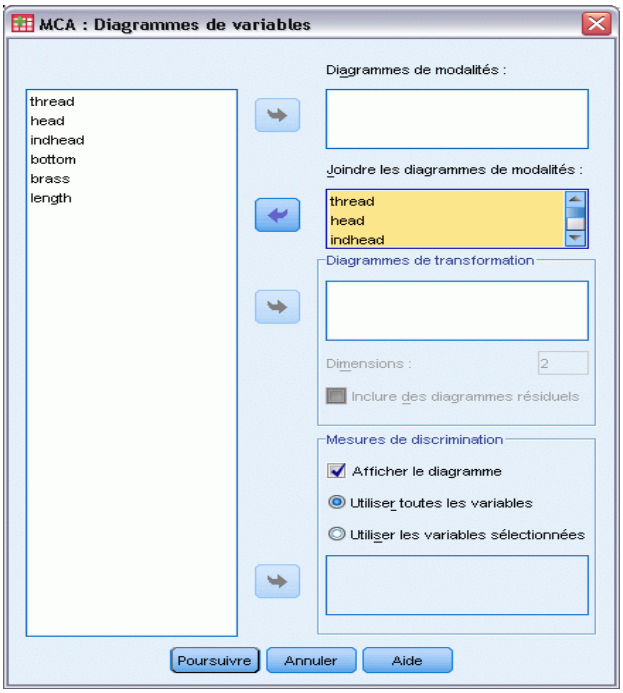
Etape 9.

Le bouton « **Diagramme d'objets** » permet d'indiquer les types de diagrammes souhaités ainsi que les variables à représenter.



Etape 10.

Le bouton « Diagrammes de variables » permet d'indiquer les types de diagrammes souhaités ainsi que les variables à représenter.



8.3.2.2 Interprétation des résultats

Les différentes étapes pour l'interprétation des résultats d'une ACM seront ponctuées par l'introduction de résultats d'un exemple permettant de mieux s'orienter lors des interprétations.

L'objectif de l'Analyse de Correspondance Multiple (ACM) est de rechercher les quantifications optimales dans la mesure où les modalités sont le plus possible séparées les unes des autres. Les objets de la même modalité doivent donc être représentés aussi éloignés que possible.

Tableau 12 : Exemple d'interprétation des resultats de l'ACM (description du matériel)

Nom de variable	Etiquette variable	Modalité
Filetage	*Filetage*	*Yes_Thread, No_Thread*
Titre	*Forme de tête*	*Plate, Creuse, Cônique, Arrondie, Cylindrique*
Indtête	*Indentation de la tête*	*Aucune, Cruciforme, Fendue*
Tige	*Forme tige*	*Pointe, plate*
Longueur	*Longueur en demi-pouces*	*1/2_in, 1_in, 1_1/2_ in, 2_in, 2_1/2_in*
Cuivre	*Cuivre*	*Yes_Br, Not_Br*
Objet	*Objet*	*broquette, clou1, clou2, clou3, clou4, clou5, clou6, clou7, clou8, vis1, vis2, vis3, vis4, vis5, boulon1, boulon2, boulon3, boulon4, boulon5, boulon6, broquette1, broquette2, cloub, visb*

Après l'exécution du processus d'exécution de l'ACM, les résultats ci-dessous.

Résultat 1 : Récapitulatif des modèles

L'ACM peut calculer une solution pour plusieurs dimensions. Le nombre maximal de dimensions est égal :

- soit au nombre de modalités moins le nombre de variables n'ayant aucune donnée manquante ;
- soit au nombre d'observations moins 1, selon le nombre qui est le plus petit.

N'utilisez toutefois que rarement le nombre maximal de dimensions. Un nombre de dimensions plus petit est plus facile à interpréter et, après un certain nombre de dimensions, le total de l'association supplémentaire représentée devient négligeable. Une solution à une, deux ou trois dimension dans une ACM est chose courante.

Récapitulatif du modèle

Presque toute la variance des données est représentée par la solution : 62,1 % par la 1ère dimension et 36,8 % par la 2ième dimension.

Dimension	Alpha de Cronbach	Variance expliquée		
		Total (valeur propre)	Inertie	Pourcentage de variance expliquée
1	,878	3,727	,621	62,123
2	,657	2,209	,368	36,809
Total		5,936	,989	
Moyenne	,796[a]	2,968	,495	49,466

a. La valeur Alpha de Cronbach moyenne est basée sur la valeur propre moyenne.

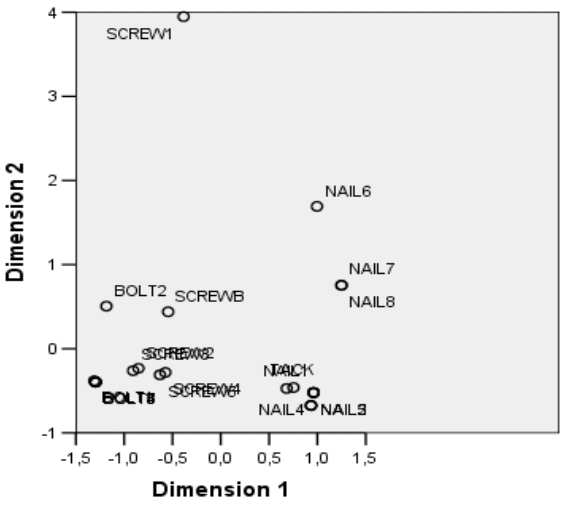
Résultat 2 : Coordonnées principales

Après avoir analysé le récapitulatif des modèles, vérifier les coordonnées des objets. Vous pouvez indiquer une ou plusieurs variables pour étiqueter le diagramme de coordonnées des objets. Chaque variable d'étiquetage génère un diagramme distinct étiqueté avec les valeurs de la variable.

Si vous observez le diagramme, vous constatez que la 1ère dimension (l'axe horizontal) distingue les VIS et BOULONS (qui ont des filetages) et des CLOUS et BROQUETTES (qui n'ont pas de filetage).

En effet, les VIS et les BOULONS se trouvent à une extrémité de l'axe horizontal alors que les CLOUS et les BROUQUETTES sont à l'autre extrémité.

Dans une moindre mesure, la 1ère dimension sépare également les BOULONS (qui ont un fond plat) de tous les autres objets (qui ont un fond pointu).



Normalisation principale de la variable.

La 2ième dimension (l'axe vertical) semble séparer VIS1 et CLOU6 de tous les autres objets. VIS1 et CLOU6 partagent des valeurs identiques en ce qui concerne la longueur des variables (ce sont les objets les plus longs des données). De plus, VIS1 est beaucoup plus loin de l'origine que les autres objets , ce qui laisse supposer que dans l'ensemble, de nombreuses descriptives de cet objet ne sont pas partagées par les autres objets.

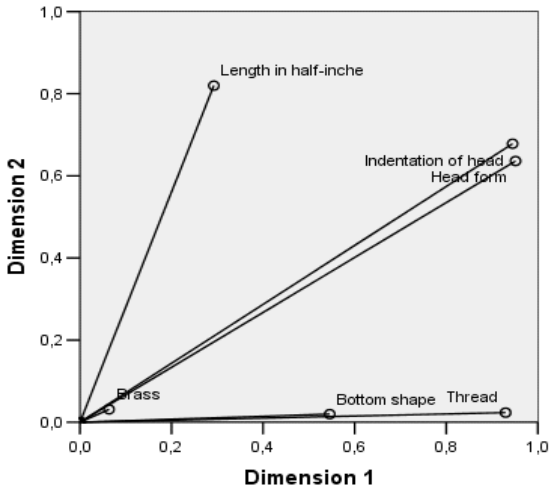
Le diagramme de coordonnées des objets est plus particulièrement utile pour rechercher les valeurs éloignées.

Résultat 3 : *Mesures de discrimination*

Avant d'étudier le reste des diagrammes de coordonnées des objets, vérifions si les mesures de discrimination sont conformes aux propos précédents.

Comme le diagramme de coordonnées des objets, le diagramme des mesures de discrimination indique que la 1ère dimension est liée aux variables FILETAGE et FORME TIGE. Ces variables disposent de mesures de discrimination élevées sur la 1ère dimension et de mesure de discrimination limitées sur la 2ième dimension. La valeur des variable LONGUEUR EN DEMI-POUCES est élevée sur la 2ième dimension, mais faible sur la 1ère dimension.

Conformément à l'observation du diagramme de coordonnées des objets, la 2ième dimension semble séparer les objets les plus longs des autres objets.



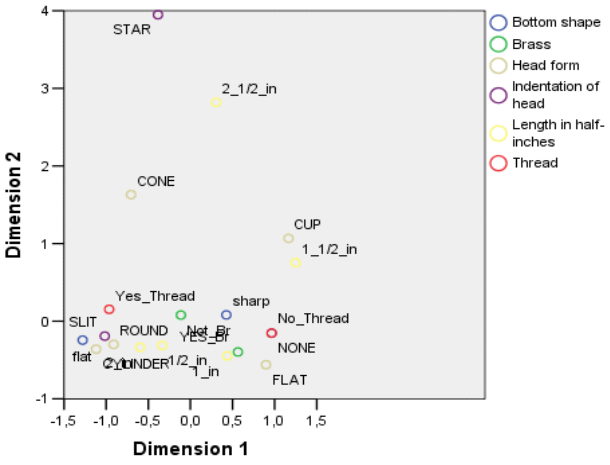
Les valeurs des variables INDENTATION DE LA TETE et FORME DE TETE sont relativement élevées sur les 2 dimensions, ce qui indique une discrimination dans les 2 premières dimensions.

La variable CUIVRE, très proche de l'origine, ne fait aucune distinction dans les 2 premières dimensions. Ceci est logique étant donné que tous les objets peuvent être en cuivre ou dans un autre matériau.

Résultat 4 : Valeurs affectées aux modalités

Les diagrammes de valeurs affectées aux modalités offrent un autre mode d'affichage de la discrimination des variables qui peut identifier les relations entre les modalités. Dans ce diagramme, les coordonnées des modalités de chaque dimension sont affichées. Vous pouvez donc déterminer les modalités similaires pour chaque variable.

La variable LONGUEUR EN DEMI-POUCES compte 5 modalités, dont 3 modalités sont regroupées près de la partie supérieure du diagramme. Les 2 autres modalités se trouvent dans la moitié inférieure du diagramme, la modalité 2_1/2_IN se trouvant très loin du groupe



La discrimination élevée de longueur le long de la dimension 2 est due à cette modalité qui est très différente des autres modalités de longueur.

De la même façon, pour la variable FORME DE TETE, la modalité CRUCIFORME est très loin des autres modalités et génère une mesure de discrimination élevée le long de la 2ième dimension. Il est impossible d'illustrer ces modèles dans un diagramme de mesures de discrimination.

Non seulement le diagramme de valeurs affectées aux modalités détermine le mode de discrimination et les dimensions le long desquelles une variable a un pouvoir discriminant, mais il compare également la discrimination des variables. Une variable ayant des modalités éloignées les unes des autres a un pouvoir discriminant plus élevé qu'une variable comportant des modalités proches les unes des autres.

Par exemple, le long de la dimension 1, les deux modalités de la variable CUIVRE sont plus proches l'une de l'autre que les deux modalités de la variable FILETAGE. Ceci indique que la variable FILETAGE a un pouvoir discriminant plus élevé que la variable CUIVRE le long de cette dimension. Cependant, le long de la dimension 2, les distances sont très similaires, ce qui laisse supposer que ces variables ont un pouvoir discriminant identique le long de cette dimension.

8.4 EXERCICES

Parmi les propositions suivantes, choisir celle(s) qui est (sont) exactes(s).

QCM1. Vous souhaitez procéder à une recherche de segmentation à partir d'une analyse multivariée portant sur des variables uniquement qualitatives. Pour ce faire, vous utilisez quelle procédure ?

A. Une procédure d'analyse en composantes principales
B. Une procédure d'analyse factorielle
C. Une procédure de test de Khi2
D. Une procédure d'analyse des correspondances multiples

QCM2. L'ACP permet de:

A. Réduire des tableaux de grandes tailles en petite nombre de variables en perdant le maximum d'information
B. Réduire des tableaux de grandes tailles en petite nombre de variables en conservant le maximum d'information
C. Construire des tableaux de grandes tailles en réduisant la perte d'informations
D. Construire des tableaux de grandes tailles en maximisant les informations

QCM3. Pour réaliser une ACP, on a besoin de :

A. Définir une hypothèse nulle uniquement
B. Définir une hypothèse alternative uniquement
C. Définir une hypothèse nulle et son alternation
D. Pas besoin de définir des hypothèses

QCM4. Les conditions pour réaliser une ACP sont :

A. Avoir moins de cinq variables
B. La forme des réponses aux questions doit être la même
C. Avoir au moins cinq variables
D. La forme des réponses aux questions doit être la même, sinon, les réduire et normaliser

QCM5. Pour faire un ACM, les variables doivent être :

A. Au moins deux variables qualitatives
B. Moins de deux variables qualitatives
C. Au moins deux variables quantitatives
D. Moins de deux variables quantitatives

QCM6. Pour faire un ACP, les variables doivent être :

A. Au moins deux variables qualitatives
B. Moins de deux variables qualitatives
C. Au moins deux variables quantitatives
D. Moins de deux variables quantitatives

QCM7. Les éléments suivant font partie de l'ACM :

A. Matrices de corrélation
B. Indice KMO
C. Tableau de BRUT
D. Tableau disjonctif

QCM8. Dans une ACP, le test de Sphéricité de Bartlett pose l'hypothèse nulle suivante :

A. La matrice de corrélation est une matrice identité
B. La matrice de corrélation n'est pas une matrice identité
C. Toutes les matrices de corrélation de l'échantillon sont les mêmes
D. Toutes les matrice de corrélation de l'échantillon ne sont pas les mêmes

QCM9. Dans une ACP, le nombre d'axes principaux à retenir se fait en fonction :

A. De la matrice de corrélation
B. De la restitution minimale d'informations que l'on souhaite
C. De la valeur du test de Durbin-Watson
D. Selon le bon vouloir du chercheur

QCM1 (D) - QCM2 (B) - QCM3 (D) - QCM4 (B C D) - QCM5 (A) - QCM6 (C) - QCM7 (C D) - QCM8 (A) - QCM9 (B).

9 APPLICATION DES MÉTHODES D'ANALYSES DES DONNÉES[3]

9.1 ANALYSE BIVARIÉE / ANOVA

9.1.1 ANOVA SANS TEST POST-HOC

L'analyse de variance permet de confronter les données d'une variable aux données d'une variable qualitative comportant deux catégories ou plus.

Question de recherche 1 : « Existe-t-il une différence d'âge entre les enfants selon leur niveau d'insécurité alimentaire ? »

Type de variable : une **variable quantitative** (Age de l'enfant = *Q63_3*) et une **variable qualitative** (niveau d'insécurité alimentaire = *ISA*)

Type d'analyse : ANOVA à 1 facteur et le risque α = 5 %

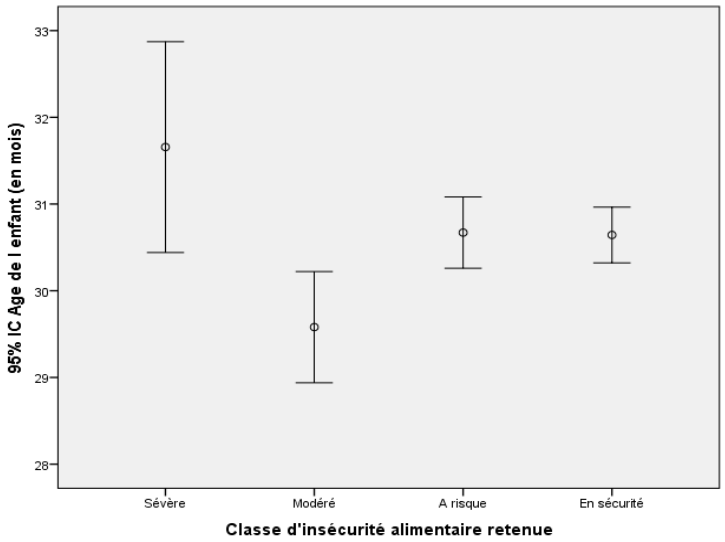
Vérifier la distribution si les variances sont égales en passant par le graphique de la barre d'erreur.

Syntaxe du graphique

```
GRAPH
/ERRORBAR(CI 95)=Q63_3 BY ISA.
```

Interprétation du graphique

La longueur des barres d'erreurs est variable ce qui laisse supposer que la variance entre les groupes d'enfants n'est pas forcément égale. Cela peut être vérifié par un test ANOVA.



Hypothèse de recherche :

- ***H0 :*** les moyennes d'âge des enfants dans les différents niveaux d'insécurité alimentaire sont égales.
- ***H1 :*** au moins une des moyennes d'âge des enfants n'est pas égale aux autres

Autre formulation

Ici on a 4 niveaux d'insécurité alimentaire (Sévère = μ1 ; Modéré = μ2 ; A risque = μ3 ; En sécurité = μ4).

[3] Base utilisée : base alimentation (base_vuln_2017/ménage/nationale/base_alimentation_finale)

- ***H0 :*** μ1 = μ2 = μ3 = μ4
- ***H1 :*** μ1 ≠ μ2 ≠ μ3 ≠ μ4

Syntaxe d réalisation du test ANOVA

```
ONEWAY Q63_3 BY ISA
/STATISTICS DESCRIPTIVES HOMOGENEITY
/MISSING ANALYSIS.
```

Interprétation des résultats de l'ANOVA

Descriptives

Age de l'enfant (en mois)

	N	Moyenne	Ecart-type	Erreur standard	Intervalle de confiance à 95 % pour la moyenne		Minimum	Maximum
					Borne inférieure	Borne supérieure		
Sévère	542	**31,66**	14,411	,619	30,44	32,87	0	59
Modéré	2085	**29,58**	14,938	,327	28,94	30,22	0	59
A risque	5128	**30,67**	15,041	,210	30,26	31,08	0	59
En sécurité	8600	**30,64**	15,203	,164	30,32	30,96	0	59
Total	16355	**30,55**	15,097	,118	30,32	30,78	0	59

Dans le tableau « Descriptives », il est indiqué les différents niveaux d'insécurité alimentaire et les moyennes et écart-types de ces niveaux. On remarque que la moyenne d'âge la plus élevée est celle des enfants ayant un niveau d'insécurité alimentaire sévère. Toutefois, les moyennes sont relativement les mêmes.

Le tableau « test d'homogénéité des variances » met en évidence le test de Levene. Le résultat indique que **Sig = 0,95 > α = 0,05. Alors on accepte l'hypothèse d'homogénéité des variances.**

Test d'homogénéité des variances

Age de l'enfant (en mois)

Statistique de Levene	ddl1	ddl2	**Signification**
2,123	3	16351	**,095**

On peut donc passer à l'interprétation du tableau ANOVA.

Attention

Si le test d'homogénéité des variances n'était pas concluant, on allait vérifier le test Brown-Forsythe ou le Welch Robust F.

ANOVA à 1 facteur

Age de l'enfant (en mois)

	Somme des carrés	ddl	Moyenne des carrés	F	**Signification**
Inter-groupes	2773,577	3	924,526	4,059	**,007**
Intra-groupes	3724694,511	16351	227,796		
Total	3727468,087	16354			

Conclusion : Le résultat du test consigné dans le tableau « ANOVA à 1 facteur » indique que : **Sig = 0,07 > α = 0,05. On accepte H0. On conclut qu'au seuil de 5 %, les moyennes d'âge des enfants dans les différents niveaux d'insécurité alimentaire sont égales.**

9.1.2 ANOVA AVEC TEST POST-HOC

Question de recherche 2 : « Existe-t-il une différence du périmètre brachial des enfants selon leur niveau d'insécurité alimentaire ? »

Type de variable : une **variable quantitative** (Age de l'enfant = *Q64_4*) et une **variable qualitative** (niveau d'insécurité alimentaire = *ISA*)

Type d'analyse : ANOVA à 1 facteur et le risque α = 5 %

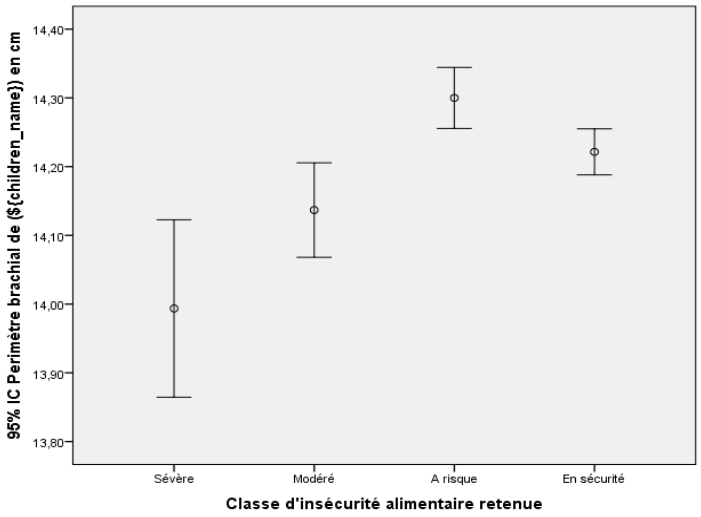
Vérifier la distribution si les variances sont égales en passant par le graphique de la barre d'erreur.

Syntaxe du graphique

```
GRAPH
/ERRORBAR(CI 95)=Q64_4 BY ISA
```

Interprétation du graphique

La longueur des barres d'erreurs est variable ce qui laisse supposer que la variance entre les groupes d'enfants n'est pas forcément égale. Cela peut être vérifié par un test ANOVA.



Hypothèse de recherche :

- ***H0 :*** les moyennes de périmètre brachial des enfants dans les différents niveaux d'insécurité alimentaire sont égales.
- ***H1 :*** au moins une des moyennes de périmètre brachial des enfants n'est pas égale aux autres

Autre formulation

Ici on a 4 niveaux d'insécurité alimentaire (Sévère = μ1 ; Modéré = μ2 ; A risque = μ3 ; En sécurité = μ4).

- ***H0 :*** μ1 = μ2 = μ3 = μ4
- ***H1 :*** μ1 ≠ μ2 ≠ μ3 ≠ μ4

Syntaxe de réalisation du test ANOVA

```
ONEWAY Q64_4 BY ISA
/STATISTICS DESCRIPTIVES HOMOGENEITY
/MISSING ANALYSIS.
```

Interprétation des résultats de l'ANOVA

Descriptives

Perimètre brachial de (${children_name}) en cm

	N	Moyenne	Ecart-type	Erreur standard	Intervalle de confiance à 95 % pour la moyenne		Minimum	Maximum
					Borne inférieure	Borne supérieure		
Sévère	429	**13,9936**	1,35974	,06565	13,8646	14,1226	10,20	19,50
Modéré	1699	**14,1368**	1,44392	,03503	14,0681	14,2055	10,00	22,00
A risque	4194	**14,3000**	1,46463	,02262	14,2557	14,3443	9,40	21,00
En sécurité	7262	**14,2215**	1,45596	,01709	14,1880	14,2550	8,00	21,80
Total	13584	**14,2279**	1,45557	,01249	14,2035	14,2524	8,00	22,00

Le tableau « Descriptives » indique les différents niveaux d'insécurité alimentaire et les moyennes et écart-types de ces niveaux. On remarque que les moyennes sont relativement les mêmes.

Le tableau « test d'homogénéité des variances » met en évidence le test de Levene. Le résultat indique que **Sig = 0,286 > α = 0,05. Alors on accepte l'hypothèse d'homogénéité des variances.**

Test d'homogénéité des variances

Perimètre brachial de (${children_name}) en cm

Statistique de Levene	ddl1	ddl2	Signification
1,259	3	13580	**,286**

On peut donc passer à l'interprétation du tableau ANOVA.

ANOVA à 1 facteur

Perimètre brachial de (${children_name}) en cm

	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	59,746	3	19,915	9,417	,000
Intra-groupes	28718,161	13580	2,115		
Total	28777,907	13583			

Conclusion 1 : Le résultat du test dans le tableau « ANOVA à 1 facteur » indique que : **Sig = 0,000 < α = 0,05 ; on rejette H0. On conclut qu'au seuil de 5 %, les moyennes d'âge des enfants dans les différents niveaux d'insécurité alimentaire sont différentes.** Cependant, ce résultat n'indique pas les moyennes qui sont différentes des autres et leur significativité statistique. Pour cela il faut faire le test post-hoc.

Syntaxe de réalisation du test ANOVA combiné au test post-hoc :

```
ONEWAY Q64_4 BY ISA
/STATISTICS DESCRIPTIVES HOMOGENEITY
/MISSING ANALYSIS
/POSTHOC=SCHEFFE ALPHA(0.05).
```

Interprétation des résultats

Comparaisons multiples

Variable dépendante: Perimètre brachial de (${children_name}) en cm

Scheffe

(I) Classe d'insécurité alimentaire retenue	(J) Classe d'insécurité alimentaire retenue	Différence de moyennes (I-J)	Erreur standard	Signification	Intervalle de confiance à 95 %	
					Borne inférieure	Borne supérieure
Sévère	Modéré	-,14322	,07858	,345	-,3629	,0765
	A risque	**-,30640***	,07371	**,001**	-,5125	-,1003
	En sécurité	**-,22789***	,07225	**,019**	-,4299	-,0259
Modéré	Sévère	,14322	,07858	,345	-,0765	,3629
	A risque	**-,16318***	,04182	**,002**	-,2801	-,0463
	En sécurité	-,08467	,03919	,198	-,1942	,0249
A risque	Sévère	,30640*	,07371	,001	,1003	,5125
	Modéré	,16318*	,04182	,002	,0463	,2801
	En sécurité	,07851	,02820	,052	-,0003	,1574
En sécurité	Sévère	,22789*	,07225	,019	,0259	,4299
	Modéré	,08467	,03919	,198	-,0249	,1942
	A risque	-,07851	,02820	,052	-,1574	,0003

*. La différence moyenne est significative au niveau 0.05.

Conclusion 2 : Selon les résultats du tableau « Comparaison multiple », le périmètre brachial des enfants qui ont un niveau d'insécurité alimentaire **à risque et en sécurité** est plus élevé que le périmètre brachial des enfants qui ont un niveau d'insécurité alimentaire **sévère**. Aussi, le périmètre brachial des enfants ayant un niveau d'insécurité alimentaire **à risque** est plus élevé que le périmètre brachial des enfants qui ont un niveau d'insécurité alimentaire **modéré.** Par ailleurs, ces différences sont statistiquement significatives **(Sig < 0,05).**

9.2 TABLEAU DE CONTINGENCE / TEST DU KHI2

Le test de Khi2 est une analyse bivariée qui consiste à déterminer s'il existe une association entre deux variables qualitatives.

9.2.1 KHI2 SANS DÉTERMINATION DE LA FORCE DE LA RELATION

Question de recherche 1 : « Existe-il un lien entre le sexe de l'enfant et la présence de diarrhée ? »

Type de variable : deux variables qualitatives : le sexe de l'enfant (Q63_2) et la présence de diarrhée (Q66a).

Type d'analyse : Khi2 et le risque α = 5 %

Vérification des conditions d'utilisation : les 2 variables sont qualitatives. Chaque variable comporte 2 modalités. Les observations sont indépendantes. La taille de l'échantillon est de 16 355 individus. Les modalités des variables s'excluent mutuellement et les effectifs théoriques sont supérieures à 5.

Hypothèse de recherche :

- **H0** : il n'existe pas de lien entre le sexe de l'enfant et le fait qu'il ait eu ou non la diarrhée ;
- **H1** : il existe un lien entre le sexe de l'enfant et le fait qu'il ait eu ou non la diarrhée.

Autre formulation

- **H0** : les variables Q63_2 et Q66a sont indépendantes ;
- **H1** : les variables Q63_2 et Q66a ne sont pas indépendantes.

Syntaxe de réalisation du test de Khi2

```
CROSSTABS
/TABLES=Q63_2 BY Q66a
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT ROW
/COUNT ROUND CELL.
```

Interprétation des résultats

Les résultats du « ***tableau croisé Sexe de l'enfant * Diarrhée*** » indiquent que dans l'échantillon, 18,90 % des enfants ont eu la diarrhée et parmi eux, 19 % sont des garçons

Tableau croisé Sexe de l'enfant * Diarhée

			Diarrhée		Total
			Pas de diarrhée	Diarrhée	
Sexe de l'enfant	Masculin	Effectif	5770	1356	7126
		% compris dans Sexe de l'enfant	81,0%	**19,0%**	100,0%
	Féminin	Effectif	5256	1216	6472
		% compris dans Sexe de l'enfant	81,2%	18,8%	100,0%
Total		Effectif	11026	2572	13598
		% compris dans Sexe de l'enfant	81,1%	**18,9%**	100,0%

.Tests du Khi-deux

	Valeur	ddl	**Signification asymptotique (bilatérale)**	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,128[a]	1	**,721**		
Correction pour la continuité[b]	,112	1	,737		
Rapport de vraisemblance	,128	1	,721		
Test exact de Fisher				,726	,369
Association linéaire par linéaire	,128	1	,721		
Nombre d'observations valides	13598				

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 1224,15.
b. Calculé uniquement pour un tableau 2x2

Conclusion 1 : Selon les résultats de l'analyse, **Sig =0,721 > α = 0,05**. On accepte H0. On conclut qu'au seuil de 5 %, il n'existe pas de lien entre le sexe de l'enfant et le fait qu'il ait eu ou non la diarrhée.

9.2.2 KHI2 AVEC DÉTERMINATION DE LA FORCE DE LA RELATION

Question de recherche 2 : « Existe-il un lien entre le niveau d'insécurité alimentaire de l'enfant et la présence de diarrhée ? »

Type de variable : deux variables qualitatives : le niveau d'insécurité alimentaire de l'enfant (ISA) et la présence de diarrhée (Q66a).

Type d'analyse : Khi2 et le risque α = 5 %

Vérification des conditions d'utilisation : les 2 variables sont qualitatives. Chaque variable comporte 2 ou plusieurs modalités. Les observations sont indépendantes. La taille de l'échantillon est de 16 355 individus. Les modalités des variables s'excluent mutuellement et les effectifs théoriques sont supérieures à 5.

Hypothèse de recherche :

- **H0** : il n'existe pas de lien entre le niveau d'insécurité alimentaire l'enfant et le fait qu'il ait eu ou non la diarrhée ;
- **H1** : il existe un lien entre le niveau d'insécurité alimentaire de l'enfant et le fait qu'il ait eu ou non la diarrhée.

Autre formulation

- **H0** : les variables ISA et Q66a sont indépendantes ;
- **H1** : les variables ISA et Q66a ne sont pas indépendantes.

Syntaxe de réalisation du test de Khi2

```
CROSSTABS
/TABLES=ISA BY Q66a
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT ROW
/COUNT ROUND CELL.
```

Interprétation des résultats

Tableau croisé Classe d'insécurité alimentaire retenue *Diarhée

			Diarhée		Total
			Pas de diarhée	Diarhée	
Classe d'insécurité alimentaire retenue	Sévère	Effectif	333	98	431
		% compris dans Classe d'insécurité alimentaire retenue	77,3%	**22,7%**	100,0%
	Modéré	Effectif	1379	321	1700
		% compris dans Classe d'insécurité alimentaire retenue	81,1%	18,9%	100,0%
	A risque	Effectif	3317	882	4199
		% compris dans Classe d'insécurité alimentaire retenue	79,0%	21,0%	100,0%
	En sécurité	Effectif	5997	1271	7268
		% compris dans Classe d'insécurité alimentaire retenue	82,5%	17,5%	100,0%
Total		Effectif	11026	2572	13598
		% compris dans Classe d'insécurité alimentaire retenue	81,1%	**18,9%**	100,0%

Le « *Tableau croisé Classe d'insécurité alimentaire retenu *Diarrhée* »* permet de décrire les 2 groupes d'enfants. Ainsi, 18,9 % des enfants issu de l'échantillon ont eu la diarrhée. Parmi eux, 22,7 % ont un niveau d'insécurité alimentaire sévère.

Tests du Khi-deux

	Valeur	ddl	Signification asymptotique (bilatérale)
Khi-deux de Pearson	25,722[a]	3	**,000**
Rapport de vraisemblance	25,405	3	,000
Association linéaire par linéaire	13,328	1	,000
Nombre d'observations valides	13598		

a. 0 cellules (0,0%) ont un effectif théorique inférieur à 5. L'effectif théorique minimum est de 81,52.

Conclusion 1 : Au vu des résultats du tableau « *Tests de Khi-deux* », on constate que **Sig = 0,000 < α = 0,05**. On rejette H0. Il existe donc un lien entre le niveau d'insécurité alimentaire de l'enfant et le fait qu'il ait eu ou non la diarrhée, au seuil de 5 %. Cependant, ce résultat n'indique pas la force de la relation. Pour cela il faut faire le test de Phi et le V de Cramer.

Syntaxe de réalisation du test de Khi2 avec le test de Phi et le V de Cramer

```
CROSSTABS
/TABLES=ISA BY Q66a
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ PHI
/CELLS=COUNT ROW
/COUNT ROUND CELL.
```

Interprétation des résultats

Etant donné que nous sommes dans un cas de tableau croisé plus grand que le format 2*2, c'est le résultat du V de Cramer qui doit être utilisé pour l'interprétation sur la force de la liaison.

Mesures symétriques

		Valeur	Signification approximée
Nominal par Nominal	Phi	,043	,000
	V de Cramer	**,043**	**,000**
Nombre d'observations valides		13598	

Conclusion 2 : il y a une forte relation entre le niveau d'insécurité alimentaire de l'enfant et le fait qu'il ait eu ou non la diarrhée (*0,40 < **0,43** < 0,80*). Et cette force de la relation est statistiquement significative (***Sig = 0,000< α = 0,05***).

9.3 CORRÉLATION

La corrélation est une quantification de la relation linéaire entre des variables continues.

Question de recherche : « Est-ce que la circonférence du bras chez les enfants est lié au nombre d'aliments complémentaires qu'ils consomment ? »

Type de variable : deux variables quantitatives : la circonférence du bras (MUAC) et le nombre d'aliments complémentaires (Complement).

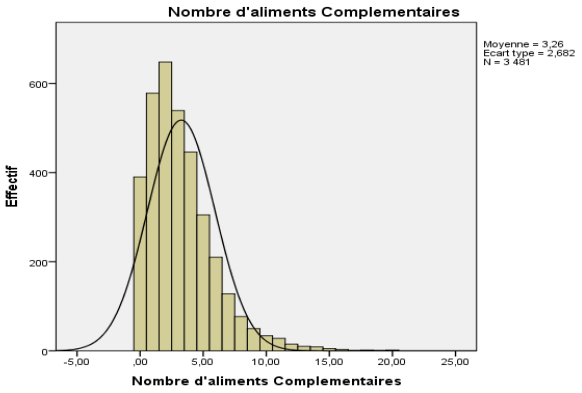
Type d'analyse : Corrélation simple et le risque α = 5 %

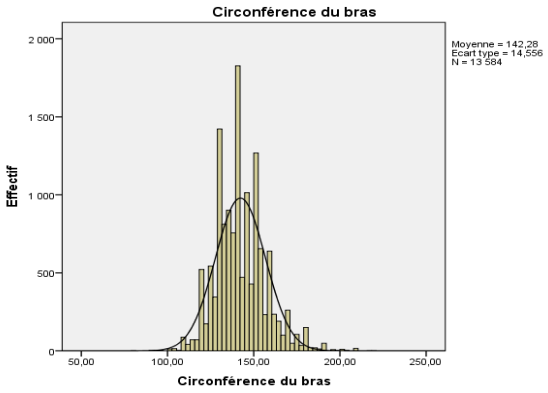
Vérifier la normalité des variables :

- **Faire un histogramme de chaque variable**

Syntaxe du graphique

```
FREQUENCIES VARIABLES=Complement MUAC
/FORMAT=NOTABLE
/HISTOGRAM NORMAL
/ORDER=ANALYSIS
```





Interprétation du graphique

Les courbes sur les deux histogrammes ressemblent à la distribution de la loi normale. Toutefois, vérifions cela grâce au test de normalité.

- **Test de normalité**

Syntaxe du Test

```
EXAMINE VARIABLES=MUAC Complement
/PLOT BOXPLOT STEMLEAF NPPLOT
/COMPARE GROUPS
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

Interprétation du graphique

Le test de normalité compare 2 hypothèses :

- **H0** : les variables suivent une loi normale
- **H1** : les variables ne suivent pas une loi normale

Tests de normalité

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistique	ddl	Signification	Statistique	ddl	Signification
Circonférence du bras	,092	3384	**,000**	,969	3384	**,000**
Nombre d'aliments Complementaires	,158	3384	**,000**	,886	3384	**,000**

a. Correction de signification de Lilliefors

On remarque que :

- Au niveau du **test de Kolmogorov-Smirnov**, le **Sig = 0,000 < α = 0,05**. On rejette donc H0. Les variables ne suivent donc pas une loi normale ;
- Au niveau du **test de Shapiro-Wilk**, le **Sig = 0,000 < α = 0,05**. On rejette aussi H0. Les variables ne suivent donc pas une loi normale.

Le test de normalité vient en contradiction avec les représentations graphiques.

- **Appliquons le test de corrélation.**

Hypothèse de recherche :

- **H0** : il n'existe pas de corrélation entre la circonférence du bras des enfants et le nombre

d'aliments complémentaires qu'ils consomment.

- **H1** : il existe une corrélation entre la circonférence du bras des enfants et le nombre d'aliments complémentaires qu'ils consomment.

Syntaxe de réalisation du test de Corrélation

```
CORRELATIONS
/VARIABLES=MUAC Complement
/PRINT=TWOTAIL NOSIG
/MISSING=PAIRWISE.
```

Interprétation des résultats

Corrélations

		Circonférence du bras	Nombre d'aliments Complementaires
Circonférence du bras	Corrélation de Pearson	1	-,013
	Sig. (bilatérale)		**,461**
	N	13584	3384
Nombre d'aliments Complementaires	Corrélation de Pearson	-,013	1
	Sig. (bilatérale)	,461	
	N	3384	3481

Conclusion : A partir des résultats du tableau « Corrélations », on constate que **Sig =0,461 > α = 0,05**. On accepte donc H0. Il n'existe donc pas de corrélation entre la circonférence du bras des enfants et le nombre d'aliments complémentaires qu'ils consomment.

9.4 RÉGRESSION SIMPLE

Le but de la régression simple est d'établir un lien entre une variable dépendante Y et une variable indépendante X pour pouvoir ensuite faire des prévisions sur Y lorsque X est mesuré (les variables X et Y doivent être quantitative)..

Question de recherche : « Est-ce que la circonférence du bras de l'enfant dépend de son âge ? »

Type de variable : deux variables quantitatives : la circonférence du bras (MUAC) et l'âge (Q63_3).

Type d'analyse : Régression linéaire simple et le risque α = 5 %.

Variable dépendante : MUAC

Variable indépendante : Q63_3

Vérification des conditions d'utilisation :

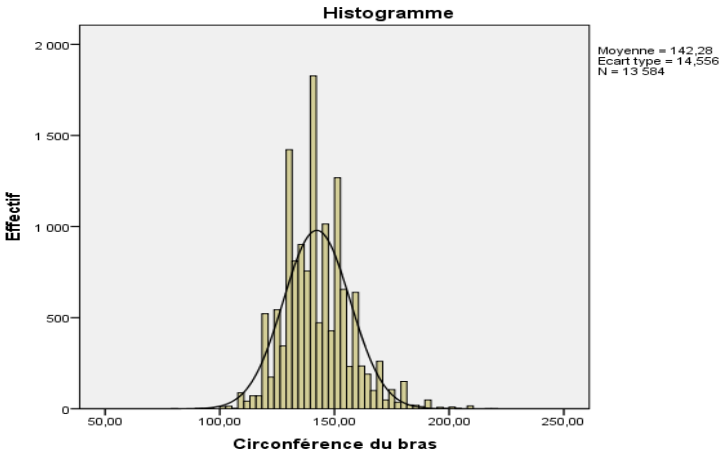
- Normalité de la variable dépendante : vérification à partir d'un histogramme

Syntaxe du graphique

```
FREQUENCIES VARIABLES=MUAC
/FORMAT=NOTABLE
/HISTOGRAM NORMAL
/ORDER=ANALYSIS.
```

Interprétation des résultats

La courbe sur l'histogramme de distribution de la variable dépendante ressemble à une loi normale.



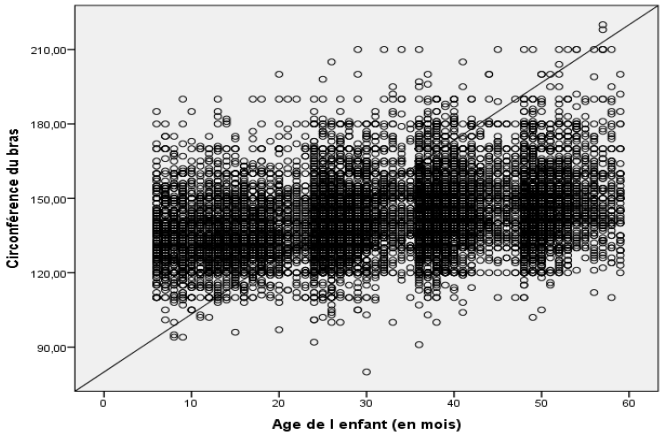
- Relation linéaire entre variable dépendante et variable indépendante : vérification à partir d'un nuage de points

Syntaxe du graphique

```
GRAPH
/SCATTERPLOT(BIVAR)=Q63_3 WITH MUAC
/MISSING=LISTWISE.
```

Interprétation du graphique

On a un nuage de points sensiblement linéaire.



Hypothèse de recherche :

- **H0** : l'âge de l'enfant ne permet pas de prédire la circonférence de son bras ;
- **H1** : l'âge de l'enfant permet de prédire la circonférence de son bras.

Syntaxe de réalisation du test de Régression linéaire simple

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING PAIRWISE
/STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHANGE
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT MUAC
/METHOD=ENTER Q63_3
/SCATTERPLOT=(*ZRESID ,*ZPRED)
/RESIDUALS DURBIN NORMPROB(ZRESID)
/CASEWISE PLOT(ZRESID) OUTLIERS(3).
```

Interprétation des résultats

Corrélations

		Circonférence du bras	Age de l'enfant (en mois)
Corrélation de Pearson	Circonférence du bras	1,000	,330
	Age de l'enfant (en mois)	**,330**	1,000
Sig. (unilatérale)	Circonférence du bras	.	,000
	Age de l'enfant (en mois)	**,000**	.
N	Circonférence du bras	13584	13584
	Age de l'enfant (en mois)	13584	13584

Il existe une corrélation d'intensité moyenne (0,20 < **r =0,33** < 0,50) entre l'âge de l'enfant et la circonférence de son bras et cette corrélation est statistiquement significative au seuil de 5 % (p = 0,000 < α = 0,05). **La condition d'existence de relation linéaire entre les 2 variables est ainsi vérifiée.**

ANOVA[a]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	313903,412	1	313903,412	1662,880	**,000[b]**
	Résidu	2563887,252	13582	188,771		
	Total	2877790,664	13583			

a. Variable dépendante : Circonférence du bras
b. Valeurs prédites : (constantes), Age de l'enfant (en mois)

Selon les résultats du tableau « ANOVA », Sig = 0,000 < α = 0,05. La variable explicative (âge de l'enfant) **contribue de manière très significative** à l'explication de la variable « circonférence du bras ».

Récapitulatif des modèles[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Changement dans les statistiques					Durbin-Watson
					Variation de R-deux	Variation de F	ddl1	ddl2	Sig. Variation de F	
1	,330[a]	**,109**	,109	13,73939	,109	1662,880	1	13582	,000	1,352

a. Valeurs prédites : (constantes), Age de l'enfant (en mois)
b. Variable dépendante : Circonférence du bras

Les résultats consignés dans le tableau « Récapitulatif des modèles », on a **R^2 = 0,109** ce qui permet de dire que seule 10 % de la dispersion est expliquée par le modèle. De ce fait, le résultat suggère que 10,9 % de la circonférence du bras de l'enfant est expliqué par son âge.

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	Statistiques de colinéarité	
		A	Erreur standard	Bêta			Tolérance	VIF
1	(Constante)	131,315	,294		447,275	,000		
	Age de l'enfant (en mois)	,346	,008	**,330**	40,778	**,000**	1,000	1,000

a. Variable dépendante : Circonférence du bras

Conclusion : Il y a une relation de faible intensité entre l'âge de l'enfant et la circonférence de son bras et cette relation est statistiquement significative. **Aussi, nous pouvons conclure que l'âge de l'enfant permet de prédire la circonférence de son bras au seuil de 5 %.**

En remplaçant les coefficients par leur valeur, on a le modèle suivant : MUAC = 131,315 + 0,346 Q63_3.

9.5 RÉGRESSION MULTIPLE

La régression linéaire multiple généralise l'approche adoptée dans la régression linéaire simple. Dans la régression linéaire multiple, la variable dépendante est toujours une variable continue tandis que les variables indépendantes peuvent être continues ou catégorielles.

Question de recherche : « Est-ce que la circonférence du bras de l'enfant dépend ses caractéristiques sociodémographiques, de la présence d'œdèmes bilatéraux, du fait qu'il ait souffert ou pas de la diarrhée et de la classe d'insécurité alimentaire dans lequel il se trouve ? »

Type de variable : variable dépendante (variable quantitative) : la circonférence du bras (MUAC). Variables indépendantes (variables qualitatives ou quantitatives) : sexe de l'enfant (Q63_2), âge de l'enfant (Q63_3), la présence d'œdèmes bilatéraux (Q65), le fait qu'il ait souffert ou pas de diarrhée (Q66a) et la classe d'insécurité alimentaire (ISA).

Type d'analyse : Régression multiple et le risque α = 5 %.

Hypothèse de recherche :

- **H0** : il n'y a pas de relation linéaire entre la combinaison des variables indépendantes (Q63_2, Q63_3, Q65, Q66a, ISA et la variable dépendante (MUAC) :
- **H1** : la combinaison des variables indépendantes est associée significativement à la variable dépendante.

Syntaxe de réalisation du test de Régression multiple

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING PAIRWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA COLLIN TOL CHANGE ZPP
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT MUAC
/METHOD=ENTER Q63_2 Q63_3 Q65 Q66a ISA
/SCATTERPLOT=(*ZRESID ,*ZPRED)
/RESIDUALS DURBIN NORMPROB(ZRESID)
/CASEWISE PLOT(ZRESID) OUTLIERS(3).
```

Les résultats du tableau de « Corrélations » indiquent qu'il existe une très forte relation positive entre la présence des œdèmes bilatéraux sur le corps de l'enfant et la circonférence de son bras. Il existe également une corrélation d'intensité moyenne et positive entre l'âge de l'enfant et la circonférence de son bras.

Par ailleurs, on note une corrélation négative, mais de faible intensité entre le fait que l'enfant ait souffert de diarrhée et la circonférence de son bras. Toutes ces relations sont statistiquement significatives.

Interprétation des résultats

Corrélations

		Circonférence du bras	Sexe de l'enfant	Age de l'enfant (en mois)	Est-ce que (${children_name}) presente des oédemes bilateraux	Diarhée	Classe d'insécurité alimentaire retenue
Corrélation de Pearson	Circonférence du bras	1,000	-,010	**,330**	**,097**	**-,146**	**,019**
	Sexe de l'enfant	-,010	1,000	-,004	-,002	-,003	-,003
	Age de l'enfant (en mois)	,330	-,004	1,000	,060	-,188	,014
	Est-ce que (${children_name}) presente des oédemes bilateraux	,097	-,002	,060	1,000	-,137	,018
	Diarhée	-,146	-,003	-,188	-,137	1,000	-,032
	Classe d'insécurité alimentaire retenue	,019	-,003	,014	,018	-,032	1,000
Sig. (unilatérale)	Circonférence du bras	.	,114	**,000**	**,000**	**,000**	**,013**
	Sexe de l'enfant	,114	.	,311	,395	,365	,361
	Age de l'enfant (en mois)	,000	,311	.	,000	,000	,048
	Est-ce que (${children_name}) presente des oédemes bilateraux	,000	,395	,000	.	,000	,017
	Diarhée	,000	,365	,000	,000	.	,000
	Classe d'insécurité alimentaire retenue	,013	,361	,048	,017	,000	.

ANOVA[a]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	348849,804	5	69769,961	374,598	**,000[b]**
	Résidu	2528940,860	13578	186,253		
	Total	2877790,664	13583			

a. Variable dépendante : Circonférence du bras
b. Valeurs prédites : (constantes), Classe d'insécurité alimentaire retenue, Age de l'enfant (en mois), Sexe de l'enfant, Est-ce que (${children_name}) presente des oédemes bilateraux, Diarhée

En observant le tableau « ANOVA », on a **Sig = 0,000 < α = 0,05**. Le modèle avec toutes les variables indépendantes explique significativement plus de variabilité que le modèle sans prédicteur (avec seulement la moyenne de la circonférence). Toutes les variables introduites contribuent à améliorer significativement la variabilité expliquée par le modèle final. **La prédiction n'est donc pas due au hasard.**

Récapitulatif des modèles[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Changement dans les statistiques					Durbin-Watson
					Variation de R-deux	Variation de F	ddl1	ddl2	Sig. Variation de F	
1	**,348[a]**	**,121**	,121	13,64745	,121	374,598	5	13578	,000	**1,339**

a. Valeurs prédites : (constantes), Classe d'insécurité alimentaire retenue, Age de l'enfant (en mois), Sexe de l'enfant, Est-ce que (${children_name}) presente des oédemes bilateraux, Diarhée
b. Variable dépendante : Circonférence du bras

Conclusion 1 : L'association de toutes les variables indépendantes est relativement faible **R=0,348**. Les variables indépendantes prédisent la variable dépendante, mais il se pourrait qu'une variable indépendante ne contribue pas au résultat. Pour le vérifier, voir le tableau « Coefficients ».

Aussi, les variables explicatives contribuent à raison de 12,1 % (R^2= 0,121) dans la variabilité de la variable à expliquer à savoir la circonférence du bras. La qualité du modèle n'est pas très bonne. Toutefois leur contribution est très significative.

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.	95,0% % intervalles de confiance pour B		Corrélations			Statistiques de colinéarité	
		A	Erreur standard	Bêta			Borne inf	Limite sup	Corrélation simple	Partielle	Partie	Tolérance	VIF
1	(Constante)	120,216	1,709		70,333	,000	116,865	123,566					
	Sexe-enfant	-,314	,235	-,011	-1,337	,181	-,773	,146	-,010	-,011	-,011	1,000	1,000
	Age_enfant	,300	,008	**,311**	37,996	**,000**	,285	,316	,330	,310	,306	,964	1,038
	Presence oédemes	6,645	,801	**,067**	8,291	**,000**	5,074	8,216	,097	,071	,067	,980	1,021
	Diarhée	-2,911	,307	**-,078**	-9,480	**,000**	-3,513	-2,309	-,146	-,081	-,076	,948	1,055
	ISA	,229	,143	,013	1,606	,108	-,051	,508	,019	,014	,013	,999	1,001

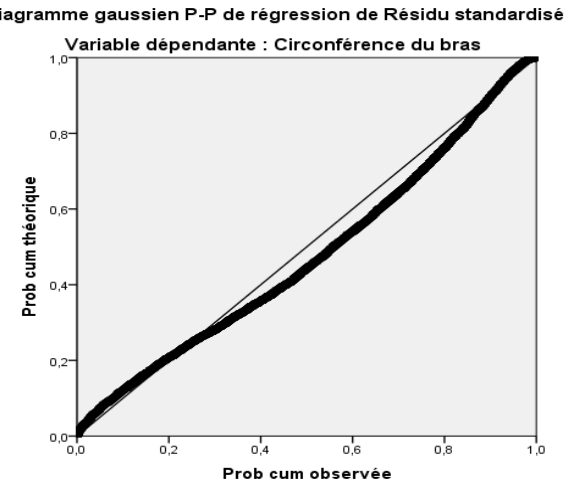
a. Variable dépendante : Circonférence du bras

Conclusion 2 : Le tableau des « Coefficients » montre que la variable « Age de l'enfant » a un effet positif d'intensité moyenne et significatif sur la variable « Circonférence du bras » ***(β = 0,311 et Sig = 0,000 < α = 0,05)***. Aussi cette variable (âge de l'enfant) contribue le plus à la dimension de la circonférence du bras de l'enfant par rapport aux autres variables (elle a le coefficient standardisé **β** le plus élevé de tous).

La variable « Présence des œdèmes sur le corps de l'enfant » a un effet positif, faible et significatif sur la variable à expliquer. La variable « Diarrhée » a, quant à elle, un effet négatif, faible mais significatif sur la variable à expliquer. Les autres variables indépendantes ne présentent pas d'effet significatif sur la variable dépendante « Circonférence du bras ».

Conclusion 3 : Le test de Durbin-Watson dans le tableau « *Récapitulatif des modèles* » permet de vérifier la validité du modèle. Dans le tableau, on a **1 < DW = 1,339 < 3** ; ce qui signifie que **les résidus ne sont pas corrélés et que le modèle de régression est valide**. Toutefois, il faut procéder à l'examen du graphique pour confirmer et valider cette conclusion.

Bien qu'il ait une légère déviation au niveau de la courbe, les résidus standardisés suivent une droite. Aussi, aucun résidu ne présente une valeur statistiquement trop élevée. Cela confirme la normalité de leur distribution ; la prédiction est donc valable et appropriée.



Conclusion Final : Les variables indépendantes prédisent la variable dépendante, toutefois, la variable « âge de l'enfant » contribue le plus à la prédiction de la dimension de la circonférence du bras de l'enfant. Mais la qualité du modèle n'est pas très bonne, ce qui n'empêche pas le modèle de régression d'être valide étant donné que les résidus ne sont pas corrélés.

9.6 RÉGRESSION LOGISTIQUE

La régression logistique propose de tester un modèle de régression dont la variable dépendante est dichotomique (codée « 0 »-« 1 ») et dont les variables indépendantes peuvent être continues ou catégorielles.

Question de recherche : « quelles sont les raisons pour lesquelles un enfant ait souffert de diarrhée au cours des 2 dernières semaines précédant l'enquête ? »

Type de variable : variable dépendante (variable qualitative) : présence ou non de diarrhée (Q66a). **Variables indépendantes** (variables qualitatives ou quantitatives) : le sexe de l'enfant (Q63_2), l'âge de l'enfant (Q63_3), la présence d'œdèmes bilatéraux (Q65), la circonférence du bras (MUAC), la classe d'insécurité alimentaire (ISA) et la classe d'insécurité alimentaire du ménage

Type d'analyse : Régression logistique et le risque α = 5 %.

Hypothèse de recherche :

- **H0** : la combinaison des variables indépendantes (Q63_2, Q63_3, Q65, MUAC, ISA, ISA_2CLASSE) ne parvient pas à mieux expliquer la présence/absence de la variable dépendante qu'un modèle sans variables indépendantes ;
- **H1** : au moins une variable indépendante est associée significativement à la variable dépendante.

Syntaxe de réalisation du test de Régression logistique

```
LOGISTIC REGRESSION VARIABLES Q66a
/METHOD=ENTER Q63_2 Q63_3 Q65 MUAC ISA ISA_2CLASSE
/SAVE=PRED PGROUP ZRESID
/CLASSPLOT
/CASEWISE OUTLIER(2)
/PRINT=GOODFIT ITER(1) CI(95)
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Interprétation des résultats

Historique des itérations[a,b,c]

Itération		-2log-vraisemblance	Coefficients Constante
Etape 0	1	13281,209	-1,243
	2	13184,004	-1,442
	3	13183,683	-1,454
	4	**13183,683**	-1,454

a. La constante est incluse dans le modèle.

b. -2log-vraisemblance initiale : 13183,683

c. L'estimation a été interrompue au numéro d'itération 4 parce que les estimations de paramètres ont changé de moins de ,001.

Récapitulatif des modèles

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	**12412,013[a]**	,055	,089

a. L'estimation a été interrompue au numéro d'itération 5 parce que les estimations de paramètres ont changé de moins de ,001.

Commentaires des tableaux : Le tableau « Historique des itérations » est celui du modèle de base. On observe que la probabilité **(-2LL) = 13183,683** et c'est cette probabilité que nous cherchons à améliorer en ajoutant les variables indépendantes.

On observe également que dans le tableau « Récapitulatif des modèles », la probabilité (-2LL) qui est = 12414,013 est inférieur à la probabilité du modèle de base et la différence qui est de **771,670** est évaluée dans une distribution de Khi2.

Tests de spécification du modèle

		Khi-Chi-deux	ddl	Sig.
Etape 1	Etape	**771,670**	6	**,000**
	Bloc	771,670	6	,000
	Modèle	**771,670**	6	**,000**

Conclusion 1 : la différence de probabilité du modèle de base et du modèle final étant significative au seuil de 5 %, nous pouvons dire que le modèle final permet de prédire significativement mieux la probabilité pour un enfant de souffrir de diarrhée que le modèle incluant seulement la constante.

A la lecture du tableau du « Test de Hosmer-Lemeshow », il n'existe pas d'écart significatif entre les valeurs prédites et celles observées (**Sig = 0,695 > α = 0,05**) et ce, dès la 1ère étape.

Test de Hosmer-Lemeshow

Etape	Khi-Chi-deux	ddl	Sig.
1	5,573	8	,695

Variables dans l'équation

		A	E.S.	Wald	ddl	Sig.	Exp(B)	IC pour Exp(B) 95%	
								Inférieur	Supérieur
Etape 1[a]	Q63_2	-,035	,045	,599	1	,439	,966	,884	1,055
	Q63_3	-,029	,002	**279,160**	1	**,000**	,971	,968	,974
	Q65	-1,448	,122	**140,205**	1	**,000**	,235	,185	,299
	MUAC	-,017	,002	**93,444**	1	**,000**	,983	,980	,987
	ISA	-,246	,047	**26,934**	1	**,000**	,782	,713	,858
	ISA_2CLASSE	,437	,106	**16,995**	1	**,000**	1,548	1,258	1,906
	Constante	4,721	,345	187,139	1	,000	112,324		

a. Variable(s) entrées à l'étape 1 : Q63_2, Q63_3, Q65, MUAC, ISA, ISA_2CLASSE.

Conclusion 2 : les variables : « âge de l'enfant », « présence des œdèmes bilatéraux », « circonférence du bras », « classe d'insécurité alimentaire » et « classe d'insécurité alimentaire du ménage » contribuent significativement à l'amélioration du modèle final. Toutefois, la variable « âge de l'enfant » à une contribution plus élevée.

Récapitulatif des modèles

Etape	-2log-vraisemblance	R-deux de Cox & Snell	R-deux de Nagelkerke
1	12412,013[a]	,055	,089

a. L'estimation a été interrompue au numéro d'itération 5 parce que les estimations de paramètres ont changé de moins de ,001.

Conclusion 3 : les valeurs des **R^2** dans le tableau « Récapitulatif des modèles » montrent que le modèle n'est pas bien ajusté aux données. En calculant le pseudo **R^2** on obtient : **R^2 = 0,0585** (**771,670 / 13 183,638**). Ainsi, les variables explicatives contribuent à raison de 5,85 % dans la variabilité de la variable à expliquer.

Conclusion Final : Le modèle final permet de mieux prédire la probabilité pour un enfant de souffrir de la diarrhée que le modèle incluant seulement la constante. Notons que seule la variable « sexe de l'enfant » ne contribue pas significativement à l'amélioration de la prédiction. Par ailleurs, le modèle final n'est pas mieux ajusté aux données.

9.7 ANALYSE EN COMPOSNTES PRINCIPALES (ACP)

L'ACP permet de traiter simultanément un nombre quelconque de variables toutes quantitatives.

Question de recherche : on veut savoir s'il y a des groupes qui se distinguent dans la base alimentation chez les enfants.

Type de variable : toutes les variables ciblées sont quantitatives.

Afin de remplir la condition sur le nombre des variables à analyser, cet exemple inclut 2 variables qui donnent la même information (MUAC et Q64_4)

Syntaxe de réalisation de l'ACP

```
FACTOR
/VARIABLES children_hh_position Q63_3 Q64_4 Complement MUAC
/MISSING LISTWISE
/ANALYSIS children_hh_position Q63_3 Q64_4 Complement MUAC
/PRINT UNIVARIATE INITIAL KMO ROTATION FSCORE
/FORMAT SORT
/PLOT EIGEN
/CRITERIA MINEIGEN(1) ITERATE(25)
/EXTRACTION PC
/CRITERIA ITERATE(25)
/ROTATION VARIMAX
/SAVE AR(ALL)
/METHOD=CORRELATION.
```

Interprétation des résultats

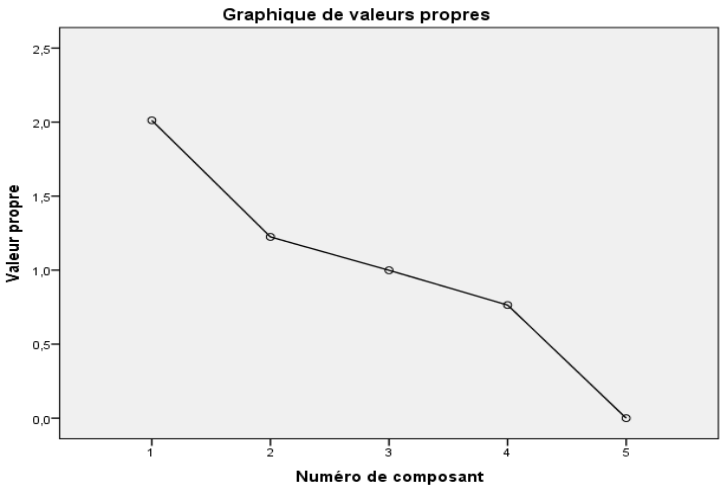
Variance totale expliquée

Composante	Valeurs propres initiales			Somme des carrés des facteurs retenus pour la rotation		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	2,012	**40,244**	40,244	2,010	40,193	40,193
2	1,225	24,491	**64,735**	1,227	24,542	64,735
3	,999	19,978	84,713			
4	,764	15,287	100,000			
5	1,094E-013	2,886E-013	100,000			

Méthode d'extraction : Analyse en composantes principales.

Le tableau de la « Variance totale expliquée » donne une idée sur le degré d'informations que représente chaque composante ou facteur. Selon les résultats, la plus grande valeur propre de la matrice de corrélation est de **2,012** et elle est associée au 1er axe qui explique **40,244 %** de la variabilité. On choisit les **2 premiers axes** qui restituent **64,735 %** de la variance.

Le graphique conforte dans ce choix des axes.



Matrice des composantes après rotation[a]

	Composante	
	1	**2**
Perimètre brachial de (${children_name}) en cm	**,999**	-,007
Circonférence du bras	**,999**	-,007
Nombre d'aliments Complementaires	-,035	**,789**
Age de l'enfant (en mois)	,113	**,773**
N° de l'enfant dans la compsition du ménage (Q12A)	,009	-,080

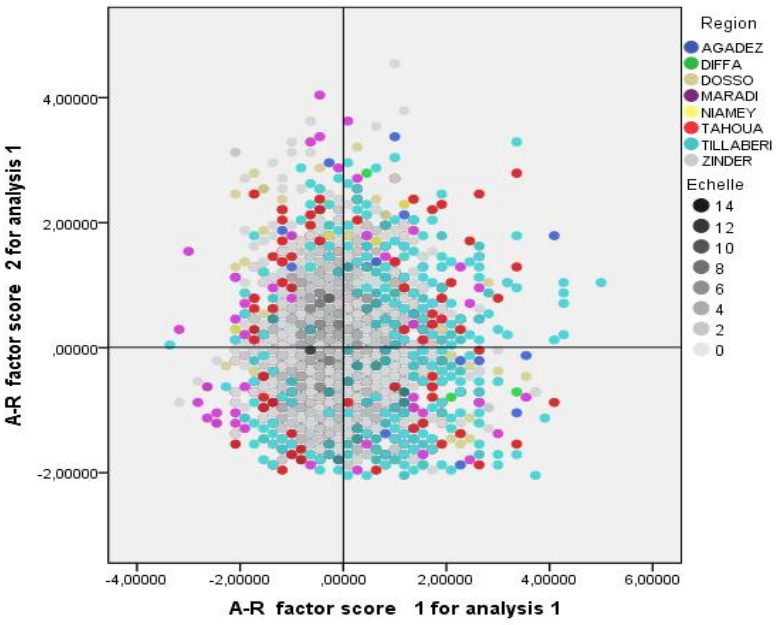
Méthode d'extraction : Analyse en composantes principales.
Méthode de rotation : Varimax avec normalisation de Kaiser.
a. La rotation a convergé en 3 itérations.

L'axe 1 est fortement et positivement corrélé avec les variables 'Périmètre brachial de l'enfant' et 'Circonférence du bras'. Aussi, il est négativement corrélé avec la variable 'Nombre d'aliment complémentaire'. Les autres corrélations sont relativement faibles.

L'axe 2 est quant à lui fortement et positivement avec les variables 'Nombre d'aliments complémentaires' et 'Age de l'enfant' ; les autres corrélations sont relativement faibles.

Syntaxe pour la représentation des individus sur les axes principaux

```
GRAPH
/SCATTERPLOT(BIVAR)=FAC1_1 WITH FAC2_1 BY region
/MISSING=LISTWISE.
```



On remarque que selon les axes principaux retenus le premier groupe qui est constitué de la mesure de la circonférence brachial est plus observé dans la région de TILLABERI.

Le deuxième groupe constitué du nombre d'aliments complémentaires consommés par l'enfant et son âge est plus observé dans la région de TAHOUA.

9.8 ANALYSE DES CORRESPONDANCES MULTIPLES (ACM)

L'Analyse des Correspondant Multiples (ACM) est une méthode qui permet d'étudier l'association entre au moins 2 variables qualitatives.

Question de recherche : on veut savoir s'il y a des groupes qui se distinguent dans la base alimentation chez les enfants.

Type de variable : toutes les variables ciblées sont qualitatives.

Syntaxe de réalisation de l'ACM

```
MULTIPLE CORRES VARIABLES=Q63_2 cons_ind cons_ind1 Q65 ISA
ISA_2CLASSE SCORE_7 TRANCHAGE Q66
/ANALYSIS=Q63_2(WEIGHT=1) cons_ind(WEIGHT=1)
cons_ind1(WEIGHT=1) Q65(WEIGHT=1) ISA(WEIGHT=1)
ISA_2CLASSE(WEIGHT=1) SCORE_7(WEIGHT=1)
TRANCHAGE(WEIGHT=1) Q66(WEIGHT=1)
/MISSING=Q63_2(PASSIVE,MODEIMPU)
cons_ind(PASSIVE,MODEIMPU) cons_ind1(PASSIVE,MODEIMPU)
Q65(PASSIVE,MODEIMPU) ISA(PASSIVE,MODEIMPU)
ISA_2CLASSE(PASSIVE,MODEIMPU) SCORE_7(PASSIVE,MODEIMPU)
TRANCHAGE(PASSIVE,MODEIMPU) Q66(PASSIVE,MODEIMPU)
/DIMENSION=2
/NORMALIZATION=VPRINCIPAL
/MAXITER=100
```

```
/CRITITER=.00001
/PRINT=CORR DISCRIM
/PLOT=OBJECT(20) JOINTCAT(Q63_2 cons_ind cons_ind1 Q65 ISA
ISA_2CLASSE SCORE_7 TRANCHAGE) (20)
DISCRIM (20).
```

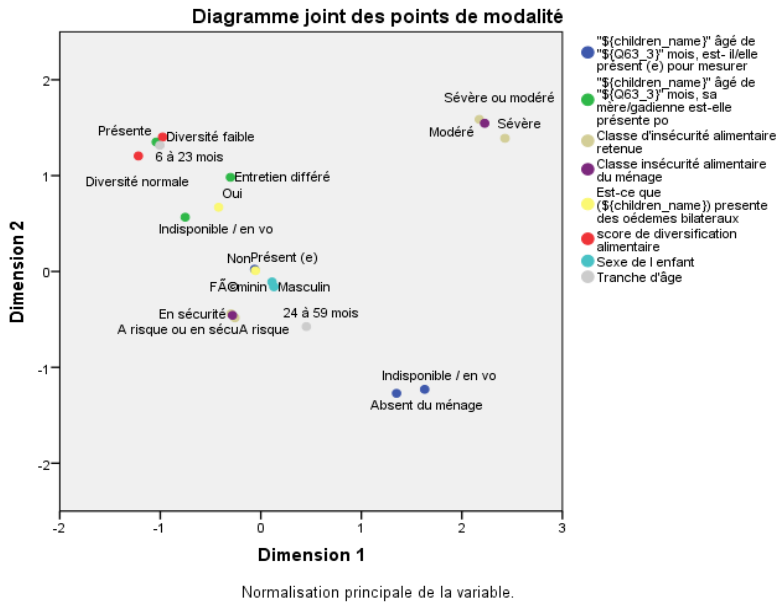
Interprétation des résultats

Récapitulatif des modèles

Dimension	Alpha de Cronbach	Variance expliquée	
		Total (valeur propre)	Inertie
1	,726	2,823	,314
2	,719	2,768	,308
Total		5,591	**,621**
Moyenne	,723[a]	2,795	,311

a. La valeur Alpha de Cronbach moyenne est basée sur la valeur propre moyenne.

Selon les résultats du tableau « Récapitulatif des modèles », l'ensemble des variables a généré 2 facteurs ou 2 dimensions d'un total qui est de 62,1 %. Ces 2 facteurs résument donc 62,1 % de l'ensemble des informations des variables qui ont été introduites dans l'analyse.



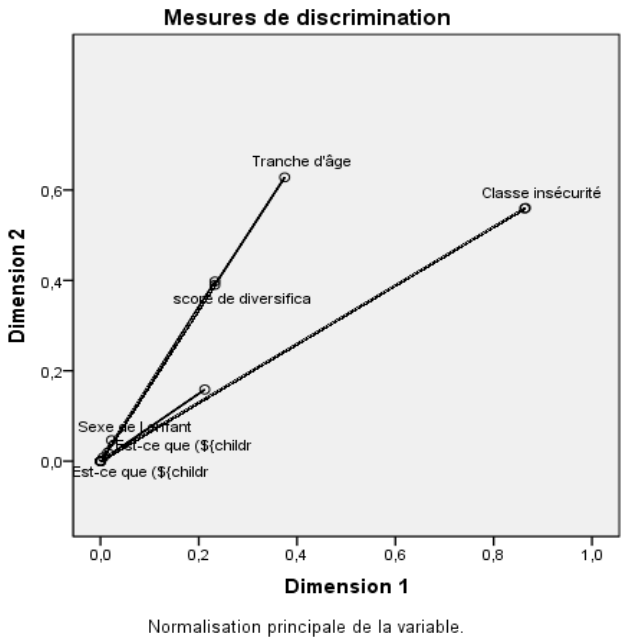
En observant le « diagramme joint des points de modalité », on remarque que la 1ère dimension distingue l'indisponibilité ou l'absence de l'enfant dans le ménage pour la prise de ses mesures et la classe des ménages qui ont un niveau d'insécurité alimentaire sévère ou modéré, ainsi que la classe des enfants ayant un niveau d'insécurité alimentaire modéré et sévère.

La 2nde dimension, distingue les enfants âgés de 6 à 23 mois ayant leur mère ou gardienne présente pour leur prise de mesure et les enfants ayant un score de diversification alimentaire faible ou normale, des autres modalités de variable.

Mesures de discrimination

	Dimension		Moyenne
	1	2	
Sexe de l'enfant	,015	,019	,017
"${children_name}" âgé de "${Q63_3}" mois, sa mère/gadienne est-elle présente po	,233	,398	,315
"${children_name}" âgé de "${Q63_3}" mois, est- il/elle présent (e) pour mesurer	,212	,159	,185
Est-ce que (${children_name}) presente des oédemes bilateraux	**,006**	**,008**	,007
Classe d'insécurité alimentaire retenue	**,865**	**,561**	,713
Classe insécurité alimentaire du ménage	**,863**	**,559**	,711
score de diversification alimentaire	,233	,390	,311
Tranche d'âge	,375	**,628**	,501
Est-ce que (${children_name}) a souffert de diarrhée au cours des 2 derniéres se	,022	,046	,034
Total actif	2,823	2,768	2,795

Les variables « Classe d'insécurité alimentaire retenue » et « Classe d'insécurité alimentaire du ménage » sont positivement fortement corrélées avec la dimension 1 cependant, cette intensité est faible au niveau de la dimension 2.



Normalisation principale de la variable.

La dimension 2 quant à elle est fortement corrélée avec la variable « Tranche d'âge ». Par ailleurs, l'intensité de la variable « Présence des œdèmes bilatéraux » est relativement très faible au niveau des 2 dimensions.

Le dernier diagramme « Mesures de discrimination » ci-dessous indique que la 1ère dimension est liée aux variables « Classe d'insécurité alimentaire » et « Présence de l'enfant pour la prise de mesure ». La 2ième dimension est liée aux variables « Tranche d'âge » et « Score de diversification alimentaire ».

Les variables « Sexe de l'enfant » et « Présence de la mère/gardienne lors de la prise de mesure de l'enfant » sont proches de l'origine et ne font aucune distinction entre les 2 dimensions.

SOFRECO


PNiN
Plateformes nationales
d'information pour la nutrition