N° d'ordre:16/2016-M/MT

REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE

Ministère De L'enseignement Supérieur Et De La Recherche

ScientifiqueUniversité Des Sciences Et De La Technologie Houari Boumediene

Faculté Des Mathématiques

Mémoire
Présenté pour l'Obtention du Diplôme de MAGISTER
En MATHEMATIQUES

Spécialité : Statistiques Mathématiques & Probabilités

Par : BOUNI Nora

THEME

L'Analyse Exploratoire et L'Analyse Confirmatoire

Aide à la Décision

Soutenu publiquement, le 12/05/2016, devant le jury composé de :

M. A.TATACHAK	Professeur à L'USTHB	Président
M. M. DJEDOUR	Professeur à L'USTHB	Directeur de Mémoire
Mme. H.SAGGOU	Maitre de conférences/A à L'USTHB	Examinatrice
M. T.MEDKOUR	Maitre de conférences/A à L'USTHB	Examinateur

Contents

Préambule

L'institut national de santé publique d'Algérie a entamé, au cours de l'année 2005, un projet de recherche portant sur l'étude de la transition sanitaire dans les pays d'Afrique du Nord dénommé TAHINA (Transition and Health Impact in North Africa), pour tenter de quantifier et mieux saisir les profils récents de morbidité et mortalité, y compris leurs causes.

L'analyse des données de TAHINA s'est essentiellement appuyée sur des travaux à caractère ordinaire. Les résultats n'on pas pour autant répondu à l'objectif assigné : augmenter l'attention à la prévention des maladies chroniques non transmissibles.

Notre intérêt dans ce mémoire est d'utiliser l'analyse factorielle exploratoire et l'analyse factoriel confirmatoire pour permettre un éclairage nouveau de l'influence des habitudes alimentaires et l'environnement socio-économique sur les différentes maladies chroniques, ce que permettre de prise des décisions à caractère économique et sociale.

Pour ce faire nous allons extraire un fichier de travail de l'enquête TAHINA nommé « tahina-ex » de 4818 individus et 147 variables, les 147 variables sont regroupées comme suit :

- Groupe 1 : localisation régionales (régions, milieu).

- Groupe 2 : État civil et différentes occupation (âge, sexe, activité, sport, tabac,...).
- Groupe 3 : produits Alimentaires (biscuits, artichaut, moutons,...).
- Groupe 4 : Fréquence Alimentaire par jour (fruit secs, protéine, ...).
- Groupe 5 :Etats morbides de l'individu (diabète, dyslipidémie, hypertension artérielle).
- Groupe 6 : pathologies des antécédents (diabète, dyslipidémie, hypertension artérielle).
- Groupe 7 : Examen clinique (surpoids, pression artérielle systolique,...).

Les variables à expliquer sont :Diabète,Dyslipidémie, Hypertension artérielle Les variables explicatives sont toutes les variables des autres groupes.

Les variables du groupe 4 et quelques variables de groupe 7 sont de types quantitatifs. Les autres groupes de« tahina-ex » sont des variables qualitatives de type Dichotomique.

Dans le but de mieux saisir les facteurs de risque des maladies chroniques sélectionnées (les variables à expliquer) et pour arriver à une solution satisfaisante, on va suivre les étapes suivantes :

1. A partir de notre fichier de travail On tire de façon aléatoire un échantillon de 50% des individus.On obtient deux fichiers: « tahina-ex1 » et « tahina-ex2 », chaque fichier des deux contient 2409 individus et 147 variables.
2. Application de l'analyse exploratoire sur « tahina-ex1 ».
3. Application de l'analyse confirmatoire sur « tahina-ex2 ».

Ce mémoire comporte quatre chapitres, dans le premier chapitre on présente l'objectif, le questionnaire et l'échantillon de l'enquête TAHINA (4818 individus et 1312 variables). Nous allons aborder aussi l'objectif de l'étude et la méthodologie suivie, ainsi qu'une présentation de logiciel LISREL utilisé pour l'analyse exploratoire et l'analyse confirmatoire.

Dans le deuxième chapitre on donne un rappel sur la théorie d'analyse en composantes principales normée, après on explique la théorie d'analyse factorielle exploratoire : l'objectif de l'analyse, les équations, l'estimation des paramètres, les types de rotation, considérations théoriques et pratiques, les étapes de l'analyse. On présente aussi dans ce chapitre l'analyse factorielle confirmatoire : l'utilisation de l'analyse, le modèle et les méthodes d'estimation.

Le troisième chapitre est une application numérique de l'analyse descriptive et de l'analyse en composantes principales sur des variables de notre fichier de travail sélectionnée, une application de l'analyse factorielle exploratoire sur « tahina-ex1 » et enfin une application de l'analyse factorielle confirmatoire sur « tahina-ex2 ».

Le quatrième chapitre est une conclusion générale de notre travail.

Pour effectués l'analyse exploratoire et l'analyse confirmatoire on a eu recours au logiciel LISREL qui a donnée des résultats significatifs dans ce domaine.

LISREL permet de travailler avec les variables ordinales largement présentes dans notre fichier de travail.

LISREL formalise les relations entre variables latentes et variables observées, il est développé par Jöreskog et Sörbom en 1970, utilisé surtout dans l'analyse factorielle confirmatoire et l'analyse des systèmes de relation.

Ces 30 dernières années, le modèle, les méthodes et le logiciel LISREL sont devenus synonymes de modélisation en équations structurelles. La modélisation en équations structurelles permet aux chercheurs en sciences sociales, en sciences de gestion, en sciences du comportement, en sciences biologiques, en sciences de l'éducation et dans

d'autres domaines encore de tester empiriquement leurs théories. Ces théories sont généralement formulées sous la forme de modèles théoriques composés de variables observées et de variables latentes (non-observées). Une fois que des données ont été collectées pour les variables observées du modèle théorique, le programme LISREL peut être utilisé pour tester l'ajustement du modèle aux données. Ce logiciel permet ainsi de réaliser des analyses factorielles confirmatoires. LISREL dispose également d'une interface permettant de dessiner des modèles et de générer directement une syntaxe à partir du dessin.

LISREL: Ce logiciel disponible notamment pour Windows, est avant tout destiné aux modèles d'équations structurelles, mais il dispose également d'un très bon module pour l'analyse multi-niveaux. Il s'agit d'un logiciel payant, mais sa version étudiante, limitée par la taille des modèles qui peuvent être définis (15 variables), est parfaitement fonctionnelle et gratuite.

Chapter 1

Introduction générale

1.1 Présentation de l'enquête TAHINA

Le Projet TAHINA « Transition Epidémiologique et Impact sur la Santé en Afrique du Nord » est un projet de recherche financé par l'Union Européenne dans le cadre du programme INCO « Confirming the international Role of Community Research ».

Il part du principe que la transition épidémiologique, caractérisée par la persistance ou la réémergence des « maladies du passé » et l'augmentation de l'importance des maladies chroniques, pose de façon accrue la problématique des (la) stratégie(s) d'intervention sanitaire à lancer sur le terrain. Les actions engagées par l'Institut National de Santé Publique dans le cadre du projet TAHINA sont des tentatives de réponse visant l'élaboration de recommandations à l'attention des acteurs du système de santé impliqués dans la gestion de cette transition. Lesquelles recommandations, réunies avec d'autres données sanitaires issues d'autres sources, apporteront une contribution à la réorientation du système de santé amorcée déjà depuis quelques mois dans le cadre des réformes sanitaires.

TAHINA a été réalisé en collaboration avec l'Office National des Statistiques, au terme de plusieurs réunions de travail, sur la base de la note méthodologique pour

l'échantillonnage. Ainsi 126 districts, ont été identifiés et délimités géographiquement, en respectant la répartition en zone rurale et urbaine. Au total 16 wilayas, 64 communes et 126 districts ont été tirés au sort.

Ce travail a été suivi par la numérisation graphique des territoires géographiques lieu d'enquête; cartes utilisées par les enquêteurs sur le terrain pour identifier les ménages à enquêter.

Le nombre de ménages est de 4818 dont 2930 ménages urbains, soit 60,8 % et 1888 ménages ruraux. La taille des ménages varie de 1 à 25 personnes avec une moyenne et une médiane sensiblement identique respectivement de 6,7 et 6,0 personnes L'écart type est de 2,8. L'enquête a couvert une population de 32 463 personnes qui se répartissent selon la dispersion « urbain 60.8%, rural 39.2% ».

L'enquête s'est déroulée sur le terrain du 11 juin au 07 juillet 2005.

Après reconnaissance géographique des districts, des avis de passage on été distribués aux ménages et un planning a été retenu en accord avec les ménages qui ont accepté de participer à l'enquête. Au niveau de chaque district, 40 ménages sont à enquêter. Une fois que les 40 ménages ont été enquêtés l'équipe est passe à l'autre district où là aussi 40 ménages ont été retenus. Ainsi, une équipe a eu à interroger au total 80 ménages.

La définition du « ménage » est : « un ménage ordinaire généralement composé d'un groupe de personnes vivant ensemble dans un même logement sous la responsabilité d'un chef de ménage, préparant et prenant en général les principaux repas ensemble. Ces personnes sont généralement liées entre elles par le sang, par le mariage ou par alliance ».

1.1.1 Objectif de l'enquête

Les objectifs généraux de l'enquête visent à :

- renforcer la capacité des services de santé à gérer les problèmes posés par l'avancée de la transition épidémiologique à travers une stratégie globale, intégrée et multisectorielle
- augmenter l'attention à la prévention des maladies chroniques non transmissibles de tous les secteurs concernés par les changements dans les modes de vie. Les objectifs spécifiques se proposent de :
- Mesurer la charge de morbidité globale et ses coûts associés liés à la transition épidémiologique
- Caractériser les déterminants alimentaires, économiques, sociaux, culturels et environnementaux de cette situation
- Identifier les représentations et l'évolution des pratiques actuelles des professionnels de santé face aux changements de la situation sanitaire et nutritionnelle
- Identifier la perception et la sensibilité des acteurs d'autres secteurs concernés sur ces changements
- Mettre en évidence l'évolution des représentations et des pratiques de la population en matière de santé, d'alimentation et de modes de vie
- Initier un processus d'élaboration conjoint de stratégies d'interventions intégrées et globales.

1.1.2 Le questionnaire

Une fois le ménage identifié, les enquêteurs vérifient sur le livret de famille que le ménage comporte au moins une personne âgée entre 35 et 70 ans. Lorsqu'il n'y a aucune personne correspondant à cette tranche d'âge, l'équipe change de ménage. Lorsqu'il y a plusieurs personnes de cette tranche d'âge et pour éviter une surreprésentation féminine, un tirage au sort est effectué sur place pour désigner la personne à enquêter. Les différents volets du questionnaire sont alors passés à la personne tirée au sort à savoir :

- Volet A « aspects socioéconomiques – morbidité - facteurs de risque »
- Volet B « nutrition : aliments consommés – fréquence de consommation alimentaire –pratiques alimentaires »
- Volet C « activité physique » • Volet D « qualité de vie »
- Volet E « échelle d'attitudes »

1.1.3 L'échantillon

4 818 ménages dont 2 930 ménages urbains et 1 888 ménages ruraux (Le nombre total d'individus les composant est de 32 463) des quels seront tirés 4 818 personnes âgées de 35 à 70 ans

La méthodologie utilisée est inspirée de celle de l'indice de développement humain(IDH) (PNUD) qui permet de positionner chaque wilaya à partir des indices de la tendance démographique (ITD), de la situation sanitaire (ISS), d'encadrement sanitaire (IES), de commodité de logement (ICL) et de la situation économique (ISE).

La moyenne de ces indices donne un Indice Global de la Situation Sanitaire et Sociale noté (IGSS). A partir de cet indice ils ont constitué les typologies des wilayas ayant un même niveau de développement sanitaire et social.

L'indice composite de développement social sanitaire a permis la répartition des 48 wilayas en 6 strates.

Répartition des wilayas par strate

- Strate 1 : Alger
- Strate 2 : Tizi-Ouzou
- Strate 3 : Oran, Tlemcen, Blida, Bejaia, Skikda, El Tarf, Sétif, Batna, Bouira, Tipaza

- Strate 4 : Constantine, Boumerdes, Annaba, Oum El Bouaghi, S.B Abbés, Médéa Guelma, Chlef, Jijel, Mascara, Ain Temouchent, Mila, Ain Defla, Bordj Bou Arérridj, Tébessa, Souk Ahras, Relizane, M'sila
- Strate 5 : Saïda, Mostaganem, Biskra, Khenchela, Tiaret, Bechar, Ghardaïa, Laghouat Ouargla, Tissemsilt
- Strate 6 : Djelfa, Naâma, El Bayadh, El-Oued, Tamanrasset, Ilizi,Tindouf, Adrar

La répartition de l'échantillon selon la strate et la dispersion géographique commune (urbaine et rurale) se présente comme suit :

Strates	Total ménages	Ménage urbain	Ménage rural	Commune urbaine	Commune rurale	Total
Strate1	432	394	38	5	1	6
Strate 2	216	77	139	1	2	3
Starte 3	1224	834	390	12	6	18
Strate 4	1800	1145	655	16	9	25
Strate 5	576	308	268	4	4	8
Strate 6	288	86	202	2	2	4
Total	4536	2844	1692	40	24	64

1.2 Le choix de notre fichier de travail

Le fichier TAHINA (4818 individus, 1312 variables) contenant excessivement de variables ayant un taux très élevé des valeurs manquantes, pour cela on sélectionne 147 variables qui contenant moins de 5% de ces valeurs, on obtient un fichier de travail « tahina-ex » de 4818 individus et 147 variables.

Les variables retenus comportent deux types de variable, des variables quantitatives et d'autre qualitatives (de type nominal et Dichotomique). Une variable est dite, selon le cas :

Quantitative : Ses valeurs sont des nombres exprimant une quantité, sur lesquels les opérations arithmétiques (somme, ...) ont un sens. La variable peut alors être discrète ou continue selon la nature de l'ensemble des valeurs qu'elle est susceptible de prendre (valeurs isolées ou intervalle de R).

Qualitative : Ses valeurs sont des modalités, (ou catégories, ou caractères) exprimées sous forme littérale ou par un codage numérique sur lequel des opérations arithmétiques n'ont aucun sens. On distingue des variables qualitatives ordinales ou nominales, selon que les modalités peuvent être naturellement ordonnées ou pas.

- Dichotomique : Une variable est dichotomique si elle n'a que 2 modalités.
- Binaire : une variable binaire ne peut prendre que deux valeurs : 0 ou 1. Cette variable est souvent obtenue par éclatement des modalités d'une variable nominale
- Ordinale : une variable ordinale prend ses valeurs dans les nombres cardinaux (nombres entiers) au sein d'un intervalle ayant une valeur minimale et maximale

1.2.1 Les variables sélectionnées

Les variables retenus sont regroupées comme suit :

Groupe 1 : localisation régionales.

Groupe 2 : État civil, différent occupation, activités diverses.

Groupe 3 : produits alimentaires.

Groupe 4 : Fréquence alimentaires.

Groupe 5 : Etats morbides de l'individu.

Groupe 6 : pathologies des antécédents.

Groupe 7 : Examen clinique.

Le tableau ci-dessous nous donne le code, les valeurs et le type de ces variables

Table 1.2.1: « tahina-ex »

Groupes	Variable	Code	Valeurs	Type
Groupe 1	Régions	régions	Nominale	Nominale
	Milieu	millieu		
Groupe 2	age	age-c4	(1)<40 ans	Ordinale
	sexe	sexe	Femme	Nominale
			Homme	
	poids	poids	(kg)	Quantitative
	Avez-vous une activité professionnelle principale?	activit 1	1:oui, 2:Non	Dichotomique

Groupes	Variable	Code	Valeurs	Type
	les trajets parcourus en marchant compotaient-il généralement une partie plus difficile,comme une série d'au moins 20 marches d'escalier	marchdit	(1)Non (2)oui=20 (3) oui>20	Ordinale
	Pratiquez-vous un sport?	sport	1:oui, 2:Non	Dichotomique
	Avez-vous l'habitude de faire une sieste dans la journée parfois ?	sieste	1:oui, 2:Non	Dichotomique
	Habituellement, vous arrive-t-il de manger quelque chose en dehors des repas?	entrepas	1:oui, 2:Non	Dichotomique
	Au cours du mois dernier, avez-vous mangé à l'extérieur de chez vous ?	exterier	1:oui, 2:Non	Dichotomique
	Fumez-vous des produits tabagiques	Smoke	1:Non fumeur 2-Fumeur actuel 3-Ancien fumeur	Nominal

Groupes	Variable	Code	Valeurs	Type
	Consommez-vous actuellement le tabac non fume tel que le tabac a priser, a snifer ou a mâcher	Tabacnf	1:oui, 2:Non	Dichotomique
Groupe 3	Pain de boulanger	Painboul	1:oui, 2:Non	Dichotomique
	Biscotte	Biscotte	"	"
	Couscous	Couscous	"	"
	Pâtes alimentaires (vermicelle, macaroni, spaghetti)	Patealim	"	"
	Haricot sec	haricsec	"	"
	Lentille	Lentille	"	"
	Pois cassé	poiscass	"	"
	Pomme de terre (sauf frites)	Pomterre	1:oui, 2:Non	Dichotomique
	Frites	Frites	"	"
	Biscuits secs	Biscuits	"	"
	Tarte aux fruits	Tarte	"	"
	Pâtisserie (gâteau à la crème)	Patisser	1:oui, 2:Non	Dichotomique
	Cake, madeleine, mouskoutchou	Cake	"	"
	Artichaut	Artichou	"	"
	Aubergine	aubergin	"	"
	Betterave rouge	Betterav	"	"
	Carotte	Carotte	"	"

Groupes	Variable	Code	Valeurs	Type
	Chou vert, blanc, rouge	chouvert	"	Dichotomique
	Chou-fleur	choufleu	"	"
	Concombre	concomb	"	"
	Courgette	courgetto	"	"
	Epinards	Epinards	"	"
	Haricot vert	haricvert	"	"
	Navet	Navet	"	"
	Petit pois	Petitpoi	"	"
	Piment	Piment	"	"
	Poivron	Poivron	"	"
	Salade verte	Saladver	"	"
	Tomate fraîche	tomatefr	"	"
	Concentré de tomate	concento	"	"
	Fruits au sirop (tous fruits)	Fruitsir	"	"
	Abricot	Abricot	1:oui, 2:Non	Dichotomique
	Datte	Datte	"	"
	Mandarine, clémentine	Mandarin	"	"
	Melon	Melon	"	"
	Orang	Orang	1:oui, 2:Non	Dichotomique
	Pastèque	Pastèque	"	"
	Raisin	Raisin	"	"
	Cacahuète	Cacahuète	"	"
	Pistache	Pistache	"	"
	Amande	Amande	"	"
	Mouton	Mouton	"	"
	veau Bœuf	Bœuf	"	"

Groupes	Variable	Code	Valeurs	Type
	Poisson frais ou surgelé	poissfra	"	Dichotomique
	Poisson en conserve (thon, sardine, anchois	Poisscon	"	"
	ouefs	oufs	"	"
	Lait de vache	laitvach	"	"
	Yaourt	Yaourt	"	"
	Fromage blanc	fromblco	"	"
	Lben	lben	"	"
	rayab	rayab	"	"
	Huile d'olive	huiloliv	"	"
	Smen	Smen	"	"
	Huile tournesol	huiltour	"	"
	Margarin	margarin	"	"
	Beurre	beurre	"	"
	Sucre	sucre	"	"
	Barres chocolatées ou biscuitées	Barrecho	1:oui, 2:Non	Dichotomique
	Bonbons	Bonbons	"	"
	Confiture	Confitur	"	"
	Café	Café	"	"
	Thé, thé à la menthe	Thé	1:oui, 2:Non	Dichotomique
	Limonade	Limonade	"	"
	Jus de fruits du commerce	Jusfruco	"	"
	Pizza	Pizza	"	"
	chips	chips	"	"

Groupes	Variable	Code	Valeurs	Type
Groupe 4	Fruits secs	Fruitjc	Fréquence moyenne/jour	Quantitatifs
	Féculente (céréales, pt, légumes secs)	fecutotj	"	"
	Céréales (pain,couscous, riz,...)	Cerealej	"	"
	protéines	Proteinj	"	"
	Protéines animales (poisson, viande, œuf)	Protainij	"	"
	Produit gras et sucrés Fruits et légumes	Grasucrj	"	"
	Matières grasses, huile	Fruilegj	"	"
	Graisses végétales	Graissej	"	"
	Graisse animales	Graisanj	Fréquence moyenne/jour	Quantitatifs
	Graisse + oléagineux	Graoleanj	"	"
	Viandes, œufs	vianoeuj	"	"
Groupe 5	Diabète	diabete	1:oui, 2:Non	Dichotomique
	Dyslipidémie	dyslipid	"	"
	Hypertension artérielle	hta	"	"

Groupes	Variable	Code	Valeurs	Type
Groupe 6	Parmi les membres de votre famille, certains souffrent ou on souffert de Diabète ?	antediab	1:oui, 2:Non	Dichotomique
	Parmi les membres de votre famille, certains souffrent ou on souffert de l'HTA ?	antehta	"	"
	Parmi les membres de votre famille, certains souffrent ou on dyslipidémies	antedys	"	"
Groupe 7	IMC>=25Kg/m2	Surpoids	1:oui, 0:Non	Binaire
	IMC>=30Kg/m2	Obese	"	"
	Hypertension :(PAS>=130 et /ou PAD>=85	Hupten1	"	"
	Pression artérielle systolique	PAS	Des valeurs	Quantitative
	5- Pression artérielle diastolique	PAD	Des valeurs	Quantitative
	Triglycerides	triglyce	"	"
	Cholesterol total	cholest	"	"
	Glycemie	glycemie	"	"

Le fichier de travail « tahina-ex » contient des variables à expliquer et des variables explicatives.

Les variables à expliqué sont : Diabète, Dyslipidémie et l'Hypertension artérielle.

Les variables explicatives sont toutes les variables des autres groupes.

Diabète:

Le diabète, qu'on réduit souvent au diabète sucré, est dans ce cas une pathologie qui apparaît à la suite d'un déficit de production d'insuline, ou de sa mauvaise utilisation par l'organisme. S'il existe des traitements pour abaisser et contrôler le taux de sucres dans le sang, le diabète reste une maladie incurable. Il touche principalement des individus au-delà de 40 ans. Il semble d'origine génétique et est souvent associé à une obésité. Longtemps asymptomatique, il est diagnostiqué lors d'une prise de sang à jeun. La dérégulation de la glycémie est due à une résistance des organes à l'insuline : bien que l'hormone soit synthétisée à des doses normales voire élevées, les cellules ne réagissent plus suffisamment. Sans traitement, la maladie abîme les vaisseaux sanguins, ce qui peut se révéler mortel à plus ou moins long terme .

Dyslipidémies:

La dyslipidémie est un terme désignant une modification du taux normal des lipides sanguins. Elle désigne en pratique courante une concentration trop élevée d'un ou des types de lipides présents dans le sang. Les dyslipidémies sont la cause de dépôts de cholestérol dans les artères et provoquent une athérosclérose qui diminue le calibre de l'artère et à long terme, une diminution du flux sanguin à travers le vaisseau atteint. Les dyslipidémies sont découvertes sur des prises de sang : le taux de LDL cholestérol, cholestérol athérogène ou « mauvais cholestérol » le plus important,

car c'est ce cholestérol qui est principalement responsable des dépôts artériels ; le HDL cholestérol, cholestérol protecteur ou « bon cholestérol » protégerait contre les maladies cardiovasculaires ; les triglycérides sont également observés et sont le reflet de l'alimentation.

Hypertension artérielle :

La distribution de la pression artérielle (PA) dans la population s'effectue de façon continue, uni modale, des plus basses au plus élevées et la définition de l' hypertension artérielle (HTA) résulte de l'attribution du risque cardiovasculaire à un niveau de PA donné. Le retentissement sur les organes cibles (cœur, rein, cerveau) doit être pris en compte ; la maladie hypertensive n'est pas uniquement une maladie "de chiffres", mais une authentique maladie générale vasculaire avec ses implications thérapeutiques. Pathologies cardiovasculaires.

La maladie hypertensive est manifestement polygénique, expliquant la difficulté de l'approche génétique. Les facteurs d'extériorisation sont connus notamment le rôle délétère d'une consommation élevée de sel, du surpoids, de la consommation d'alcool, et peut être de l'exposition au stress et du contexte socioprofessionnel.

1.2.2 Traitements des valeurs manquantes

L'appréhension des données manquantes est un problème délicat. Non pas à cause de sa gestion informatique mais plutôt à cause des conséquences de leur traitement (suppression des individus ayant une mesure manquante ; ou remplacement par une valeur plausible à partir des observations disponibles : On parle d'imputation) sur les résultats d'analyse ou sur les paramètres d'intérêt. Les données manquantes peuvent se retrouver dans les variables à expliquer ou les variables indépendantes.

Il existe plusieurs solutions aux problèmes des données manquantes. La méthode d'élimination est le mode de gestion le plus couramment utilisée (c'est la méthode par défaut de tous les logiciels statistiques usuels). Cette technique est raisonnable pour une proportion de données manquantes au moins égale à 5% ([8]).

1.3 L'objectif de l'étude

L'objectif de notre travail est le suivant :

- Présentation de la méthode d'analyse factorielle exploratoire.
- Présentation de la méthode d'analyse factorielle confirmatoire.
- Application de l'analyse factorielle exploratoire et l'analyse factoriel confirmatoire sur un fichier extrais de TAHINA ; L'utilisation de ces méthodes nous permettre un éclairage nouveau sur l'influence des habitudes alimentaires et l'environnement socio-économique sur les maladies chroniques sélectionnées(les variables à expliquer), ce que nous aide à prendre des décisions à caractère économique et sociale pour augmenter l'attention à la prévention de ces maladies chroniques non transmissibles.

Pour arriver à une solution satisfaisante, on va suivre les étapes suivantes :

1. de façon aléatoire, on tire un échantillon de 50% des individus à partir de notre fichier de travail. On obtient deux fichiers : « tahina-ex1 » et « tahina-ex2 », chaque fichier des deux contient 2409 individus et 147 variables.
2. sur « tahina-ex1 » on applique l'analyse factorielle exploratoire.
3. sur « tahina-ex2 » on applique l'analyse factorielle confirmatoire.

Pour ce faire nous allons suivre la méthodologie suivante :

1.4 Méthodologie

Dans le cadre des objectifs fixé, la méthodologie de notre travail est la suivante :

1.4.1 Analyse descriptives des données

Les méthodes de statistique descriptive mise en œuvre sont celles d'une première exploration graphique et numérique permettant de caractériser l'échantillon. Cette première étape est très importante car elle permet de bien comprendre le profil des participants à l'enquête et aussi les associations entre variables [11].

Les outils descriptifs d'une variable qualitative se limitent habituellement à la fréquence absolue (les effectifs absolus) et relative de ses modalités.

o Les résultats des analyses réalisées avec le logiciel STATA sont repris tels quels dans des encadrés et les histogrammes.

1. Analyse univariée : l'analyse univariée permet de préparer les données qui seront utilisées ultérieurement dans l'analyse bi- et multi variée. Le diagramme en bâtons (barchart) est un outil graphique qui se prête bien à la visualisation de la distribution des fréquences (absolues ou relatives) de variables nominales. Son avantage est de repérer, d'un seul coup d'œil, les modalités les plus fréquentes et celles qui ne concernent que très peu d'individus. Les paramètres d'une variable quantitative c'est la moyenne et l'écart-type, et sa distribution est représentée par un « Histogramme ». L'objectif de cette première application pratique est d'explorer les variables à expliqué et les variables explicatifs concernant l'identification régionale, occupation et activité divers.

2. Analyse bivariée : en statistique classique, l'indice utilisé pour mesurer la dépendance entre variables qualitatives est le Chi-deux de contingence.

On se propose de ne mesurer que les liaisons qui peuvent exister entre les variables a expliquer , et quelque variable explicatifs [11]

1.4.2 Analyse en composantes principales (ACP)

L'analyse en composantes principales qui est un cas particulier d'analyse factorielle est un outil d'analyse des relations s'établissant entre plusieurs variables quantitatives sans leur attribuer de rôles de dépendante et d'indépendante(s) [11].

Les variables retenus pour l'ACP sont des variables quantitatives de groupe 4 qui représente la fréquence moyenne de consommation /jour et des varibles de groupe 7 qui représente les examens cliniques (« tahina-ex »).

Les points-variables sont d'autant plus proches des axes que leur saturation sur les composantes est élevée. Par ailleurs, plus les points-variables sont distants de l'origine (l'origine correspond à une saturation = 0) et se rapprochent d'une saturation = 1 (qui correspond à une corrélation parfaite), mieux ces variables sont représentées par les dimensions.

o Les résultats sont réalisés avec le logiciel R version 3.0.2, (package FactoMineR).

1.4.3 Analyse factorielle exploratoire

L'analyse factorielle exploratoire est une méthode statistique utilisée dans le traitement des questionnaires, pour déterminer les influences non observées dans une collection de variables. La base empirique de cette technique est l'observation que les variables d'un domaine soigneusement choisi sont souvent corrélées entre elles ([2]).

Il s'agit d'une méthode qui permet de découvrir l'existence éventuelle des facteurs sous-jacents synthétisant l'information contenue dans un plus grand nombre de variables mesurées Les variables inclues dans l'analyse sont des variables qualitatives de type dichotomique, pour cela nous allons utiliser le logiciel LISREL .

Pour arriver à une solution satisfaisante, On va suivre les étapes suivantes :

- On tir au hasard 50% des individus inclus dans « tahina-ex » a l'aide de logiciel R ; on obtient deux échantillons ; 50% des individus dans chacun.

library (stats)

library (utils)

tahina.index<-c(sample(1 :4818,2409,replace=F))

tahina.ex1<-data[tahina.index,]

tahina.ex2<-data[-tahina.index,]

Exporter les deux fichiers (tahina.ex1,tahina.ex2) sous forme texte

- Application de l'analyse exploratoire sur le premier échantillon (tahina.ex1). Cette étapes est faite en familiarisons avec le logiciel LISREL, les démarche sont les suivants :

File →Import Data→ Fichier de type : Tab Delimited Data (,txt). On ouvre la boite de dialogue .

Pour définie le type de variable :

Data → Define variables → Variable types → Ordinal, Après on click Save.

Pour appliqué l'analyse exploratoire :

Statistics → EFA of Ordinal Variables.

La version LISREL utiliser est la version étudiant, elle nous limite à une analyse de 15 variables.

Les résultats contiennent une première solution qui est une solution non réorientée normalisée qui est une transformation de la solution normalisée obtenue par la procédure FIML. La seconde solution est la solution varimax de Kaiser (1958). Ces deux éléments sont orthogonaux, c'est à dire, les facteurs sont pas corrélés. La troisième

solution est la solution Promax de Hendrickson& White (1964) .Il s'agit d'une solution oblique, c'est à dire, les facteurs sont corrélés. Les solutions varimax et la promax sont des transformations de la solution normalisée sans rotation et en tant que tels ils sont aussi solutions de maximum de vraisemblance. La quatrième solution est une solution des variables de référence. Les variables de référence sont choisis dans la solution de Promax, ceux qui ont les plus grands coefficients de saturation en chaque colonne [13].

1.4.4 Analyse factorielle confirmatoire

En analyse factorielle confirmatoire tout doit s'élaborer à partir d'une théorie explicite. Le chercheur doit formuler a priori un ensemble d'hypothèses concernant les données qu'il s'apprête à analyser ou à recueillir. Il doit identifier à l'avance les variables latentes (les facteurs) faisant partie du modèle théorique qu'il veut mettre à l'épreuve et il doit sélectionner en conséquence les variables observées qu'il utilisera comme reflets de l'influence des variables latentes [10].

Dans cette étude une analyse factorielle confirmatoire de type « (Moindres carrés non pondérés - ULS) avec rotation Oblimin fut réaliser sur les même variables sur les quelles on a effectué l'analyse exploratoire mais cette fois ci pour « tahina.ex2 ».

- Les résultats sont exprimés par un schéma obtenus à l'aide de logiciel LISREL .

File →New→Path Diagram, après la sélection des variables observés et la nomination des variables lattent on clic OK

La valeur de χ^2 représente un indice du niveau de correspondance entre une structure factorielle proposée. La valeur du χ^2 obtenu permet d'évaluer l'hypothèse nulle, soit l'hypothèse étant que la matrice des corrélations données n'est pas différente de la structure factorielle proposée, un χ^2 non-significatif nous indique que l'hypothèse nulle peut-être retenue. Par contre, il est important de noter qu'un χ^2 non-significatif

ne nous indique pas nécessairement que les données ne représentent pas adéquatement le modèle proposé.

1.5 Description sur le logiciel LISREL

LISREL est le produit vedette commercialisé par Scientific Software International. Une section de leur site web est d'ailleurs consacrée exclusivement à la dernière version de ce produit. Ce logiciel est passablement dispendieux, mais ceux qui aimeraient se familiariser avec le produit peuvent le faire en téléchargeant une version étudiante gratuite. Une fois installée cette version allégée vous permettre d'explorer toutes les facettes de LISREL, mais ne fonctionnera qu'avec des modèles qui comportant pas plus de 15 variables observées. La popularité des modèles LISREL est remarquable.si vous aves l'abitude de fouiller régulièrement la documentation scientifique dans l'un ou l'autre des champs de la psychologie, vous avez surement noté l'apparition de ces attrayants schémas composés de cercles et rectangles reliés les uns aux autres par des flèches uni-ou bidirectionnelles,il y a de fortes chances que ces analyses aient produites à l'aide de LISREL[14].

Le fichier de données du Système LISREL [13]: Ces fichiers sont généralement stockés dans un format binaire de sorte que la lecture et l'écriture de données sur le disque dur est aussi rapide que possible. Des exemples de formats de fichiers sont : Microsoft Excel (* .xls), Fichier SPSS Data (* .sav), Fichier de données SAS (* .sas7bdat), STATA (* .dta), STATISTICA (* .sta) Ces fichiers système de données contiennent généralement toutes les informations connues sur un ensemble de données spécifique. LISREL peut importer des formats de fichiers ci-dessus et d'autre et les convertir en un fichier * .lsf.

Informations générales: Cette partie du fichier * .lsf contient le nombre de case dans l'ensemble de données, le nombre de variables, valeurs manquantes et la position (nombre) de variable poids [13].

Les informations des variables: Le nom de la variable (8 caractères maximum); et

le type de variable (par exemple, continu ou ordinal); le code la valeur de variable manquante.

Pour une variable ordinale l'information suivante est également enregistrée: le nombre de catégories et étiquettes de catégorie (actuellement 4 caractères maximum) ou des valeurs numériques attribué à chaque catégorie [13].

Le type de variable qui traite LISREL:

Data → Define variables → Variable types .On sélectionne le type confort avec nos donnés et on clic OK.

Chapter 2

L'Analyse exploratoire et L'Analyse confirmatoire

2.1 Introduction

L'analyse factorielle (A.F.) est une technique statistique aujourd'hui surtout utilisée pour dépouiller des enquêtes : elle permet, quand on dispose d'une population d'individus pour lesquelles on possède de nombreux renseignements concernant les opinions, le pratiques et le statut (sexe, âge, etc.), d'en donner une représentation géométrique, c' est-à-dire en utilisant un graphique qui permet de voir les rapprochements et les oppositions entre les caractéristiques des individus.

l'A.F. vise à écrire chaque variable aléatoire du problème en fonction de facteurs sous-jacents communs à toutes les variables, et d'un facteur spécifique ou unique à la variable aléatoire considérée. Il repose sur différentes hypothèses dont principalement la non corrélation des facteurs communs. Différentes méthodes d'estimation existent, les plus courantes sont l'estimation via les composantes principales, la méthode du facteur principal et le maximum de vraisemblance [6].

On distingue aujourd'hui deux grands types d'analyse factorielle :

L'analyse factorielle exploratoire (en anglais, exploratory factor analysis ou EFA).

L'analyse factorielle confirmatoire (en anglais, confirmatory factor analysis ou CFA).

Avant de commencer l'examen des techniques d'analyse de données, nous avons besoin de définir quelques termes qui seront utilisés [1].

2.1.1 Définition des variables observées

Les variables observées sont exactement ce qu'ils sonnent comme bits d'informations qui sont effectivement observées, comme la réponse d'une personne à une question, ou un attribut mesuré, comme le poids en livres. Variables observées sont également appelés «indicateurs» ou «articles».

2.1.2 Définition des variables latentes

Les variables latentes (sous-jacents) ne sont pas observées (et sont parfois appelées «variables non observées" ou "constructions"), mais ils sont généralement les plus intéressés par la mesure. Par exemple, les participants de la recherche ou les clients peuvent nous dire s'ils se sont été sentis gênés, ou heureux. La dépression, ou la construction sous-jacente, est une variable latente parce que nous n'observons pas directement; plutôt, nous observons ses symptômes.

Les variables latentes sont des variables qui existent sur le plan conceptuel seulement et qui ne sont pas mesurées

2.2 Rappels sur l'analyse en composantes principales normée(ACP)

2.2.1 L'objectif de l'ACP

L'analyse en composantes principales (ACP) est une méthode de statistique exploratoire permettant de décrire un grand tableau de données de type individus

/ variables, quantitative (n, p), par une matrice de même dimensions mais de rang q $<$ p ; q étant souvent de petite valeur 2, 3, représente les composant principale qui sont une combinaison linaire des variables initial. L'analyse en composantes principales est une technique de réduction de données utilisée pour identifier un plus petit nombre de composants sous-jacents à un jeu de variables ou des objets observés [1].

L'A.C.P est vue comme une technique visant à représenter de façon optimale des données, selon certains critères géométriques et algébriques. Nous adopterons une présentation de l'A.C.P. qui nous permettra de faire le lien avec l'analyse factorielle exploratoire.

L'A.C.P présente deux variantes, elle peut être réalisée à partir des données centrées ou des données centrées réduites. Dans le premier cas, on parle d'A.C.P. non normée ou A.C.P sur matrice des covariances. Dans le second cas, on parle d'A.C.P normée ou A.C.P. sur matrice des corrélations. Nous présentons ici l'A.C.P. normée, la plus utilisée.

2.2.2 Les données

L'objectif de l'analyse du nuage des individus $\{x_1, ..., x_n\}$ en A.C.P. est de déterminer q nouvelles variables $\psi^1, ..., \psi^q$avec $q \leq p$, permettant de résumer (au mieux) les p variables$\bar{x}^1, ...,\bar{x}^p$. Ces q nouvelles variables sont appelées les composantes principales des individus. Elles sont définies comme des combinaisons linéaires des p variables$\bar{x}^1, ...,\bar{x}^p$

On a donc, pour $\alpha = 1, ..., q$:

$$\psi^\alpha = \upsilon_1^\alpha \bar{x}^1 + ... + \upsilon_p^\alpha \bar{x}^p = \bar{X}\upsilon^\alpha \tag{2.2.1}$$

υ^α est le vecteur propre associé a la α éme plus grande valeur propre de la matrice des corrélations R

On suppose que l'espace $\mathbb{R}^n$ est muni de la métrique M, matrice de dimension$(n \times n)$, avec $M = diag(1/\sqrt{m}, ..., 1/\sqrt{m})$, ou $m = n$ ou $m = n - 1$ selon l'estimateur de la variance choisi.

On veut que ces composantes soient de variance maximale et deux à deux orthogonales. Par construction, les colonnes de $\bar{X}$ sont centrées et donc les composantes principales le sont aussi. On a donc :

$$V(\psi^\alpha) = (\psi^\alpha)' M \psi^\alpha = (\upsilon^\alpha)' R \upsilon^\alpha \tag{2.2.2}$$

ou $R = \bar{X}' M \bar{X}$ est la matrice des corrélations empiriques entre les variables initiales $x^1, ..., x^p$.

En ajoutant la contrainte $(\nu^\alpha)' \nu^\alpha = 1$, pour $\alpha = 1, ..., q$, υ^αest le vecteur propre associé à la α éme plus grande valeur propre de la matrice des corrélations R.

on construit ainsi la matrice Ψ_q dont les colonnes sont les composantes principales des individus ψ_q

$$\Psi_q = \bar{X} V_q \tag{2.2.3}$$

ou $V_q$est la matrice $(p \times q)$ dont les colonnes sont les vecteurs propres ν^α,$\alpha = 1, ..., q$, associés aux q plus grandes valeurs propres de la matrice$R$.

2.3 Théorie d'analyse factorielle exploratoire

L'analyse factorielle exploratoire (en anglais, exploratory factor analysis ou EFA) est utilisée pour identifier les facteurs sous-jacents ou les variables latentes pour un ensemble de variables. Elle est basé sur le modèle de facteur commun, où chaque variable observée est une fonction linéaire d'un ou de plusieurs facteurs communs.

L'analyse factorielle exploratoire est souvent considérée comme une approche axée sur les données à l'identification d'un plus petit nombre de facteurs sous-jacents ou variables latentes [1].

L'analyse factorielle est une collection des méthodes pour expliquer la corrélation entre les variables en termes plus fondamentales appelées facteurs. Les tests de capacité humaine et des mesures de fonction interpersonnelle, sont souvent corrélés entre eux, cela signifie que les scores sur chaque information de la part d'un variable contenue dans les autres. L'objectif scientifique est de comprendre pourquoi il est ainsi. Selon l'analyse factorielle perspective, les variables sont corrélées car elles sont déterminées par l'influence commune non observées [2].

2.3.1 L'objectif de L'EFA

Les objectifs de l'analyse factorielle sont à déterminer le nombre d'influences fondamentales des variable, pour quantifier la mesure dans laquelle chaque variable est associée au facteur, et d'obtenir des informations sur leur nature par l'observation sur quelles variables la contribution des facteurs est performant [2].

L'analyse factorielle cherche à réduire un nombre important d'informations à quelques grandes dimensions. Elle tente d'expliquer la plus forte proportion de la covariance par un nombre aussi restreint que possible de variables (appelées facteurs) [2].

2.3.2 Les équations

Soit,

$X = (x^1, x^2, \ldots, x^p)^{'}$ Un vecteur aléatoire de$\mathbb{R}^p$d'espérance $\mu \in \mathbb{R}^p$.

On note$\bar{X} = X - \mu$ La version centrée de X

$$\bar{X}_{(p,1)} = A_{q(p,q)} F_{(q,1)} + e_{(p,1)} \tag{2.3.1}$$

Ou :

- A_q est une matrice $(p \times q)$ de coefficient $a_j^\alpha, j = 1, \ldots, p, \alpha = 1, \ldots, q$(« loadings » en anglais).elle est appelée matrice de saturation (« factor loadings matrix » ou « factor pattern matrix »).
- $F = (f^1, \ldots, f^q)'$est un vecteur aléatoire de $\mathbb{R}^q$, composé des q facteurs communs (« common factors ») aux p variables aléatoires $\bar{x}^1 ... \bar{x}^p$.
- $e = (e^1, \ldots, e^p)'$est un vecteur aléatoire centré de$\mathbb{R}^p$, composé des p facteurs spécifiques (ou uniques) (« unique factors ») à chaque variable $\bar{x}^j, j = 1, \ldots, p$.

a partir de (2.3.1) et si on prend $E(e) = 0$ on a :

$E(\bar{X}) = A_{q(p,q)} E(F) + E(e)$

$E(\bar{X}) = E(X - \mu) = E(X) - \mu = 0$

donc : $A_{q(p,q)} E(F)$=0

d'ou on a la propriété suivante

$E(F) = 0$

Pour tout $j = 1, ..., p$, on a

$$\bar{x}^j = \sum_{\alpha=1}^{q} a_j^\alpha f^\alpha + e^j \tag{2.3.2}$$

Chaque variable $\bar{x}^j$ s'écrit comme la somme d'une combinaison linaire de facteur $f^1, f^2, \ldots, f^q$communs a toutes les variables $\bar{x}^1 ... \bar{x}^p$,et d'un facteur e^j spécifique à la variable considérée $\bar{x}^j$.

Les facteurs communs $f^1, f^2, \ldots, f^q$sont aléatoires.

Le modèle (2.3.1) repose sur plusieurs hypothèses :

(H1) : $E(FF^{'}) = I_q$, ou I_q est la matrice identité $(q \times q)$.

L'hypothèse $(H1)$ signifie que les facteurs communs$f^\alpha, \alpha = 1, ..., q$, sont non corrélé et de variance 1. Cette hypothèse de non corrélation des facteurs s'explique par le fait que l'on souhaite exprimer les variables aléatoires $\bar{x}^j$en fonction du plus petit nombre de facteurs possible, et donc éviter des redondances.

(H2) : $E(ee') = \Xi$, ou $\Xi = diag(\xi,^j j = 1, \ldots, p)$

ξ^j : la variance spécifique

L'hypothèse$(H2)$ signifie que les facteurs uniques e^j,$j = 1, ..., p$, ne sont pas corrélés. Ils expriment pour chaque variable la part non expliquée par les facteurs communs. Ils ont chacun une variance spécifiqueξ^j .

(H3) : $E\{eF'\} = 0$

L'hypothèse$(H3)$ traduit le fait que chaque variable e^j, $j = 1, \ldots, p$, traduit la part spécifique à la variable $\bar{x}^j$qui n'a pu être exprimée par les facteurs communsf^α , $\alpha = 1, \ldots, q$ Donc les variablese^j et f^α , ne sont pas corrélées.

On noteΣ la matrice de variance covariance de X. On déduit du modèle (2.3.1) que :

$E(\bar{x}\bar{x}') = A_q E(FF') A_q' + E(ee')$

$$\sum = A_q A_q' + \Xi \tag{2.3.3}$$

L'´equation (2.3.3) est appelée modèle de structure de covariance.

D'après (2.3.1) ou (2.3.2), on peut écrire pour tout$j = 1, ..., p$:

$V(x^j) = (a_j^1)^2 + (a_j^2)^2 + \cdots (a_j^q)^2 + \xi^j$

$V(x^j) = \sum_{\alpha=1}^{q} (a_j^\alpha)^2 + \xi^j$

est on obtient

$$V(x^j) = h_j^2 + \xi^j \tag{2.3.4}$$

De même pour j $\neq$ k :

$cov(x^j, x^k) = \sum_{\alpha=1}^{q} a_j^\alpha a_k^\alpha + 0$

On voit ainsi que les covariances des variables aléatoirex^j, $j = 1, \ldots, p$, sont complètement reconstituées par la matrice de saturation $A_q$tandis que les variances se décomposent en une part aux facteurs communs, appelée variance commune, et une part aux facteurs spécifiques, appelée variance spécifique ou résiduelle.

On remarque également que $A_q$est la matrice des covariances entre les variables aléatoires x^j , $j = 1, ..., p$, et les facteurs communs f^α , $\alpha = 1, ..., q$. En effet :

$cov(XF) = E(XF')$

$cov(X, F) = E((A_qF + e + \mu)F')$

$cov(X, F) = AqE(FF') + E(eF') + \mu E(F')$

$cov(X, F) = A_q$

On travaille maintenant sur les variables standardisées, c'est-à-dire que $\bar{X}$ correspond au vecteur X centrée réduit

$\bar{X} = \Sigma^{-1/2}(X - \mu)$

Dans ce cas, la matrice A_q devient la matrice des corrélations linéaires entre les variables x^jet les facteurs f^α, et l'équation (2.3.3) s'écrit :

$Y = A_qA_q' + \Xi$

Ou Y est la matrice de corrélation linéaire de X

De façon analogue à (2.3.4), on a

$1 = h_j^2 + \xi^j$

[4]

2.3.3 L'estimation des paramètres

On veut estimer A_q et F dans le modèle (2.3.1). Rigoureusement, on ne devrait pas parler d'estimation pour F car il s'agit d'un vecteur aléatoire, on va donc obtenir une réalisation et non une estimation de F .

Pour cela, on dispose d'un échantillon $\{X_1, ..., X_n\}$de n réalisations indépendantes et identiquement distribuées du vecteur aléatoire X de $\mathbb{R}^p$

D'après (2.3.1), on peut écrire pour tout $i = 1, ..., n$:

$\bar{X}_i = A_q F_i + e_i$

On note :

- $\bar{X}$ la matrice $(n \times p)$ des données centrées réduites.

-F_q la matrice $(n \times q)$ correspondant aux$n$ réalisations des q facteurs communs. Elle est appelée matrice des scores des facteurs communs ("factor scores matrix").

-E_q la matrice$(n \times p)$ des erreurs spécifiques.

Le modèle sur échantillon s'écrit alors :

$$\bar{X}_{(n\times p)} = F_{q_{(n\times q)}} A'_{q_{(q\times p)}} + E_{q_{(n\times p)}} \tag{2.3.5}$$

Nous présentons ici trois méthodes d'estimation de A_q et F_q. Pour toutes les méthode d'estimation, il faut ensuite choisir le nombre q (avec q $\leq$ p) de facteurs communs que l'on retient.

[4]

2.3.4 Méthodes d'estimation

2.3.4.1 Estimation via les composantes principales

Cette technique utilise l'A.C.P. comme méthode d'estimation du modèle d'analyse factorielle exploratoire.

On suppose que l'espace $\mathbb{R}^n$ est muni de la métrique M, matrice de dimension$(n \times n)$, avec $M = diag(1/\sqrt{m}, ..., 1/\sqrt{m})$, ou $m = n$ ou $m = n - 1$ selon l'estimateur de la variance choisi.

Le lien entre l'A.C.P. et l'A.F. s'obtient à partir de la décomposition en valeurs singulières de $Z = M\bar{X}$, on note r (avec $r \leq p < n$) le rang de la matrice Z et on écrit sa décomposition en valeurs singulières (D.V.S) :

$Z = U\Lambda V^{\backprime}$

ou

- $\Lambda = diag\left(\sqrt{\lambda_1}, ..., \sqrt{\lambda_r}\right)$ des valeurs singulières des matrices ZZ' et $Z'Z$ rangées par ordre décroissant.$\left(\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq .. \geq \sqrt{\lambda_r}\right)$

-U est la matrice orthonormée$(n \times r)$ dont les colonnes sont les vecteurs propres de ZZ' associés aux r valeurs propres.

- V est la matrice orthonormée $(p \times r)$ dont les colonnes sont les vecteurs propres de $Z^{'}Z$ associés aux r valeurs propres.

On a donc :

$\bar{X} = M^{-1}Z = M^{-1}U\Lambda V^{\backprime}$

On noteU_q, Λ_q et V_q les matrices contenant respectivement les q premières colonnes deU, Λ et V.

Pour se ramener au modèle d'analyse factorielle exploratoire (2.3.5) :

Avec q = r, on pose

$$\hat{F}_q = M^{-1}U_q \tag{2.3.6}$$

$$\hat{A}_q = V_q\Lambda_q \tag{2.3.7}$$

dans ce cas on a : $\hat{E}_q = 0$.

Comme $\mathrm{U}_q^{\backprime}U_q = I_r$, on obtient :

$\hat{A}_q = Z^{\backprime}U_q = X^{\backprime}\tilde{M}^{-1}U_q$

Cette écriture est utilisée pour démontrer que les éléments de la matrice $\hat{A}_q$, notés $\hat{a_j^{\alpha}}$ sont les corrélations empiriques entre les variables x^j et les facteursf^{α} .

Comme $V_qV_q^{'} = I_P$ on montre également que :

$$\hat{F}_q = \bar{X}V_qV_q^{-1} \tag{2.3.8}$$

Cette écriture de $\hat{F}_q$en fonction de $\bar{X}$ fait ainsi apparaitre la matrice $V_qV_q^{-1} = V_q^{\star}$ des coefficients des scores des facteurs communs.

En ACP

$\Psi_q = \bar{X}V_q$

donc $\hat{F}_q = \Psi_qV_q^{-1}$ (lien de l'EFA avec l'ACP)

• Avec q $<$ r.

En ne retenant que les vecteurs propres associés aux q plus grandes valeurs propres, on a l'approximation de $\overline{X}$suivante :

$\bar{X} = \hat{F}_q \hat{A}_q + \hat{E}_q$

Ou

–$\hat{F}_q$contient les$q$ premières colonnes de $\hat{F}_q$ définie dans (2.3.6).

-$\hat{A}_q$contient les q premières colonnes de $\hat{A}_q$définie dans (2.3.7).

- $\hat{E}_q$est la matrice des erreurs associée à cette approximation.

[4]

2.3.4.2 Maximum de vraisemblance (ML pour Maximum Likelihood):

Maximise la probabilité que la solution factorielle reflète une distribution dans la population. Cette méthode produit aussi un test de $\chi 2$ de maximum de vraisemblance qui indique si la solution factorielle retenue est généralisable à l'ensemble de la population. La probabilité de ce test doit être supérieure à 0,05, c'est-à-dire que l'on ne doit pas rejeter l'hypothèse nulle qui veut que le modèle soit compatible avec les données .Cette méthode est toutefois sensible aux déviations à la normalité des distributions.([2])

2.3.4.3 Moindres carrés non pondérés (ULS ou Unweighted Least Square):

Minimise les résidus. Cette méthode est privilégiée lorsque les échelles de mesure sont ordinales ou que la distribution des variables n'est pas normale. Cette situation se présente fréquemment en sciences sociales.([2])

2.3.5 Nombre de facteurs

La décision sur le nombre de facteurs est peut-être la partie la plus difficile d'une analyse. Chaque aspect des résultats est affecté par ce choix. Même si les estimations des paramètres de deux méthodes d'estimation différentes sont comparables de sorte

que les différences d'estimation peuvent être ignorées, le choix du nombre de facteurs peut être difficile à faire avec assurance [2].

Il ya deux critères utilisés pour décider le nombre de facteurs :

- Le critère de Kaiser :Nous pouvons ne retenir que les facteurs ayant une valeur propre supérieure à 1. C'est le critère proposé par Kaiser (1960), et c'est sans doute le critère le plus couramment utilisé [5].
- Le test des valeurs propres:une méthode graphique est le test des valeurs propres qui a été proposé par Cattell (1966). Cattell suggère de trouver l'endroit où les valeurs propres semblent s'équilibrer [5].

Quel critère utiliser ?

Les deux critères ont été étudiés en détail (Browne, 1968 ; Cattell et Jaspers, 1967 ; Hakstian, Rogers, et Cattell, 1982 ; Linn, 1968 ; Tucker, Koopman et Linn, 1969). Théoriquement, on peut évaluer ces critères en générant des données aléatoires basées sur un nombre particulier de facteurs. On peut voir si le nombre de facteurs est précisément détecté par ces critères. En utilisant cette technique générale, la première méthode (Critère de Kaiser) retient parfois trop de facteurs, tandis que la seconde (test des valeurs propres) en retient parfois trop peu ; toutefois, l'une et l'autre donnent de bons résultats dans des conditions normales, c'est-à-dire avec relativement peu de facteurs et de nombreuses observations. En pratique, un aspect important à prendre en compte est le degré auquel une solution est interprétable. C'est pourquoi, nous examinons souvent plusieurs solutions avec plus ou moins de facteurs, et choisissons celle qui est la plus facilement interprétable [5].

2.3.6 Les type de rotation

La rotation est le processus mathématique qui permet de faciliter l'interprétation des facteurs en maximisant les saturations les plus fortes et en minimisant les plus faibles

de sorte que chaque facteur apparaisse influencer un ensemble restreint et unique de variables. Ce processus est effectué par rotation, soit un repositionnement des axes. Il existe deux principaux types de rotations [5].

2.3.6.1 La rotation orthogonale

on utilise cette rotation lorsque l'on a de bonnes raisons de croire qu'il est possible d'extraire des facteurs qui soient indépendants les un des autres. Une solution orthogonale est toujours préférable parce qu'une telle solution indique que chaque facteur apporte une information unique, non partagée par un autre facteur. Toutefois, ce type de solution est rarement possible en sciences sociales puisque habituellement, il existe des liens conceptuels entre les facteurs, ce qui entraîne que les facteurs sont corrélés entre eux. Il existe plusieurs méthodes pour produire une rotation orthogonale; la plus fréquemment utilisée et probablement la plus stable est VARIMAX .

2.3.6.2 La rotation oblique :

la rotation oblique permet qu'il y ait corrélation entre les facteurs. Comme elle correspond habituellement mieux à la "réalité", elle est fréquemment utilisée en sciences sociales. La méthode utilisée est OBLIMIN.

2.3.7 Considérations théoriques et pratiques

- Pour qu'une variable soit intégrée dans l'analyse, sa distribution doit montrer une certaine variance : elle doit discriminer les positions des individus. Il faut donc examiner la distribution des variables avant de décider des variables à intégrer à l'analyse. La mesure n'est pas toujours conforme à nos attentes!
- Idéalement, on cherche une structure simple, c'est-à-dire une solution où chaque variable est influencée fortement par un et un seul facteur.

- Lorsqu'une variable est corrélée à plus d'un facteur, on dit que il s'agit d'une variable complexe; ce qui signifie que les réponses à cette variable s'interprètent selon deux dimensions ou même plus.

- La structure factorielle peut être différente pour différentes populations. Il ne faut pas regrouper dans l'analyse des populations trop différentes.

- Les variables utilisées pour l'analyse devraient se distribuer normalement. Toutefois, lorsqu'on utilise l'analyse factorielle uniquement comme outil exploratoire, il est possible de "transgresser" cette règle et donc de faire une analyse factorielle sur des variables dont la distribution est très peu normale ou même avec des variables binaires de type (0,1). Il faut alors utiliser une procédure d'extraction (Moindres carrés non pondérés - ULS) qui tient compte du fait que la distribution des variables n'est pas normale.

- La relation entre les paires de variables est présumée linéaire.

- On devrait idéalement repérer et éliminer les cas ayant des patrons de réponses atypiques

- La matrice de corrélation ne peut pas être singulière. Ceci signifie que les variables ne peuvent pas être à ce point corrélé entre elles qu'une variable constitue une combinaison linéaire d'une ou de plusieurs autres variables; il y a alors "redite", c'est-à-dire que la même information est présente à deux reprises.

- La matrice de corrélation doit contenir un patron, une solution factorielle. Certains ensembles de variables doivent être corrélés entre eux suffisamment pour qu'on puisse dire qu'ils dépendent d'un même facteur.

- La solution factorielle doit aussi expliquer une proportion suffisamment importante de la variance pour que la réduction à un nombre restreint de facteurs ne se fasse pas au prix d'une perte importante d'information

- Toutes les variables doivent faire partie de la solution c'est-à-dire être corrélées substantiellement avec au moins une autre variable, sinon elles doivent être retirées de l'analyse puisqu'elles n'appartiennent pas à la solution factorielle [5].

2.3.8 Les étapes de l'analyse factorielle exploratoire

- Sélectionner l'ensemble des variables qui seront analysées conjointement
- examiner cet ensemble de façon conceptuelle et déterminer la solution qui apparaîtrait plausible quant au nombre de facteurs et au regroupement des variables.
- Effectuer une analyse factorielle avec une rotation oblique (oblimin). Il s'agit ici de voir s'il est possible d'obtenir une structure simple ou chaque variable est liée à un et un seul facteur. La rotation oblique permet de savoir si les facteurs sont corrélés entre eux. Si les corrélations sont faibles, moins de 0,32 selon Tabachnick et Fidell (2013:651), on peut passer à une rotation orthogonale.
- Comparer la solution proposée par l'analyse avec l'hypothèse de regroupement faite au départ.
- Refaire l'analyse de façon itérative jusqu'à arriver à une solution simple satisfaisante. Il faut enlever les variables problématiques une par une et non toutes ensemble, d'un seul coup car le fait de retirer une variable règle d'autres problèmes et permette de garder plus de variables dans l'analyse [5].

2.3.9 L'analyse en composantes principales et l'analyse factorielle exploratoire

La différence entre ces deux types d'analyse n'est pas toujours évidente, ceci d'autant plus que, suivant les "habitudes" disciplinaires et culturelles, certains ont tendance à utiliser systématiquement un ou l'autre de ces types d'analyse.

- L'analyse en composantes principales (ACP) cherche une solution à l'ensemble de la variance des variables mesurées. De plus, elle cherche une solution où les composantes sont orthogonales (c'est-à-dire indépendantes) entre elles. Quelque soit la matrice de corrélations, il y a toujours une solution en ACP. L'ACP maximise la variance expliquée. Les composantes sont en quelque sorte une agrégation des variables corrélées .

- L'analyse factorielle (A.F.) cherche une solution à la covariance entre les variables mesurées. Elle tente d'expliquer seulement la variance qui est commune à au moins deux variables et présume que chaque variable possède aussi une variance unique représentant son apport propre.

L'analyse en composantes principales (ACP) est une technique qui résume l'information contenue dans plusieurs variables dans un petit nombre de composants pondéré. L'ACP est une méthode d'analyse de données axé sur une collection particulière de variables. Il est souvent employé à tort comme une sorte d'analyse des facteurs, et de nombreuses études publiées par erreur présents les résultats de l'ACP comme une variété d'analyse factorielle. Pour faire empirer la situation, les commerçants des certains logiciels statistiques produisent l'ACP comme la méthode de leurs programmes d'analyse du facteur par défaut. Les deux méthodes ne font pas partie de la même approche analytique, ils ont un objectif scientifique différent, et les formes algébriques sont distinctes.

Une différence pratique clé est que les composants principaux ne sont pas conçus pour tenir compte des corrélations entre les scores observés, mais ils sont construits pour résumer au maximum l'information entre les variables dans un ensemble de données. C'est précisément le résumé des corrélations que l'analyse factorielle est conçue pour accomplir.

2.4 L'analyse factorielle confirmatoire

L'analyse factorielle confirmatoire (en anglais, confirmatory factor analysis ou CFA) est une technique statistique qui se situe évidemment dans le prolongement de l'analyse factorielle exploratoire. En ce sens, les deux techniques partagent certaines ressemblances : elles s'intéressent toutes deux à la structure latente d'un ensemble de données complexes et permettent d'expliquer les corrélations observées entre les variables à l'aide d'un nombre réduit de variables latentes, communément appelées « facteurs ». Dans les deux cas on peut ajouter que la mise en évidence des facteurs latents constitue aussi une forme de réduction des données. Cependant, comme son nom le laisse clairement sous-entendre, l'analyse confirmatoire se situe à une étape beaucoup plus avancée dans la démarche de recherche que l'analyse exploratoire [1].

La CFA dans sa version traditionnelle s'appuie sur un modèle identique à celui de l'EFA, mais la structure liant les facteurs sous-jacents aux variables mesurées est supposée connue [10].

L'analyse factorielle confirmatoire est un type de modèle d'équations structurelles (SEM) qui traite spécifiquement des modèles de mesure, c'est à dire les relations entre les mesures ou les indicateurs et les variables latentes [3].

La modélisation par équation structurelle (SEM) est une méthodologie statistique qui représente un ensemble de procédures, comme la régression multiple, l'analyse factorielle et l'analyse de covariance. Elle permet de tester un modèle théorique à l'aide d'une série d'équations de régression et son utilisation donne la possibilité d'examiner des modèles explicatifs sur des phénomènes sociaux , cette méthodologie teste des hypothèses sur les relations entre des variables observées et des variables latentes et a deux grands objectifs : valider le modèle de mesure et ajuster le modèle structurel.

2.4.1 CFA pour la recherche en travail social

Les chercheurs en travail social doivent avoir des mesures avec une bonne fiabilité et validité qui sont appropriées pour une utilisation dans diverses populations. Développement de mesures sonores psychométrique est un processus coûteux et fastidieux, et CFA peut être une étape de ce processus de développement.

Parce que les chercheurs n'ont souvent pas le temps ni les ressources pour développer une nouvelle mesure, ils peuvent avoir besoin d'utiliser des mesures existantes. En plus des économies de temps et de coûts, l'utilisant des mesures existantes contribue également à rendre les conclusions de recherche comparables avec les autre études lorsque la même mesure est utilisée dans plus d'une étude. Cependant, lors de l'utilisation d'une mesure existante, il est important d'examiner si la mesure est appropriée pour la Population inclus dans la présente étude. Dans ces circonstances, CFA peut être utilisé pour examiner si la structure originale de la mesure fonctionne bien dans la nouvelle population [1].

2.4.2 Utilisations de CFA

Dans le travail social, CFA peut être utilisé à des fins multiples, y compris mais non limité à la mise au point de nouvelles mesures, l'évaluation des propriétés des mesures psychométriques nouvelles et existantes, et l'examen des effets de la méthode EFA. CFA peut également être utilisée pour examiner la validation de construction et si une mesure est invariante ou immuable dans tous les groupes, les populations, ou de temps. Il est important de noter que ces usages se recoupent plutôt que de réellement distinctes [1].

2.4.3 La création d'un modèle CFA

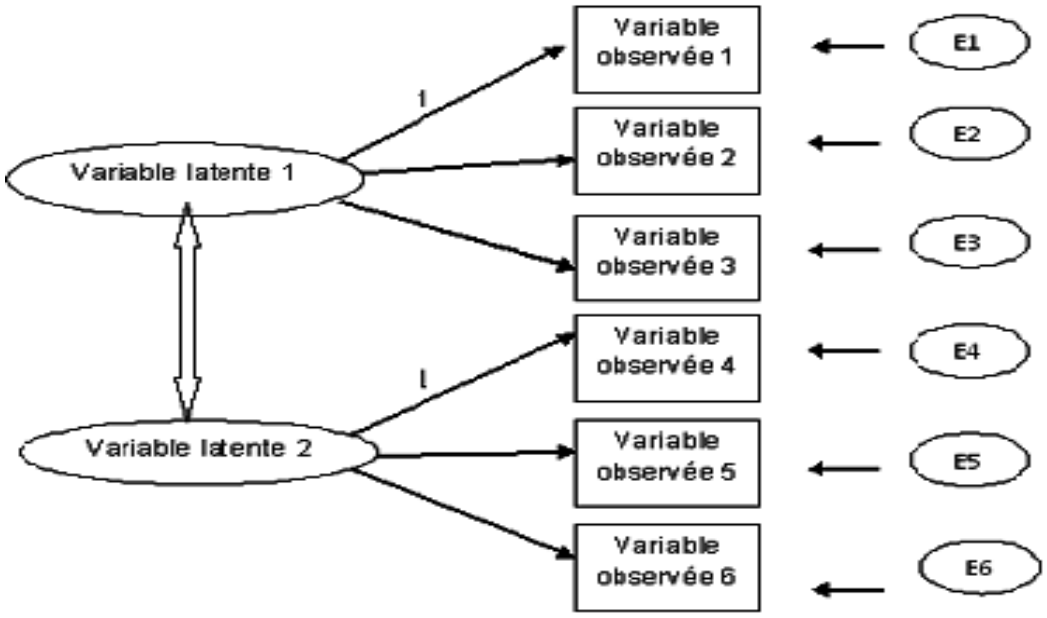
2.4.3.1 Spécification du modèle

Au début du processus de développement de mesure, les chercheurs peuvent compter entièrement sur la théorie de développer un modèle CFA. Cependant, en tant que

mesure est utilisé au fil du temps, CFA peut être utilisé pour reproduire l'EFA ou d'autres analyses qui ont été menées sur la mesure.

L'analyse factorielle confirmatoire peut ne pas être une analyse appropriée à utiliser si il n'ya pas de base solide sous-jacent sur le quel le modèle est basée [1].

Figure 2.4.1: CFA modèle avec des paramètres étiquetés



"E": l'erreur de mesure pour chaque variable observée
Flèche à seule tête: facteur de chargement ou de coefficient de régression a partir de variable latente vers la variable observée
Flèche à double tête (peut être courbe): corrélation / covariance entre variables latentes "1": variable a été mise à l'échelle pour ce variable observée

2.4.3.2 Les Paramètres de modèle CFA

Les paramètres du modèle sont les caractéristiques de la population qui seront estimés et testés dans le CFA. Les relations entre les variables observées et les variables latentes sont indiqués dans les modèles CFA par des flèches allant des variables

latentes aux variables observées. La direction de la variable latente vers la variable observée indique l'espoir que le concept sous-jacent (par exemple, la dépression) provoque les variables observées (par exemple, les symptômes de tristesse, déprime, modification de l'appétit, etc.). Les coefficients de saturation sont les coefficients de régression (c'est à dire, les pentes) pour prédire les indicateurs à partir du facteur latent.

Notez qu'une charge de 0,71 carré serait de 50% de la variance expliquée, alors que 0,32 carré serait de 10% de la variance expliquée. En CFA, l'interprétation des coefficients de saturation ou coefficients de régression est un peu plus complexe si il ya plus d'une variable latente dans le modèle, mais cette interprétation de base va travailler pour nos fins.

Considérant que chaque indicateur est censé être causé par le facteur latent, il peut aussi y avoir une différence unique dans un indicateur qui n'est pas expliquée par le facteur (s) latente. Cette variance unique est également connue comme l'erreur de mesure, la variance d'erreur, ou un indicateur de fiabilité (voir E1 à E6 à la figure (2.4.1)).

Autres paramètres dans un modèle CFA comprennent le facteur variance, qui est la variance pour un facteur dans les exemples de données (dans la solution non standardisé), et covariances d'erreur, qui sont des erreurs qui démontrent que les indicateurs sont liés corrélée à cause de quelque chose d'autre que de l'influence partagée du facteur latent. Erreurs corrélées pourraient résulter d'effets de la méthode (c.-à-méthode de mesure commun, tels que le rapport de soi) ou un libellé similaire des éléments (par exemple, le phrasé positive ou négative).

Covariances des facteurs ou des corrélations sont présentés dans les modèles CFA comme des flèches à deux têtes (généralement courbes) entre deux variables latentes.

2.4.3.3 Identification du modèle

Les modèles d'analyse factorielle confirmatoire doivent être identifié pour exécuter le modèle et estimer les paramètres. Quand un modèle est identifié, il est possible

de trouver une seule estimation pour chaque paramètre avec des valeurs inconnues dans le modèle, comme les saturations factorielles et corrélations. Par exemple, si on a une équation telle que a + b = 44, il existe un certain nombre de combinaisons de valeurs de a et b qui peuvent être utilisés pour résoudre cette équation, tels que a = 3 et b = 41 ou a = -8 et b = 52 dans ce cas, le modèle (ou l'équation) est n'est pas identifié, car il n'y a pas assez de paramètres connus pour permettre une solution unique. Les modèles doivent ont des degrés de liberté supérieur à 0 (ce qui signifie que nous avons plus des paramètres connue que des paramètres inconnus), et toutes les variables latentes doivent être mis à l'échelle pour les modèles à identifié (Kline, 2005). Lorsque nous rencontrons ces deux conditions, le modèle peut être résolu et un ensemble unique de paramètres a estimé.

- Modèles non identifié :

Les modèles sont non identifiés lorsque le nombre de paramètres estimés librement (à savoir, inconnues) dans le modèle est plus grand que le nombre d'éléments connus. Comme l'équation a + b = 44 exemple donné plus haut, ne peu pas être résolus, car il ya un nombre infini des estimations des paramètres qui produira un ajustement parfait (Brown, 2006). Dans cette situation, nous avons un degré de liberté (dl) négative, ce qui indique que le modèle ne peut pas parvenir à une solution unique parce que trop de variables sont laissées à varier par rapport au nombre de variable qui sont connues. Le nombre d'inconnues peut être réduit par la fixation de certains paramètres à des valeurs spécifiques. Par exemple, si nous fixons b = 4 dans l'équation ci-dessus, donc elle peut être résolue.

- Modèles Just-identifiés :

Les modèles sont juste identifiés lorsque le nombre d'inconnues est égal au nombre de variable connues et dl = 0. Dans cette situation, il ya un ensemble unique de paramètres qui s'adaptent parfaitement et reproduire les données. Bien que cela

puisse d'abord sembler une bonne idée (Ce qui pourrait être un problème avec un modèle parfaitement raccord?), Dans la pratique, les modèles parfaitement ajustées ne sont pas très instructif car ils ne permettent pas d'essai de modèle.

- Mise à l'échelle des variables latentes :

Comme indiqué plus haut, de plus d'avoir le dl supérieur à 0, la seconde condition pour identifier le modèle est que les variables latentes doivent être mises à l'échelle. Mise à l'échelle des variables latente réduire les inconnus. Parce que les variables latentes ne sont pas observés, et ils n'ont pas une unité pré de mesure, le chercheur doit définir l'unité de mesure. Il ya deux façons de le faire. Une option est de la faire comme un des variables indicatrices. La deuxième option consiste à définir la variance égale à 1 pour la variable latente. En général, la première option est le plus populaire (Brown, 2006).

2.4.4 Méthodes d'estimation

L'objectif de CFA est d'obtenir des estimations pour chaque paramètre de modèle de mesure (c.-à-saturations factorielles, variances des facteurs et covariances, variances d'erreur d'indicateur) que de produire une matrice de variance-covariance prédite (symbolisé par Σ) qui représente la matrice variance-covariance de l'échantillon. En d'autres termes, en CFA, nous testons si le modèle correspond aux données. Il existe plusieurs méthodes d'estimation disponibles pour tester l'ajustement d'un modèle CFA : maximum de vraisemblance (ML), les moindres carrés pondérés (WLS), moindres carrés généralisé (GLS), et non pondérées des moindres carrés (ULS). Bien que GLS et ULS sont disponibles dans Amos 7.0 et peuvent apparaître dans la littérature, les deux sont utilisés avec des données normales multi variées (Kline, 2005), et si les données sont normale multi variée, alors ML est une meilleure méthode d'estimation à utiliser.

2.4.5 Comparaison des techniques CFA avec les autres analyses

L'analyse factorielle confirmatoire est fortement liée à trois autres techniques d'analyse de données commune: EFA, CFA, et SEM. Bien qu'il existe des similitudes entre ces analyses, il ya aussi quelques différences importantes qui seront discutées ci-dessous [1].

2.4.5.1 L'analyse factorielle exploratoire

Toute fois, les tests CFA est nécessaire pour confirmer les résultats de l'EFA (Haig, 2005). Les deux EFA et CFA sont basés sur le modèle de facteur commun . EFA peut être utilisée en tant que première étape d'exploration au cours de l'élaboration d'une mesure, puis CFA peut être utilisée comme une seconde étape pour examiner si la structure identifiée dans l'EFA fonctionne dans un nouvel échantillon. En d'autres termes, CFA peut être utilisé pour confirmer la structure de facteur identifié dans l'EFA. Contrairement à l'EFA, CFA nécessite une condition prédéfinie de tous les aspects du modèle à testé . Si une nouvelle mesure est en cours d'élaboration d'un cadre théorique très forte, alors il peut être possible de sauter la première étape EFA et aller directement à la CFA.

2.4.5.2 Analyse en composantes principales

Contrairement à l'EFA et CFA, ACP n'est pas basée sur le modèle de facteur commun, et par conséquent, CFA peut ne pas fonctionner correctement en essayant de reproduire les structures identifiées par l'ACP.

On utilise l'ACP a la place de l'EFA pour plusieurs raisons, y compris le modèle mathématique relativement simple utilisé dans l'ACP et l'absence du problème de facteur de l'indétermination trouvé dans l'analyse factorielle (analyse des facteurs peut donner un nombre infini d'ensembles de scores factoriels qui sont également compatibles avec les mêmes coefficients de saturation, et il n'y a aucun moyen pour déterminer quel jeu est le plus précis). Cependant, d'autres ont fait valoir que l'ACP

ne doit pas être utilisé à la place de l'EFA . Dans les applications pratiques avec de grands échantillons et un grand nombre d'articles, l'ACP et l'EFA donnent souvent similaire Résultat, même si les charges peuvent être un peu plus petites dans l'EFA par rapport à l'ACP Pour nos besoins, il est très important de noter que l'ACP peut être utilisé à des fins similaires à celles de l'EFA (par exemple, réduction des données), mais il s'appuie sur un modèle mathématique différent et donc ne peut pas fournir une fondation pour CFA que l'EFA. Enfin, il est important de noter qu'il est souvent difficile de dire à partir d'articles de journaux si une ACP ou un EFA a été réalisée parce que les auteurs disent souvent faire une analyse de facteur, mais pas ce type d'extraction ils utilisés (par exemple, des composantes principales, qui se traduit par une ACP, ou une autre forme d'extraction telle que l'axe principal, ce qui résulte en une analyse factorielle). Une partie de la difficulté peut être l'étiquetage utilisé par des logiciels populaires tels que SPSS, où les composants principaux est la forme d'extraction par défaut sous le régime du facteur .

Comme mentionné précédemment, en raison de l'EFA et CFA sont tous deux basés sur le modèle de facteur commun, les résultats provenant d'un EFA peut être une base solide pour CFA que les résultats d'un ACP.

L'EFA est une méthode de variable latente, ainsi à distance de la méthode de réduction de données d'une analyse en composantes principales. De cela, il s'ensuit évidemment que l'EFA doit toujours être utilisé de préférence à analyse en composantes principales quand on étudie la structure d'un domaine de causalité sous-jacent commun.

2.4.5.3 Modélisation par équation structurelle

La modélisation par équation structurels est une famille large et générale d'analyses utilisé pour tester des modèles de mesure (c'est à dire, les relations entre les indicateurs et les variables latentes) et d'examiner le modèle structurel des relations entre les variables latentes. Modélisation par équation structurelle est largement utilisée car elle fournit une méthode quantitative pour tester des théories de fond, et il prend

explicitement en compte l'erreur de mesure, qui est toujours présente dans la plupart des disciplines , y compris le travail social. Modélisation par équation structurelle est un terme générique qui comprend de nombreux modèles communs qui peuvent inclure des constructions qui ne peuvent pas être mesurés directement (c.-à variables latentes) et les erreurs possibles de mesure .

Un modèle CFA est parfois décrit comme un type de modèle de mesure, et, comme tel, il est un type d'analyse qui relève de la famille SEM. Cependant, ce qui distingue un CFA à partir d'un modèle SEM est que le CFA met l'accent sur les relations entre les indicateurs et les variables latentes, alors qu'une SEM comprend pistes structurelles ou de causalité entre les variables latentes.

Chapter 3

Résultats numériques

3.1 Analyse descriptifs

3.1.1 Analyse univarié

Les variables de groupe 1 : ce groupe contient la région et le milieu des individus (voir « tahina-ex »)

Table 3.1.1: Fréquences des variables localisations régionales

	Fréquence	Percent	Cum.
Région			
Hautes plaines	1558	32,34	32,34
Sud	307	6,37	38,71
tell	2953	61,29	100,00
Total	4818	100,00	
Milieu			
rural	1849	38,38	38,38
urbain	2969	61,62	100,00
Total	4818	100,00	

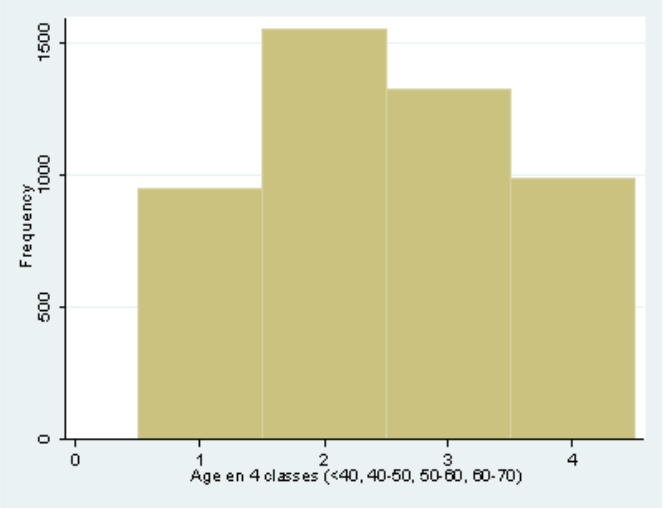
Les variables de groupe 2 : ce groupe comporte l'état civil, occupation et les diverses activités des individus (« tahina-ex »)

Table 3.1.2: Fréquences des variables occupation, activités diverses

	Fréquence	Percent	Cum.
Activit 1			
non	3267	67,81	67,81
oui	1551	32,19	100,00
Total	4818	100,00	
Smoke			
Ancien fumeur	538	11,30	11,30
Fumeur actuel	552	11,59	22,89
Non fumeur	3671	77,11	100,00
Total	4761	100,00	
Sport			
non	4631	96,12	96,12
oui	187	3,88	100,00
Total	4818	100,00	
Marchdit			
Non	2678	55,81	55,81
$\approx$ 20 marches	1067	22,94	78,05
$>$20 marches	1053	21,95	100,00
Total	4798	100,00	
Sieste			
Non	1203	24,97	24,97
Oui	3615	75,03	100,00
Total	4818	100,00	
entrepas			
Non	4186	87,21	87,21
oui	614	12,79	100,00
Total	4800	100,00	
exterier			
Non	2673	55,63	55,63
Oui	2132	44,37	100,00
Total	4805	100,00	
tabacnf			58
Non	4290	90,05	90,05
Oui	474	9,95	100,00
Total	4764	100,00	

Le variable « âge » est un variable ordinal comporte 4 classe , on visualise l'age des individus par l'histogramme suivant :

Figure 3.1.1: Histogramme d'âge

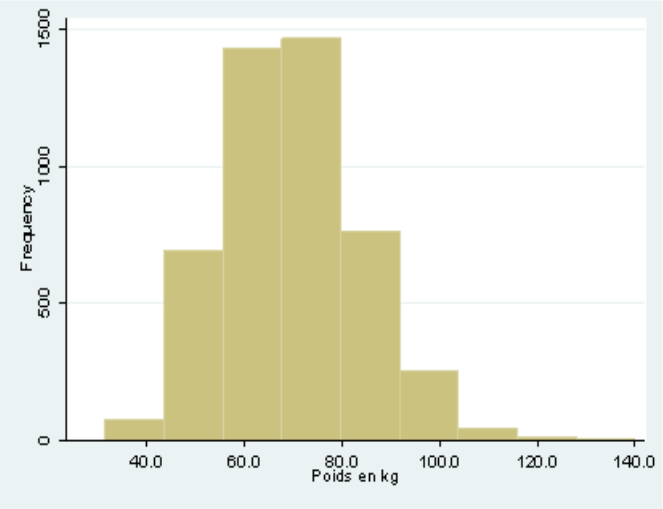


Le variable « poids » et un variable quantitative (voir « tahina-ex ») , pour bien voire le poids moyen de la population interrogé on donne l'histogramme suivant :

Table 3.1.4: Statistique de poids

Valeurs Manquante	67
Moyenne	69,591
Ecart-type	13,9617

Figure 3.1.2: Histogramme de poids



Les variables de groupe 5 : les maladies sélectionné sont l'hypertension, diabète et la dyslipidémie.

Table 3.1.5: Fréquence des maladies

	Fréquence	Percent	Cum.
HTA			
non	4077	84,62	84,62
oui	741	15,38	100,00
Total	4818	100,00	
Diabète			
non	4437	92,09	92,09
oui	381	7,91	100,00
Total	4818	100,00	
Dyslipidémie			
non	4674	97,01	97,01
oui	144	2,99	100,00
Total	4818	100,00	

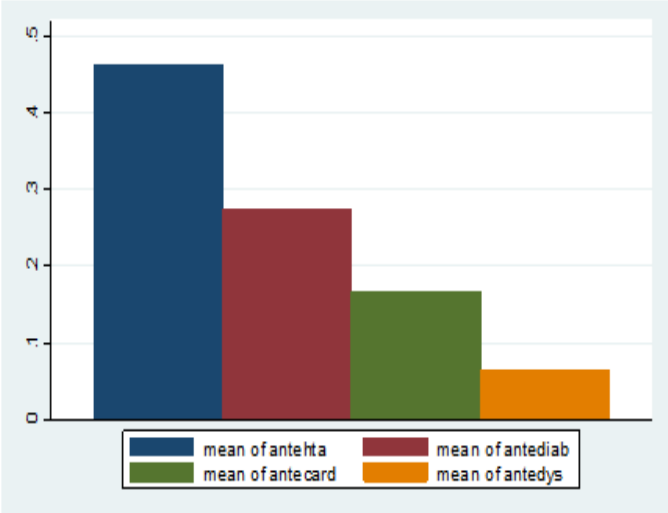
les variables de groupe 7 : les variable de ce groupe représente les examens clinique (voire « tahina-ex »).

Table 3.1.7: Les variables d'examens Clinique

	Fréquence	Percent	Cum.
Surpoids			
non	2136	44,97	44,97
oui	2614	55,03	100,00
Total	4750	100,00	
Obese			
non	3722	78,36	78,36
oui	1028	21,64	100,00
Total	4750	100,00	
Hyperten 1			
non	3470	72,14	72,14
oui	1340	27,86	100,00
Total	4810	100,00	

les variables de groupe 6 : représente les pathologies des antécédents (voire « tahina-ex »)

Figure 3.1.3: Histogramme des pathologies des antécédents



3.1.2 Analyse bivariée

- La statistique du chi-deux est la plus populaire pour tester le non association entre les lignes et les colonnes d'un tableau croisé

On a croisé l'HTA avec les variables « région », « pain de boulanger » et le variable « biscuits », les résultats sont les suivant :

Table 3.1.9: Test de chi2 « HTA/région, pain de boulanger, biscuits»

HTA	non	oui	Total
région			
tell	2472	481	2953
Hautes plaines	1330	228	1558
sud	275	32	307
Total	4077	741	4818
Pearson chi2(1)=83341 Pr=0,015			
pain de boulanger			
non	1374	248	1595
oui	248	491	3217
Total	4073	7399	4812
Pearson chi2(1)=0.0671 Pr=0,796			
biscuits			
non	3825	250	4075
oui	694	46	740
Total	4519	296	4815
Pearson chi2(1)=0.0072 Pr=0,933			

Le croisement de variable « Diabète » avec les variables : « région », « pain de boulanger »,« confiture » et le variable « yaourt» nous donne résultats sont les suivant

Table 3.1.11: Test de chi2 « Diabète/pain de boulanger, confiture, yaourt »

Diabète	non	oui	Total
région			
tell	2709	244	2953
Hautes plaines	1434	124	1558
sud	294	13	307
Total	4437	381	4818
Pearson chi2(1)=6,2046 Pr=0,045			
pain de boulanger			
non	1494	101	1595
oui	2938	279	3217
Total	4432	380	4812
Pearson chi2(1)=8.0303 Pr=0,005			
confiture			
non	3936	367	4303
oui	496	14	510
Total	43432	381	4813
Pearson chi2(1)=20,9250 Pr=0,000			
yaourt			
non	3988	336	4324
oui	442	45	487
Total	4430	381	4811
Pearson chi2(1)=1,2965 Pr=0,255			

- Le croisement de variable « Dyslipidémie » avec les variables : « région », « surpoids » et le variable « pain de boulanger » nous donne les résultats sont les suivant :

Table 3.1.13: Test de chi2 « Dyslipidémie/région, surpoids, pain de boulanger »

Dyslipidémie	non	oui	Total
région			
tell	2843	110	2953
Hautes plaines	1524	34	1558
sud	307	0	307
Total	4674	144	4818
Pearson chi2(1)=18,4739 Pr=0,000			
Surpoids			
non	2115	21	2136
oui	2492	122	2614
Total	4607	143	4750
Pearson chi2(1)=54,6377 Pr=0,000			
pain de boulanger			
non	1566	29	1595
oui	3102	115	3217
Total	4668	144	4812
Pearson chi2(1)=11,3339 Pr=0,001			

- Le tableau de contingence « Diabète » et « Dyslipidémies » :

Table 3.1.14: Test de chi2 « Diabète/Dyslipidémies »

Diabète/Dyslipi	non	oui	Total
non	4355	82	4437
oui	319	62	381
Total	4674	144	4818

Pearson chi2 (1) =251,7989 Pr=0,000

- Le tableau de contingence « Diabète » et « HTA» :

Table 3.1.15: Test de chi2 « Diabète/ HTA »

Diabète/HTA	non	oui	Total
non	3873	564	4437
oui	204	177	381
Total	4077	741	4818

Pearson chi2 (1) =307,0093 Pr=0,000.

- Le tableau de contingence « Dyslipidémie» et « HTA» :

Table 3.1.16: Test de chi2 « HTA/ Dyslipidémie »

HTA/Dyslipi	non	oui	Total
non	4017	60	4077
oui	657	84	741
Total	4674	144	4818

Pearson chi2 (1) =251,7989 Pr=0,000

3.1.3 Conclusion

- 61.29% des individus situant sur le tell, et 62,62% vie dans un milieu urbain.
- l'âge de la majorité des individus interrogé est entre 40 et 50 ans, le poids moyen et de 67 kg, 32,19% ont une activité professionnelle, 75,03% font la sieste et 44,37% prenne des repas en dehors de chez eux, le taux des individus pratiquant de sport et de 3,88%, 44,89% parcourus des trajets difficile en marchant, 77,11%

fume pas, 90,05% prenne pas des tabacs non fume et 87,21% mange pas en d'hors des repas.

- Le taux de l'hypertension artérielle 15,38%, le Diabète 7,91%, et la Dyslipidémies de 2,99%, le même enchainement on le trouve à l'état morbide des antécédents, les examens clinique confirme ces résultats tel que on a trouve 27,86 ont l'HTA de premier degré. Le taux de surpoids est de 55,03%, obésité est de 21,64%, les autres examens clinique ont un taux faible.

- Le taux de l'HTA est très élevé dans le tell par rapport aux autres régions ; 16% ont l'HTA parmi 2953 ménages situé sur le tell, 14% parmi 1558 ménages des haute plaines et 10% parmi 307 ménages du sud. La statistique de Pearson fournie, donne le seuil avec lequel : L'HTA est donc distribués indépendamment de pain de boulanger et des biscuits.

- Concernant le diabète, elle n'est pas distribuée indépendamment de la région de pain de boulanger et de confiture, mais elle est distribuée indépendamment de yaourt (statistique de Pearson : Pr=0,215).

- La dyslipidémie n'est pas distribuée indépendamment de la région, de surpoids, et de pain de boulanger ($Pr < 0.05$).

- On ce qui concerne la relation entre le diabète, la dyslipidémie et l'HTA, la statistique de Pearson rejet l'hypothèse nulle ($Pr < 0.05$) donc les variables sont pas distribuée indépendamment.

3.2 Analyse en composantes principales normée avec R :

Les variables retenus pour ces analyse sont des variables quantitatives de groupe 4 qui représente la fréquence moyenne de consommation /jour et des varibles de groupe 7 qui représente les examens cliniques (« tahina-ex »):

- Pain
- Produits laitiers
- Œufs
- Viande
- Desserts sucrés, limonade
- glycémie
- cholestérol
- tri glycémie
- pas (Pression artérielle systolique)
- pad (Pression artérielle diastolique)

Table 3.2.1: Variance total expliquée d'ACP normée

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7
Standard deviation	1.364	1.222	1.105	1.033	0.978	0.923	0.891
Proportion of Variance	0.186	0.149	0.122	0.106	0.095	0.085	0.079
Cumulative Proportion	0.186	0.335	0.457	0.564	0.660	0.745	0.825

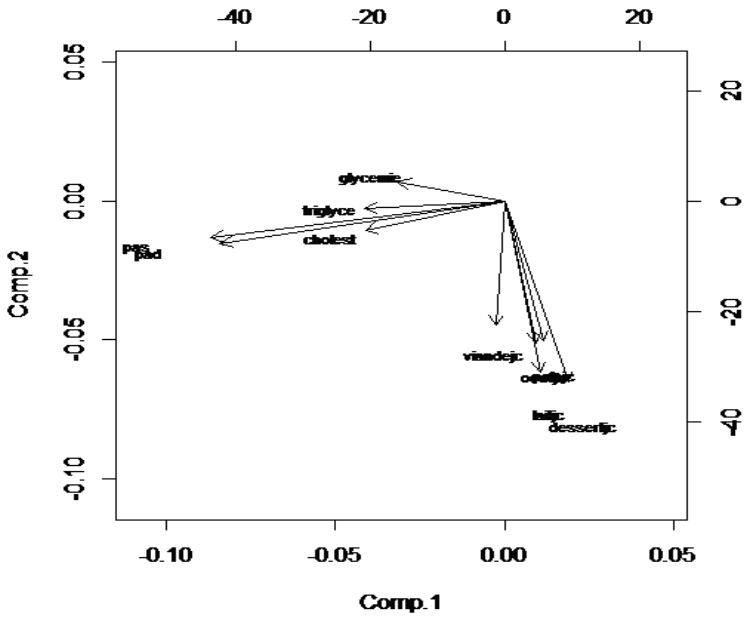
	Dim.8	Dim.9	Dim.10
Standard deviation	0.864	0.840	0.542
Proportion of Variance	0.074	0.070	0.029
Cumulative Proportion	0.899	0.970	1.000

Table 3.2.2: Matrice des composantes d'ACP normée

	comp.1	comp.2	comp.3	comp.4	comp.5	comp.6
painjc	0.111	-0.491	0.199	0.591	0.027	-0.224
laitjc	0.101	-0.599	0.065	0.376	-0.173	0.065
oeufjc	0.089	-0.496	-0.152	-0.369	-0.402	0.572
viandejc	-0.026	-0.435	-0.081	-0.580	0.335	-0.428
dessertjc	0.182	-0.634	0.019	-0.177	0.101	-0.122
pas	-0.844	-0.130	-0.328	0.087	-0.019	-0.047
pad	-0.817	-0.149	-0.374	0.127	0.069	0.050
glycemie	-0.307	0.065	0.326	-0.183	-0.751	-0.434
cholest	-0.400	-0.102	0.634	-0.077	0.164	0.147
triglyce	-0.401	-0.026	0.623	-0.091	0.208	0.236

	comp.7	comp.8	comp.9	comp.10
painjc	-0.183	0.430	-0.292	0.006
laitjc	0.546	-0.341	0.185	-0.011
oeufjc	-0.030	0.273	-0.143	-0.006
viandejc	0.352	0.174	-0.122	0.013
dessertjc	-0.574	-0.380	0.178	-0.003
pas	-0.051	0.000	0.006	-0.385
pad	-0.045	-0.028	0.002	0.377
glycemie	-0.033	-0.045	-0.030	0.054
cholest	-0.003	0.311	0.523	0.005
triglyce	0.025	-0.309	-0.494	-0.011

Figure 3.2.1: Cercle des corrélations pour le premier plan factoriel ACP



Toutes les variables sont bien représentées dans ce plan factoriel puisque leurs corrélations avec les axes sont relativement importantes (les projections sont proches du cercle de corrélation). L'interprétation que l'on peut faire des deux premiers axes factoriels est la suivante :

1- Le premier axe représente bien les variables : glycémie, cholestérol, tri glycémie, pas , pad .

2- Le deuxième axe représente bien les variables: Pain, Produits laitiers, Œufs, Viande, Desserts sucrés et limonade

3.3 Analyse factorielle exploratoire sur l'échantillon1

On applique l'analyse exploratoire sur « tahina-ex1 ». Les variables sélectionnées pour l'analyse sont : pain de boulangé, biscuit, cake, abricot, mouton, œufs, yaourt, confitures, limonade, antécédent diabétique, hypertension, diabète, dyslipidémie, l'hyperglycémie, surpoids.

Table 3.3.1: Unrotated Factor Loadings

	Factor 1	Factor 2	Unique var
pain de boulangé	0.511	0.000	0.739
biscuit	0.423	-0.173	0.791
cake	0.490	-0.087	0.752
abricot	0.217	-0.090	0.945
mouton	0.350	-0.022	0.877
oeufs	0.323	-0.205	0.853
yaourt	0.537	-0.293	0.626
confitures	0.466	-0.299	0.694
limonade	0.321	-0.352	0.773
Antécédent diabétique	0.332	0.417	0.716
hypertension	0.130	0.509	0.724
diabète	0.579	0.812	0.005
dyslipidémie	0.250	0.543	0.642
l'hyperglycémie	0.332	0.753	0.323
surpoids	0.137	0.273	0.906

Table 3.3.2: Varimax-Rotated Factor Loadings

	Factor 1	Factor 2	Unique var
pain de boulangé	0.462	0.218	0.739
biscuit	0.457	0.024	0.791
cake	0.480	0.131	0.752
abricot	0.235	0.011	0.945
mouton	0.326	0.130	0.877
oeufs	0.380	-0.047	0.853
yaourt	0.610	-0.036	0.626
confitures	0.549	-0.071	0.694
limonade	0.440	-0.182	0.773
Antécédent diabétique	0.122	0.519	0.716
hypertension	-0.100	0.516	0.724
diabète	0.177	0.982	0.005
dyslipidémie	-0.006	0.598	0.642
l'hyperglycémie	-0.021	0.823	0.323
surpoids	0.007	0.306	0.906

Table 3.3.3: Promax-Rotated Factor Loadings

	Factor 1	Factor 2	Unique var
pain de boulangé	0.447	0.228	0.739
biscuit	0.454	0.034	0.791
cake	0.471	0.141	0.752
abricot	0.234	0.016	0.945
mouton	0.317	0.137	0.877
oeufs	0.383	-0.039	0.853
yaourt	0.612	-0.023	0.626
confitures	0.553	-0.060	0.694
limonade	0.452	-0.173	0.773
Antécédent diabétique	0.088	0.522	0.716
hypertension	-0.134	0.514	0.724
diabète	0.112	0.986	0.005
dyslipidémie	-0.046	0.599	0.642
l'hyperglycémie	-0.076	0.823	0.323
surpoids	-0.013	0.306	0.906

Table 3.3.4: Factor Correlations

	Factor 1	Factor 2
Factor 1	1,000	
Factor 2	0.046	1,000

Table 3.3.6: Reference Variables Factor Loadings

	Factor 1	Factor 2	Unique var
pain de boulangé	0.419	0.247	0.739
biscuit	0.448	0.051	0.791
cake	0.453	0.160	0.752
abricot	0.231	0.025	0.945
mouton	0.300	0.150	0.877
oeufs	0.385	-0.025	0.853
yaourt	0.611	0.000	0.626
confitures	0.557	-0.039	0.694
limonade	0.469	-0.157	0.773
Antécédent diabétique	0.029	0.529	0.716
hypertension	-0.191	0.513	0.724
diabète	0.000	0.998	0.005
dyslipidémie	-0.113	0.601	0.642
l'hyperglycémie	-0.168	0.826	0.323
surpoids	-0.047	0.308	0.906

Table 3.3.7: Factor Correlations

	Factor 1	Factor 2
Factor 1	1,000	
Factor 2	0.119	1,000

3.3.1 Conclusion

Les résultats contiennent une première solution qui est une solution non réorientée normalisée qui est une transformation de la solution normalisée obtenue par la

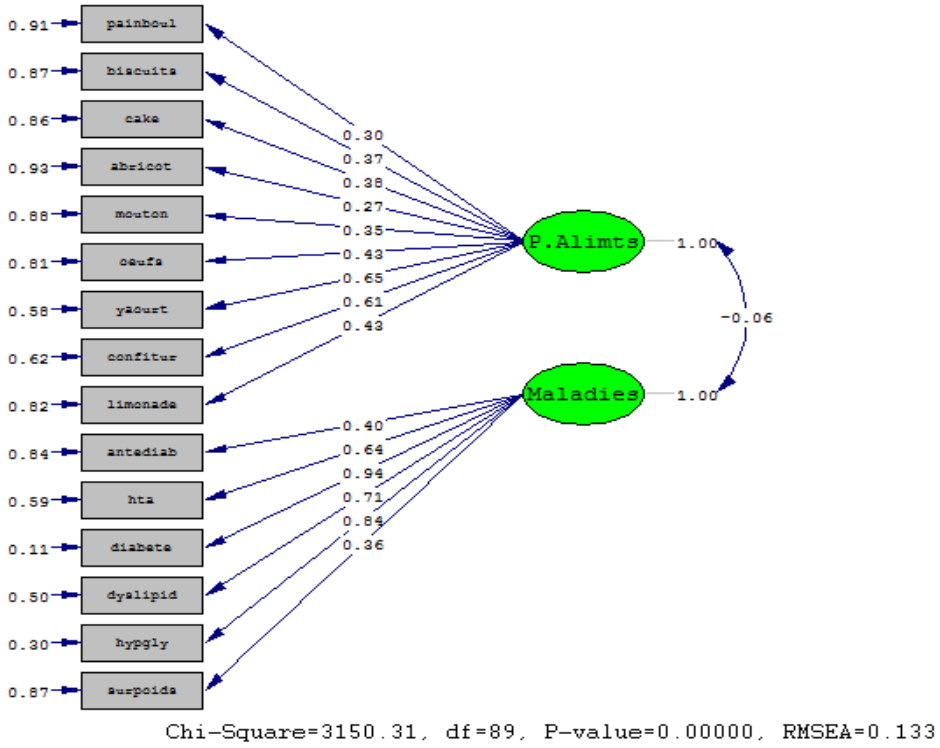
procédure FIML. La seconde solution est la solution varimax . Ces deux éléments sont orthogonaux , donc les facteurs sont pas corrélés. La troisième solution est la solution Promax .Il s'agit d'une solution oblique, c'est à dire, les facteurs sont corrélés. Les solutions varimax et la promax sont des transformations de la solution normalisée sans rotation et en tant que tels ils sont aussi solutions de maximum de vraisemblance. La quatrième solution est une solution des variables de référence. Les variables de référence sont choisis dans la solution de Promax, Ceux qui ont les plus grands coefficients de saturation en chaque colonne.

- Promax-Rotated Factor et les variables de référence suggère qu'il existe deux facteurs pratiquement non corrélées (0,046) et que le facteur 1 ayant de grandes charges sur : Pain de boulangé, biscuit, cake, abricot, mouton, œufs, yaourt, confitures, et le facteur 2 est un facteur ayant de grandes charges sur : Antécédent diabétique, l'hyperglycémie, hypertension, diabète, dyslipidémie, Surpoids.
- Le premier facteur représente les déférents produits alimentaire , le deuxième facteur représente les maladies.

3.4 Analyse factorielle confirmatoire sur l'échantillon2

pour les mêmes variables de l'analyse exploratoire de « tahina-ex1 » , on applique l'analyse confirmatoire sur « tahina-ex2 » .les résultats sont les suivantes :

Figure 3.4.1: Analyse confirmatoire



3.4.1 Conclusion

Weston et Gore (2006) ont proposé de valider tout modèle structurel par les indicateurs suivants qui sont moins sensibles aux tailles d'échantillons et aux complexités des modèles .

- Le khi-carré mesure l'écart entre la matrice des variances covariances observées et la matrice des covariances prédite par le modèle . Sa valeur normée par les degrés de liberté du modèle pour tenir compte de la taille de l'échantillon doit être comprise dans l'intervalle de 95% de la distribution de la loi du Khi-carré.

- Le Godness-of-Fit Index (GFI) est comparable au R^2 dans une régression multiple

et représente la variance du modèle expliquée par les données. GFI doit être $\geq$0.90.

- Le RMSEA ou Root Mean Square Error of Approximation sa valeur doit être $\leq$ 0.1 (Stanley et al., 1989).

- Le Root Mean Square Residual (RMSR) compare le modèle aux données à partir des résidus de toutes les variables. Sa valeur doit être $\leq$ 0.08.

[15]

- Tous les indices retenus sont dans les normes pour valider le modèle de l'étude.

Degrees of Freedom for (C1)-(C2) : 89.

Goodness of Fit Index (GFI) : 0.934.

Root Mean Square Error of Approximation (RMSEA) : 0.133.

Root Mean Square Residual (RMR) : 0.0832.

les résultats obtenus nous a confirmé les résultats de l'analyse exploratoire ; la corrélation entre les deux variables lattent est faible (-0,06), est elle nous a permet d'avoir un graphique causal des variables sélectionné.

Chapter 4

conclusion générale

L'objectif de ce travail, était de présenté deux méthode d'analyse des données, l'analyse factorielle exploratoire et l'analyse factorielle confirmatoire, ainsi une application de ces deux méthodes, qui aide a la décision, à partir du fichier Tahina (Transition Epidémiologique et Impact sur la Santé en Afrique du Nord) après avoir sélectionné un certain nombre de maladies .

Nous avons commencé par mettre en valeur le fichier TAHINA, a partir de le quel nous avons sélectionner l'ensemble des variables analysées (fichier de travail : « tahina-ex ») et nous avons rappelé quelques résultats théoriques qui nous ont servi de base pour l'analyse factorielle.

Ce travail porte principalement sur l'analyse exploratoire et l'analyse confirmatoire, ces deux technique statistique aujourd'hui surtout utilisée pour dépouiller des enquêtes : elle permet, quand on dispose d'une population d'individus pour lesquelles on possède de nombreux renseignements concernant les opinions, les maladies, le pratiques et le statut , d'en donner une représentation géométrique .

Dans la littérature anglo-saxon l'analyse exploratoire et l'analyses confirmatoire ont un lien avec l'analyse en composantes principales, principalement à la réponse des questions en sciences sociales, dans une partie de ce mémoire nous avons donnée tout

d'abord un rappel sur l'analyse en composantes principales, après nous avons présenté la théorie d'analyse exploratoire, ou on a parlé sur les objectifs, les considérations et les différents étapes nécessaire pour effectué cette analyse, nous avons aussi présenté les méthode d'estimation utilisé suivant la nature des variable inclus dans l'analyse, on a constaté que l'analyse exploratoire nous aide à faire remarqué et donc extraire des variables latentes qui nous éclairassent l'explication de la corrélation existe entre les variables initiale.

Afin de confirmer les hypothèses conclu par l'analyse exploratoire on a présenté l'analyse factorielle confirmatoire, nous avons présenté la théorie de cette analyse et on donnée une comparaison de cette technique par rapport aux autre techniques d'analyse (ACP, EFA et SEM).L'analyse factorielle exploratoire et une étape préliminaire avant d'entamé l'analyse confirmatoire.

Dans la partie pratique, on a choisi le fichier de travail « tahina-ex »(4818,147) depuis TAHINA(4818,1312), « tahina-ex » contiens des variables a expliquer qui présente les maladies chronique : L'Hypertension, Diabète et la Dyslipidémie ainsi que des variables explicatifs qui indiquent la situation socioprofessionnelle des individus. Dans le but de chercher les facteurs de risque des maladies chronique sélectionné, on a commencé par une analyse descriptive des données : les méthodes de statistique descriptive mise en œuvre sont celles d'une première exploration graphique et numérique permettant de caractériser l'échantillon. Cette première étape est très importante car elle nous a permet de bien comprendre le profil des participants à l'enquête et aussi les associations entre variables. On conclu d'après cette première analyse que la maladie la plus connu dans la population algérienne est bien l'hypertension artérielle, d'autre part on ce qui concerne la vie socioprofessionnelle, on remarque que le poids moyen des individus et de 67 kg, 32,19% ont une activité professionnelle, 75,03% font la sieste et 44,37% prenne des repas en dehors de chez eux. On parlants des autres occupations, le taux des individus pratiquant de sport et de 3,88%, 44,89% parcourus des trajets difficile en marchant, 11,59% fume, 9,95% prenne des tabacs non fume.

Pour les variables quantitatives qui représentent la fréquence moyen de consommation /jour et les examens cliniques on a appliqué une analyse en composantes principales normée ou on a remarqué que :

Le premier axe représente les variables : glycémie, cholestérol, tri glycémie, pas, pad.

Le deuxième axe représente les variables: Pain, Produits laitiers, Œufs, Viande, Desserts sucrés et limonade.

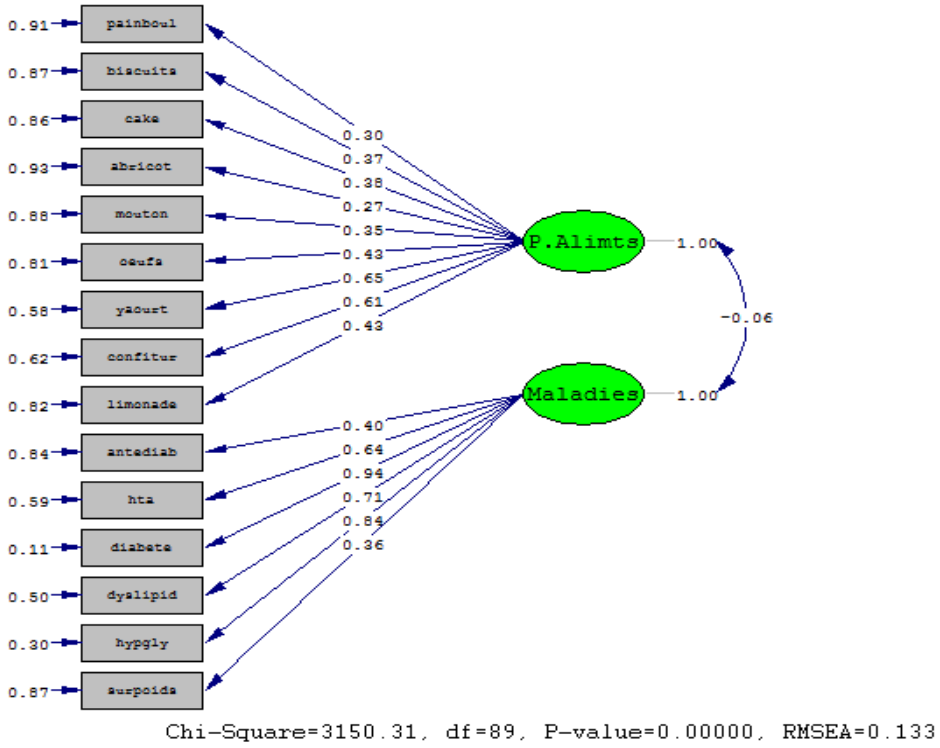
A l'aide de logiciel R on a tiré deux échantillons au hasard « tahina-ex1 » (2409,147) et « tahina-ex2 » (2409,147). nous avons effectué une analyse exploratoire sur « tahina-ex1 »

	Factor 1	Factor 2	Unique var
pain de boulangé	0.447	0.228	0.739
biscuit	0.454	0.034	0.791
cake	0.471	0.141	0.752
abricot	0.234	0.016	0.945
mouton	0.317	0.137	0.877
oeufs	0.383	-0.039	0.853
yaourt	0.612	-0.023	0.626
confitures	0.553	-0.060	0.694
limonade	0.452	-0.173	0.773
Antécédent diabétique	0.088	0.522	0.716
hypertension	-0.134	0.514	0.724
diabète	0.112	0.986	0.005
dyslipidémie	-0.046	0.599	0.642
l'hyperglycémie	-0.076	0.823	0.323
surpoids	-0.013	0.306	0.906

on distingue deux facteur pratiquement non corrélées (0,046) ; le premier représente les différent produits alimentaire et le deuxième représente les maladies inclus dans

l'analyse, de cela on a conclue deux variables latente : produits alimentaire et maladies chronique.

Dans le but de confirmé les résultats obtenus on a applique l'analyse confirmatoire sur le deuxième échantillon « tahina-ex2 » :



les résultats obtenus nous a confirmé les résultats de l'analyse exploratoire ; la corrélation entre les deux variables lattent est faible (-0,06), est elle nous a permet d'avoir un graphique causal des variables sélectionné.

la non corrélations entre les maladies et les produits alimentaire selectionné est expliquée par la predence des malades en vers les maladies, par exemple, un diabitique il peut manger tout ce qui il veut mais avec des quantité controlé .

La corrélation entre les maladies est expliqué par le lien qui existe entre le surpoids et l'hypertension ainsi entre l'hypertension et la diabète.

Comme nous venons de le voir, l'analyse exploratoire nous a aidé à clarifié et a comprendre la corrélation entre les variables et l'analyse confirmatoire comme le nom indique, nous aide a confirmé les résultats obtenus.

Selon Tabachnik et Fidell (2013),

"Le choix entre l'analyse en composantes principales et l'analyse factorielle dépend de votre évaluation de l'adéquation entre les modèles, les données et le but de la recherche. Si vous êtes intéressés par une solution théorique non contaminée par la variance spécifique et la variance d'erreur et que vous avez élaboré votre recherche en vous basant sur des concepts précis qui devraient donner lieu à des scores spécifiques sur les variables observées, l'analyse factorielle est le choix approprié. Par contre, si vous voulez simplement un sommaire empirique de vos données, l'analyse en composantes principales est le choix approprié". P. 640.

List of Figures

List of Tables

Bibliography

[1] DONNA HARRINGTON, Confirmatory factor analysis, OXFORD, University press, 2009

[2] ROBERT CUDECK, Exploratory factor analysis, Department of psychology, University of Minnesota, Minneapolis, Minnesota.

[3] TIMOTHY A. BROWN, Confirmatory factor analysis for applied research, series editor's note by David A .Kenny, The Guilford press New York London.

[4] MARIE CHAVENT&VANESSA KUENTZ&JEROME SARACCO, Analyse en facteurs : présentation et comparaison des logiciels, SAS, SPAD et SPSS, Universités Bordeaux 1 et 2 , Revue Modulad, 2007

[5] CLAIRE DURAN, L'analyse factorielle et l'analyse de fidélité, notes de cours et exemples, Université de Montréal, département de sociologie, 2013.

[6] JEREMY J. ALBRIGHT, Confirmatoryfactor analysis using Amos, Lisrel, and Mplus, the Trustees of Indiana University, 2006.

[7] MICHAEL E.TIPPING& CHRISTOPHER M.BISHOP, Probabilistic principal component analysis, 27 September 1999.

[8] EMMANUEL JAKOBOWICZ, contributions aux modèles d'équations structurelles à variables latentes, conservatoire national des arts et métiers, PARIS, 2007.

[9] RAFAEL COSTA & G.MASY-STROOBANT, Pratique de l'analyse de données, Spss appliqué à l'enquête « Identités et Capital social en Wallonie », centre de recherche en démographie et sociétés, UCL/IACCHOS/DEMO, Louvain-la-Neuve 2013.

[10] JACQUES BAILLARGEON, Cours, Applications et interprétation des techniques statistiques avancées, Université du Québec à Trois-Rivières, 2008.

[11] SAMUEL AMBAPOURS, Applications de l'analyse des données au traitement d'enquêtes ; Mesure de satisfaction de clientèle pour les grands services publics : le cas de la société nationale d'électricité, bureau d'application des méthodes statistiques et informatiques, 05/2003.

[12] LEDYARD TUCKER & ROBERT MACCALLUM,Exploratory Factor Analysis.

[13] LISREL For Windows, Help, New features in LISREL 9.

[14] PHILIPPE CIBOIS, Professeur à l'université de Versailles –St-Quentin, L'analyse Factorielle Confirmatoire.

[15] WESTON, REBECCA AND GORE, "A Brief Guide to Structural Equation Modeling" The Counseling Psychologist, Vol. 34, n°5, pp: 719-751.

[16] Internet: Santé-Médecine : les définitions des maladies chroniques. (recherche).