



Série de Fiches Méthodologiques en Recherche et en Rédaction Scientifique

Fiche Méthodologique n: 3/2021

Analyse en Composantes Principales (ACP)

Principal Component Analysis (PCA)

تحليل المكونات الرئيسية

Kamel Ben Salem¹, Ahmed Ben Abdelaziz², Réseau Maghrébin PRP2S*

* Réseau- Maghrébin: Pédagogie- Recherche- Publication en Sciences de Santé (PRP2S)

1. Professeur de Médecine Préventive et Communautaire. Faculté de Médecine de Monastir. Université de Monastir (Tunisie).

2. Professeur de Médecine Préventive et Communautaire. Laboratoire de Recherche LR19SP01. Université de Sousse (Tunisie).

Cette série...

Cette série...

Le Réseau Maghrébin PRP2S et la Rédaction de la revue «La Tunisie Médicale» ont l'honneur de continuer d'une manière régulière, à partir du numéro de mars 2021, et pour la deuxième année successive, la série des fiches techniques en épidémiologie, en bio statistique et en rédaction médicale scientifique.

Cette série a eu un grand succès au cours de sa première année d'édition en 2020, comme indique le nombre de téléchargements dépassant significativement celui des articles originaux et illustrant un besoin très manifeste des jeunes chercheurs, au renforcement de leurs capacités en méthodologie de recherche scientifique en sciences de santé, selon une pédagogie centrée sur l'acquisition des compétences pratiques de recherche biomédicale.

En effet, nos fiches méthodologiques décrivent, d'une manière standardisée, les modes d'usage des concepts, des outils et des méthodes, utilisés d'une part lors du continuum de la recherche biomédicale scientifique, dès la phase conceptuelle jusqu'à la phase rédactionnelle et d'autre part lors des différentes phases de la rédaction médicale scientifique, depuis l'étape de la recherche documentaire jusqu'à l'étape de la communication médicale scientifique.

Cette série est rédigée par les experts du Réseau Maghrébin PRP2S, en méthodologie de recherche, exerçant dans les universités du Grand Maghreb et les facultés sœurs au Nord de la Méditerranée. Chaque fiche répond à trois questions essentielles (Quoi ? Pourquoi ? Comment) du concept étudié, en se basant sur un article publié dans la revue «La Tunisie Médicale».

Le coordinateur de la série «Fiches Méthodologiques»

Professeur Ahmed Ben Abdelaziz (Président du Réseau Maghrébin PRP2S)

Email : ahmedbenabdelaziz.prp2s@gmail.com

Série des Fiches méthodologiques

Sommaire

Année 2020

Fiche n°1 (janvier 2020):

Comment calculer la taille d'un échantillon pour une étude observationnelle

Serhier Z, et al. (Faculté de Médecine et de Pharmacie de Casablanca. Maroc)

Fiche n°2 (février 2020):

La recherche qualitative: méthodes, outils, analyse

Soulimane A. (Faculté de Médecine, Université Djillali Liabes, Sidi Bel Abbes, Algérie)

Fiche n°3 (mars 2020)

Et Allah ...créa la variabilité

Barhoumi T, et al (Réseau Maghrébin PRP2S)

Fiche n°4 (mai 2020)

Réussir votre recherche bibliographique sur PubMed

Ben Abdelaziz A, et al (Réseau Maghrébin PRP2S)

Fiche n°5 (juin 2020)

Réussir la rédaction de votre «Protocole de Recherche» en sciences de la santé

Ben Abdelaziz A, et al (Réseau Maghrébin PRP2S)

Fiche n°6 (juillet 2020)

Analyse multi variée par régression logistique

Ben Salem K, et al (Réseau Maghrébin PRP2S)

Fiche n°7 (août 2020)

Tests non paramétriques pour comparer deux ou plusieurs moyennes sur des échantillons indépendants

Bezzaoucha A, et al (Réseau Maghrébin PRP2S)

Fiche n°8 (septembre 2020)

Comment évaluer la concordance entre deux mesures qualitatives par le test Kappa?

Mellakh R, et al (Réseau Maghrébin PRP2S)

Fiche n°9 (octobre 2020)

Comment comparer plusieurs moyennes par le test d'Analyse de Variance (ANOVA) ?

Khiri H, et al (Réseau Maghrébin PRP2S)

Fiche n°10 (novembre 2020)

Tests non paramétriques sur SPSS pour comparer deux ou plusieurs moyennes sur des échantillons appariés. (test de Wilcoxon et test de Friedman)

Bezzaoucha A et al (Réseau Maghrébin PRP2S)

Année 2021

Fiche n°1 (mars 2021):

Comment calculer et interpréter la valeur de «p» dans une étude épidémiologique

Ladner J et al. (Faculté de Médecine et de Pharmacie de Rouen. France)

Fiche n°2 (avril 2021)

La Charge globale de morbidité (Global Burden of Diseases): c'est Quoi? Pourquoi? Comment?

Hsairi M et al (Faculté de Médecine de Tunis, Tunisie)

Fiche n°3 (mars 2021)

Analyse en Composantes Principales (ACP)

Ben Salem et al (Faculté de Médecine de Monastir)

Correspondance

Professeur Kamel Ben Salem

Email: kbsalem@gmail.com

ETUDE DE CAS

L'équipe du Service de Pédiopsychiatrie de l'hôpital Razi, de La Manouba (Tunisie) a publié en 2017, dans la revue «Encéphale» (1), un article dont l'objectif a été la validation en population générale et clinique, chez les enfants, d'une échelle de Signes Neurologiques Mineurs (SNM), appelée échelle de Krebs et al. (2000). Les auteurs ont examiné les propriétés psychométriques de l'échelle chez 86 enfants âgés de 6 à 12 ans dont 26 enfants atteints des Troubles du Spectre Autistique (TSA) et 60 enfants ayant un développement normal. En plus de la cohérence interne de l'échelle qui était bonne avec un alpha de Cronbach de 0,87 et de la fiabilité inter juges ayant donné un coefficient de corrélation intra classe de 0,91, l'Analyse en Composantes Principales (ACP) rapportait cinq facteurs représentant 63% de la variance totale. L'évaluation des SNM a montré des différences significatives entre patients et témoins pour le score total de SNM ($p < 10^{-3}$) et dans différents sous-scores. A la fin de leur étude, les auteurs ont conclu que «la validité de l'échelle est confirmée pour l'évaluation des SNM chez les enfants, en particulier pour les enfants atteints de TSA».

QUIZZ

1. L'Analyse en Composantes Principales (ACP) est une technique multi variée dite d'interdépendance, car:
 - a. Il n'y a pas de variables dépendantes ou indépendantes au préalable.
 - b. Il n'y a pas d'hypothèse nulle à tester ou à vérifier.
2. L'Analyse en Composantes Principales (ACP), vise deux objectifs principaux:
 - a. Comprendre les associations des variables dans un questionnaire,
 - b. Condenser l'information contenue à l'intérieur d'un questionnaire ou d'une échelle, en un ensemble restreint de nouvelles dimensions composites, tout en assurant une perte minimale d'informations
3. Lors de l'Analyse en Composantes Principales (ACP), il est recommandé d'avoir, préférentiellement:
 - a. Des variables quantitatives
 - b. Un échantillon relativement grand (10 sujets par variable)
4. Parmi les postulats de la conduite d'une l'Analyse en

Composantes Principales (ACP) :

- a. L'existence des corrélations minimales entre les items ou les variables qui feront l'objet de l'analyse.
- b. Un indice KMO, mesurant l'adéquation de l'échantillonnage (qualité des corrélations inter-items) de 0,5 ou plus

5. L'Analyse en Composantes Principales (ACP) est :

- a. basée sur la variance spécifique des variables et permet d'extraire un minimum de ses facteurs explicatifs
- b. différente de l'analyse des facteurs communs (Principal axis factoring) qui est basée sur la variance commune partagée par les variables analysées.

6. Lors de l'Analyse en Composantes Principales (ACP), le critère utilisé fréquemment pour déterminer le nombre de facteurs à extraire, est la valeur dite de «eigen» (eigenvalue), traduite en français par «Valeur propre initiale»

- a. Plus la valeur propre initiale est élevée, plus le facteur explique une portion significative de la variance totale
- b. Tout facteur avec une valeur propre initiale supérieure à 1 est considéré comme facteur significatif.

INTRODUCTION

Nous avons appris que pour décrire un état de santé donné, nous devons recueillir des variables, de les synthétiser, une à une, entre autre sous forme numérique et de rechercher des corrélations entre ces variables quand elles sont quantitatives (2). Cette synthèse, bien qu'importante dans l'analyse descriptive, ne prend pas en compte l'ensemble des variables de **façon conjointe**. Elle reste parcellaire et donne une image incomplète sur l'état de santé d'intérêt. Afin de remédier à cette insuffisance, le recours à la statistique descriptive multidimensionnelle permet de décrire plus fidèlement cet état. L'Analyse en Composante Principale (ACP), **variante de l'analyse factorielle**, répond à cette insuffisance et permet de synthétiser conjointement plusieurs variables quand elles sont **toutes quantitatives** afin de décrire le mieux possible l'ensemble des individus définis par ces variables objets de l'analyse descriptive. L'objectif de cette fiche méthodologique est d'expliquer l'intérêt de l'ACP, son vocabulaire, d'apporter les éléments nécessaires pour sa compréhension et comment interpréter ses résultats à partir d'un exemple.

L'ACP: POURQUOI ?

Pour répondre à cette question, prenons un exemple d'une base (fictive) de données relative à 12 individus de trois différentes ethnies (A, B, C). Les variables collectées pour les décrire, outre l'origine ethnique, sont au nombre de quatre et sont toutes quantitatives. Nous savons déjà comment résumer une à une ces variables quantitatives (2) ou résumer simultanément deux variables de façon graphique (nuage de points), de calculer le coefficient de corrélation entre ces deux variables ou faire une régression d'une variable par rapport à une autre. Les informations ainsi obtenues resteront parcellaires. Alors comment s'y prendre pour avoir une image globale tenant compte des quatre variables de façon simultanée. La solution est de trouver le moyen mathématique pour réduire l'**ensemble** des quatre variables en un espace au mieux de deux dimensions sans beaucoup de perte de l'information recueillie au départ.

L'ACP: C'EST QUOI ?

Ainsi réaliser une ACP c'est **réduire** le nombre de variables initiales et **restituer** une **quantité d'information**

maximale (optimale) dont on dispose dans un espace de dimension élevée à n individus et p variables à un espace de dimension réduite par exemple 2 dimensions. Cette réduction se fera par l'exploration des liaisons entre variables (recherche de corrélations) d'une part et des ressemblances entre individus (écarts entre individus) d'autre part. Sur le plan théorique, l'ACP fait appel à des notions mathématiques avancées (calcul matriciel, valeur propres, calcul vectoriel ...) mais non nécessaires à connaître pour comprendre le mécanisme d'une ACP et l'utiliser correctement. L'ACP crée de nouvelles variables théoriques (deux au mieux ou trois combinaisons entre l'ensemble des variables prises pour cette analyse). Deux types de graphiques seront générés par l'ACP, un pour visualiser les relations entre les variables et l'autre pour identifier des similitudes/non similitudes entre les individus (3).

1) Vocabulaire de l'ACP

Des termes propres à l'ACP doivent être expliqués pour la comprendre. Nous rappelons que réaliser une ACP c'est réduire des variables. De cette réduction, résulte la construction de nouvelles variables caractérisées par **les valeurs propres** «*eigen values*»; le symbole utilisé est la lettre grecque « λ » (calcul matriciel): valeurs permettant de mesurer les quantités de variance prise en compte par la **composante principale et les vecteurs propres** «*eigen vectors*», support mathématique de la composante principale (calcul matriciel également) seront calculés pour chaque nouvelle variable. Une **composante principale** est la somme des produits des modalités (normalisées) pondérées par les vecteurs propres. Il s'agit d'une nouvelle variable appelée **facteur**, résultat de la projection des nuages de points (inertie totale) sur un **axe théorique** (premier plan factoriel) qui prend le maximum de cette inertie totale. Rappelons que «l'étalement» d'un nuage de point reflète la variance des variables, c'est-à-dire, les distances «**écarts**» entre individus. Nous avons appris que les écarts par rapport à la moyenne nous permettent de calculer la variance (2). La somme des variances de toutes les variables s'appelle **inertie totale**. Le reste de l'inertie, inertie résiduelle, sera projeté sur un **deuxième axe perpendiculaire au premier** et constituera le deuxième facteur qui prendra lui aussi le maximum de l'inertie résiduelle. Ces deux facteurs perpendiculaires sont indépendants et ainsi de suite et de façon séquentielle, se construisent le reste des facteurs.

Une composante principale permet de calculer le score de chaque individu sur la composante. La représentation graphique des scores se fera sur les **axes principaux**, en prenant la première composante en abscisses qui prend la plus grande variabilité et la deuxième qui prend le maximum du reste de la variance en ordonnées. Le graphe obtenu, représente **l'espace des individus**. Il permet d'identifier les similitudes/non similitudes dans les groupes d'individus.

L'ACP permet également de construire **l'espace des variables**. Cette représentation est le résultat du calcul du coefficient de corrélation de la variable avec la composante principale. Le coefficient de corrélation, entre une composante et une variable propre, lui attribue ses coordonnées sur les axes de ces composantes rapportées à un cercle de rayon 1 (pour des données normalisées). Ces axes sont deux diamètres orthogonaux du cercle. Ce cercle s'appelle également cercle des corrélations. Cette représentation permet de voir les corrélations, si elles existent entre les variables.

2) Rationnel de ces termes

Les composantes principales sont le résultat de projections orthogonales des modalités numériques des variables de chaque individu sur les axes représentant cette variable. Un premier nuage de point est obtenu sur deux axes, ce nuage sera lui-même projeté orthogonalement sur le troisième axe de la variable suivante; un nouveau nuage est ainsi obtenu et ainsi de suite. Les nuages de points représentent en fait la dispersion des modalités des variables utilisées. Ils se définissent par **leur inertie** qui est l'équivalent de la variance. Les écarts globaux entre les différents individus sont quantifiés par le calcul de la **variance**. Les détails de son calcul ne sont pas nécessaires, leur valeur se trouve sur la diagonale de la matrice variance covariance qu'il suffit de la demander au logiciel que vous utilisez pour réaliser une ACP.

Cependant un léger bémol, les variables utilisées dans une ACP sont quantitatives et par conséquent de différentes unités, la variance va refléter surtout la variabilité entre ces unités. La solution est de réaliser l'ACP sur des variables normalisées quand celles-ci sont de nature et d'unités différentes. La normalisation se fait par la transformation de toutes les variables quantitatives en variables centrées réduites. Cette transformation nous débarrasse des unités et mieux encore et pour toute variable quantitative, la moyenne de la transformée est

nulle et sa variance est égale à 1 (l'unité). Ainsi la variance totale utilisée pour sélectionner les composantes est égale au nombre de variables utilisées pour l'ACP. Les logiciels font automatiquement cette transformation en choisissant de faire l'ACP sur **la matrice des corrélations**. Cette normalisation permet également de construire le cercle des corrélations pour les variables. **Le cosinus²** de l'angle entre deux variables à partir du centre du cercle est égal au coefficient de corrélation entre ces deux variables. Les variables proches du cercle sont bien représentées; celles proches du centre sont mal représentées. Les variables se trouvant sur des rayons orthogonaux et proches du cercle sont indépendantes et les variables diamétralement opposées et proches du cercle sont corrélées négativement.

Le nombre d'axes à retenir sera fonction de la quantité de la variance utilisée pour leur construction (valeur propre ou *eigen value*). D'une façon générale, les axes dont la valeur propre est supérieure à 1 seront retenus et quand la somme cumulée des variances utilisées sur ces axes, dépasse les 75% de la variance totale. Dans les sciences de la santé, on se contentera de deux axes au maximum trois, au delà l'interprétation devient difficile. On se contentera de deux axes au lieu de quatre dans notre exemple. Ces deux axes seront les axes principaux.

L'ACP: COMMENT ?

Afin d'illustrer comment réaliser une ACP (annexes), nous allons prendre comme base de données l'exemple des 12 individus et des quatre variables quantitatives et une variable qualitative, cité plus haut. Quelques soit le logiciel d'analyse que vous utilisez, vous respecterez les étapes suivantes:

- Choisir, dans le menu principal, «analyse factorielle », dans le menu suivant choisir ACP,
- Choisir la matrice des corrélations pour que cette ACP se fasse sur des données normalisées, les autres possibilités sont la rotation «varimax», la carte factorielle, et facteur pour enregistrer, sur votre base, les nouvelles variables créées en choisissant «régression».

Dans le menu de départ «descriptif» il est important de vérifier l'adéquation de vos données pour une ACP, à partir de l'indice Kaiser-Meyer-Olkin (**KMO**) (tableau 1) qui doit être supérieur à 0,5 pour une adéquation acceptable des données à une ACP. Pour notre exemple, les tableaux à lire seront :

Tableau 1. Indice Kaiser-Meyer-Olkin (KMO). Sortie SPSS

Indice KMO et test de Bartlett		
Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,520
	Khi-deux approximé	22,822
Test de sphéricité de Bartlett	Ddl	6
	Signification de Bartlett	,001

Le tableau suivant est celui des valeurs propres des composantes (tableau 2). La première composante a la valeur propre la plus élevée soit 2,322 et «prend» 58,059% de la variance totale suivie par la deuxième dont la valeur propre est de 1,340 et prend 33,488% de la variance résiduelle. La variance cumulée prise par ces deux composantes est 91,547%. Ces deux composantes (nouvelles variables) serviront à décrire les quatre variables prises simultanément. Remarquez que la troisième partie du tableau a affinée la distribution des

variances entre les deux premières composantes, suite à la rotation «*varimax*», sans perdre de l'information en absence de rotation.

Le dernier tableau (tableau 3) est celui des vecteurs propres des variables originales qui permet l'écriture mathématique de la composante et d'accorder à chaque individu ses coordonnées sur les nouvelles variables créées.

Ainsi la première composante s'écrit sur les données normalisées:

$$C1 = -0,082 \times \text{poids} + 0,426 \times \text{taille} - 0,938 \times \text{revenu} + 0,186 \times \text{nombre d'enfants}$$

Nous remarquons que les composantes retenues sont celles dont les valeurs propres sont **supérieures à 1** (>1).

Tableau 2. Valeurs propres des composantes et pourcentage de variance utilisée (sortie SPSS).

Composante	Variance totale expliquée								
	Valeurs propres initiales			Extraction Sommes des carrés des facteurs retenus			Somme des carrés des facteurs retenus pour la rotation		
	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés	Total	% de la variance	% cumulés
1	2,322	58,059	58,059	2,322	58,059	58,059	1,916	47,907	47,907
2	1,340	33,488	91,547	1,340	33,488	91,547	1,746	43,640	91,547
3	,235	5,869	97,415						
4	,103	2,585	100,000						

Tableau 3. Matrice des composantes retenues pour la réduction des variables (sortie SPSS).

Matrice des composantes			Matrice des composantes après rotation		
	Composante			Composante	
	1	2		1	2
poids	,563	,799	Poids	-,082	,974
taille	,887	,394	Taille	,426	,873
revenu	-,733	,586	Revenu	-,938	-,022
Nombre d'enfant	,825	-,449	Nombre d'enfant	,921	,186
Vecteurs propres avant rotation					

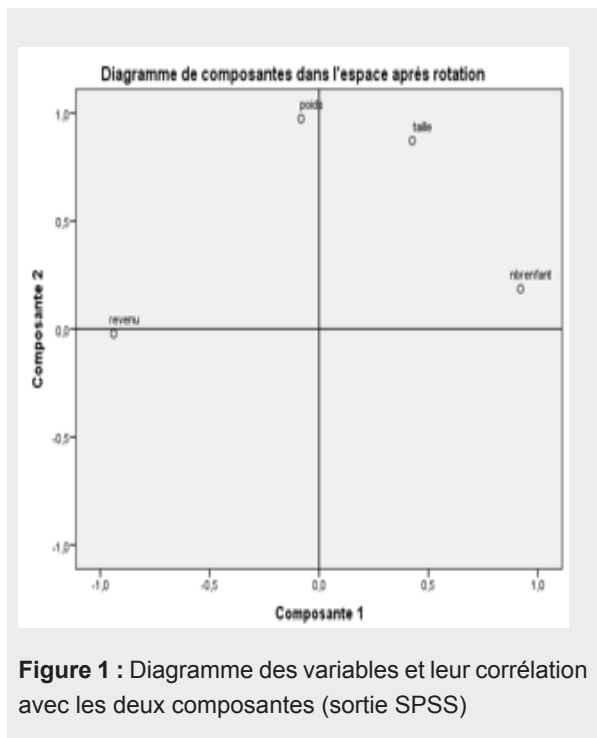


Figure 1 : Diagramme des variables et leur corrélation avec les deux composantes (sortie SPSS)

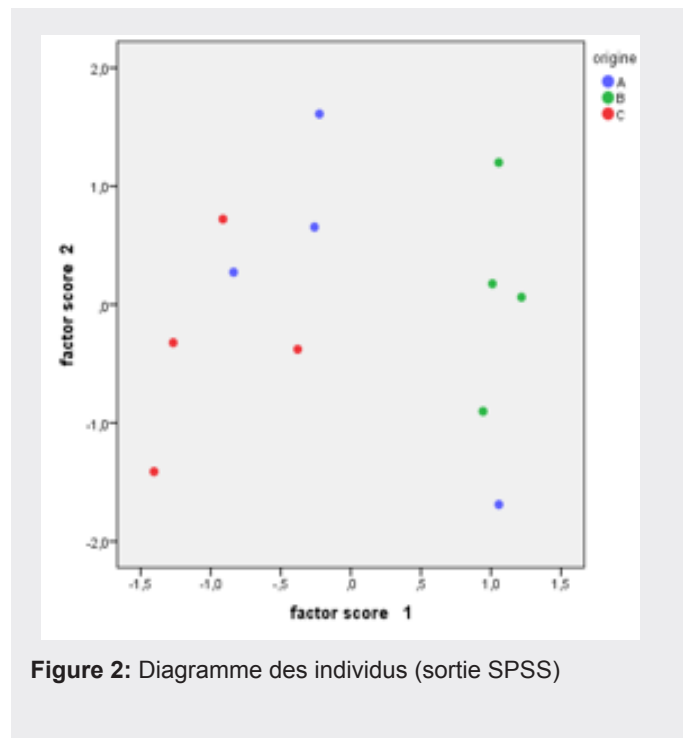


Figure 2: Diagramme des individus (sortie SPSS)

En plus de ces trois tableaux, notons les deux graphiques suivants pour décrire les variables d'origines (figure 1) et un pour décrire les individus (figure 2). Le diagramme des variables montre qu'elles sont toutes bien corrélées avec les composantes. Elles sont toutes loin du centre. Les deux variables (taille et poids) sont proches et corrélées positivement (ce qui est logique, elles décrivent la corpulence) alors que les variables poids et nombre d'enfants sont indépendants et se trouvent sur des rayons orthogonaux. Les variables (revenu et nombre d'enfants) sont diamétralement opposées et corrélées de façon négative. Elles reflètent la dimension socio-économique des individus étudiés.

Le diagramme des individus (figure 2) nous montre que l'ethnie B se distingue par ses caractéristiques des ethnies A et C qui sont proches ce qui nous pousse à chercher plus d'information pour expliquer les similitudes ethnie A ethnie C et les spécificités de l'ethnie B.

Réponses au Quizz

1. a et b
2. a et b
3. a et b

4. a et b
5. a et b
6. a et b

L'essentiel à retenir

- L'Analyse en Composante Principale (ACP) permet de synthétiser conjointement plusieurs variables quand elles sont toutes quantitatives afin de décrire le mieux possible l'ensemble des individus définis par ces variables objets de l'étude descriptive.
- L'ACP permet de *réduire* le nombre de variables initiales et *restituer* une *quantité d'information maximale* (optimale)
- Deux types de graphiques seront générés par l'ACP, un pour visualiser les relations entre les variables et l'autre pour identifier des similitudes/non similitudes entre les individus.
- Il est important de vérifier, dans une phase préliminaire, l'adéquation des données pour une ACP, à partir de l'indice Kaiser-Meyer-Olkin (KMO), qui doit être supérieur à 0,5 pour une adéquation acceptable des données à une ACP.
- Deux graphiques sont générées par l'ACP pour décrire les variables d'origines et les individus.

RÉFÉRENCES

1. Halayem S, Hammami M, Fakhfakh R, Gaddour N, Tabbane K, Amado I, Krebs MO, Bouden A. Adaptation et validation d'une échelle des signes neurologiques mineurs chez l'enfant. *Encephale*. 2017;43(2):128-134.
2. Barhoumi T, Ben Abdelaziz A, Sakly N, Ben Salem K, Ben Abdelaziz A. Methodological sheet n° 3: And Allah... created variability. *Tunis Med*. 2020;98(3):191-205.
3. Bruno F. Comprendre et utiliser les statistiques dans les sciences de la vie. Masson: Paris ; 1996.

Annexe 1: Base (fictive) ayant servi pour l'exemple

Ethnie	Nombre d'enfants	Poids	Taille	Revenu
A	3	75	176	500
A	2	66	168	700
A	2	70	170	400
A	4	55	160	300
B	4	67	170	200
B	5	70	180	350
B	5	65	171	300
B	3	60	165	150
C	1	65	160	700
C	1	57	155	850
C	2	69	170	700
C	3	66	159	600

Annexe 2 : L'Analyse par Composantes Principales (ACP), sous SPSS

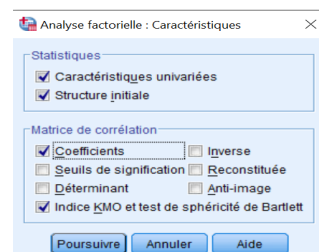
	origine	poids	taille	revenu	nbrenfant
1	A	75	176	500	3
2	A	66	168	700	2
3	A	70	170	400	2
4	A	55	160	300	4
5	B	67	170	200	4
6	B	70	180	350	5
7	B	65	171	300	5
8	B	60	165	150	3
9	C	65	160	700	1
10	C	57	155	850	1
11	C	69	170	700	2
12	C	66	159	600	3

1

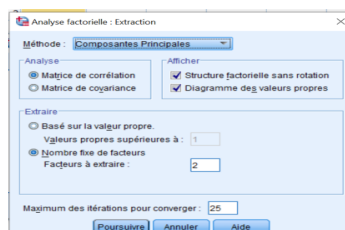
2



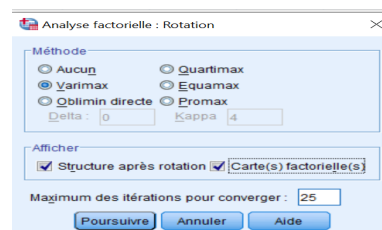
3



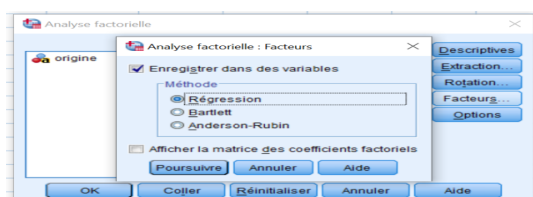
4



5



6



7

	origine	poids	taille	revenu	nbrenfant	FAC1_1	FAC2_1
1	A	75	176	500	3	-.22417	1,61088
2	A	66	168	700	2	-.63480	-.27343
3	A	70	170	400	2	-.25934	-.65540
4	A	55	160	300	4	1,05659	-1,68920
5	B	67	170	200	4	1,01038	1,7604
6	B	70	180	350	5	1,05589	1,20035
7	B	65	171	300	5	1,21789	-.06266
8	B	60	165	150	3	-.94369	-.90137
9	C	65	160	700	1	-1,26753	-.32194
10	C	57	155	850	1	-1,40455	-1,41102
11	C	69	170	700	2	-.91265	-.72248
12	C	66	159	600	3	-.37941	-.37760

8