

Ministère De L'enseignement Supérieur Et De La Recherche Scientifique

Université D'ORAN 2- Mohamed Ben Ahmed

Faculté Des Sciences Economiques Commercial Et Des Sciences De Gestions

Département Sciences Economiques



## Polycopié du module

# Analyse Des Données

Master 1

Economie Et Gestion Des Entreprises

Economie International

Semestres 2

Réaliser par :

**Mme. SABEUR CHOUIREF MOKHTARIA**

Maître de conférences- A-

Département des sciences économiques

**Année Universitaire**

2023/2024

# Avant-propos

L'analyse des données est un terme regroupant plusieurs méthodes permettant d'extraire ou de synthétiser de l'information contenue dans un jeu de données. En effet, face à la capacité toujours plus grande de stockage des données, il est de plus en plus fréquent pour le statisticien de se retrouver confronté à des jeux de données de grande dimension, dont il faudra extraire des informations pertinentes. En fonction de l'objectif de l'étude, différentes méthodes pourront être utilisées, mais elle présente en général des caractéristiques communes.

Ce cours décrit le processus d'analyse des données ainsi que la manière de gérer ce processus. Nous décrivons la nature répétitive de l'analyse des données et le rôle de la formulation d'une question précise, l'analyse exploratoire des données, l'inférence, la modélisation statistique formelle, l'interprétation et la communication.

# Fiche De Contact

- **Enseignants chargés de cours et TD** : SABEUR CHOUIREF  
MOKHTARIA
- **Public cible** : le cours s'adresse aux étudiants de première année Master (M1)
- **Coefficient** : 2
- **Crédits** : 2
- **Unité d'enseignement** : transversale\_ S2
- **Modalité d'évaluation** : Examen100%
- **Contact**: [sabeur00031@gmail.com](mailto:sabeur00031@gmail.com)

# SOMMAIRE

## Section 1 : Définition Et Rappel

Définition D'analyse Des Données .....	01
Corrélation Et Régression Linéaire : Liaison Entre Deux Variables.....	02

## Section 2 : Analyse Des Données Sous SPSS

Introduction.....	16
Les Fenêtres De SPSS.....	17
Editeur Des Données.....	21
Récupérer Un Fichier De Données.....	28
Analyse Descriptive Simple.....	29
Saisie De Données.....	32
Statistique Descriptive Et Graphique.....	37
• Cas D'une Variable Discrète.....	37
• Cas D'une Variable Continue.....	44
Analyse En Composantes Principales Sous SPSS.....	62

## Section 3 : Les Tests D'hypothèses avec SPSS

Introduction.....	91
Principe D'un Test D'hypothèses.....	91

Définition Des Concepts Utiles A L'élaboration Des Tests D'hypothèse.....	92
Les Teste D'hypothèses Sous Spss .....	94
La Corrélacion .....	94
Analyse De Variance.....	96
Test T.....	102
Test Khi-Deux.....	105
Bibliographie	
Résumé	

## Section 1 : Définition Et Rappel

### Introduction :

L'analyse des données est une technique relativement ancienne 1930 (PEARSON, SPEARMAN, HOTELLING). Elle a connu cependant des développements récents 1960 1970 du fait de l'expansion de l'informatique. L'analyse des données est une technique d'analyse statistique d'ensemble de données. Elle cherche à décrire des tableaux et à en exhiber des relations pertinentes. Elle se distingue de l'analyse exploratoire des données. L'objectif de la démarche statistique est de faire apparaître ces liaisons. Les deux types de relations fondamentales sont les relations d'équivalence et les relations d'ordre. Ainsi, une population peut elle être décomposée en classes hiérarchisées.

### Analyse des données :

#### Définition 1 :

Ensemble de méthodes descriptives ayant pour objectif de résumer et visualiser l'information pertinente contenue dans un grand tableau de données

#### Définition 2 :

Pour J-P. Fénelon « l'analyse des données est un ensemble de techniques pour découvrir la structure, éventuellement compliquée, d'un tableau de nombres à plusieurs dimensions et de traduire par une structure plus simple et qui la résume au mieux. Cette structure peut le plus souvent, être représentée graphiquement »<sup>1</sup>. Ces techniques qui sont essentiellement descriptives, ont pour but de décrire, de réduire, de classer et de clarifier les données en tenant compte de nombreux points de vue et d'étudier, en dégagant les grands traits, les liaisons, les ressemblances ou les différences entre les variables ou groupes de variables. Les documents fournis sont qualifiés de « synthétiques et percutants et valent souvent mieux

---

<sup>1</sup> 1 Fénelon, J.P., (1981). Qu'est-ce que l'analyse des données ?

qu'un long discours ». Cette approche descriptive et multidimensionnelle permet de dire que l'Analyse des Données, c'est de la statistique descriptive perfectionnée.

Lors de toute étude statistique, il est nécessaire de décrire et explorer les données avant d'en tirer de quelconques lois ou modèles prédictifs.

Dans beaucoup de situations, les données sont trop nombreuses pour pouvoir être visualisables (nombre de caractéristiques trop élevées). Il est alors nécessaire d'extraire l'information pertinente qu'elles contiennent ; Les techniques d'ADD répondent à ce besoin.

**Tableau N°01** : Trois grandes familles de méthodes.

<b>Objectif</b>	<b>V. Quantitatifs</b>	<b>Variables Qualitatifs/Mixtes</b>
Repérer et visualiser les corrélations multiples entre variables et/ou les ressemblances entre individus)	Analyse en composantes principales (ACP)	Analyse factorielle des correspondances (AFC AFCM)
Réaliser une typologie des individus	Méthodes de classification (CAH,..)	AFC ou AFCM et classification
Caractériser de groupes d'individus à l'aide de variables	Analyse discriminante (AFD,..)	Analyse discriminante (AFD,..)

## I. **Corrélation et régression linéaire : (Liaison entre deux variables)**

On constate très souvent dans la pratique qu'il existe une liaison entre deux (ou plusieurs) variables. Par exemple : la relation entre le poids et la taille des hommes adultes sont liés d'une certaine façon ; entre le revenus et les dépenses en nourriture.

Pour étudier la liaison entre deux variables quantitatives (discrètes), on commence par faire un graphique du type nuage de points. La forme générale de ce graphique indique s'il existe ou non une liaison entre les deux variables.

Pour préciser les choses, on calcule ensuite un indicateur de liaison. Pour cela, il faut d'abord introduire la covariance, généralisation bidimensionnelle de la variance. Comme elle dépend des unités de mesure des deux variables considérées, on doit la rendre intrinsèque en la divisant par le produit des écarts-types. On définit ainsi le coefficient de corrélation linéaire, indicateur de liaison cherché. Il est toujours compris entre -1 et +1, son signe indique le sens de la liaison, tandis que sa valeur absolue en indique l'intensité.

En complément, on explique ce qu'est la régression linéaire d'une variable sur une autre. Lorsqu'il existe une liaison causale entre les deux variables considérées, la régression linéaire permet d'approcher la variable réponse par une fonction de la variable causale.

Le test de corrélation (contrairement à la régression simple) ne propose pas d'identifier une variable dépendante et une variable indépendante. On ne cherche qu'à déterminer l'absence ou la présence d'une relation linéaire significative entre les variables.

**1) Détermination de coefficient de corrélation de Person:** Pour le calcul du coefficient de corrélation noté  $r$ , nous utilisons la formule la plus simple :

$$r = \frac{Cov(xy)}{\delta_x \delta_y}$$

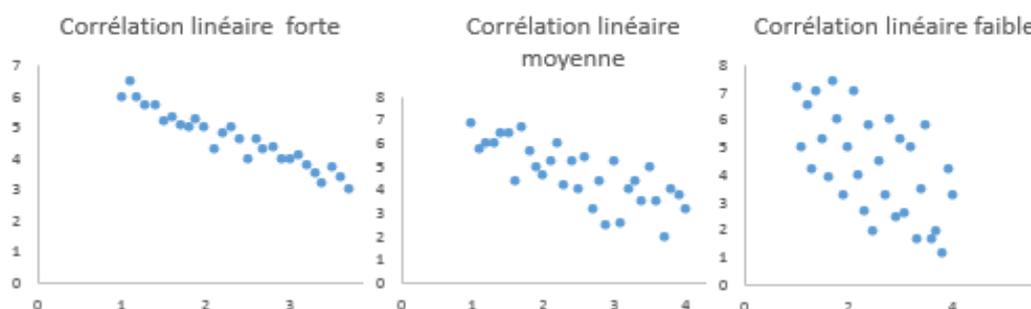
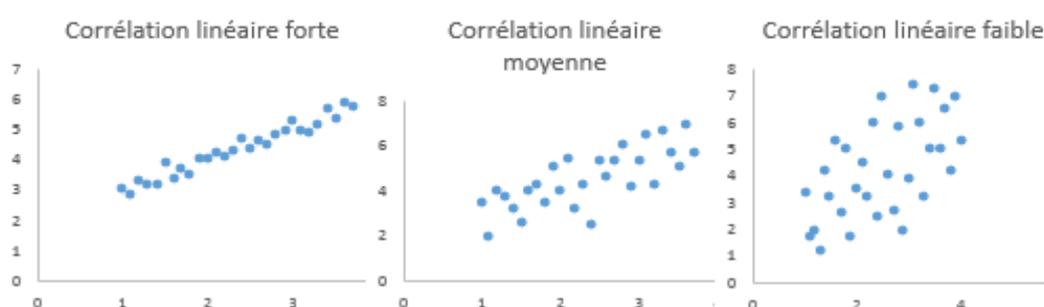
Où :

- $\bar{X}$ : la moyenne de la première distribution
- $\bar{Y}$ : la moyenne de la deuxième distribution
- $\delta(x)$  : l'écart-type de la variable x.
- $\delta(y)$  : l'écart-type de la variable y.
- $Cov$  : la covariance.

Généralement, les valeurs suivantes seront utilisées pour qualifier la corrélation linéaire:

Valeur de r	Force du lien linéaire
Près de 0	Nulle
$0 < r \leq 0.2$	Faible
$0.2 < r \leq 0.5$	Moyenne
$0.5 < r \leq 0.8$	Forte
$r > 0.8$	Très forte
1	parfaite

Afin de bien voir la différence entre chacun des qualificatifs de corrélation, voici des nuages de points qui les représentent:

**Figure N°01** : les types de corrélation entre deux variables.**Corrélation linéaire négative****Corrélation linéaire positive**

**Source** : Philippe Tassi : Méthodes statistiques 2e Eddition , Economica, 1989,p.85

**Qu'est-ce que la covariance :**

La corrélation est une quantification de la relation linéaire entre des variables continues. Le calcul du coefficient de corrélation de Pearson repose sur le calcul de la covariance entre deux variables continues. Le coefficient de corrélation est en fait la standardisation de la covariance. Cette standardisation permet d'obtenir une valeur qui variera toujours entre -1 et +1, peu importe l'échelle de mesure des variables mises en relation.

La covariance est une mesure de l'association ou du lien qui existe entre deux variables. Pour comprendre la covariance, revenons à la notion de variance. La variance d'une variable est une mesure qui quantifie la dispersion moyenne des valeurs prises par cette variable autour de sa moyenne.

Deux variables covariants ensemble lorsqu'un écart à la moyenne d'une variable est accompagné par un écart dans le même sens ou dans le sens opposé de l'autre pour le même sujet. Plus ce pattern est présent dans l'ensemble des observations, plus les deux variables semblent partager une association entre elles. Autrement dit, deux variables covariants lorsque la variation d'une des variables autour de sa moyenne semble influencer la manière dont l'autre variable varie autour de sa moyenne. La covariance exprime donc une quantité de variance partagée entre deux variables. En effet, tout comme la variance, la covariance peut se quantifier. Plus la valeur de la covariance est élevée, plus les deux variables partagent une portion importante de variance.

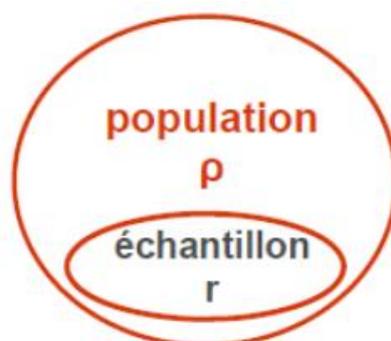
Voici la formule permettant de calculer la covariance entre deux variables continues :

$$Cov(X,Y) = \frac{\sum[(x_i - \bar{X})(y_i - \bar{Y})]}{N} = Cov(X,Y) = \frac{\sum(x_i y_i)}{N} - (\bar{X}\bar{Y})$$

### Test du coefficient de corrélation :

Après le calcul du coefficient de corrélation il faut déterminer si le coefficient de corrélation  $\rho$  de 0.  $r \approx \rho$

### Figure N°01 :



- L'hypothèse nulle (**H0**) :  $\rho=0$  (il n'y a pas une relation entre **X** et **Y**).
- L'hypothèse alternative (bilatérale) (**H1**) :  $\rho \neq 0$  (il y'a une relation entre **X** et **Y**).

Le test du coefficient de corrélation consiste à calculer la grandeur  $t_0$  et à la comparer à la valeur seuil  $t_\alpha$  sur la table de la loi de Student à  $(n-2)$  degrés de libertés.

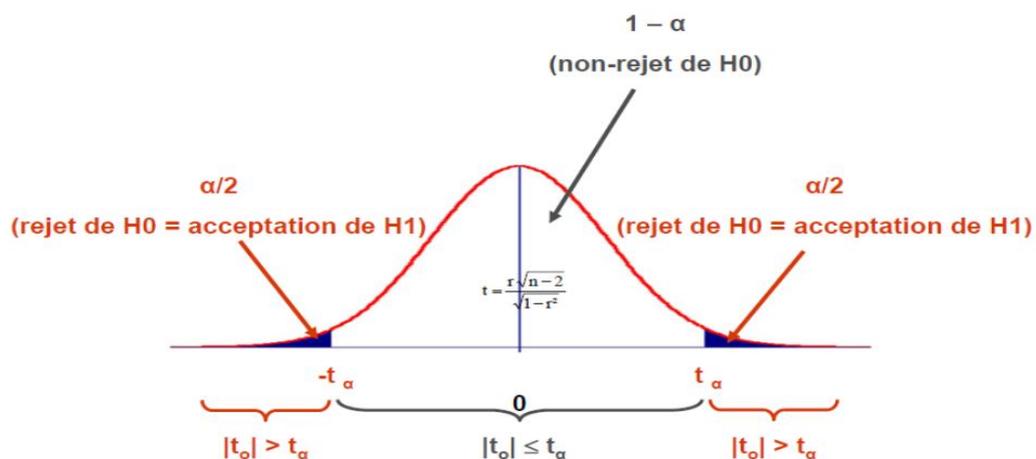
$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- $t_\alpha$ : tirez de la table du Student.
- $t_0$ : Valeur observée/calculée de  $t$  sur l'échantillon.
- $r$  : coefficient de corrélation.
- $n$  : la taille de l'échantillon.

### Conditions d'application :

- Indépendance des observations.
- Liaison linéaire entre X et Y.
- Distribution conditionnelle normale et de variance constante.

Figure N°02 :



Abscisses : valeurs possibles de  $t$  sous  $H_0$  ( $\rho = 0$ )

Détermination du degré de signification associé à  $t_0$  ( $P$  - value) :

**$p$  -value**= probabilité d'observer une valeur plus grande que  $t_0$  sous l'hypothèse nulle  $H_0$ .

**Exemple01 :**

$$t_0 = 2.12 \quad n = 20 \quad (n-2) = 18 \text{ ddl}$$

$$0.02 < p < 0.05 \rightarrow P < \alpha \rightarrow \text{rejet de } H_0$$

**Remarque :**

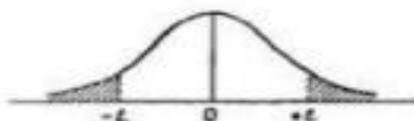
On accepte l'hypothèse alternative  $H_1$  si seulement si  $H_0$  est inférieur de **0.05**.

On rejette l'hypothèse alternative  $H_1$  si seulement si  $H_0$  est supérieur de **0.05**.

**Figure N°03 :** Table du Student.

Table de  $t$  (\*).

La table donne la probabilité  $\alpha$  pour que  $t$  égale ou dépasse, en valeur absolue, une valeur donnée, en fonction du nombre de degrés de liberté (d.d.l.).



d.d.l. \ $\alpha$	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
$\infty$	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

**Exemple :** Une entreprise veut mener une étude sur la liaison entre les dépenses (hebdomadaires) mensuelle en publicité et le volume des ventes qu'elle réalise. Nous avons obtenu au cours des six derniers mois les données suivantes :

<b>les dépenses publicitaires(en milliers de DA)</b>	70	80	30	50	35	45
<b>Le volume des ventes (en milliers de DA)</b>	580	380	200	310	400	450

1. Poser L'hypothèse nulle et l'hypothèse alternative
2. Calculer coefficient de corrélation  $r$  de  $X$  et  $Y$
3. Donner la droite de régression de  $X$  et  $Y$ .
4. Commentez les résultats obtenus.

**Solution:**

- ✓ L'hypothèse nulle ( $H_0$ ) : il n'y a pas une relation entre les dépenses publicitaires  $X$  et le volume des ventes  $Y$ .
- ✓ L'hypothèse alternative (bilatérale) ( $H_1$ ) : il y'a une relation entre les dépenses publicitaires  $X$  et le volume des ventes  $Y$ .

les mois	les dépenses publicitaires X	Le volume des ventes Y	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$	$(y_i - \bar{Y})$	$(y_i - \bar{Y})^2$	$(x_i - \bar{X})(y_i - \bar{Y})$
1	70	580	18	324	193	37249	3474
2	80	380	28	784	-7	49	-196
3	30	200	-22	484	-187	34969	4114
4	50	310	-2	4	-77	5929	154
5	35	400	-17	289	13	169	-221
6	45	450	-7	49	63	3969	-441
$\Sigma$	<b>310</b>	<b>2320</b>		<b>1934</b>		<b>82334</b>	<b>6884</b>

**La moyenne :**

La taille de l'échantillon  $N=6$ (cas des petits échantillons)

On calcule les moyennes arithmétiques **X** et **Y**. On trouve :

$$\bar{X} = \frac{\sum x_i}{N}$$

$$\bar{X} = 310 / 6 = 51.6 \approx 52$$

$$\bar{Y} = 2320 / 6 = 386.66 \approx 387$$

**La variance :**

$$V(x) = \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{N}$$

$$V(x) = 1934 / 6 = 322.33$$

$$V(y) = 82334 / 6 = 13722.33$$

**L'écart type :**

$$\delta(x) = \sqrt{V(x)} = 17.95$$

$$\delta(y) = \sqrt{V(y)} = 117.14$$

**La covariance :**

$$Cov(X, Y) = \frac{\sum[(x_i - \bar{X})(y_i - \bar{Y})]}{N}$$

$$Cov(X, Y) = 6884 / 6 = 1147.33$$

**Le coefficient de corrélation :**

$$r = \frac{Cov(xy)}{\delta_x \delta_y}$$

$$r = 1147.33 / (17.95 * 117.14) = +0.55 \rightarrow \text{Corrélation positive forte.}$$

**Le coefficient R<sup>2</sup> :**

$$r^2 = (0.55)^2 = 0.30$$

$$R^2 = 0.30 * 100 = 30\%.$$

**La droite de régression de Y en X :**

Elle est de la forme :  $Y' = ax + b$

$$a = \text{Cov}(X, Y) / V(x) = 1147.33 / 17.95 = 3.56$$

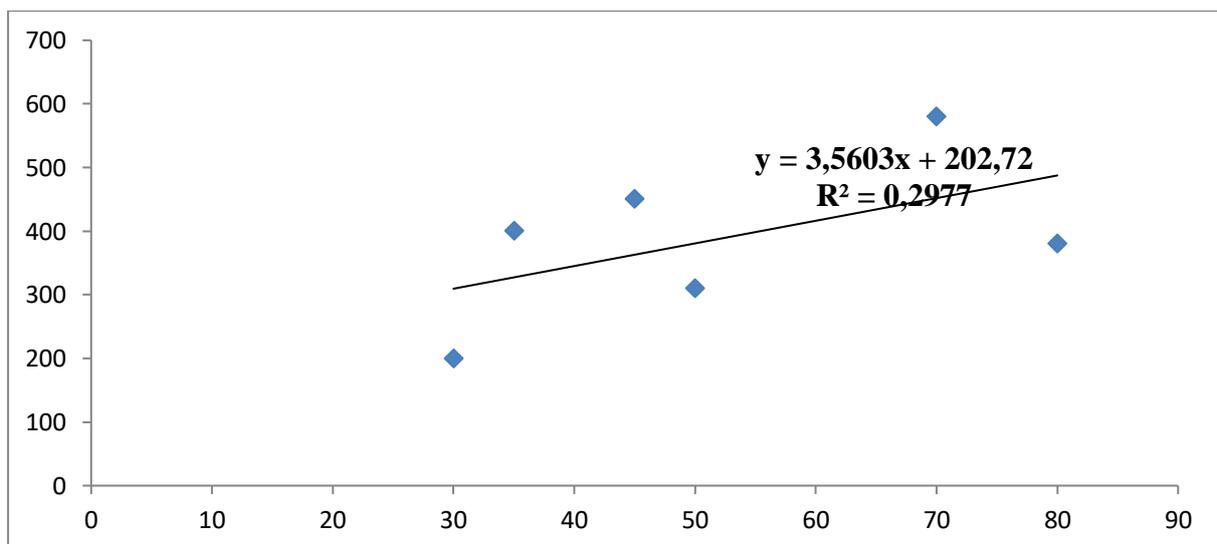
$$b = \bar{Y} - a\bar{X} \rightarrow b = 387 - (3.56) * (52) = 1403.06$$

C'est-à-dire :

$$Y' = -3.56x + 201.91$$

**Graphique:** corrélation entre les dépenses publicitaire (X) et le volume des ventes.

(Y).



**Calcul  $t_0$ :** 
$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t_0 = [0.55 * \sqrt{6-2}] / \sqrt{(1-0.30)} = [(0.55*2)/\sqrt{0.7}] = 1.33$$

$$t_0 = 1.31$$

$$dll \text{ (degré de liberté)} = n - 2 = 6 - 2 = 4$$

**Conclusion :**

Nous pouvons conclure qu'il ya une corrélation positive forte entre les dépenses publicitaire (X) et le volume des ventes ( $r= +0.55$ ).

30% de la variation du volume des ventes (Y) se trouvent expliquée par le lien entre les dépenses publicitaire (X) et le volume des ventes ( $R^2= 0.30\%$ )

Concernant les hypothèses :

On a :

$$\left\{ \begin{array}{l} t_0=1.31 \\ n=6 \\ dll=4 \end{array} \right.$$

$0.30 < \rho < 0.20 \rightarrow P < \alpha \rightarrow$  accepter  $H_0$  et rejeter  $H_1$ .

Donc il n'y a pas une relation entre les dépenses publicitaire et le volume des ventes.

Analyse des  
données avec  
SPSS

## **Introduction :**

SPSS est l'abréviation à l'origine de « **S**tatistical **P**ackage for the **S**ocial **S**ciences», mais maintenant de « **S**tatistical **P**roduct and **S**ervice **S**olutions ». Il est le principal programme pour contrôler et analyser des données<sup>i</sup>.

IBM SPSS Statistics est le premier logiciel statistique au monde utilisé pour résoudre les problèmes métier et les enjeux de la recherche au moyen d'analyses ad hoc, de tests d'hypothèses, de l'analyse géospatiale et de l'analyse prédictive. Les entreprises utilisent IBM SPSS Statistics pour comprendre les données, analyser les tendances, réaliser des prévisions et définir des plans afin de valider les hypothèses et de tirer des conclusions exactes.

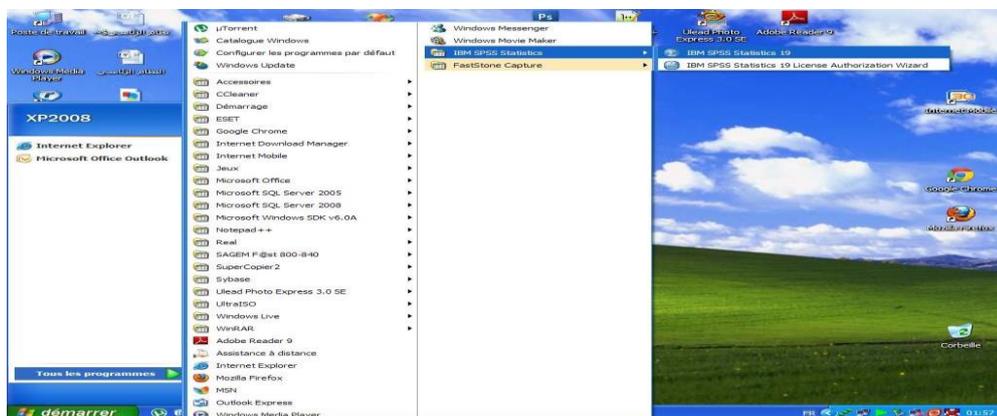
Cependant, c'est un logiciel dont le noyau central est âgé. Certaines des normes de syntaxes que nous verrons plus loin était habituelles pour les programmeurs en langage Fortran des années soixante. Il ne reste plus beaucoup de ces programmeurs et aujourd'hui, ces règles de syntaxes semblent arbitraires. De plus, il existe maintenant des langages de manipulation de fichier de données (par exemple, SQL) qui sont nettement plus performants et plus intuitifs que les commandes offertes par SPSS. Malgré ces défauts, SPSS reste un logiciel utilisé dans presque toutes les universités du monde, et il est tellement complet qu'on utilise rarement un autre logiciel<sup>ii</sup>.

D'autres produits qui remplissent le même rôle que SPSS existent, tel SAS, SYSTAT, statistica ou encore S-PLUS et son clone gratuit, le R-project. Tous sont très différents d'utilisation mais atteignent le même but: faire des tests statistiques sans devoir connaître les formules par cœur. D'autres logiciels, tel le chiffrier Excel, peuvent aussi faire quelques tests statistiques, mais ils sont limités quant au nombre de données permises, et ne sont pas plus faciles à utiliser (malgré les apparences...).

### 1) Les fenêtres de SPSS

Lorsque SPSS démarre, il ouvre une fenêtre principale qui ressemble un peu à une fenêtre du **tableur** (tel Excel). Il peut aussi ouvrir d'autres fenêtres, comme celles des résultats et de syntaxe.

Une session typique sur SPSS aura toujours ces trois fenêtres. Elles sont:



#### a. fenêtre de données

Cette fenêtre permet d'entrer des données, de les modifier ou de les effacer.

Quand vous fermez cette fenêtre, vous quittez SPSS. SPSS vous demande toujours si vous voulez sauvegarder les données: Répondez toujours non. Si vous dites oui, il va créer une copie supplémentaire de votre fichier de données, copie inutile qui en plus ne peut plus être lue par un traitement de texte, tel Bloc-notes.

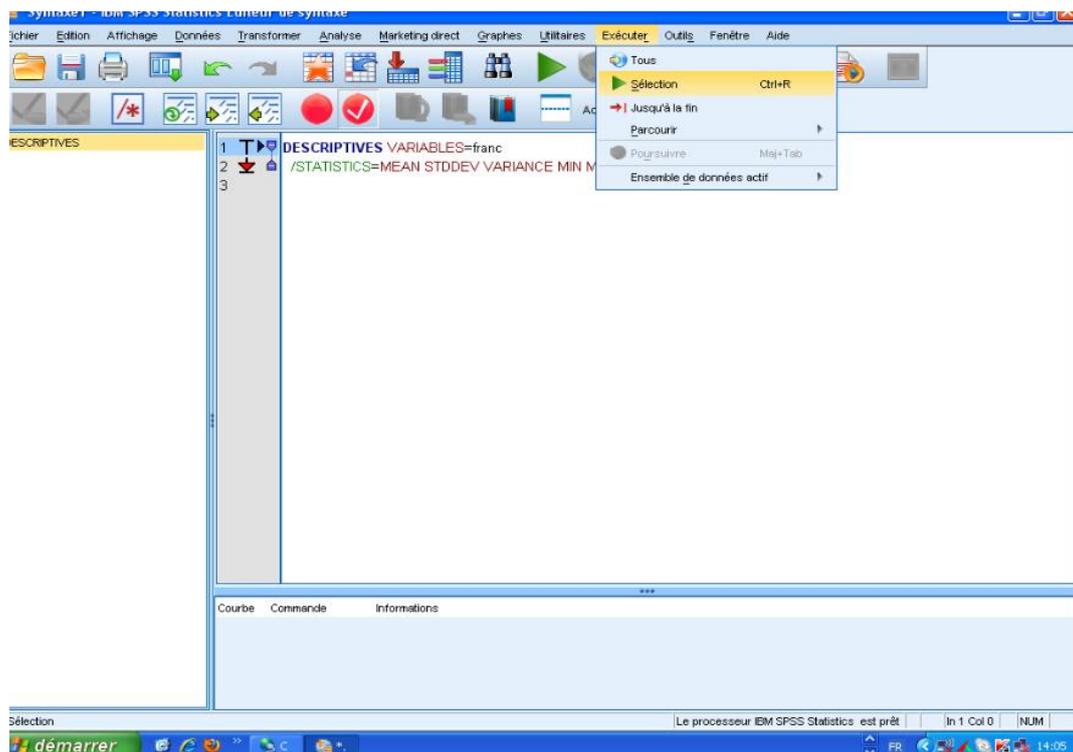
	Nom	Type	Largeur	Décimales	Etiquette	Valeurs	Manquant	Colonnes	Align	Mesur
1	ind	Numérique	9	0		Aucun	Aucun	8	Droite	Echelle
2	sexe	Numérique	9	0		{0, masculin...	Aucun	8	Droite	Ordinale
3	typEns	Numérique	9	0	type d'enseignement	{1, public}...	Aucun	8	Droite	Ordinale
4	franc	Numérique	9	2	la note de français	Aucun	Aucun	8	Droite	Echelle
5	angl	Numérique	9	2	la note d'anglais	Aucun	Aucun	8	Droite	Echelle
6	math	Numérique	9	2	la note de mathématiques	Aucun	Aucun	8	Droite	Echelle
7	phys	Numérique	9	2	la note de physique	Aucun	Aucun	8	Droite	Echelle
8	hist_geo	Numérique	9	2	la note d'histoire et géographie	Aucun	Aucun	8	Droite	Ordinale
9										
10										
11										
12										
13										
14										
15										
16										
17										
18										
19										
20										
21										
22										
23										
24										
25										

Affichage des données    Affichage des variables

Le processeur IBM SPSS Statistics est prêt

## b. fenêtre de syntaxe

Cette fenêtre permet d'écrire les commandes d'analyses statistiques. Elle fonctionne comme un traitement de texte simple. Lorsqu'une commande est complète, on peut l'exécuter en allant dans le menu "Exécuter" (ou encore en tapant Ctrl-R). Pour obtenir une fenêtre de syntaxe vide, aller dans le menu "Fichier : Nouveau : syntaxe".



### c. Fenêtre des résultats :

Cette fenêtre apparaît après qu'une commande d'analyse a été effectuée, et contient les résultats de cette analyse. Les résultats proprement dit apparaissent à droite alors qu'à gauche, on voit une table des matières des résultats générée par SPSS. Les résultats peuvent être imprimés tels quels, ou encore, on peut faire copier-coller vers un autre logiciel (tel votre traitement de texte).

The screenshot shows the IBM SPSS Statistics Viewer interface. The main window displays the following content:

**Statistiques**  
la note de mathématiques

N	Valide	200
	Manquante	0

**la note de mathématiques**

	Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide 33,00	1	,5	,5	,5
35,00	1	,5	,5	1,0
37,00	1	,5	,5	1,5
38,00	2	1,0	1,0	2,5
39,00	6	3,0	3,0	5,5
40,00	10	5,0	5,0	10,5
41,00	7	3,5	3,5	14,0
42,00	7	3,5	3,5	17,5
43,00	7	3,5	3,5	21,0
44,00	4	2,0	2,0	23,0
45,00	8	4,0	4,0	27,0
46,00	8	4,0	4,0	31,0
47,00	3	1,5	1,5	32,5

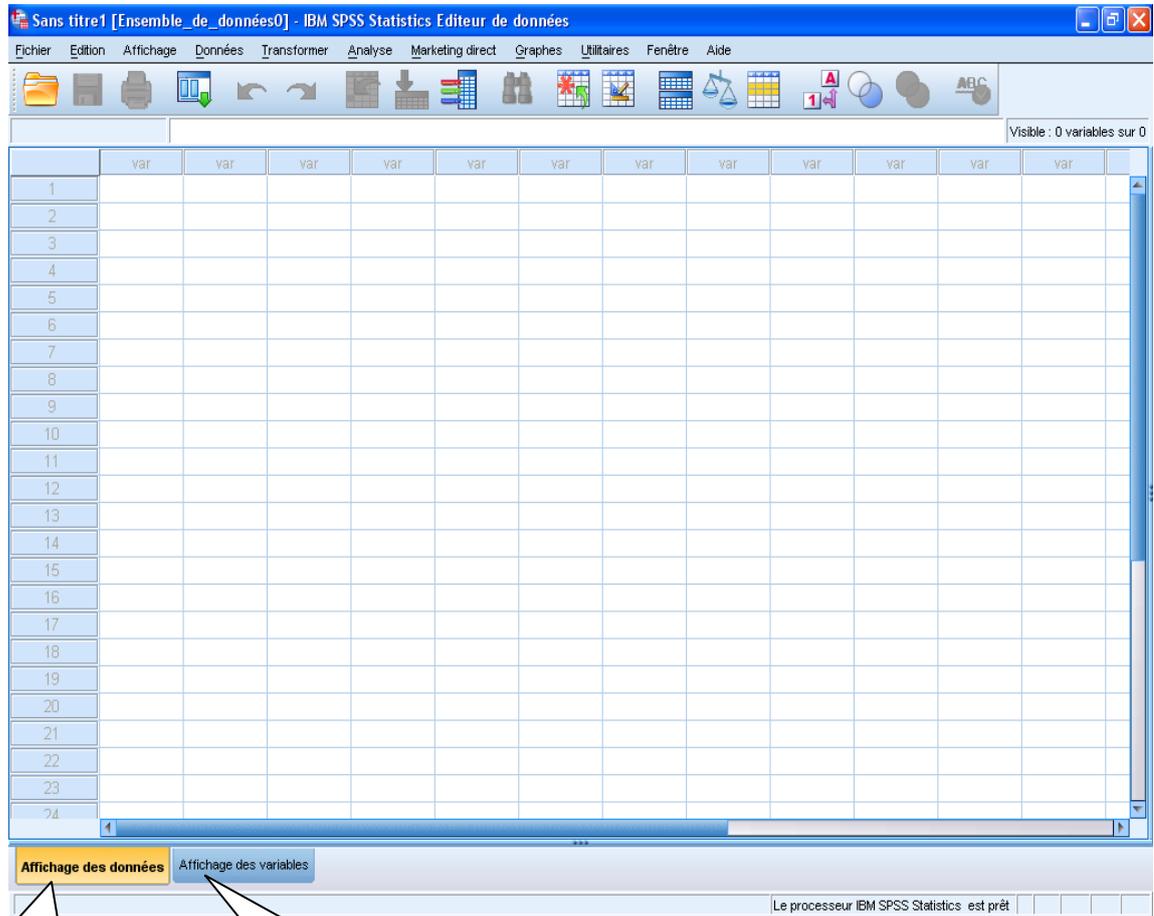
At the bottom right of the window, a status bar indicates: "Le processeur IBM SPSS Statistics est prêt".

## 2) Editeur des données

L'éditeur de données fournit une méthode pratique, semblable à celle d'un tableur, permettant de créer et de modifier des fichiers de données. La fenêtre de l'éditeur de données s'ouvre automatiquement lorsque vous lancez une session.

L'éditeur de données permet d'afficher les données de deux façons :

- ✓ **Affichage des données** :Affiche les valeurs réelles des données ou les étiquettes de valeurs définies.
- ✓ **Affichage des variables** :Affiche les informations de définition des variables,à savoir les étiquettes de valeurs et de variables définies, le type des données (par exemple, chaîne, date ou valeur numérique), le niveau de mesure (nominale, ordinale ou échelle) et les valeurs utilisateur manquantes.



**Affichage des données**

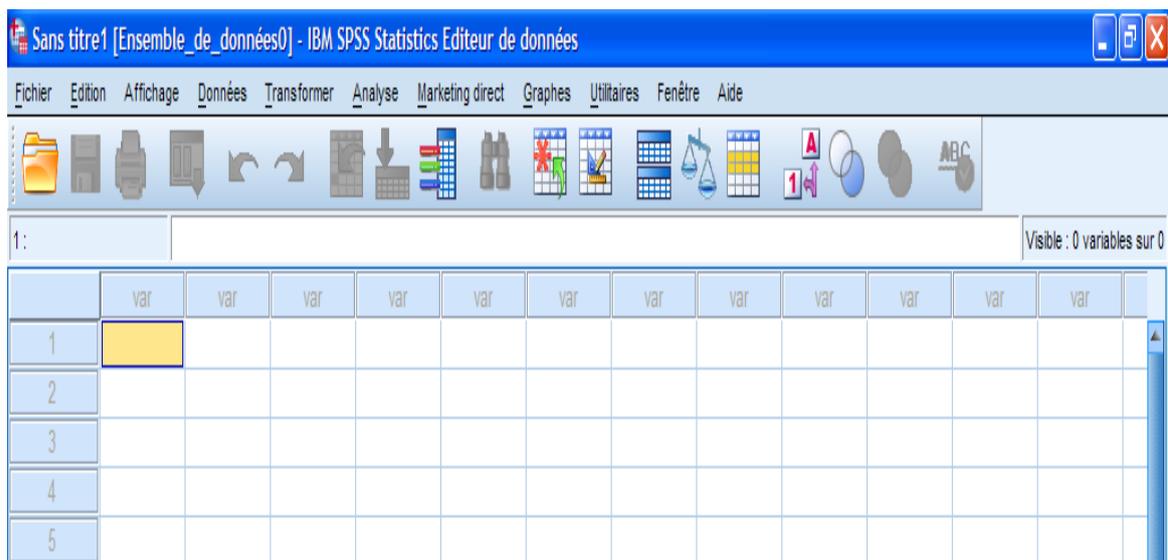
**Affichage des variables**

Dans les deux affichages, vous pouvez ajouter, modifier et supprimer les informations contenues dans le fichier de données.

### a- Affichage des données

L'éditeur des données est constitué :

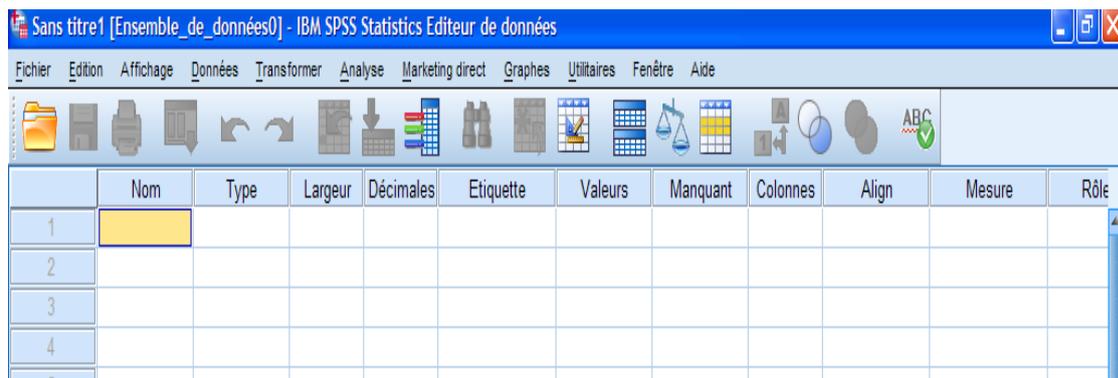
- d'un **menu de fonction** (Fichier, Edition, Affichage, Données, Transformer, Analyse, Graphes, Outils, Fenêtre, Aide) ;
- **d'une barre d'outils affichant des icônes associés aux fonctions les plus utilisées** ; et d'une **grille des données (Affichage des données)**.
- 



### b- Affichage des variables

L'affichage des variables présente les descriptions des attributs de chaque variable du fichier de données. Dans l'Affichage des variables :

- ✓ Les lignes sont des variables.
- ✓ Les colonnes sont des attributs de variable



- **Nom** : Donner un nom à la variable : 8 caractères au maximum, sans accent ni les signes ( ?, !, \*) ni espace. Le signe \_ est autorisé.
- **Type** : Définir le type de la variable (numérique, alphabétique, date, monétaire...).
- **Largeur** : nombre de chiffres accordés à la donnée (laisser 8).
- **Décimales** : Nombre de décimales possibles.
- **Etiquette** : L'étiquette permet de taper un texte plus explicite que les 8 caractères du nom de la variable.



	Nom	Type	Largeur	Décimales	Etiquette	Valeurs
1	ind	Numérique	9	0		Aucun
2	sexe	Numérique	9	0		{0, masculin...

**Exemple :**

3	typEns	Numérique	9	0	type d'enseignement	{1, public}...
---	--------	-----------	---	---	---------------------	----------------

Nom	Type	Largeur	Décimales	Etiquette	Valeurs
pays				Pays	Aucun
nbpop				Population en milliers (x 1000)	
densite				Nb d'individus / Km²	
urbain				Personnes habitant en ville (%)	Aucun
religion				Religion prédominante	
espvief				Espérance de vie moyenne des femmes	{33,00, 33,0...
espvieh				Espérance de vie moyenne des hommes	
litterat				% de personnes lettrées	
pop_aug				Augmentation de la population (% par an)	Aucun
mortbebe				Mortalité infantile (nb de morts pour 1000 vivants)	
pdb_tet				Produit domestique brut / par tête	Aucun
region				Région ou groupe économique	
calories				Consommation calorique par jour	
sida				Cas de SIDA	
tx_nais				Nb de naissances pour 1000 personnes	
tx_mort				Nb de morts pour 1000 personnes	
tx_sida				Nombre de cas de SIDA pour 100000 personnes	
log_pdb				Log (base 10) de pdb_tet	
lg_sidat				Log (base 10) de tx_sida	
nais_mor				ratio nb de naissances/nb de morts	
fertilite				Fertilité : moyenne du nombre d'enfants	
log_pop				Log (base 10) de Population	
augrecol					
lit_male				Hommes lecteurs (%)	
lit_femm				Femmes lectrices (%)	
climat				Climat prédominant	

- **Valeur** : Les étiquettes de valeur permettent de rendre les tableaux de résultats plus lisibles et plus explicites.

Largeur	Décimales	Etiquette	Valeurs	Manquant	Colonnes	A
3	0		Aucun	Aucun	8	☰ Drc
3	0		{0, masculin...	Aucun	8	☰ Drc
3	0	type d'enseign...	{1, public}...	Aucun	8	☰ Drc
3	2	la note de franç...	Aucun	Aucun	8	☰ Drc
3	2	la note d'anglais	Aucun	Aucun	8	☰ Drc
3						☰ Drc
3						☰ Drc
3						☰ Drc

**Etiquettes de valeurs**

Etiquettes de valeurs :

Valeur :

Etiquette :

Orthographe...

Ajouter

Changer

Eliminer bloc

0 = "masculin"  
1 = "féminin"

OK Annuler Aide

**Manquant :** Les valeurs manquantes permettent de traiter le cas où un répondant n'a pas répondu à une question.

**Colonnes :** La largeur d'affichage de la colonne permet d'élargir une colonne de données de façon à voir l'ensemble des données saisies.

**Aligner :** aligner à droite, à gauche ou au centre les valeurs dans le fichier des données.

**Mesure :** permet de sélectionner le type de mesure correspondant à la nature de la variable (Echelle, Ordinale ou Nominale).

Vous pouvez spécifier un niveau de mesure d'échelle (données numériques sur un intervalle ou une échelle de rapport), ordinal ou nominal. Les données nominales et ordinales peuvent être des chaînes de caractères (alphanumériques) ou numériques.

■ **Nominal.** Une variable peut être traitée comme étant nominale si ses valeurs représentent des modalités sans classement intrinsèque (par exemple, le service de la société dans lequel travaille un employé). La région, le lieu d'habitat (Oujda, Rabat, Casa...), le code postal ou l'appartenance religieuse sont des exemples de variables nominales.

■ **Ordinal.** Une variable peut être traitée comme étant ordinale si ses valeurs représentent des modalités associées à un classement intrinsèque (par exemple, des niveaux de satisfaction allant de Très mécontent à Très satisfait). Exemples de variable ordinale : des scores d'attitude représentant le degré de satisfaction ou de confiance, et des scores de classement des préférences.

■ **Echelle.** Une variable peut être traitée comme une variable d'échelle (continue) si ses valeurs représentent des modalités ordonnées avec une mesure

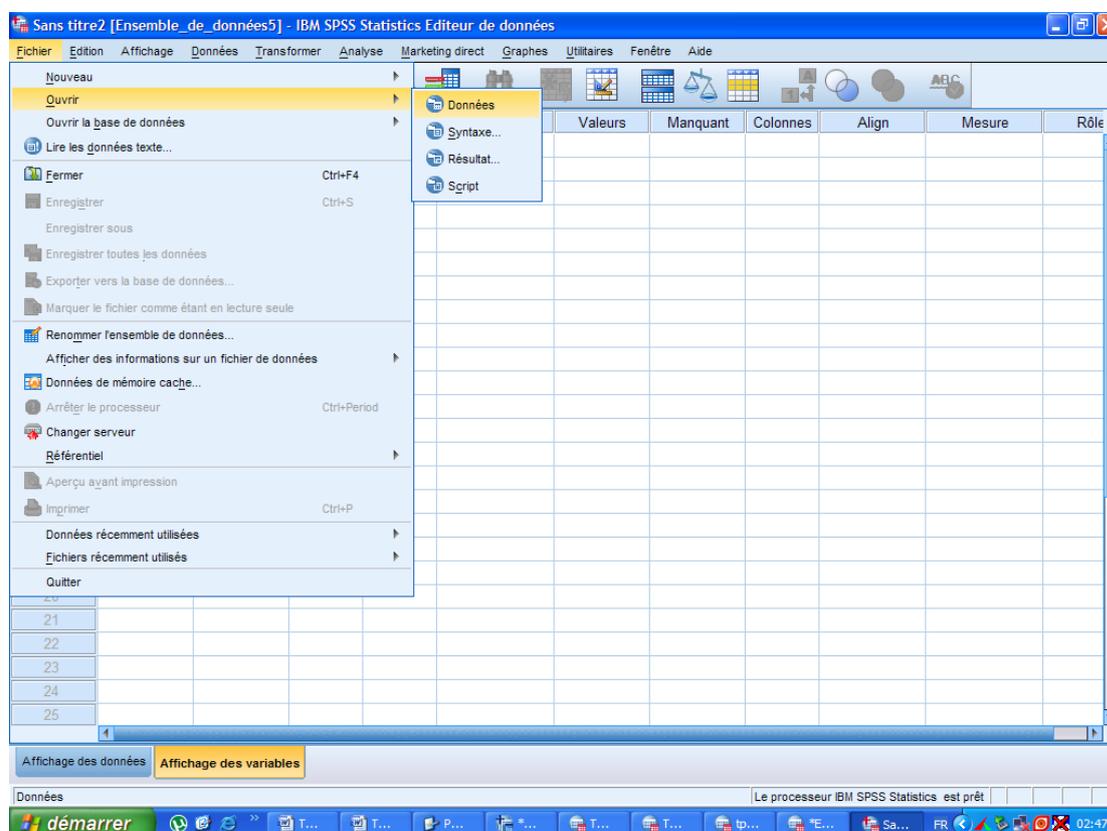
significative, de sorte que les comparaisons de distance entre les valeurs soient adéquates. L'âge en années et le revenu en milliers de dollars sont des exemples de variable d'échelle.

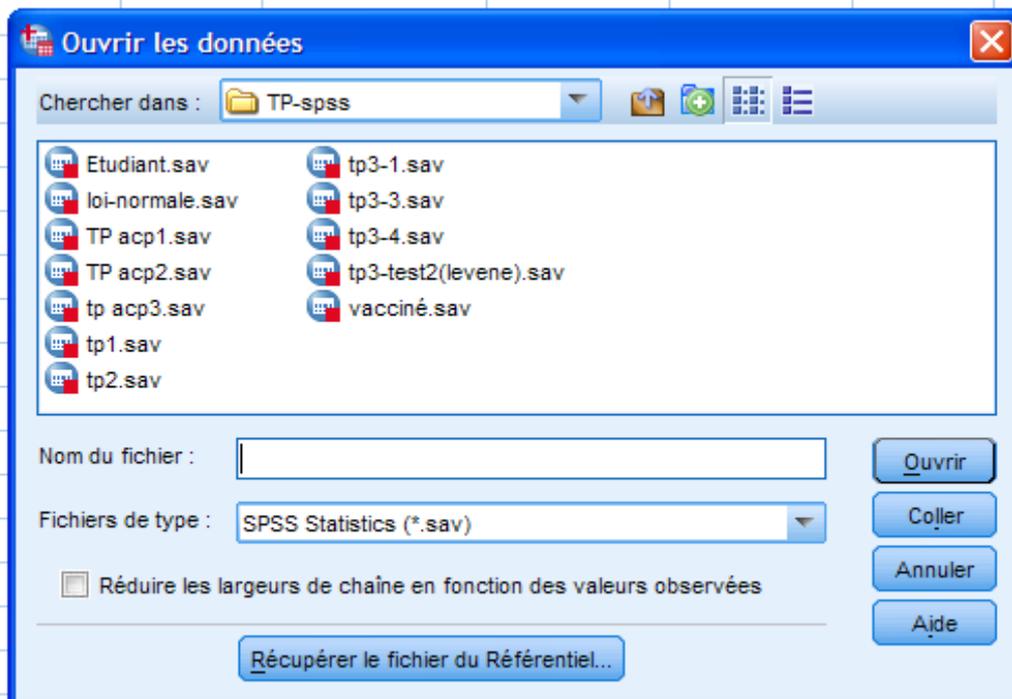
Remarque : Pour les variables chaîne ordinales, l'ordre alphabétique des valeurs chaîne est supposé refléter l'ordre des modalités. Par exemple, pour une variable chaîne comportant des valeurs Faible, Moyen, Elevé, l'ordre des modalités est interprété comme Elevé, Faible ou Moyen, ce qui ne correspond pas à l'ordre correct. En règle générale, il est recommandé d'utiliser les codes numériques pour représenter les données ordinales.

### 3) Récupérer un fichier de données

Ouvrir le fichier « **Etudiant.sav** » :

**Question** : nombre d'individu et nombre de variables.





\*Etudiant.sav [Ensemble\_de\_données1] - IBM SPSS Statistics Editeur de données

Fichier Edition Affichage Données Transformer Analyse Marketing direct Graphes Utilitaires Fenêtre Aide

	ind	sexe	typEns	franc	angl	math	phys	hist_geo	var	var
1	70	0	1	57,00	52,00	41,00	47,00	57,00		
2	121	1	1	68,00	59,00	53,00	63,00	61,00		
3	86	0	1	44,00	33,00	54,00	58,00	31,00		
4	141	0	1	63,00	44,00	47,00	53,00	56,00		
5	172	0	1	47,00	52,00	57,00	53,00	61,00		
6	113	0	1	44,00	52,00	51,00	63,00	61,00		
7	50	0	1	50,00	59,00	42,00	53,00	61,00		
8	11	0	1	34,00	46,00	45,00	39,00	36,00		
9	84	0	1	63,00	57,00	54,00	58,00	51,00		
10	48	0	1	57,00	55,00	52,00	50,00	51,00		
11	75	0	1	60,00	46,00	51,00	53,00	61,00		
12	60	0	1	57,00	65,00	51,00	63,00	61,00		
13	95	0	1	73,00	60,00	71,00	61,00	71,00		
14	104	0	1	54,00	63,00	57,00	55,00	46,00		
15	38	0	1	45,00	57,00	50,00	31,00	56,00		
16	115	0	1	42,00	49,00	43,00	50,00	56,00		

#### 4) Analyse descriptive simple : fréquence et pourcentage

Pour le calcul des fréquences et des pourcentages d'une variable, on utilise la commande :

**Analyse + Statistique Descriptive +Fréquences...**

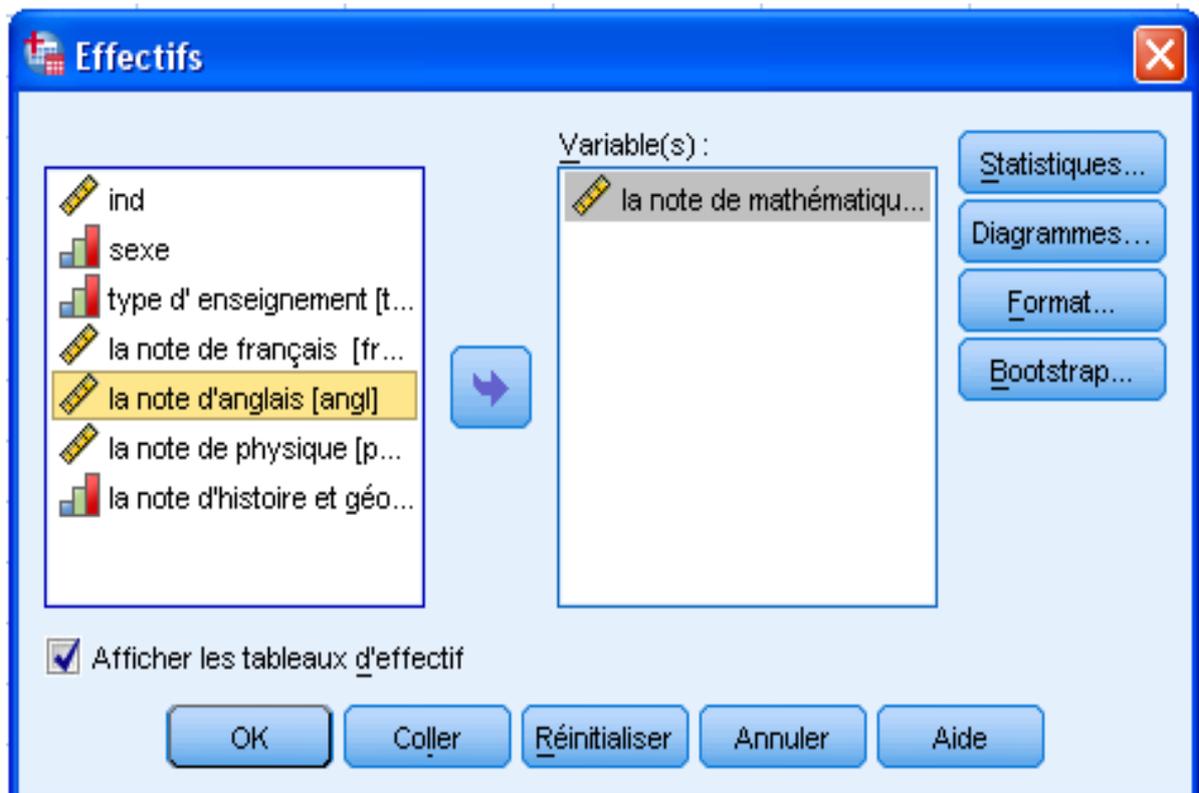
The screenshot shows the IBM SPSS Statistics interface. The 'Analyse' menu is open, and the path 'Statistiques descriptives' > 'Effectifs' is highlighted. The data table in the background has the following structure:

	ind	sexe	typE	st_geo	var	var	var	va
1	70	0		57,00				
2	121	1		61,00				
3	86	0		31,00				
4	141	0		56,00				
5	172	0		61,00				
6	113	0		51,00	63,00	61,00		
7	50	0		42,00	53,00	61,00		
8	11	0		45,00	39,00	36,00		
9	84	0		54,00	58,00	51,00		
10	48	0		52,00	50,00	51,00		
11	75	0		51,00	53,00	61,00		
12	60	0		51,00	63,00	61,00		
13	95	0		71,00	61,00	71,00		
14	104	0		57,00	55,00	46,00		
15	38	0		50,00	31,00	56,00		
16	115	0		43,00	50,00	56,00		
17	76	0		51,00	50,00	56,00		
18	195	0		60,00	58,00	56,00		
19	114	0	1	68,00	65,00	62,00	55,00	61,00
20	85	0	1	55,00	39,00	57,00	53,00	46,00
21	167	0	1	63,00	49,00	35,00	66,00	41,00
22	143	0	1	63,00	63,00	75,00	72,00	66,00
23	41	0	1	50,00	40,00	45,00	55,00	56,00
24	20	0	1	60,00	52,00	57,00	61,00	61,00

Sélectionner la variable, objet de l'étude, en cliquant sur le bouton 



Par exemple, on veut étudier la variable **math**(la note de mathématiques), on a la figure suivante :



Puis, cliquer sur **Ok**.

/ORDER=ANALYSIS.  
**Effectifs**  
 [Ensemble\_de\_données1] E:\ENSA2010\enseignement\GC\TP\Etudiant.sav

**Statistiques**  
 la note de mathématiques

N	Valide	200
	Manquante	0

**la note de mathématiques**

	Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide 33,00	1	,5	,5	,5
35,00	1	,5	,5	1,0
37,00	1	,5	,5	1,5
38,00	2	1,0	1,0	2,5
39,00	6	3,0	3,0	5,5
40,00	10	5,0	5,0	10,5
41,00	7	3,5	3,5	14,0
42,00	7	3,5	3,5	17,5
43,00	7	3,5	3,5	21,0

C'est le fichier de résultats. Il peut être sauvegardé ou imprimé. **L'extension du fichier est (.spo).**

## 5) Saisie de données

Affichage des données : les données d'une variable sont écrites dans une colonne.

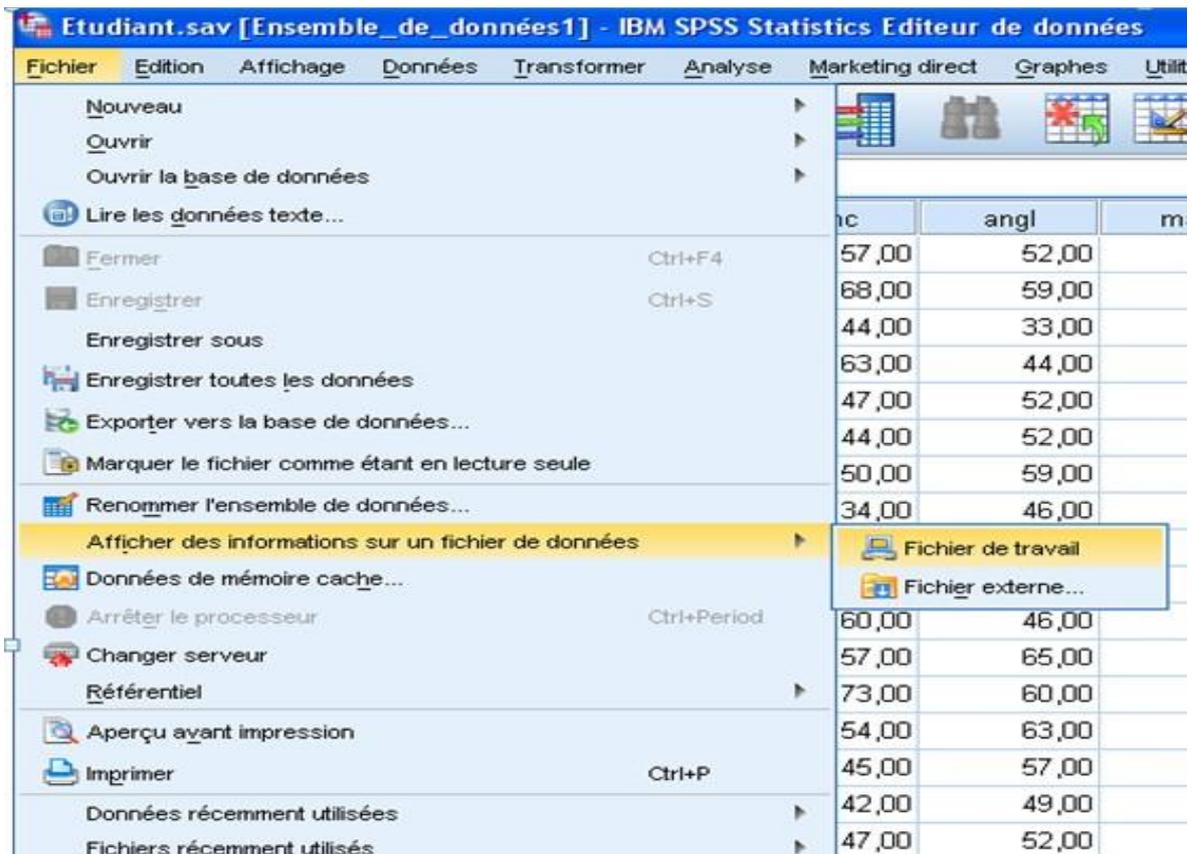
Affichage des variables : utilisation des caractéristiques (nom, type,)

### Rapport préliminaire sur les variables :

Lorsqu'on saisi soit même les données, il y a une fonction qui permet de vérifier le travail et d'obtenir une synthèse facile à lire.

La commande est :

**Fichier + Afficher des informations sur un fichier de données + Fichier de travail**



Résultats

- Log
- Informations sur les fichiers
  - Titre
  - Remarques
  - Ensemble de données
  - Informations de la variable
  - Valeurs des variables

DISPLAY DICTIONARY.

### Informations sur les fichiers

[Ensemble\_de\_données1] E:\ENSA2010\enseignement\GC\TP\Etudiant.sav

#### Informations de la variable

Variable	Position	Etiquette	Niveau de mesure	Rôle	Largeur des colonnes	Alignement	Format d'impression	Format d'écriture
ind	1	<aucune>	Echelle	Entrée	8	Droite	F9	F9
sexe	2	<aucune>	Ordinale	Entrée	8	Droite	F9	F9
typEns	3	type d'enseignement	Ordinale	Entrée	8	Droite	F9	F9
franc	4	la note de français	Echelle	Entrée	8	Droite	F9.2	F9.2
angl	5	la note d'anglais	Echelle	Entrée	8	Droite	F9.2	F9.2
math	6	la note de mathématiques	Echelle	Entrée	8	Droite	F9.2	F9.2
phys	7	la note de physique	Echelle	Entrée	8	Droite	F9.2	F9.2
hist_geo	8	la note d'histoire et géographie	Ordinale	Entrée	8	Droite	F9.2	F9.2

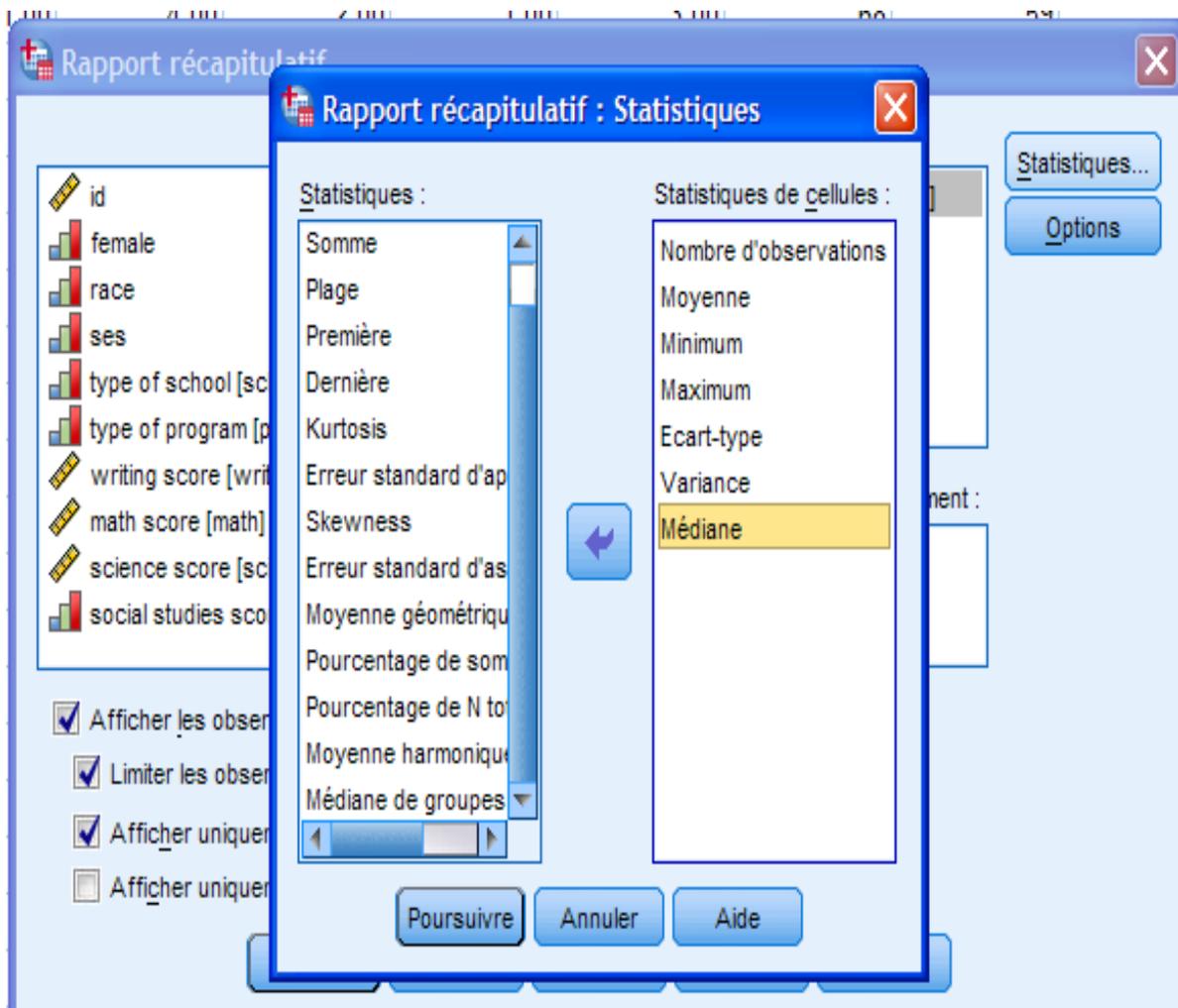
Variables du fichier de travail

## Rapports préliminaires sur les observations

Il y a une fonction qui permet de repérer les éventuelles erreurs de saisie des données.

La commande est : **Analyse + Rapports +Récapitulatifs des observations.**

On coche des statistiques simples, par exemple **minimum** et **maximum**, qui désignent respectivement la valeur minimale et la valeur maximale des observations.



## Récapituler

[Ensemble\_de\_données6] D:\GC\TP-spss\Etudiant.sav

### Observation Calculer Récapituler

	Observations					
	Inclus		Exclu(s)		Total	
	N	Pourcentage	N	Pourcentage	N	Pourcentage
reading score	200	100,0%	0	,0%	200	100,0%
math score	200	100,0%	0	,0%	200	100,0%

### Récapitulatif des observations

Lignes 1 à 100 de 206



	Numéro de l'observation	reading score	math score
1	1	57	41
2	2	68	53
3	3	44	54
4	4	63	47
5	5	47	57
6	6	44	51
7	7	50	42
8	8	34	45
9	9	63	54
10	10	57	52
11	11	60	51

		reading score	math score
Total	Moyenne	52,23	52,64
	Minimum	28	33
	Maximum	76	75
	Ecart-type	10,253	9,368
	Variance	105,123	87,768
	Médiane	50,00	52,00

### 6) Exercice :

Ici les notes des étudiants masculins et des étudiantes féminins, obtenues dans un test noté sur 100 :

**Masculins** : 30 50 95 80 45 25 30 33 44 52 55 65 75 85 80 78 87 84 48 52 55 57  
40 60 75 72 77 90 88 95 75 73 77

**Féminin** : 85 45 84 38 79 75 77 44 65 40 44 30 25 30 77 85 95 61 60 48 36 95  
98 65 66 77 42 44 62 76 75 93 98 80 70

- 1- Saisir les deux variables en leur donnant les noms (**note\_mas**) et (**note\_fem**).
- 2- Vérifier le travail de la saisie à l'aide des rapports préliminaires sur les variables et sur les observations.
- 3- Sauvegarder les données dans le fichier que vous donnez le nom (**note.sav**) dans un répertoire personnel. (c://nom étudiant/note.sav)
- 4- Quel est la fréquence de la note 75 chez les étudiants masculins, et c'est quoi son pourcentage.
- 5- Trier les données de la variable note\_mas de la plus petite valeur à la plus grande (ordre croissant). Utiliser la commande : **Données + Trier les observations**

## CHAPITRE 2 : Statistique descriptive et graphique

On peut utiliser SPSS de deux façons.

(1) SPSS est piloté par des menus descriptifs et des boîtes de dialogue simples pour la majorité des tâches. C'est le mode interactif (ou **le mode Windows**).

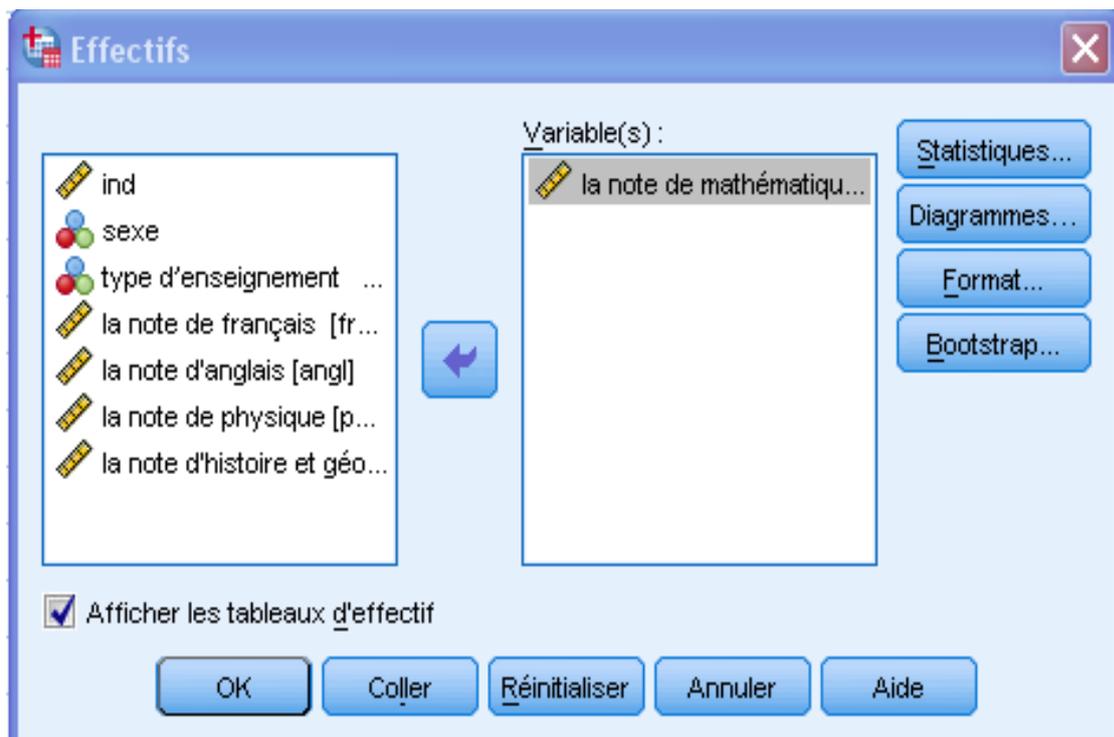
(2) Il permet aussi de composer des ensembles de commandes (programmes). C'est le mode programmation (ou **le mode Syntaxe**).

### 1- Cas d'une variable discrète

**Pratique :** ouvrir le fichier (**Etudiant.sav**), et analyser la variable (**math**=la note de mathématiques).

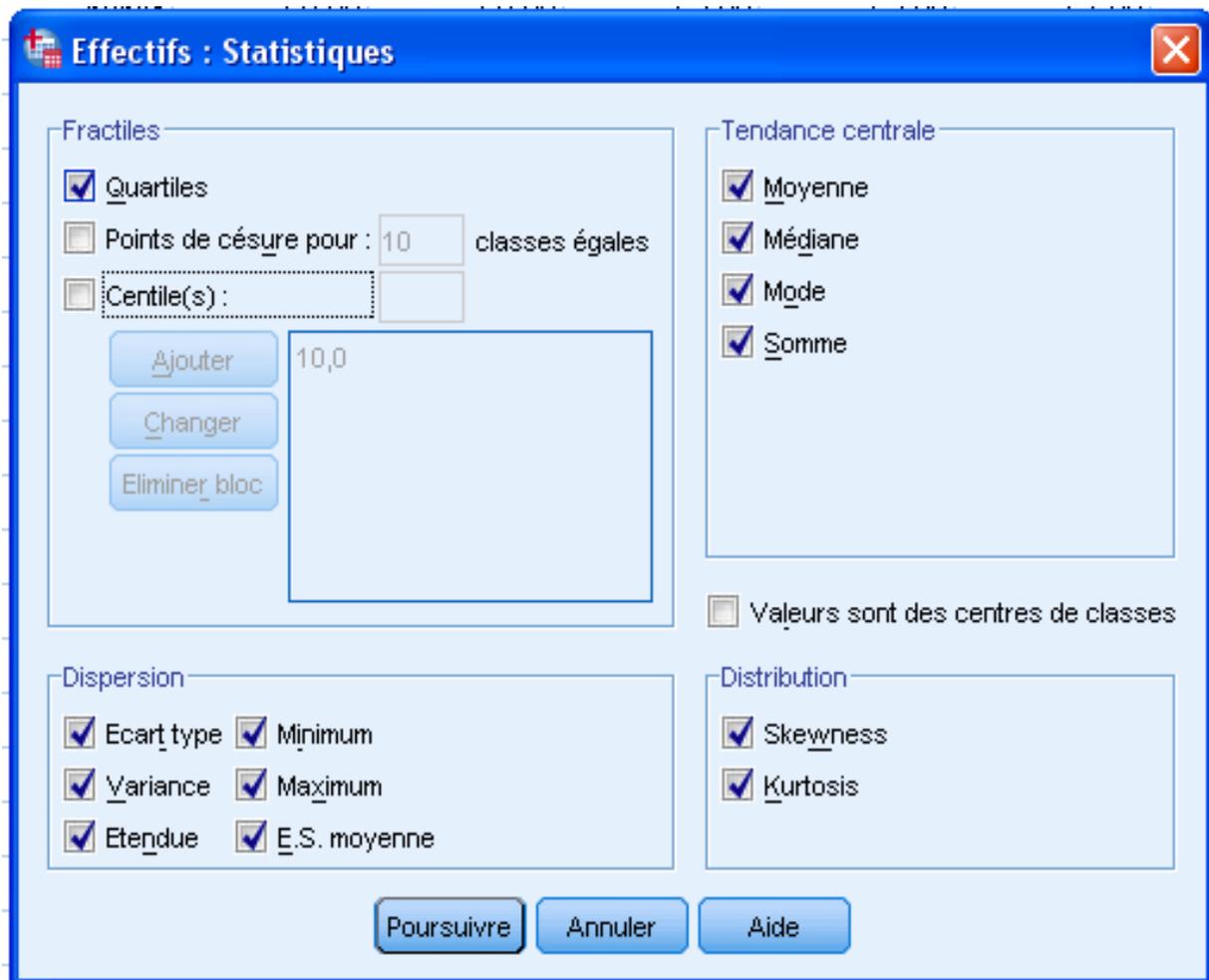
Utiliser la commande : **Analyse + Statistiques descriptives +Fréquences**

Sélectionner la variable pour laquelle on souhaite connaître les caractéristiques statistiques, puis la déplacer à l'aide de la case flèche.



Cliquer sur **statistiques** et sélectionner les éléments désirés :

- Pour une variable nominale : mode, distribution des fréquences, minimum, maximum ;
- Pour une variable ordinale : mode, distribution des fréquences, minimum, maximum, médiane ;
- Pour une variable métrique : écart-type, moyenne, minimum, maximum.



skewness =Asymétrie

$S = \frac{m_3}{\sigma_X^3}$  Le coefficient de Skewness mesure le degré d'asymétrie de la distribution.

Kurtosis= Aplatissement

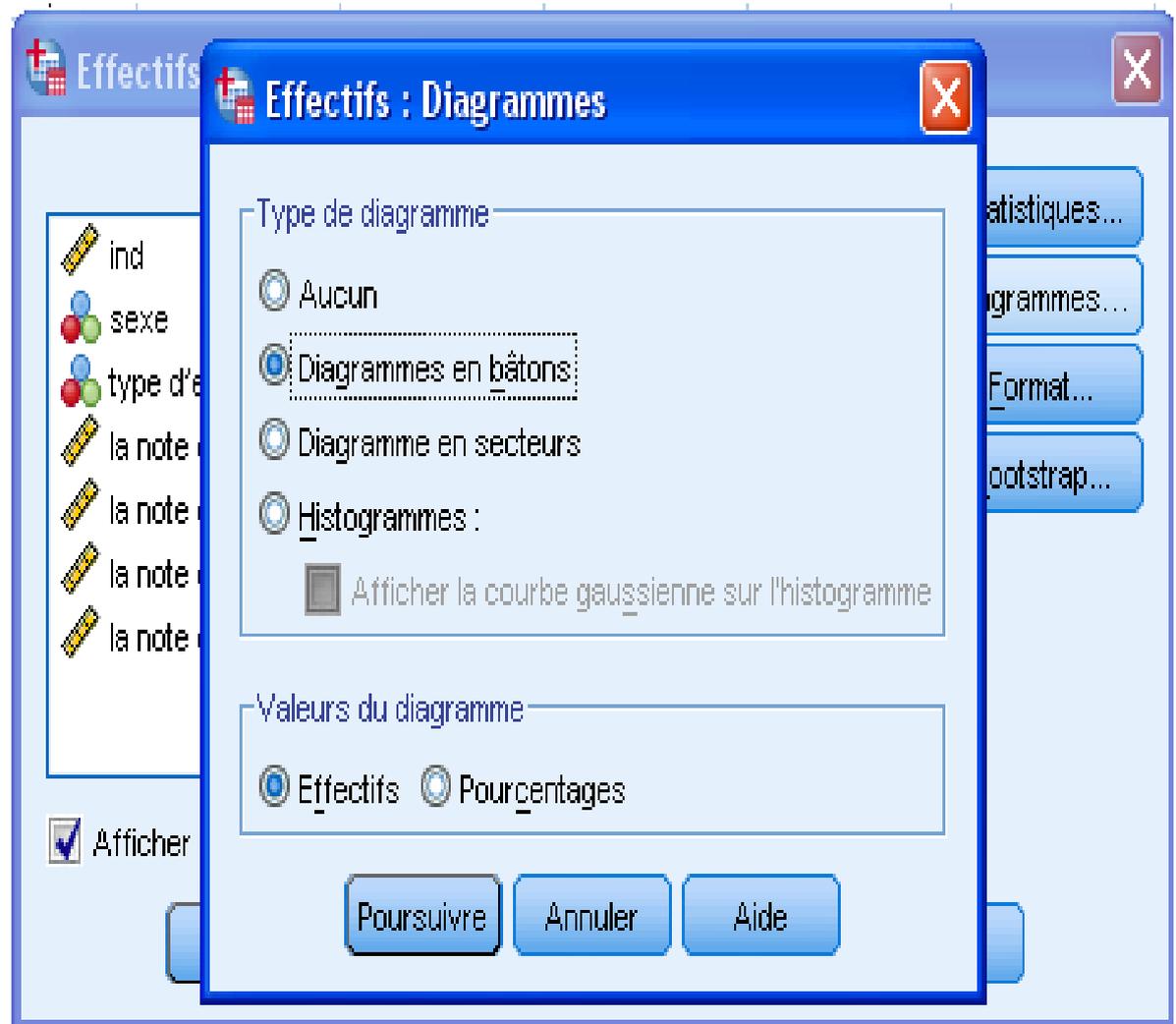
$K = \frac{m_4}{V(X)^2}$  Le coefficient de Kurtosis mesure le degré d'aplatissement de la distribution.

Avec :  $m_q$  moment d'ordre q par rapport à  $\bar{X}$

Cliquer sur **Poursuivre**.

Cliquer sur **Diagrammes** et sélectionner les éléments désirés :

- diagramme en bâtons : variable discrète ;
- graphique en secteur : caractère qualitatif ;
- histogrammes : variable continue.



On obtient le tableau des résultats suivants :

➔ **Effectifs**

[Ensemble\_de\_données1] E:\ENSA2010\enseignement\GC\TP\Etudiant.sav

**Statistiques**

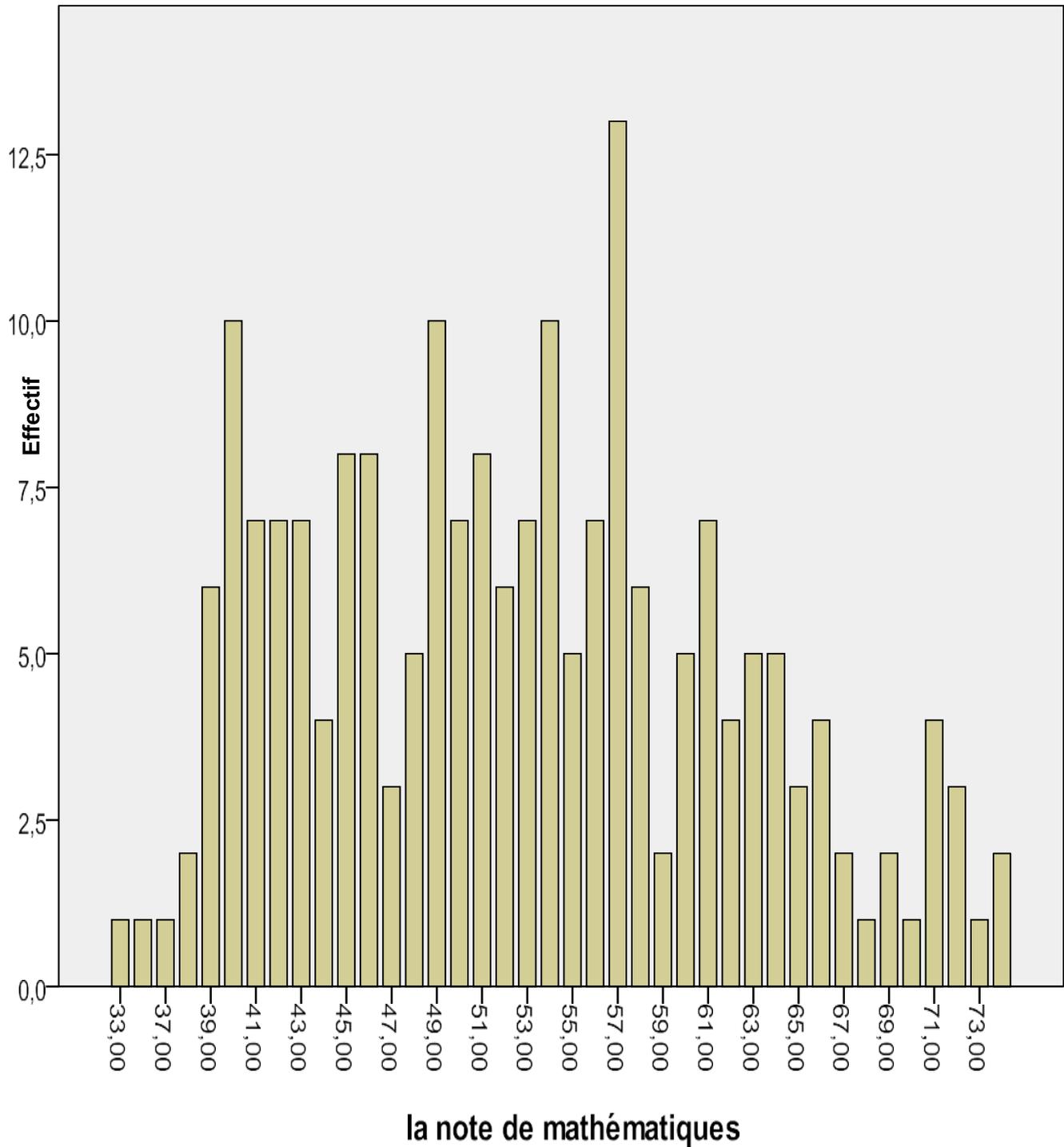
la note de mathématiques

N	Valide	200
	Manquante	0
Moyenne		52,6450
Erreur std. de la moyenne		,66245
Médiane		52,0000
Mode		57,00
Ecart-type		9,36845
Variance		87,768
Asymétrie		,287
Erreur std. d'asymétrie		,172
Aplatissement		-,649
Erreur std. d'aplatissement		,342
Intervalle		42,00
Minimum		33,00
Maximum		75,00
Somme		10529,00
Centiles	25	45,0000
	50	52,0000
	75	59,0000

### La note de mathématiques

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	33,00	1,0	0,5	0,5	0,5
	35,00	1,0	0,5	0,5	1,0
	37,00	1,0	0,5	0,5	1,5
	38,00	2,0	1,0	1,0	2,5
	39,00	6,0	3,0	3,0	5,5
	40,00	10,0	5,0	5,0	10,5
	41,00	7,0	3,5	3,5	14,0
	42,00	7,0	3,5	3,5	17,5
	43,00	7,0	3,5	3,5	21,0
	44,00	4,0	2,0	2,0	23,0
	45,00	8,0	4,0	4,0	27,0
	46,00	8,0	4,0	4,0	31,0
	47,00	3,0	1,5	1,5	32,5
	48,00	5,0	2,5	2,5	35,0
	49,00	10,0	5,0	5,0	40,0
	50,00	7,0	3,5	3,5	43,5
	51,00	8,0	4,0	4,0	47,5
	52,00	6,0	3,0	3,0	50,5
	53,00	7,0	3,5	3,5	54,0
	54,00	10,0	5,0	5,0	59,0
	55,00	5,0	2,5	2,5	61,5
	56,00	7,0	3,5	3,5	65,0
	57,00	13,0	6,5	6,5	71,5
	58,00	6,0	3,0	3,0	74,5
	59,00	2,0	1,0	1,0	75,5
	60,00	5,0	2,5	2,5	78,0
	61,00	7,0	3,5	3,5	81,5
	62,00	4,0	2,0	2,0	83,5
	63,00	5,0	2,5	2,5	86,0
	64,00	5,0	2,5	2,5	88,5
	65,00	3,0	1,5	1,5	90,0
	66,00	4,0	2,0	2,0	92,0
	67,00	2,0	1,0	1,0	93,0
68,00	1,0	0,5	0,5	93,5	
69,00	2,0	1,0	1,0	94,5	
70,00	1,0	0,5	0,5	95,0	
71,00	4,0	2,0	2,0	97,0	
72,00	3,0	1,5	1,5	98,5	
73,00	1,0	0,5	0,5	99,0	
75,00	2,0	1,0	1,0	100,0	
Total	200,0	100,0	100,0		

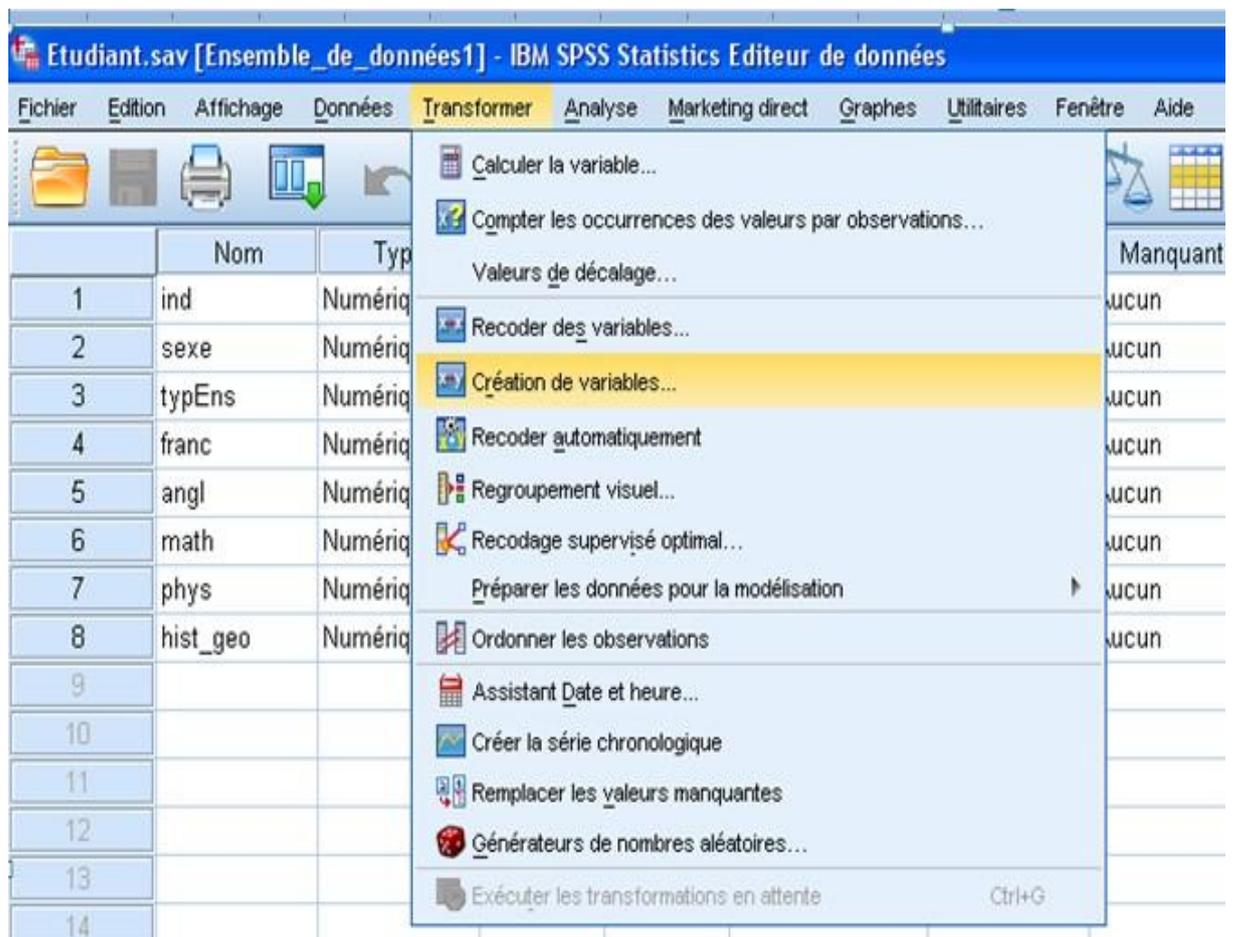
## la note de mathématiques



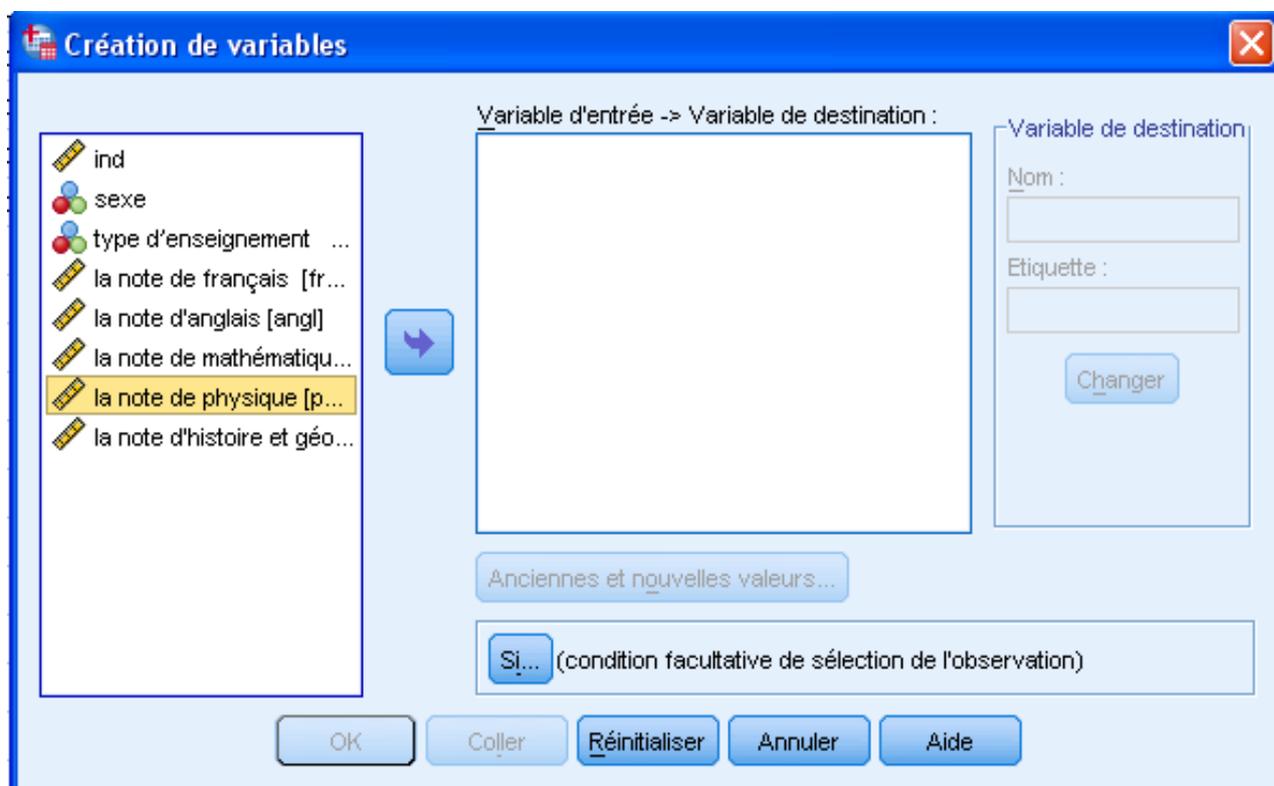
## 2- Cas d'une variable continue

Le but est de répartir les données d'une variable continue en classes. La commande est : **Transformer + Recorder + création des variables**.

**Pratique** : Considérons le fichier (**Etudiant.sav**). La variable continue qu'on peut choisir est **la note de physique**.



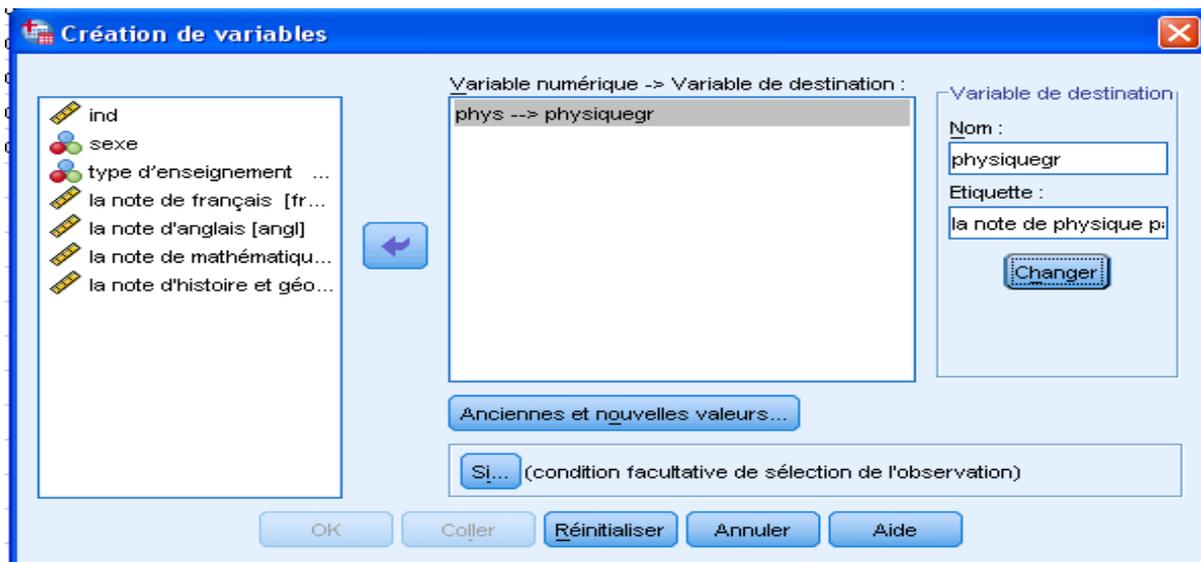
On obtient la boîte de dialogue suivante



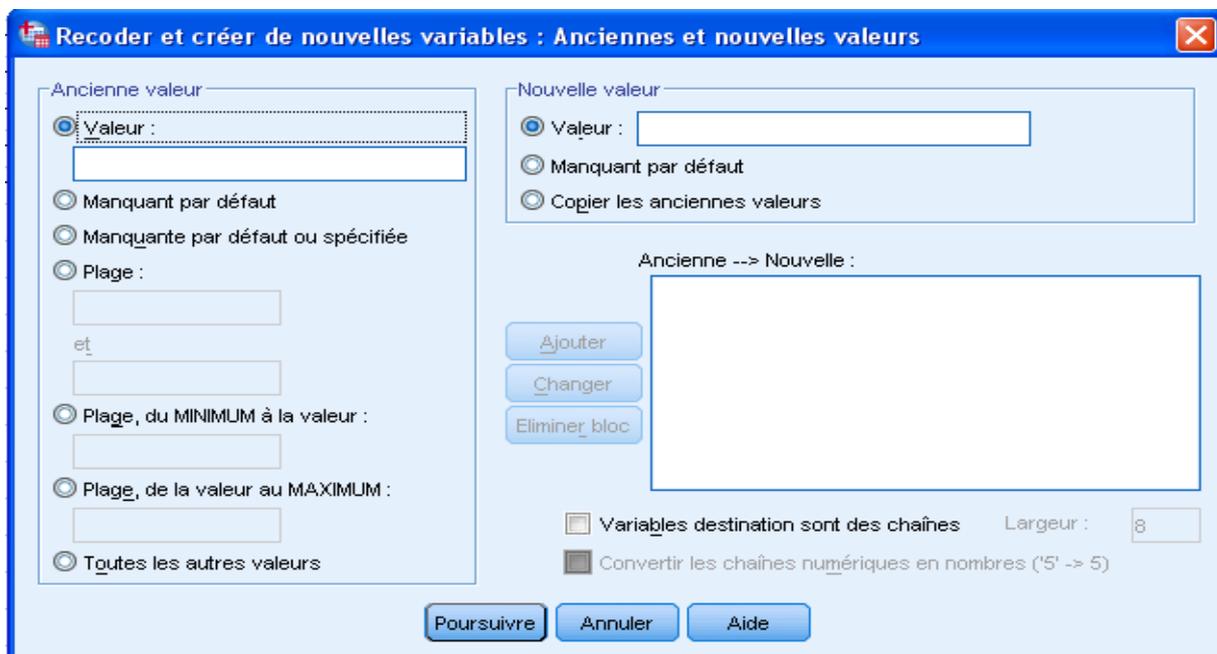
:

Sélectionner la variable **la note de physique**, puis la déplacer dans la zone sous Variable entrée->Variable, en cliquant sur le bouton flèche.

Dans la zone **Nom**, taper le nom de la nouvelle variable qui comportera les données groupées, on choisit le nom **physiquegr**, et taper sa signification dans la zone **Etiquette**, c-à-d **la note de physique par classes**. Sélectionner le bouton **Changer** pour activer la nouvelle variable :



Cliquer sur le bouton **Anciennes et nouvelles valeurs** afin de définir les classes. On obtient la boîte du dialogue :



A gauche, on choisit les bornes des classes : **plage**  
 A droite, on donne des codes aux classes : **Valeur**

0	20
1	

Recoder et créer de nouvelles variables : Anciennes et nouvelles valeurs

Ancienne valeur

Valeur :

Manquant par défaut

Manquante par défaut ou spécifiée

Plage :

21

et

40

Plage, du MINIMUM à la valeur :

Plage, de la valeur au MAXIMUM :

Toutes les autres valeurs

Nouvelle valeur

Valeur : 2

Manquant par défaut

Copier les anciennes valeurs

Ancienne --> Nouvelle :

0 thru 20 --> 1

Ajouter

Changer

Eliminer bloc

Variables destination sont des chaînes Largeur : 8

Convertir les chaînes numériques en nombres ('5' -> 5)

Poursuivre Annuler Aide

On choisit les classes suivantes : [0,20], [21,40], [41,60], [61, 80], [81, 100]

Les introduire chaque fois en cliquant sur **Ajouter**.

- A la fin ou aura la boite du dialogue :

The dialog box is titled "Recoder et créer de nouvelles variables : Anciennes et nouvelles valeurs". It is divided into two main sections: "Ancienne valeur" and "Nouvelle valeur".

**Ancienne valeur:**

- Valeur : [ ]
- Manquant par défaut
- Manquante par défaut ou spécifiée
- Plage : [ ] et [ ]
- Plage, du MINIMUM à la valeur : [ ]
- Plage, de la valeur au MAXIMUM : [ ]
- Toutes les autres valeurs

**Nouvelle valeur:**

- Valeur : [ ]
- Manquant par défaut
- Copier les anciennes valeurs

**Ancienne --> Nouvelle :**

[ Ajouter ]  
[ Changer ]  
[ Eliminer\_bloc ]

0 thru 20 --> 1  
21 thru 40 --> 2  
41 thru 60 --> 3  
61 thru 80 --> 4  
81 thru 100 --> 5

Variables destination sont des chaînes    Largeur : [ 8 ]  
 Convertir les chaînes numériques en nombres ('5' -> 5)

[ Poursuivre ]   [ Annuler ]   [ Aide ]

**Recoder et créer de nouvelles variables : Anciennes et nouvelles valeurs**

**Ancienne valeur**

- Valeur :
- Manquant par défaut
- Manquante par défaut ou spécifiée
- Plage :
- et
- Plage, du MINIMUM à la valeur :
- Plage, de la valeur au MAXIMUM :
- Toutes les autres valeurs

**Nouvelle valeur**

- Valeur :
- Manquant par défaut
- Copier les anciennes valeurs

Ancienne --> Nouvelle :

```

0 thru 20 --> 1
21 thru 40 --> 2
41 thru 60 --> 3
61 thru 80 --> 4
81 thru 100 --> 5

```

Variables destination sont des chaînes    Largeur :

Convertir les chaînes numériques en nombres ('5' -> 5)

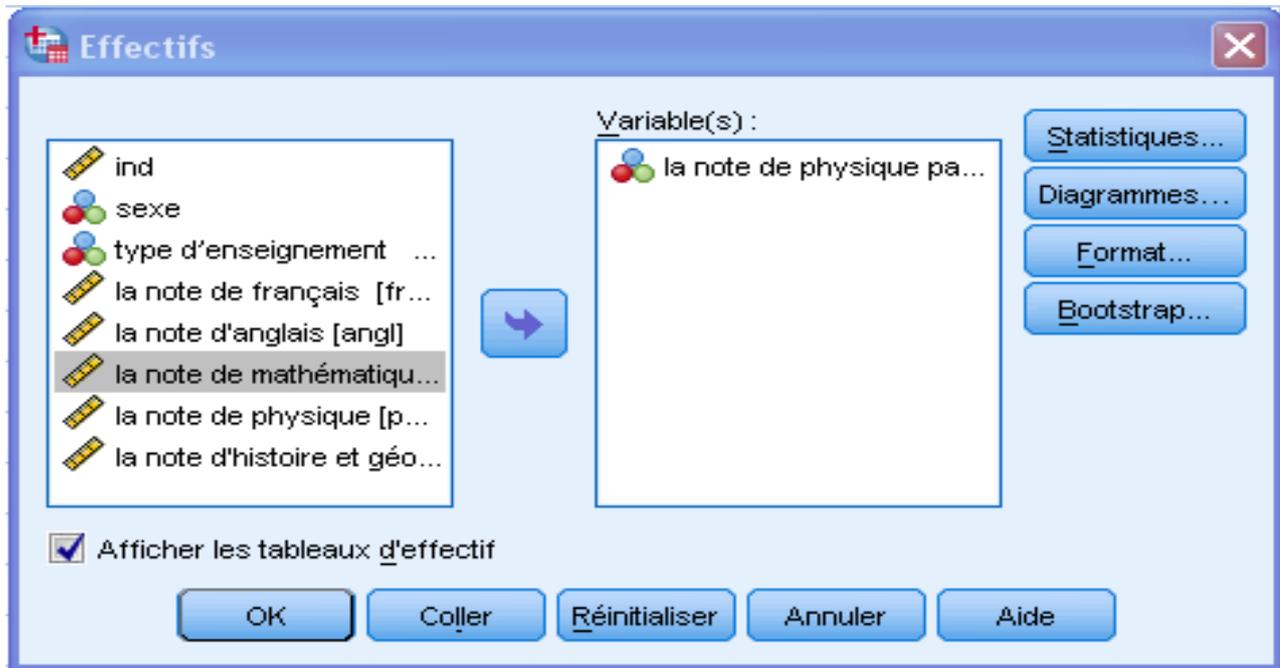
Cliquer sur le bouton **poursuivre**, puis sur **Ok** de la fenêtre **Recorder et créer de nouvelles des variables**.

Dans le fichier des données, on remarque que la variable **physiquegr** s'ajoute à la liste des variables.

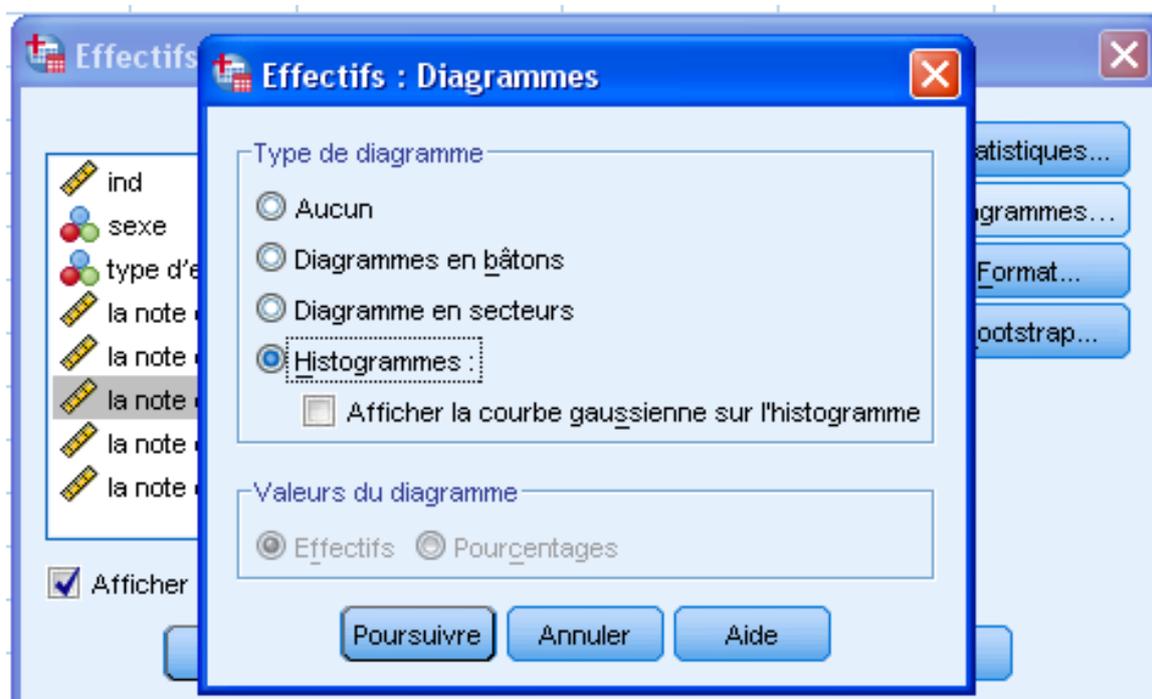
### **Distribution des fréquences et histogramme**

On utilise la commande : **Analyse + Statistiques descriptives + Fréquences**.

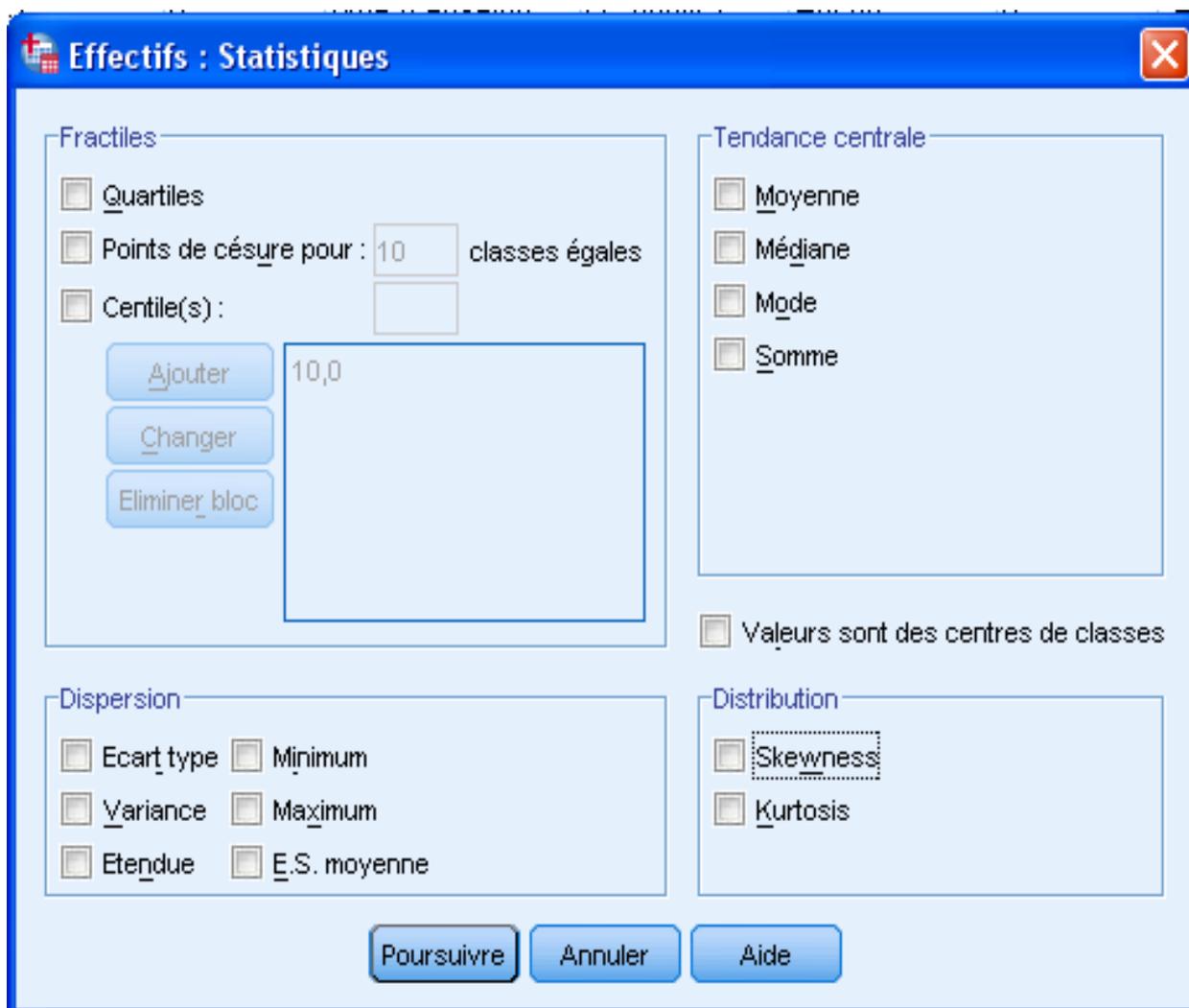
On déplace la variable **physiquegr**, pour avoir :



On clique sur le bouton **Diagrammes**, puis on sélectionne **Histogrammes**.



Cliquer sur **Poursuivre**. Désactiver toutes les commandes présentées dans la fenêtre **Statistiques**. Le calcul de ses statistiques sur les données de la variable **physiquegr** n'est pas pertinent.



On a le fichier des résultats suivant :

### Effectifs

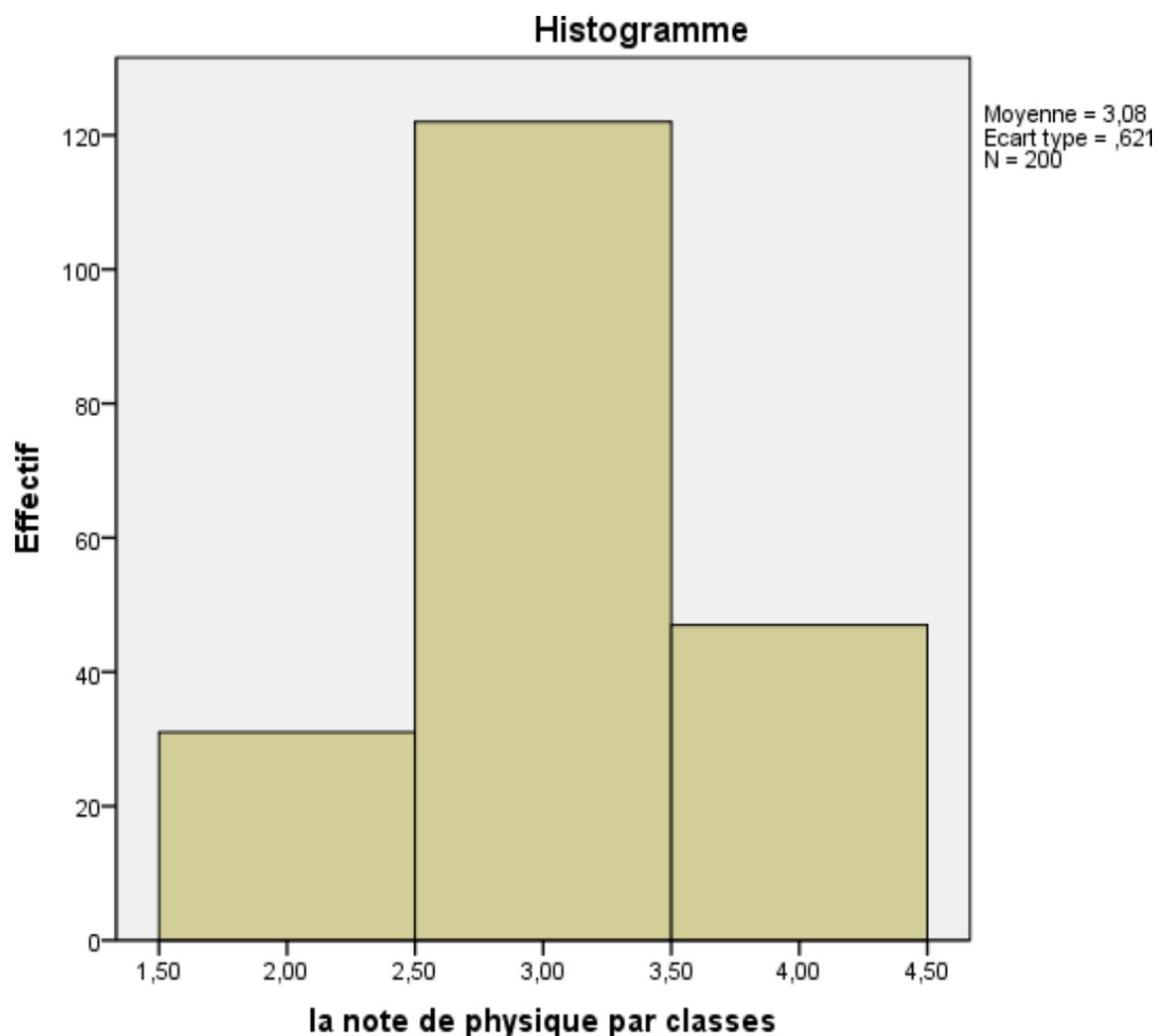
[Ensemble\_de\_données2] E:\ENSA2010\enseignement\GC\TP\Etudiant.sav

### Statistiques

### La note de physique par classes

N	Valide	200
	Manquante	0

	Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide 2,00	31	15,5	15,5	15,5
3,00	122	61,0	61,0	76,5
4,00	47	23,5	23,5	100,0
Total	200	100,0	100,0	



### 3- Cas d'un caractère qualitatif

Traitons l'exemple du fichier (**Etudiant.sav**). Considérons la variable type d'enseignement.

Utiliser la commande : **Analyse + Statistiques descriptives + Fréquences**.

Déplacer la variable à étudier (**type d'enseignement**), puis cliquer sur le bouton **Diagrammes**, et sélectionner la case **graphique en secteur**.



On a les résultats suivants :

## Effectifs

[Ensemble\_de\_données2] E:\ENSA2010\enseignement\GC\TP\Etudiant2.sav

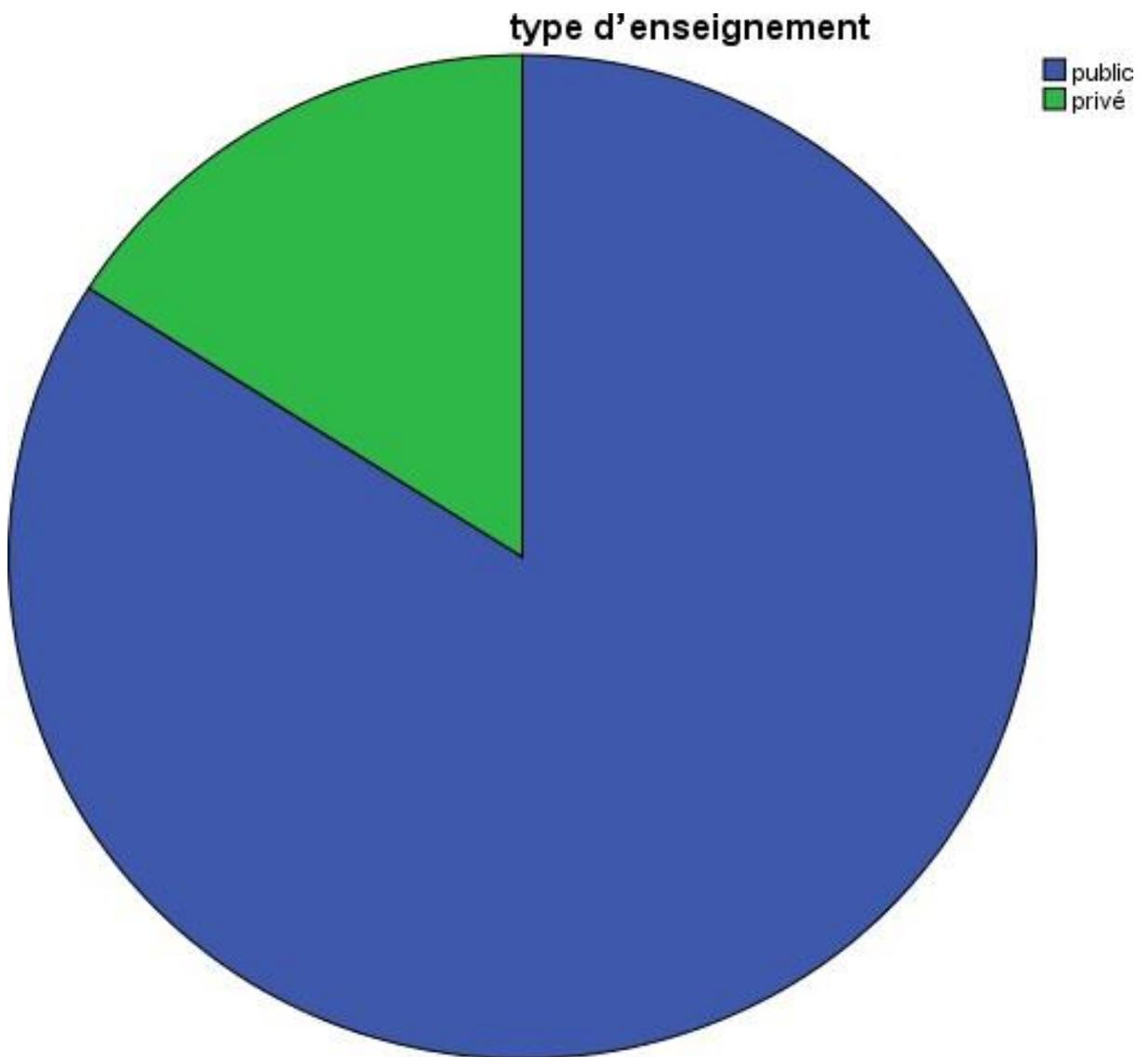
### Statistiques

type d'enseignement

N	Valide	200
	Manquante	0

### type d'enseignement

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
Valide	public	168	84,0	84,0	84,0
	privé	32	16,0	16,0	100,0
	Total	200	100,0	100,0	



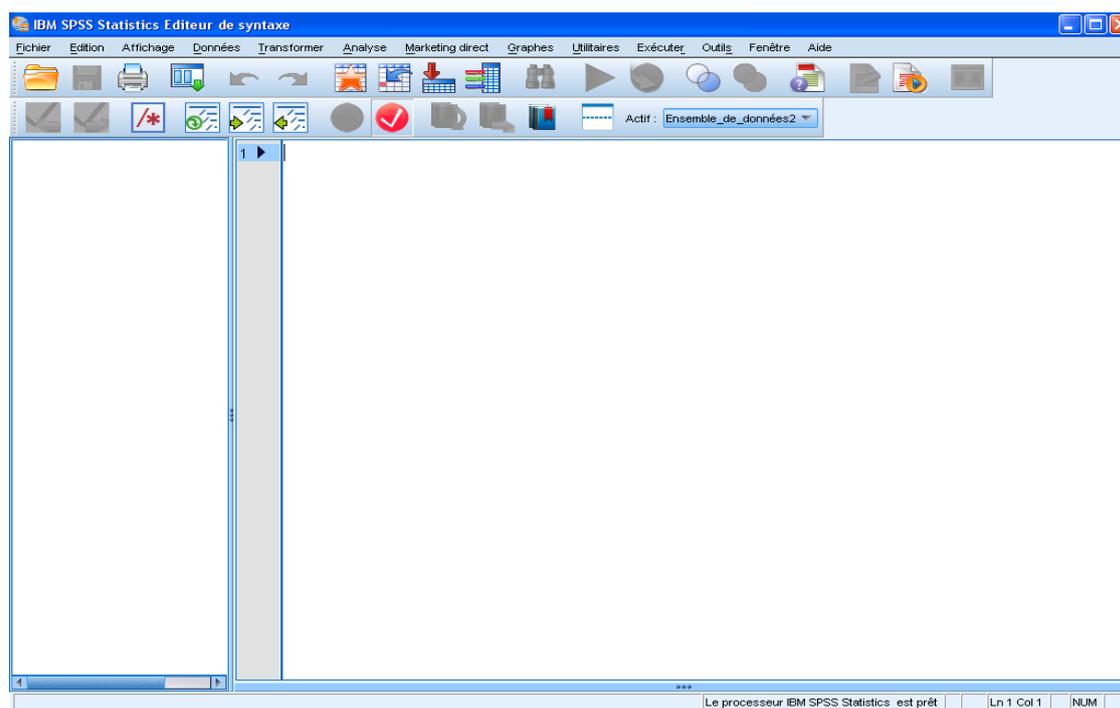
#### 4. Mode Windows et mode Syntaxe

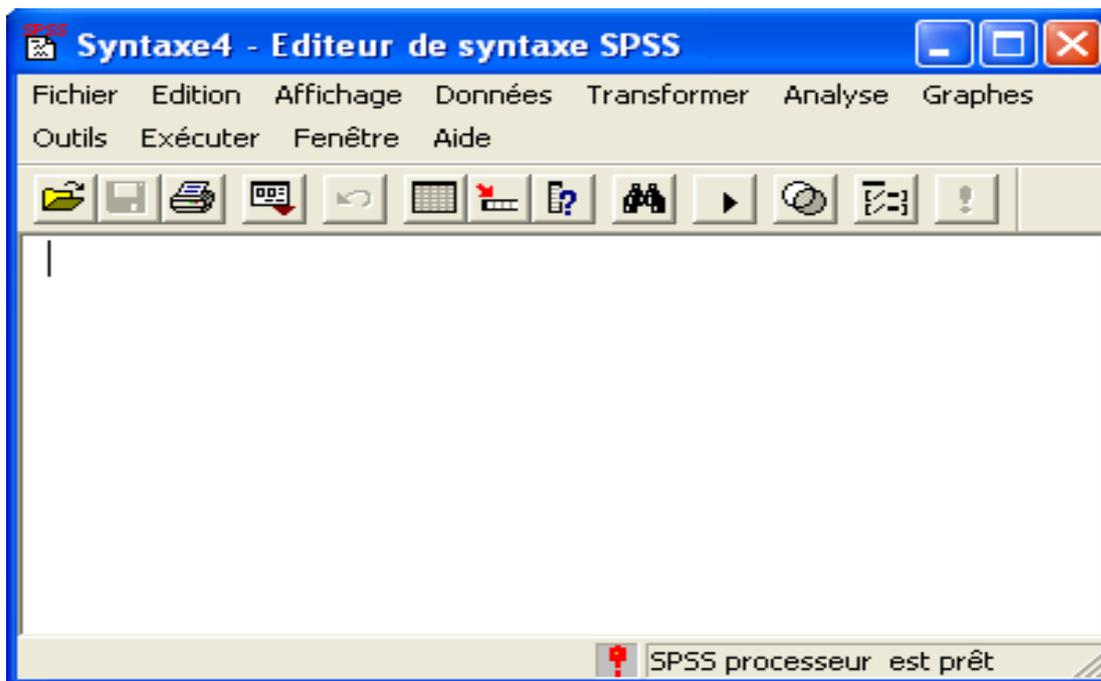
« Le mode Windows » utilise des menus descriptifs et des boîtes de dialogue simples pour effectuer les différents traitements du logiciel SPSS. Alors que « le mode syntaxe » utilise le style de commande :

Par exemple, ouvrir le fichier (**lemonde1995.sav**), et analyser la variable (**espvief**=espérance de vie des femmes) en utilisant le mode syntaxe.

Pour cela ou clique sur **Fichier + Nouveau + Syntaxe**, on aura la fenêtr suivante

:

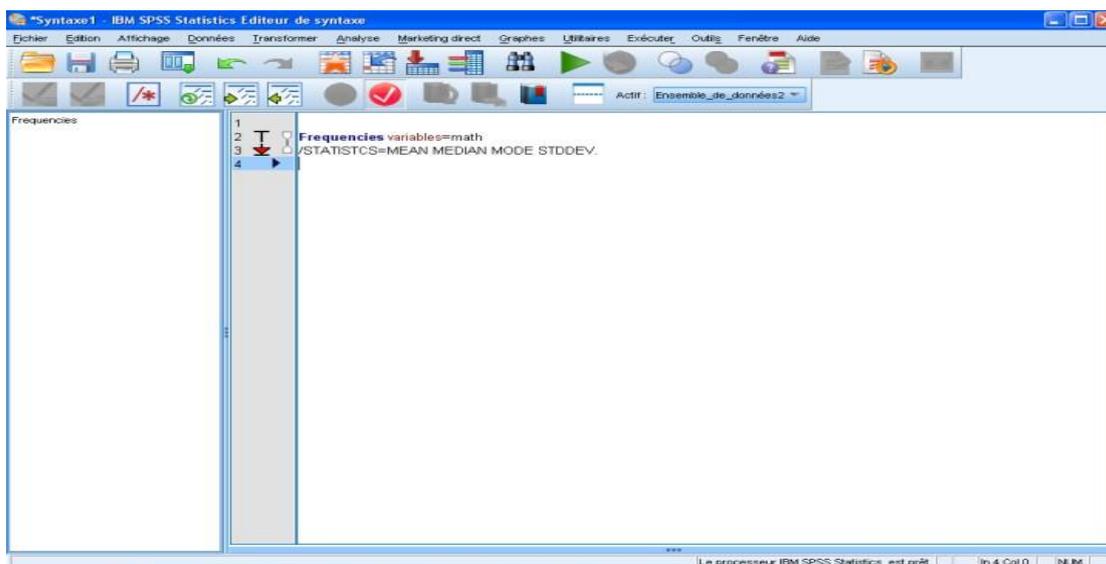




Pour déterminer la moyenne, la médiane, le mode et l'écart-type de la variable « **math**=la note de mathématiques», on utilise la commande :

**Frequencies variables=math**

**/STATISTICS=MEAN MEDIAN MODE STDDEV.**



Puis, on clique sur“Exécuter”.

On aura le tableau des résultats suivant :

## Effectifs

[Ensemble\_de\_données2] E:\ENSA2010\enseignement\GC\TP\Etudiant2.sav

### Statistiques

la note de mathématiques

<b>N</b>	<b>Valide</b>		200
	<b>Manquante</b>		0
<b>Moyenne</b>			52,6450
<b>Médiane</b>			52,0000
<b>Mode</b>			57,00
<b>Ecart-type</b>			9,36845

## La note de mathématiques

		Effectifs	Pourcentage	Pourcentage valide	Pourcentage cumulé
	33,00	1,0	0,5	0,5	0,5
	35,00	1,0	0,5	0,5	1,0
	37,00	1,0	0,5	0,5	1,5
	38,00	2,0	1,0	1,0	2,5
	39,00	6,0	3,0	3,0	5,5
	40,00	10,0	5,0	5,0	10,5
	41,00	7,0	3,5	3,5	14,0
	42,00	7,0	3,5	3,5	17,5
	43,00	7,0	3,5	3,5	21,0
	44,00	4,0	2,0	2,0	23,0
	45,00	8,0	4,0	4,0	27,0
	46,00	8,0	4,0	4,0	31,0
	47,00	3,0	1,5	1,5	32,5
	48,00	5,0	2,5	2,5	35,0
	49,00	10,0	5,0	5,0	40,0
	50,00	7,0	3,5	3,5	43,5
	51,00	8,0	4,0	4,0	47,5
	52,00	6,0	3,0	3,0	50,5
	53,00	7,0	3,5	3,5	54,0
	54,00	10,0	5,0	5,0	59,0
<b>Valide</b>	55,00	5,0	2,5	2,5	61,5
	56,00	7,0	3,5	3,5	65,0
	57,00	13,0	6,5	6,5	71,5
	58,00	6,0	3,0	3,0	74,5
	59,00	2,0	1,0	1,0	75,5
	60,00	5,0	2,5	2,5	78,0
	61,00	7,0	3,5	3,5	81,5
	62,00	4,0	2,0	2,0	83,5
	63,00	5,0	2,5	2,5	86,0
	64,00	5,0	2,5	2,5	88,5
	65,00	3,0	1,5	1,5	90,0
	66,00	4,0	2,0	2,0	92,0
	67,00	2,0	1,0	1,0	93,0
	68,00	1,0	0,5	0,5	93,5
	69,00	2,0	1,0	1,0	94,5
	70,00	1,0	0,5	0,5	95,0
	71,00	4,0	2,0	2,0	97,0
	72,00	3,0	1,5	1,5	98,5
	73,00	1,0	0,5	0,5	99,0
	75,00	2,0	1,0	1,0	100,0
	Total	200,0	100,0	100,0	

**Exercice** : ici les notes des étudiants masculins et des étudiantes féminins, obtenues dans un test noté sur 100 :

**Masculins** : 30 50 95 80 45 25 30 33 44 52 55 65 75 85 80 78 87 84 48 52 55 57  
40 60 75 72 77 90 88 95 75 73 77

**Féminin** 85 45 84 38 79 75 77 44 65 40 44 30 25 30 77 85 95 61 60 48 36 95  
98 65 66 77 42 44 62 76 75 93 98 80 70

- 6- Saisir les deux variables en leur donnant les noms (**note\_mas**) et (**note\_fem**).
- 7- Donner les caractéristiques statistiques (par les deux modes Windows et Syntaxe) et les représentations graphiques des deux variables.

## TP d'analyse en composantes principales sous SPSS

### Sujet

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

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

### Première partie : ACP simple

Dans l'ACP simple (i.e. sur données centrées), la matrice de covariance constitue la matrice dont on va extraire les valeurs propres et les vecteurs propres.

Aller dans **Analyse > Réduction des dimensions > Analyse factorielle**.

On choisit les variables et on les sélectionne puis on clique sur la flèche qui pointe vers la droite.

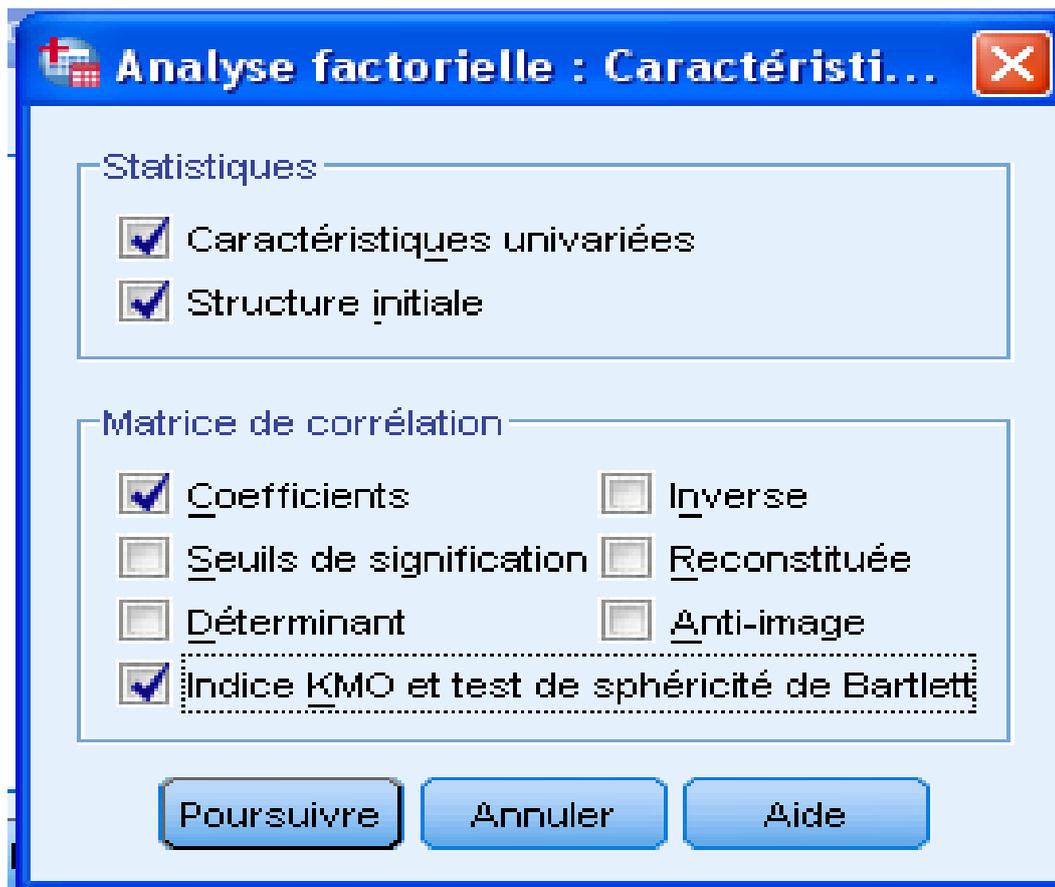


Cinq boîtes de dialogue d'options : Caractéristiques ; Extraction ; Rotation ; Facteurs ; Options.

#### 1. Descriptives

La boîte de dialogue «Analyse factorielle : Caractéristiques » apparaît.

Dans « **Statistiques** », cliquer sur « Caractéristiques univariées » et sur « Structure initiale ». Dans « **Matrice des Corrélations** », cliquer sur « Coefficients » et « Indice KMO et test de Bartlett ».

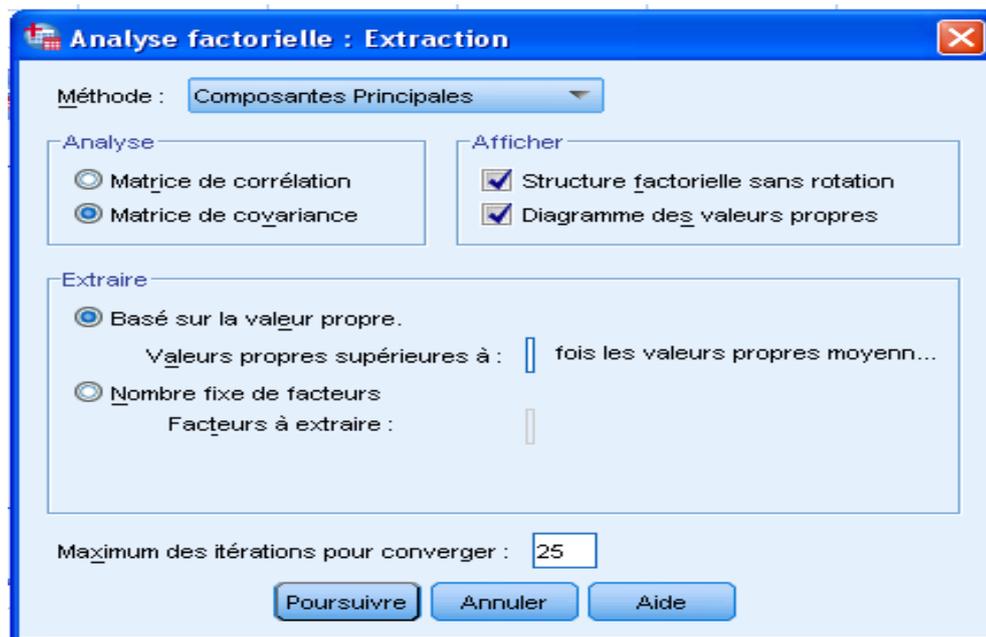


## 2. « Extraction... »

La boîte de dialogue «Analyse factorielle : Extraction » apparaît.

Cliquer sur «Graphique des valeurs propres » et sur « Structure factorielle sans rotation ». Cliquer sur « matrice de covariance » ; c'est l'ACP simple.

Si on veut utiliser une ACP normée, on doit cliquer sur « matrice de corrélation ».



### 3. « Rotation... »

La boîte de dialogue «Analyse factorielle: Rotation » apparaît.

Cocher l'option « Carte(s) factorielle(s) ». Cette option permet d'avoir une représentation des différents axes.



#### 4. Facteurs

La boîte de dialogue «Analyse factorielle: Facteurs » apparaît.

Cliquer sur « Enregistrer dans des variables » (méthode Régression) et sur « Afficher la matrice des coefficients factoriels ».



#### 5. « Options... »

La boîte de dialogue «Analyse factorielle: Options » apparaît.

Choisir l'option « Classement des variables par taille » dans Affichage des projections.



### **Analyse des résultats**

#### **1. Statistiques descriptives**

##### **Statistiques descriptives**

<b>Statistiques descriptives</b>			
	Moyenne	Ecart-type	n analyse
PVP	12,213	2,2383	24
AGR	1,996	1,6812	24
CMI	3,979	4,5507	24
TRA	8,321	2,5209	24
LOG	4,000	4,2424	24
EDU	9,942	5,3356	24
ACS	4,817	3,4821	24
ANC	4,275	4,2442	24

DEF	30,258	7,4667	24
DET	19,142	12,4560	24
DIV	1,183	1,0478	24

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

## 2. Test de Bartlett

### Indice KMO et test de Bartlett(a)

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

a. Basé sur les corrélations

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

### 3. Valeur propre et variance totale expliquée

#### Variance totale expliquée

Composante		Valeurs propres initiales <sup>a</sup>			Extraction Sommes des carrés des facteurs retenus		
		Total	% de la variance	cumulés %	Total	% de la variance	cumulés %
Non normée	1	213,162	65,891	65,891	213,162	65,891	65,891
	2	58,245	18,004	83,895	58,245	18,004	83,895
	3	22,771	7,039	90,934			
	4	17,065	5,275	96,209			
	5	5,757	1,78	97,989			
	6	3,362	1,039	99,028			
	7	1,163	0,36	99,388			
	8	1,082	0,335	99,722			
	9	0,646	0,2	99,922			
	10	0,249	0,077	99,999			
	11	0,004	0,001	100			
Redimensionné	1	213,162	65,891	65,891	4,689	42,624	42,624
	2	58,245	18,004	83,895			
	3	22,771	7,039	90,934			
	4	17,065	5,275	96,209			
	5	5,757	1,78	97,989			
	6	3,362	1,039	99,028			
	7	1,163	0,36	99,388	1,462	13,29	55,914
	8	1,082	0,335	99,722			
	9	0,646	0,2	99,922			
	10	0,249	0,077	99,999			
	11	0,004	0,001	100			

Méthode d'extraction : Analyse en composantes principales.

a. Lors de l'analyse d'une matrice de covariance, les valeurs propres initiales des solutions directe et centrée sont les mêmes.

Les valeurs propres sont données dans la colonne « **total** » ; elles représentent l'inertie portées par les axes principaux.

L'inertie totale est égale à la somme des valeurs propres ; ainsi  $I=213,586+\dots+0,000=324,025$ .

La contribution de chaque axe principal à l'inertie totale est donnée par le pourcentage d'inertie, qui est le rapport entre la valeur propre et l'inertie totale, soit  $\lambda/I$ .

La plus grande valeur propre 213,586 est associée à l'axe principale D1. Cet axe explique 65,917 de la variabilité. De même, la valeur propre 58,252 est associée à l'axe principal D2 qui explique 17,978 de la variance. Le plan constitué des deux axes D1 et D2 explique 83,895 de la variabilité.

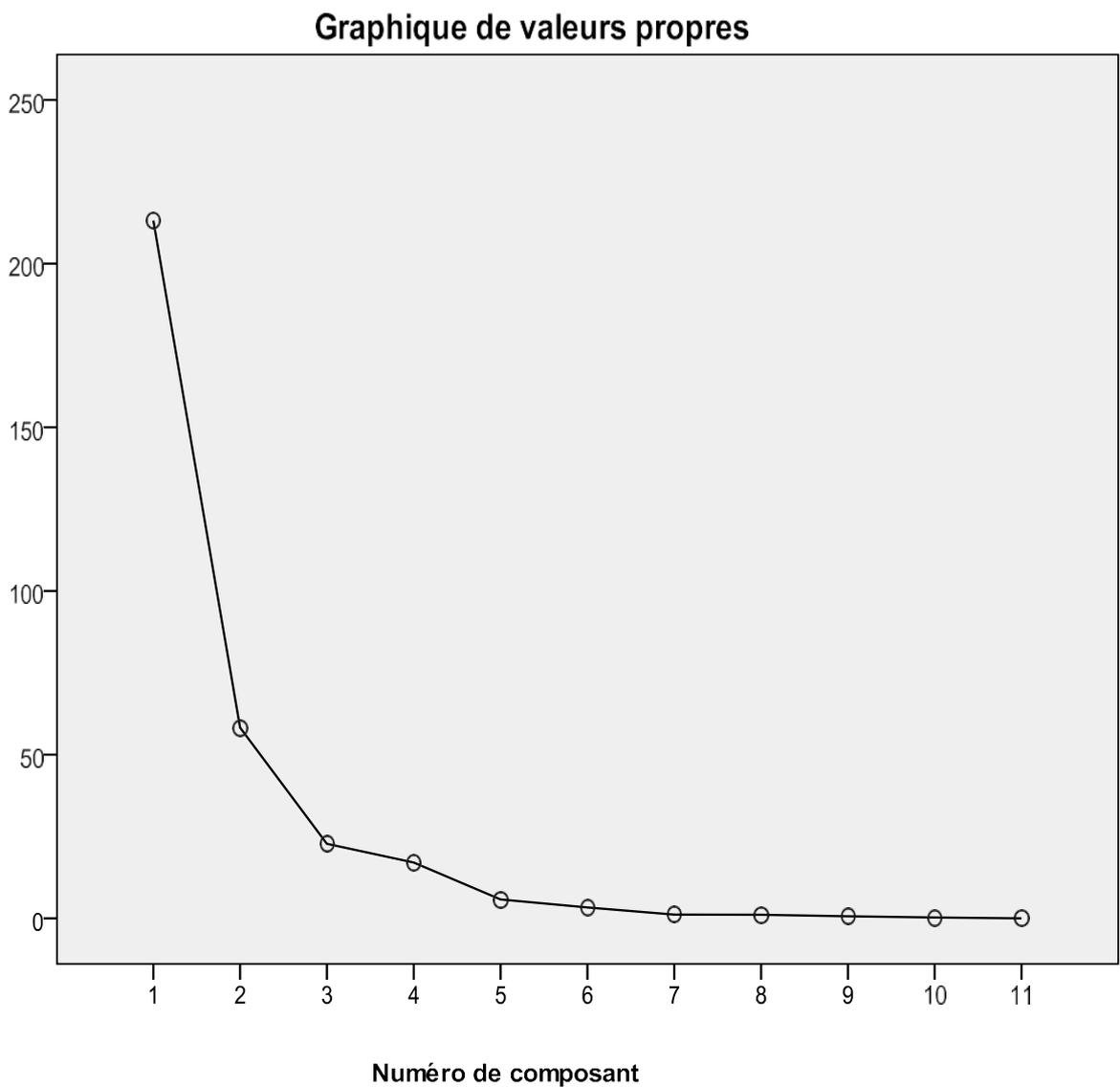
#### 4 . Nombres d'axes principaux à retenir

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

- Première règle : on choisit le nombre d'axe en fonction de la restitution minimale d'information que l'on souhaite. Par exemple, on veut que le modèle restitue au moins 80% de l'information.
- Deuxième règle : On observe le graphique des valeurs propres et on ne retient que les valeurs qui se trouvent à gauche du point d'inflexion. Graphiquement, on part des composants qui apportent le moins

d'information (qui se trouvent à droite), on relie par une droite les points presque alignés et on ne retient que les axes qui sont au-dessus de cette ligne.

-



Dans notre exemple, on ne retient que les deux premiers axes qui permettent de prendre en compte environ 83,895 de l'inertie totale.

#### 4. Qualité de représentation

##### Qualité de représentation

Qualité de représentation

	Non normée		Redimensionné	
	Initial	Extraction	Initial	Extraction
PVP	5,010	,303	1,000	,061
AGR	2,827	1,648	1,000	,583
CMI	20,709	15,482	1,000	,748
TRA	6,355	,822	1,000	,129
LOG	17,998	10,978	1,000	,610
EDU	28,469	18,029	1,000	,633
ACS	12,125	9,995	1,000	,824
ANC	18,013	5,650	1,000	,314
DEF	55,752	53,513	1,000	,960
DET	155,151	154,665	1,000	,997
DIV	1,098	,320	1,000	,292

Méthode d'extraction : Analyse en composantes principales.

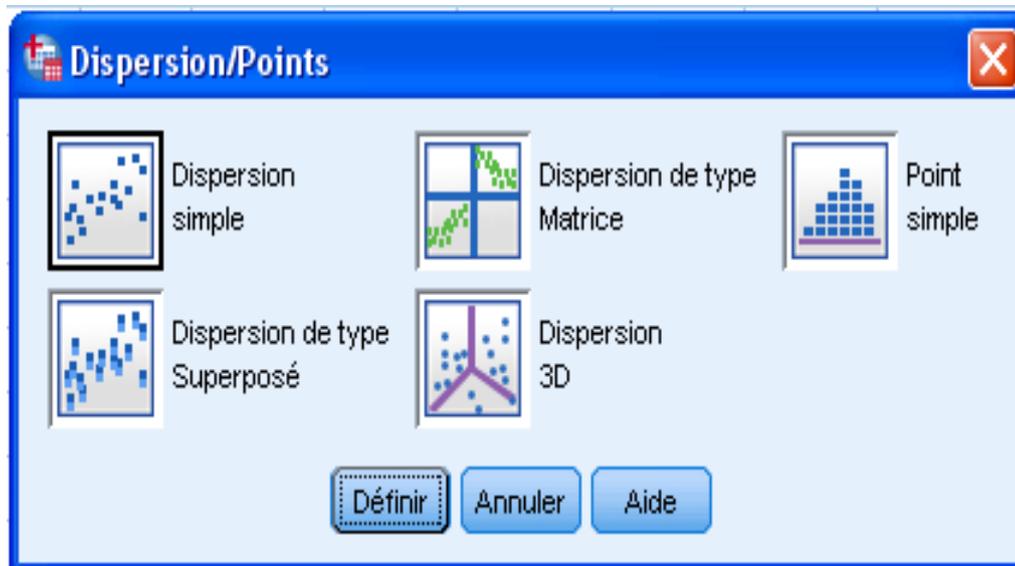
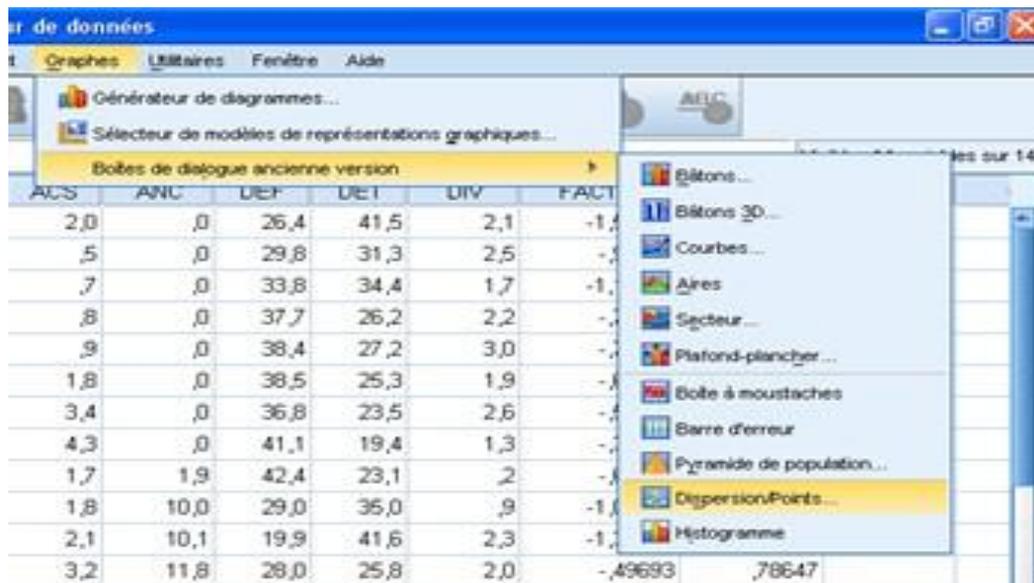
La Qualité de représentation permet de voir dans quelle mesure les variables initiales sont-elles prises en compte par les nouvelles variables extraites (i.e. les composantes principales). Par exemple, la qualité de représentation de la variable « **det** » est 0.997 ; cela veut dire que 99.7% de la variance de la variable est prise en compte par l'un des deux axes principaux. Par contre la variable « **pvp** » est mal représentée par les deux axes, et donc sera mal prise en compte par les nouvelles variables.

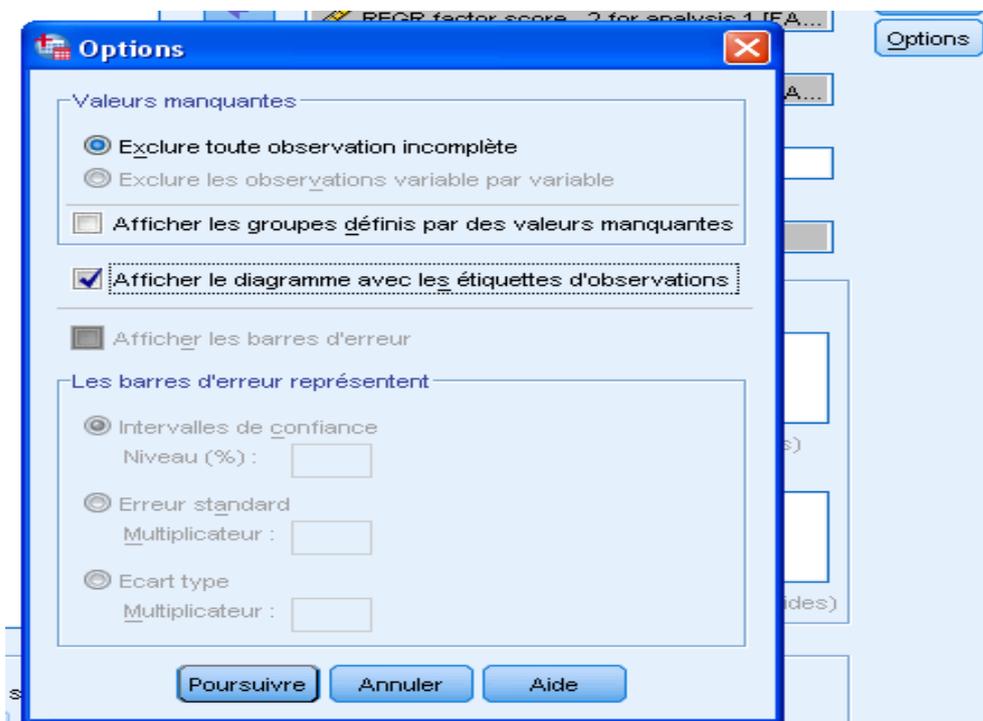
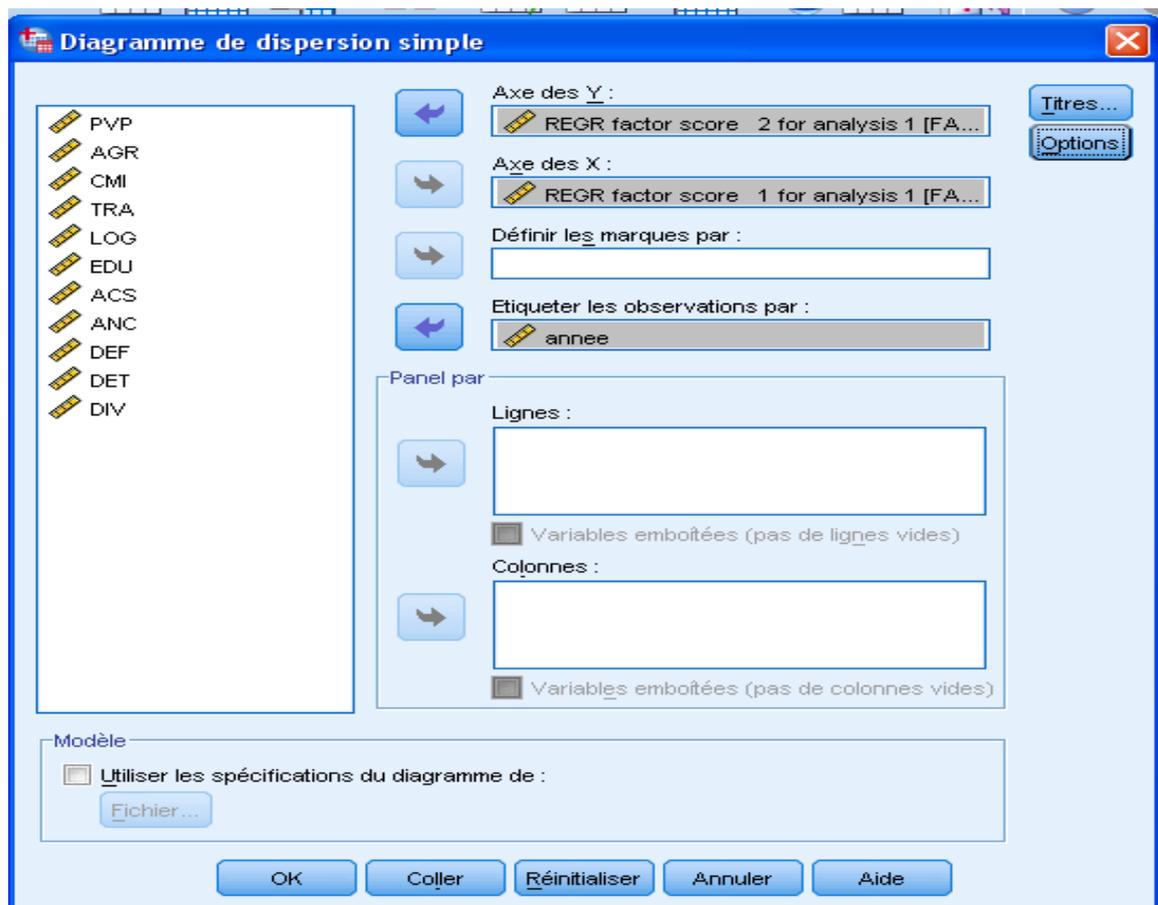
### 5. Coordonnées des individus dans les axes principaux

Dans l'éditeur des données, on retrouve les coordonnées des individus dans le premier axe principal et dans le deuxième axe principal.

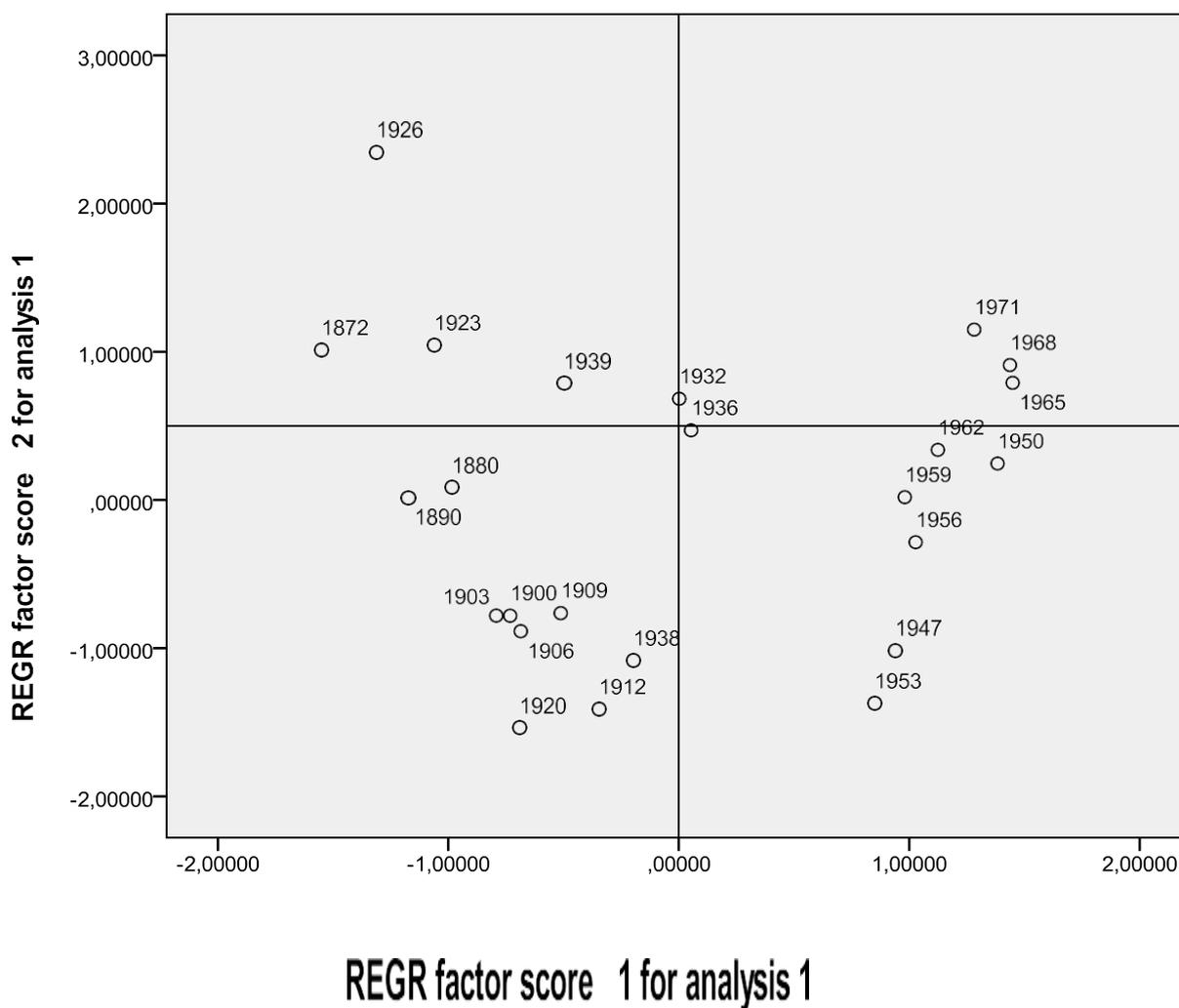
	annee	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV	FAC1_1	FAC2_1
1	1872	18,0	,5	,1	6,7	,5	2,1	2,0	,0	26,4	41,5	2,1	-1,55063	1,01329
2	1880	14,1	,8	,1	15,3	1,9	3,7	,5	,0	29,8	31,3	2,5	-,98450	,08468
3	1890	13,6	,7	,7	6,8	,6	7,1	,7	,0	33,8	34,4	1,7	-1,17421	,01662
4	1900	14,3	1,7	1,7	6,9	1,2	7,4	,8	,0	37,7	26,2	2,2	-,73179	-,78309
5	1903	10,3	1,5	1,4	9,3	,6	8,5	,9	,0	38,4	27,2	3,0	-,79146	-,78240
6	1906	13,4	1,4	,5	8,1	,7	8,6	1,8	,0	38,5	25,3	1,9	-,68606	-,88562
7	1909	13,5	1,1	,5	9,0	1,6	9,0	3,4	,0	36,8	23,5	2,6	-,51438	-,76615

Pour la représentation graphique, on clique sur « **Graphes + Diagramme de dispersion +simple** ».





En cliquant deux fois sur le graphe et en ajoutant les deux axes des **X** et **Y** dans le graphe, on ale graphe suivant :



## 6. Coordonnées des variables

Les coordonnées des variables du tableau sont données par la matrice des corrélations variables/facteurs.

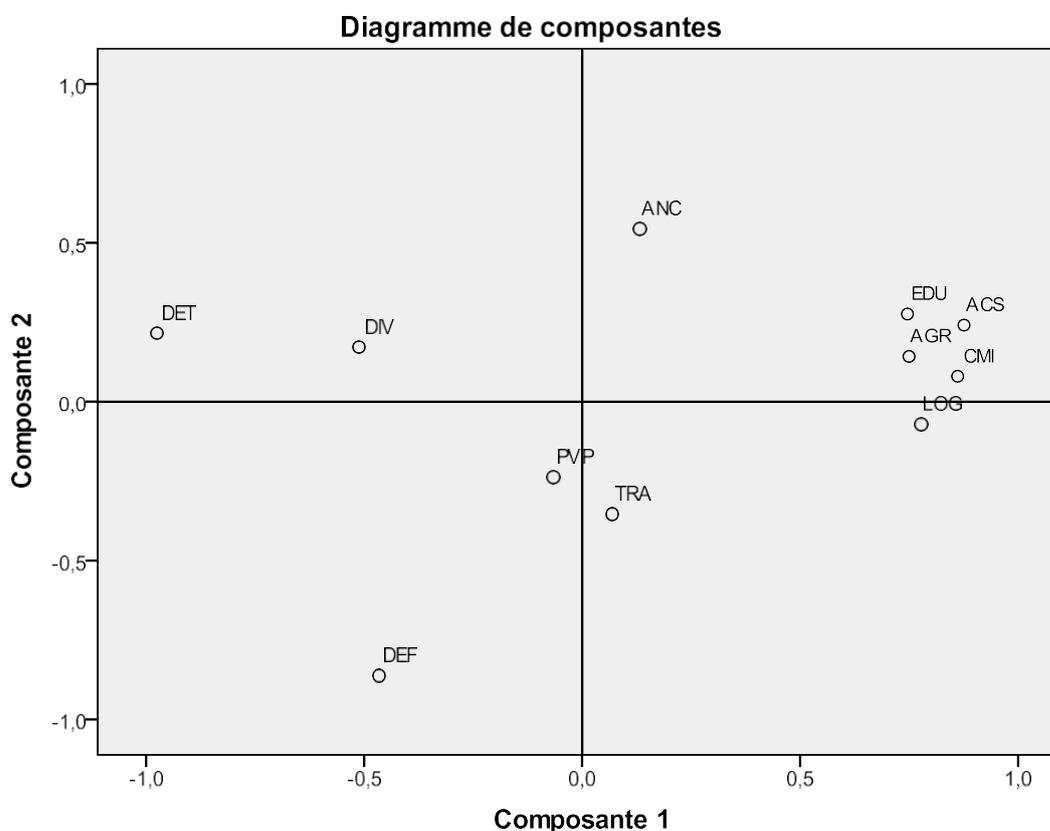
## Matrice des composantes<sup>a</sup>

Méthode d'extraction : Analyse en composantes principales.

a. 2 composantes extraites.

	Non normée		Redimensionné	
	Composante		Composante	
	1	2	1	2
<b>DET</b>	-12,142	2,69	-0,975	0,216
<b>ACS</b>	3,048	0,838	0,875	0,241
<b>CMI</b>	3,918	0,366	0,861	0,08
<b>LOG</b>	3,3	-0,3	0,778	-0,071
<b>AGR</b>	1,261	0,24	0,75	0,143
<b>EDU</b>	3,983	1,471	0,747	0,276
<b>DIV</b>	-0,537	0,18	-0,512	0,172
<b>DEF</b>	-3,478	-6,436	-0,466	-0,862
<b>ANC</b>	0,561	2,31	0,132	0,544
<b>TRA</b>	0,174	-0,89	0,069	-0,353
<b>PVP</b>	-0,147	-0,531	-0,066	-0,237

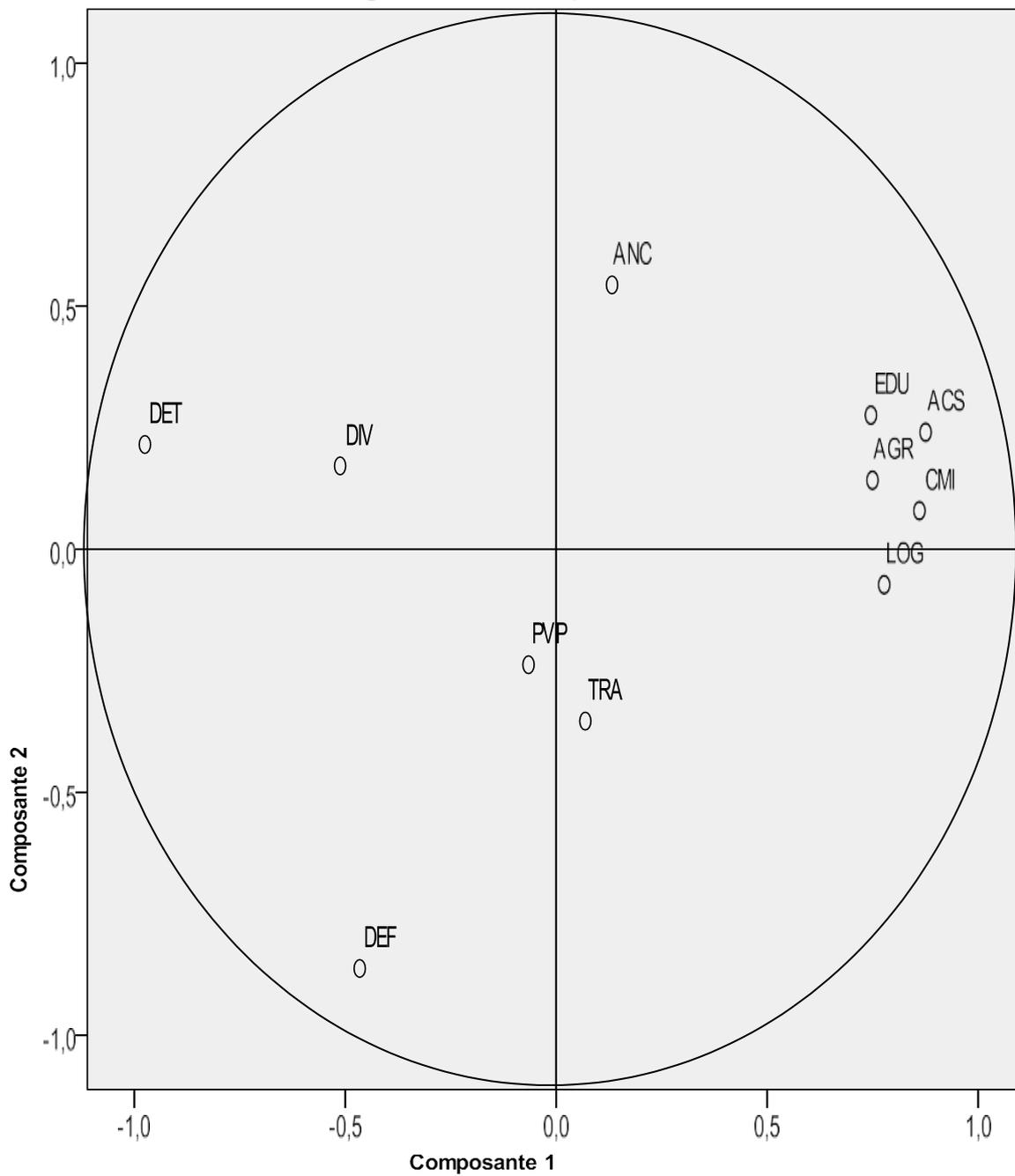
## Représentations graphiques



Il y a de fortes corrélations entre le premier axe et la Dette (négative), le Commerce, le Logement et l'Agriculture (positives). L'axe  $\phi_2$  est lui très lié à la variable DEF.

La proximité des variables EDU, ACS, AGR, CMI et LOG sur le graphe permet de conclure qu'il existe un lien fort entre ces variables mais cette interprétation ne vaut que parce ces variables sont bien représentées dans le plan 1-2 (les points correspondants sont proches du cercle). La proximité des variables PVP et TRA ne peut pas être interprétée : ces variables sont mal représentées puisque les points correspondants sont éloignés du cercle.

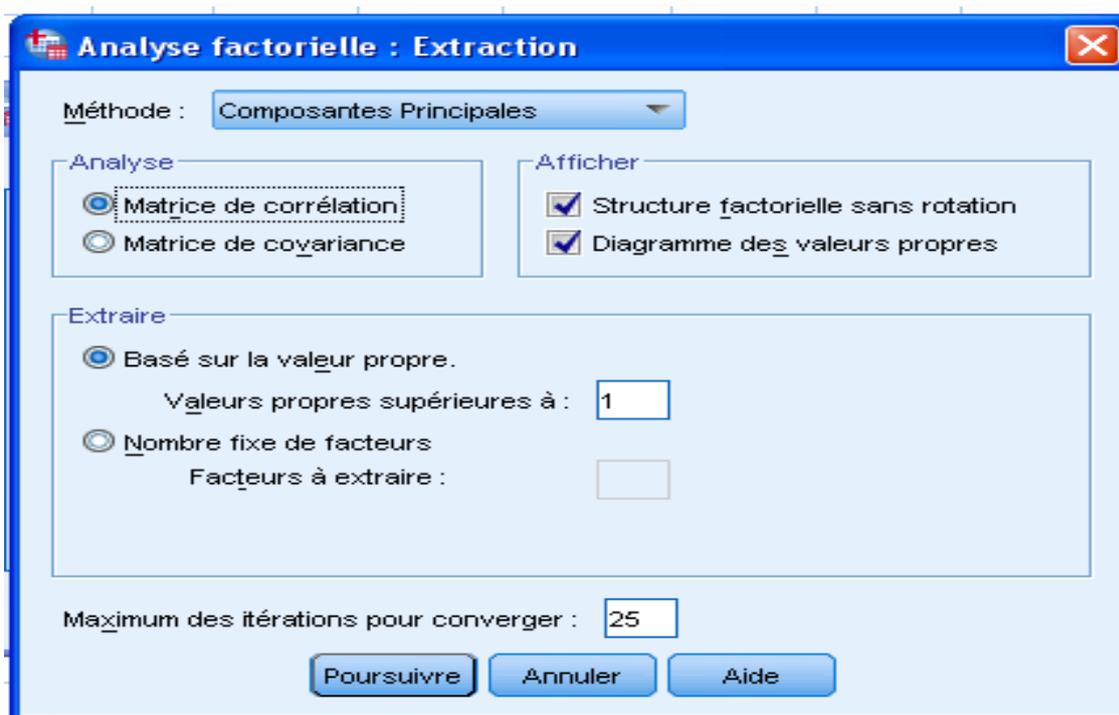
### Diagramme de composantes



**CHAPITRE 3 : ACP normée**

Dans l'ACP normée (i.e. sur données centrées réduites), la matrice des corrélations sera utilisée pour l'extraction des valeurs propres et vecteurs propres.

Ainsi, on doit cliquer sur « **Matrice de corrélation** » pour faire l'analyse.



Les résultats de l'ACP normée sont donnés comme suit :

## 1. Variance totale expliquée

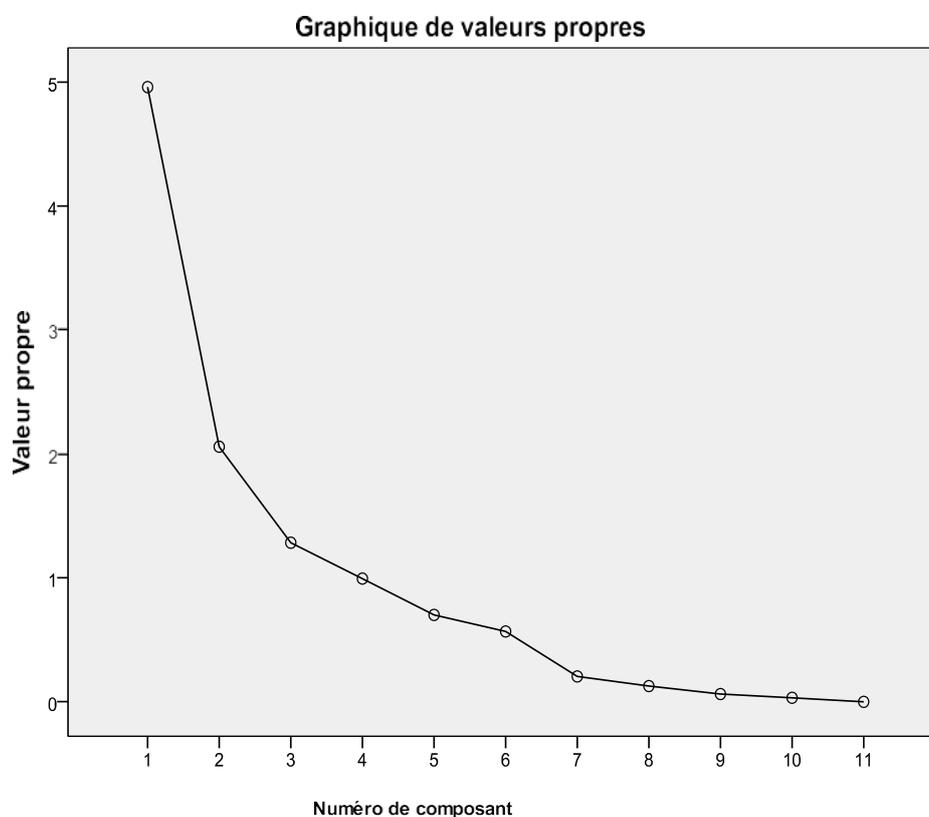
Méthode d'extraction : Analyse en composantes principales.

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

La plus grande valeur propre de la matrice de corrélation est 4.973 ; elle est associée au premier axe principal qui explique 45,212 de la variabilité.

On choisit les trois premiers axes principaux qui expliquent 75,575% de la variance. Ce choix se voit clairement dans le Graphique des valeurs propres.

## 2. Graphique des valeurs propres



### 3. Matrice des composantes

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

**Matrice des composantes<sup>a</sup>**

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

Méthode d'extraction : Analyse en composantes principales.

a. 3 composantes extraites

**Matrice des composantes(a)**

	Composante		
	1	2	3
<b>acs</b>	<b>,933</b>	-,100	,166
<b>det</b>	-,889	-,301	,161
<b>cmi</b>	<b>,834</b>	,340	-,141
<b>agr</b>	<b>,819</b>	,006	,366
<b>edu</b>	,787	-,137	,425
<b>log</b>	,722	,398	-,385
<b>def</b>	-,612	,216	-,260
<b>div</b>	-,548	,112	,537
<b>aco</b>	,289	<b>-,808</b>	-,375
<b>pvp</b>	-,174	<b>,740</b>	,341
<b>tra</b>	-,137	,631	-,376

Méthode d'extraction : Analyse en composantes principales.  
a 3 composantes extraites.

La première composante principale est corrélée positivement avec les variables action sociale, commerce-industrie, agriculture; en revanche elle est négativement corrélée avec les variables défense et remboursement de la dette. Cette opposition explique déjà près de la moitié, de la variance.

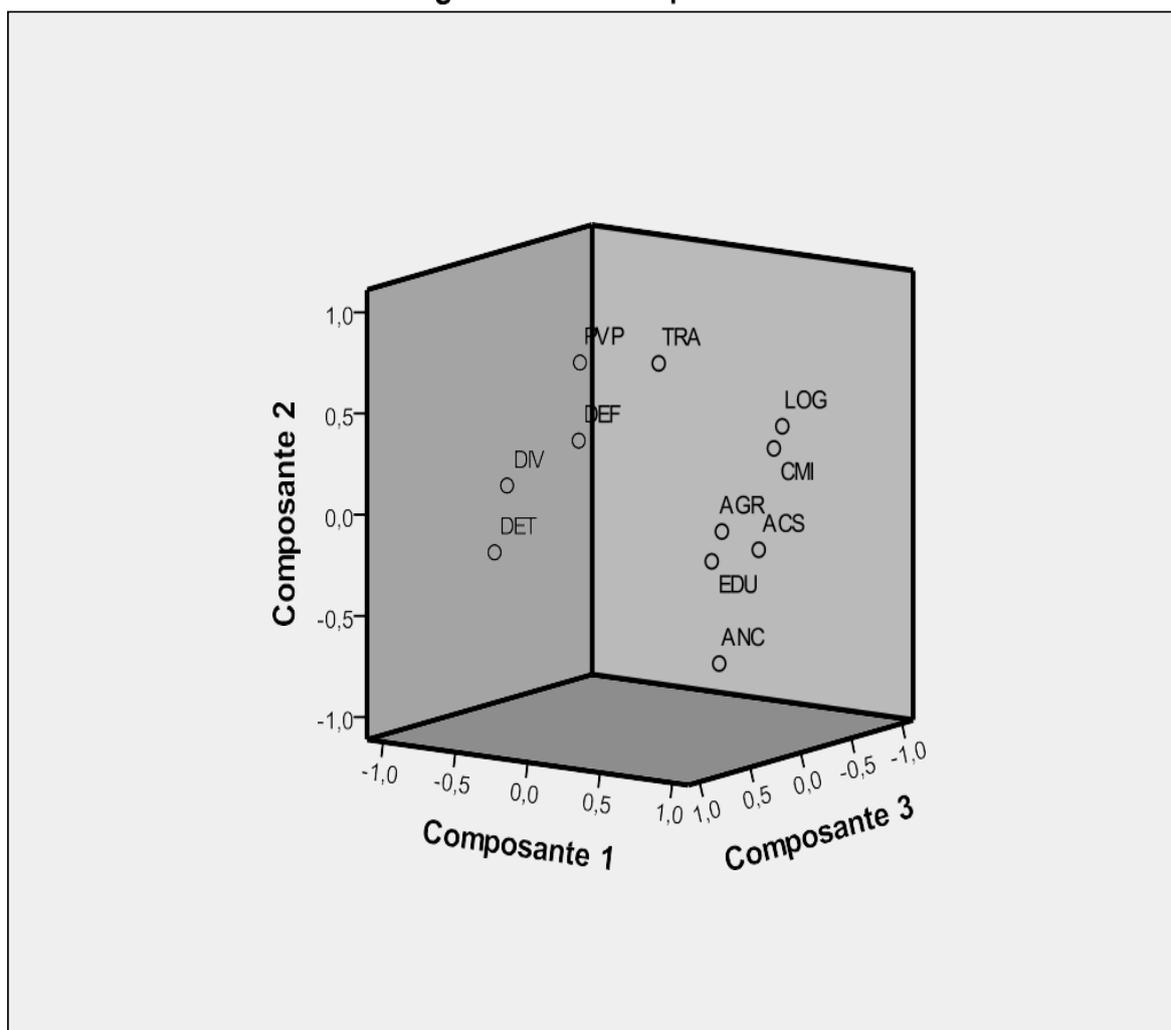
La deuxième composante est positivement corrélée avec la variable « anciens combattants » et négativement avec les variables « pouvoirs publics et transports ». Les autres corrélations sont plus faibles.

La troisième composante présente une corrélation assez importante (comparée aux autres valeurs) avec la variable « divers ».

#### 4. Diagramme des variables

Le diagramme des variables dans l'espace formé par les trois axes est le suivant :

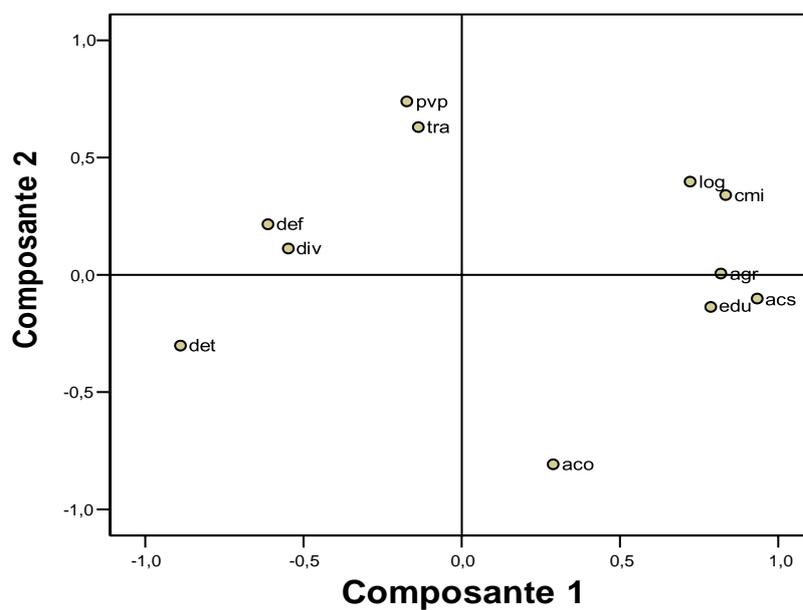
Diagramme de composantes



5.  
6.

**Remarque :** Si on se limite à deux axes principaux, on a une représentation des variables dans le plan.

## Diagramme de composantes

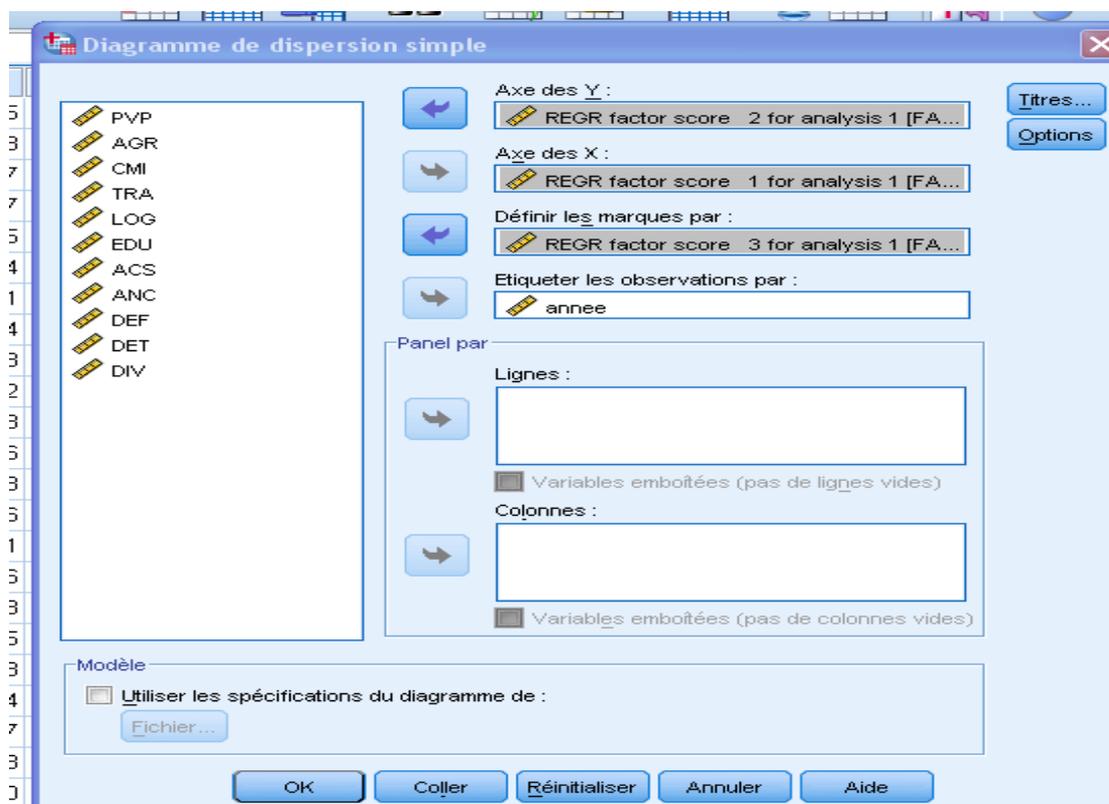
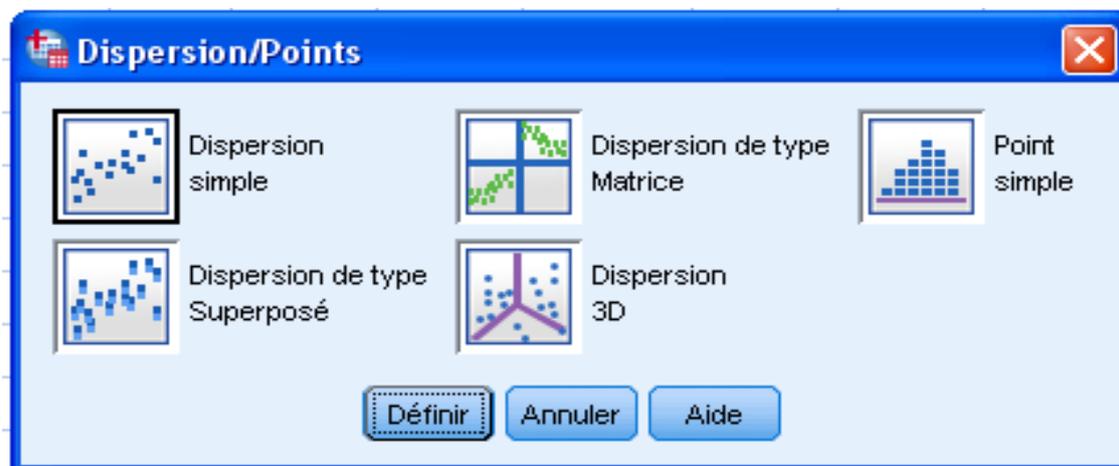


## 7. Représentation des individus

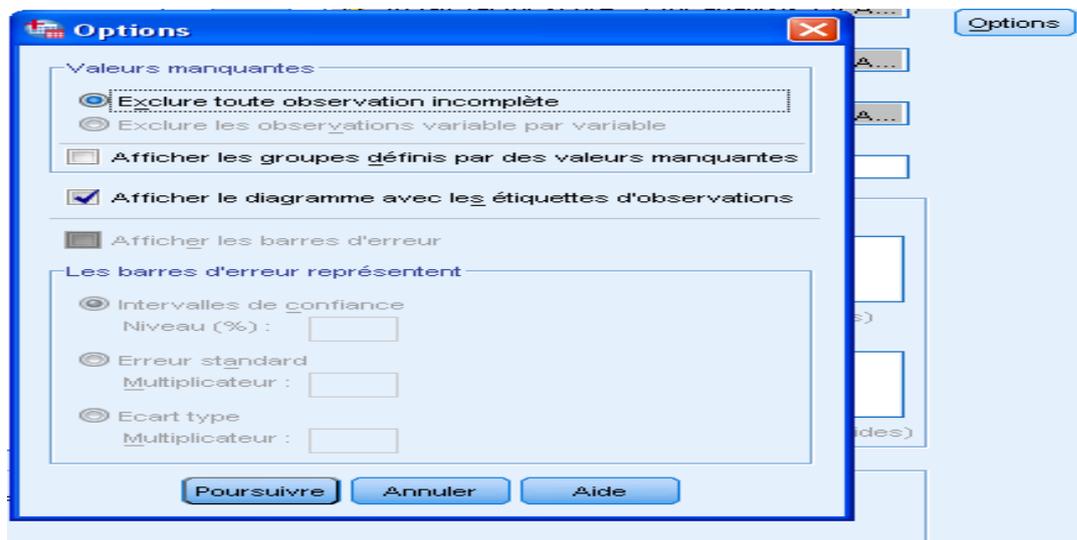
	annee	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV	FAC1_1	FAC2_1	FAC3_1
1	1872	18,0	,5	,1	6,7	,5	2,1	2,0	,0	26,4	41,5	2,1	-1,28019	,68894	1,38485
2	1880	14,1	,8	,1	15,3	1,9	3,7	,5	,0	29,8	31,3	2,5	-1,21970	1,36854	-,13178
3	1890	13,6	,7	,7	6,8	,6	7,1	,7	,0	33,8	34,4	1,7	-1,06640	,14725	,67669
4	1900	14,3	1,7	1,7	6,9	1,2	7,4	,8	,0	37,7	26,2	2,2	-,90714	,51047	,88072
5	1903	10,3	1,5	1,4	9,3	,6	8,5	,9	,0	38,4	27,2	3,0	-,99323	,15122	,51585
6	1906	13,4	1,4	,5	8,1	,7	8,6	1,8	,0	38,5	25,3	1,9	-,87579	,42088	,60277
7	1909	13,5	1,1	,5	9,0	1,6	9,0	3,4	,0	36,8	23,5	2,6	-,80687	,59507	,78695
8	1912	12,9	1,4	,3	9,4	,6	9,3	4,3	,0	41,1	19,4	1,3	-,63201	,51519	,16257
9	1920	12,3	,3	,1	11,9	2,4	3,7	1,7	1,9	42,4	23,1	,2	-,94493	,64548	-,151000

Utiliser le diagramme en trois dimensions : « **Graphes +Diagramme de dispersion + 3D**

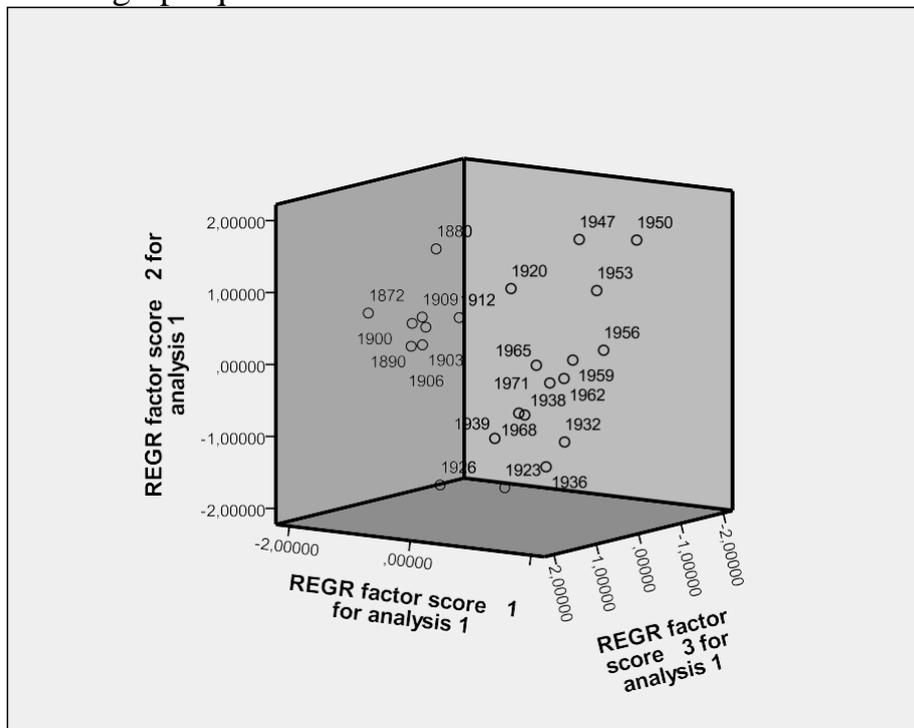
».



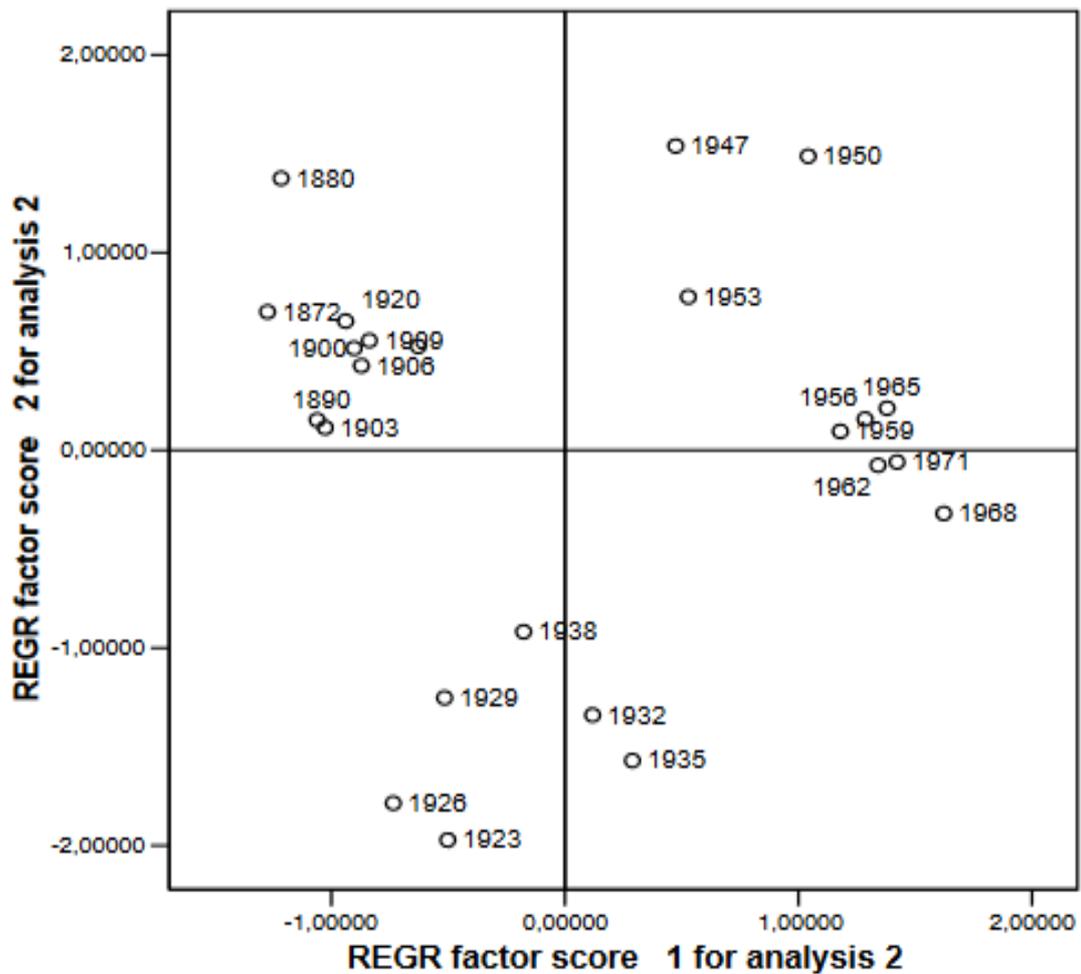
Cliquer sur « options » puis cocher « Afficher le diagramme avec les étiquettes d'observations ».



On a le graphique suivant :



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



On remarque que les années se répartissent en trois groupes (avant la première guerre mondiale, entre les deux guerres, après la seconde guerre mondiale). Seule l'année 1920, première année où il apparaît un poste de dépenses consacré aux anciens combattants est placée avec le premier groupe, alors qu'elle appartient au second.

---

<sup>i</sup> IBM SPSS Statistics Base 25 , Copyright IBM Corporation 1989, 2018, p.03.

<sup>ii</sup> IBM SPSS Statistics Core System Users Guide 25, Copyright IBM Corporation 1989, 2018, p.05.

**Les tests  
d'hypothèse  
appliqué en  
SPSS**

## Section 3 : Les Tests D'hypothèse Appliqué En SPSS

### Introduction :

Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique. Il est bien évident que la statistique (c'est-à-dire la variable d'échantillonnage) servant d'estimateur au paramètre de la population ne prendra pas une valeur rigoureusement égale à la valeur théorique proposée dans l'hypothèse. Cette variable aléatoire comporte des fluctuations d'échantillonnage qui sont régies par des distributions connues. Pour décider si l'hypothèse formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celle du paramètre spécifiée dans l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.

L'objectif de SPSS est d'offrir un logiciel intégré pour réaliser la totalité des tests statistiques habituellement utilisés en sciences sociales et en psychologie. De fait, SPSS est un logiciel très complet. Dans cet article, nous ne verrons qu'une faible partie de ses possibilités.

Grâce à SPSS, l'analyse statistique est plus accessible aux débutants et plus pratique pour les utilisateurs chevronnés. Grâce à des menus et des sélections de boîte de dialogue simples, vous pouvez exécuter des analyses complexes sans saisir une seule ligne de syntaxe de commande .L'éditeur de données offre une fonctionnalité de type tableur simple et efficace pour la saisie des données et la lecture du fichier de travail.

### 1-Généralités :

**1.1. Principe d'un test d'hypothèses :** Les tests d'hypothèse constituent un autre aspect important de l'inférence statistique. Le principe général d'un test d'hypothèse peut s'énoncer comme suit:

- On étudie une population dont les éléments possèdent un caractère (mesurable ou qualitatif) et dont la valeur du paramètre relative au caractère étudié est inconnue.
- Une hypothèse est formulée sur la valeur du paramètre : cette formulation résulte de considérations théoriques, pratiques ou encore elle est simplement basée sur un pressentiment.
- On veut porter un jugement sur la base des résultats d'un échantillon prélevé de cette population.

Il est bien évident que la statistique (c'est-à-dire la variable d'échantillonnage) servant d'estimateur au paramètre de la population ne prendra pas une valeur rigoureusement égale à la valeur théorique proposée dans l'hypothèse. Cette variable aléatoire comporte des fluctuations d'échantillonnage qui sont régies par des distributions connues.

Pour décider si l'hypothèse formulée est supportée ou non par les observations, il faut une méthode qui permettra de conclure si l'écart observé entre la valeur de la statistique obtenue dans l'échantillon et celle du paramètre spécifiée dans l'hypothèse est trop important pour être uniquement imputable au hasard de l'échantillonnage.

*La construction d'un test d'hypothèse consiste en fait à déterminer entre quelles valeurs peut varier la variable aléatoire, en supposant l'hypothèse vraie, sur la seule considération du hasard de l'échantillonnage.*

## **1-Définition des concepts utiles a l'élaboration des tests d'hypothèse :**

**1-1-Hypothèse statistique :** Une hypothèse statistique est un énoncé (une affirmation) concernant les caractéristiques (valeurs des paramètres, forme de la distribution des observations) d'une population

**1-2-Test d'hypothèse :** Un test d'hypothèse (ou test statistique) est une démarche qui a pour but de fournir une règle de décision permettant, sur la base de résultats d'échantillon, de faire un choix entre deux hypothèses statistiques.

**Hypothèse nulle (H0) et hypothèse alternative (H1) :** L'hypothèse selon laquelle on fixe à priori un paramètre de la population à une valeur particulière s'appelle l'hypothèse nulle et est notée H0. N'importe quelle autre hypothèse qui diffère de l'hypothèse H0 s'appelle l'hypothèse alternative (ou contre-hypothèse) et est notée H1.

**L'hypothèse Ho** postule que la relation entre X et Y est due au hasard, autrement dit qu'il n'y a pas de relation entre X et Y (nulle= absence de relation).

➤ **H0 : A = B (il ya pas une relation entre l'Age et le revenu mensuel )**.

Contrairement à l'hypothèse nulle, **H1** suggère que la relation entre X et Y ne peut être attribuée au hasard; il existe donc un lien entre X et Y au sein de la population.

➤ **H1 : A différend B ( il ya une relation entre l'Age et le revenu mensuel )**  
**H1 : A>B ou A<B**

**1-3-Seuil de signification du test :** Le risque, consenti à l'avance et que nous notons  $\alpha$  de rejeter à tort l'hypothèse nulle H0 alors qu'elle est vraie, s'appelle le seuil de signification du test et s'énonce en probabilité ainsi :  **$\alpha = P(\text{rejeter } H_0 | H_0 \text{ vraie})$**  .

A ce seuil de signification, on fait correspondre sur la distribution d'échantillonnage de la statistique une région de rejet de l'hypothèse nulle (appelée également région critique). L'aire de cette région correspond à la probabilité  $\alpha$ . Si par exemple, on choisit  $\alpha = 0.05$ , cela signifie que l'on admet d'avance que la variable d'échantillonnage peut prendre, dans 5% des cas, une valeur se situant dans la zone de rejet de H0, bien que H0 soit vraie et ceci uniquement d'après le hasard de l'échantillonnage.

Sur la distribution d'échantillonnage correspondra aussi une région complémentaire, dite région d'acceptation de H0 (ou région de non-rejet) de probabilité **1-  $\alpha$**

### Remarque :

- Les seuils de signification les plus utilisés sont  $\alpha = 0.05$  et  $\alpha = 0.01$  , dépendant des conséquences de rejeter à tort l'hypothèse H0.

- La statistique qui convient pour le test est donc une variable aléatoire dont la valeur observée sera utilisée pour décider du « rejet » ou du « non-rejet » de  $H_0$ . La distribution d'échantillonnage de cette statistique sera déterminée en supposant que l'hypothèse  $H_0$  est vraie.

## 2- Les tests d'hypothèses sous SPSS :

**2-1- LA CORRELATION :** Comment calculer un coefficient de corrélation? Quand et pourquoi calculer un coefficient de corrélation?

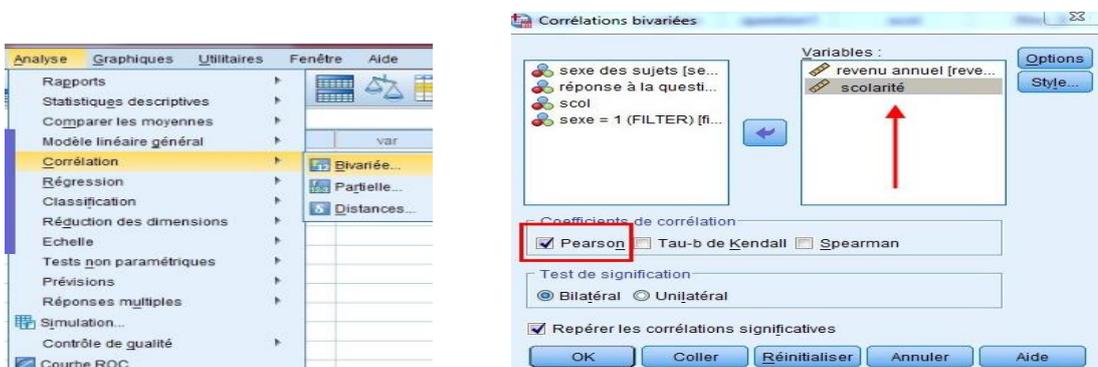
**Quand ?** Si votre recherche comporte une variable indépendante quantitative (X) et une variable dépendante quantitative (Y).

### Pourquoi?

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

### Les étapes à suivre pour calculer un coefficient de corrélation:

- Ouvrez votre matrice de données SPSS.
- Choisir ensuite le menu ➡ ANALYSE + CORRELATION + BIVARIÉE.
- Suivre les étapes selon les images suivantes :



4. Cliquez sur **OK**.
5. Une fenêtre s'ouvre...

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

1. Voici le résultat de votre test de corrélation

### Comment analyser les résultats d'un coefficient de corrélation?

- **H<sub>0</sub>** : il n' y a pas une relation entre la scolarité des participants et leur revenu annuel.
- **H<sub>1</sub>** : il ya une relation entre la scolarité des participants et leur revenu annuel.

Le résultat du test de corrélation ou Pearson Corrélation (r), ici 0,179.

La valeur de **Sig** dans ce cas-ci 0,345.

Le premier résultat 0,179 mesure le degré de liaison linéaire entre les variables dépendante (Y) et indépendante (X) de votre échantillon (**corrélation moyenne**).

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

- Parfaite si  $r = 1$
- Très forte si  $r > 0,8$ .
- Forte si r se situe entre 0,5 et 0,8.
- D'intensité moyenne si r se situe entre 0,2 et 0,5.
- Faible si r se situe entre 0 et 0.2.
- Nul si  $r = 0$

L'analyse des données de la présente recherche montre qu'il n'existe aucune relation entre la scolarité des participants et leur revenu annuel ( $r = 0,179$ ,  $p = 0,345$ ).

**2-2- ANALYSE DE VARIANCE : L'ANOVA à un facteur** (analyse de variance à un facteur) est utilisée dans le cas des tests de comparaison de moyennes sur plusieurs (plus de deux) échantillons indépendant ou groupes.

**Quand ?** Si votre recherche comporte deux groupes ou plus (indépendants ou non) et que votre variable dépendante est quantitative.

**Pourquoi** faut-il faire une analyse de variance (ANOVA) ?

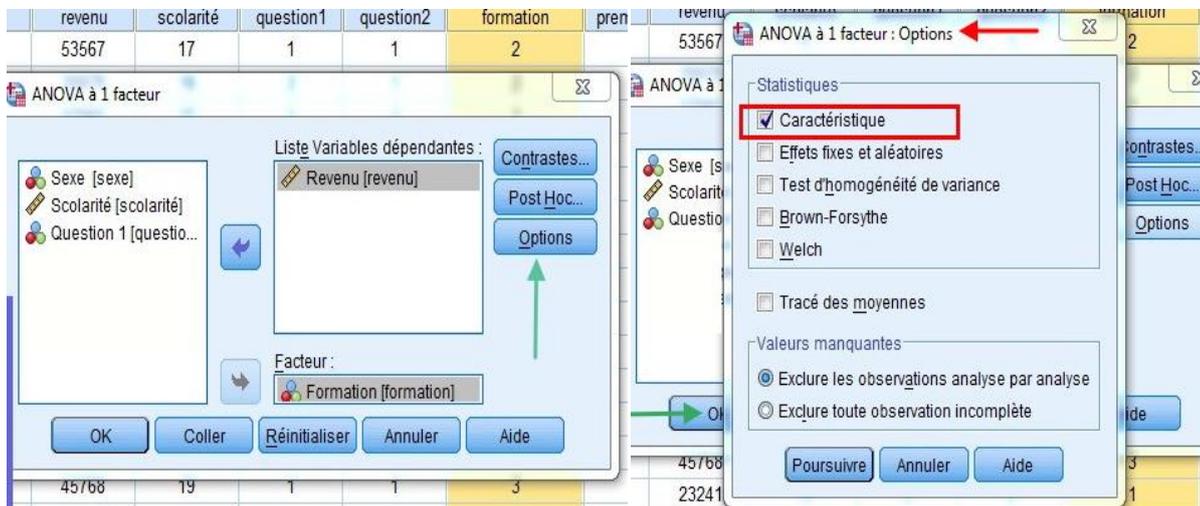
- Pour comparer la variance de ces groupes (variable x) afin d'inférer une relation entre X (Ex: le type de formation : science, technique, autres) et Y (Ex: le revenu).
- Les tests statistiques comme l'analyse de variance permettent au chercheur de rejeter ou non l'hypothèse nulle.
- Donc de prendre une décision quant à la valeur de cette hypothèse (est-elle vraie ou fausse ?).
- Avant de procéder à une analyse de variance (Analysis Of Variance), il faut formuler vos hypothèses statistiques ( $H_0$  et  $H_1$ ).

**Les étapes à suivre pour faire une ANOVA :**

1. Ouvrez votre matrice de données **SPSS**.
2. Choisir ensuite le menu ➡ **ANALYSE + COMPARER LES MOYENNES + ANOVA À UN FACTEUR**.
3. Suivre les étapes selon les images suivantes :
- 4.

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

5. Choisir ensuite en fonction de quelle variable indépendante ou Facteur vous désirez analyser le revenu. Dans cet exemple, il s'agit de [formation].
6. Le but de ce test est de comparer la variance des revenus de trois types de formation (science, technique et autre).
7. Le chercheur se pose la question suivante : La formation d'une personne influence-t-elle son revenu ?
8. Sur le plan statistique, il s'agit d'une hypothèse bilatérale.
9. Cliquez ensuite sur **Options** pour choisir vos options.
10. Une fenêtre s'ouvre...



11. Cochez Caractéristiques pour obtenir les moyennes et les variances de vos trois groupes.
12. Cliquez ensuite sur **Poursuivre**.
13. De retour à la première fenêtre, cliquez sur **OK**.
14. Voici le résultat final :

## A 1 facteur

## Descriptives

revenu annuel

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

## ANOVA

revenu annuel

2	Somme des carrés	ddl	Moyenne des carrés	F	Signification
Inter-groupes	2,710E8	2	1,355E8	2,591	,559
Intra-groupes	9,638E9	42	2,295E8		
Total	9,909E9	144			

15. Le premier tableau = analyse descriptive (moyenne) de votre échantillon; le second = analyse comparative (ou inférentielle).

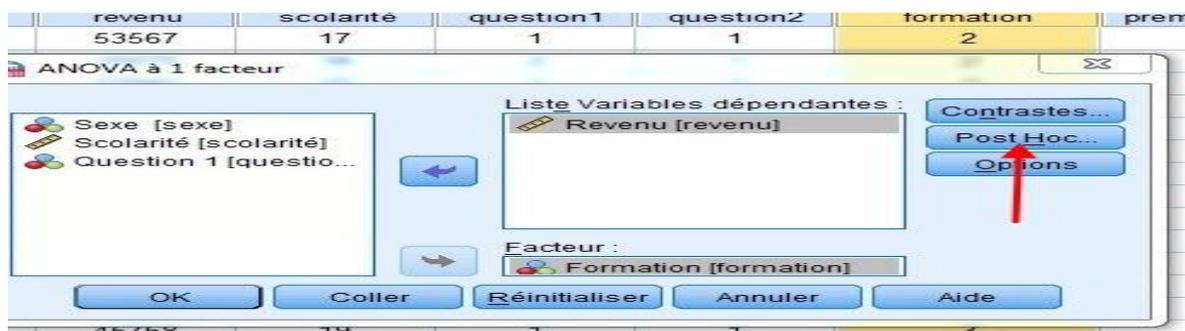
16. S'il y a d'autres tableaux, ils sont inutiles; ne les copiez pas dans votre exercice ou dans le fichier d'objectifs.

## Le cas d'une différence significative : le test POST-HOC

Que faire maintenant lorsque le niveau de signification du F est inférieur à 5 % ? Déclarer que les trois groupes est différents alors qu'il est possible que seulement deux des trois groupes le soient?

- Pour résoudre ce problème, il faut faire un test post-hoc qui nous indiquera lesquels des trois groupes sont différents.
- Ce test compare les trois groupes deux à deux.
- Pour choisir ce test, cliquez sur le bouton POST HOC.

1. Pour choisir ce test, cliquez sur le bouton **Post Hoc**, à droite.



2. Une nouvelle fenêtre apparaît: Comme vous pouvez le constater, il existe un vaste de choix de test post-hoc.



**Comparaisons multiples**  
revenu annuel  
Scheffe

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

- Voilà, les résultats de cette fenêtre - la colonne **Signification**- indique qu'il n'y a pas de différence significative entre les trois groupes comparés deux à deux (Sig. < 5 %).
- La première ligne compare la formation scientifique à la formation technique et autre.
- La seconde ligne compare la formation technique à la formation scientifique et autre.
- Finalement, les autres formations sont comparées à la formation scientifique et technique.

### Comment analyser les résultats de votre ANOVA :

- Dans une analyse de variance non significative, il y a trois données importantes :

Le **ddl** ou degré de liberté, ici 44.

Le résultat du **F** (= ,591 dans le tableau ci-bas).

La valeur de **p** ou **Signification**, dans ce cas-ci ,559.

► **A 1 facteur**

Descriptives

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

ANOVA

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

- Le **F** et le **ddl** permettent de calculer la valeur de p ou **Signification**.
- Donc 1 et 2 donne 3.
- La valeur de p ou Signification est l'erreur alpha, soit la probabilité ou le risque de commettre une erreur en déclarant qu'il existe une différence entre les 3 groupes ou les 3 mesures.
- La valeur de p ou Sig. permet de **confirmer** ou **d'infirmier** votre hypothèse alternative (H1).
- Rappelons qu'en sciences humaines, le seuil de signification est de 0,05.
- Si Signification ou valeur de p est **supérieure à 0,05**, vous devez accepter l'hypothèse nulle. Et conclure qu'il n'y a pas de différence significative entre vos deux groupes (ou mesures).
- Si la signification ou la valeur de p de votre test est **inférieure à 0,05**, il faut faire deux choses :
  1. d'abord rejetez l'hypothèse nulle et conclure qu'il y a une différence significative entre les trois groupes.
  2. procédez à un **test post-hoc** afin de savoir laquelle des comparaisons de groupes, pris deux à deux, est significativement différente.

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

**Dans le tableau** de notre rapport final, nous devons inscrire les 5 informations suivantes :

- Le nombre de participants de chacun des groupe ou **n** =
- La moyenne de chacun des trois groupes ou
- Le résultat de l'analyse de variance, soit le test **F**, dans le tableau ci- bas = ,591
- Le Sig. ou **Valeur de p** = probabilité de commettre l'erreur alpha. (=0,559).
- Si la valeur de **p** de votre test est inférieure à 0,05 (5%) donc résultat significatif.

Voici un **exemple de tableau : Analyse principale**

**Tableau 1**  
**Comparaisons entre les sujets sur le plan de la scolarité.**

Indicateurs	Groupes	n =	$\bar{X}$	F	Valeur de p =	<0,05 =*
Revenu	Science	15	50030	0,591	0,559	
	Technique	15	47297			
	Autres	15	53205			

**Analyse de données d'une analyse de variance :**

L'analyse des données de la présente recherche indique que les individus qui ont une formation en science gagnent 50 030 \$ comparativement à 47 297 \$ pour ceux qui ont une formation en techniques et 53 205 \$ pour le groupe d'individus autres (arts, métiers, etc). La différence entre les trois groupes n'est cependant pas significative ( $F = 0,591$ , ddl = 44,  $p = 0,559$ ). On peut donc conclure que le type de formation n'influence pas le salaire des individus. Et ainsi de suite avec les autres indicateurs...

**Analyse de données d'une analyse post-hoc :**

L'analyse des données de la présente recherche indique que les individus qui ont une formation en science gagnent 56 985 \$, comparativement à 43 416 \$ pour ceux qui ont une formation en techniques et 51 261 \$ pour le groupe d'individus autres (arts, métiers, etc). La différence entre ces trois groupes n'est cependant pas significative ( $F = 0,591$ , ddl = 44,

$p = 0,559$ ), sauf pour les formations universitaires et techniques (Scheffé test,  $p = 0,03$ ). Et ainsi de suite avec les autres indicateurs...

### 2-3- TEST T :

**Quand ?** À deux moments :

- Si votre recherche comporte deux groupes/échantillons indépendants et que votre variable dépendante est quantitative;
- Ou si dans votre recherche les participants ont été l'objet de deux mesures (= un groupe à mesures répétées) et que votre variable dépendante est quantitative.

**Pourquoi faire un test T ?** Pour comparer les moyennes de ces deux groupes (ou mesures) afin d'inférer une relation entre X et Y (le sexe et la scolarité, par exemple).

- Les tests statistiques comme le test t permettent au chercheur de rejeter ou non l'hypothèse nulle, donc de prendre une décision statistique.
- Avant de procéder à un test t, il faut formuler vos hypothèses statistiques ( $H_0$  et  $H_1$ ).

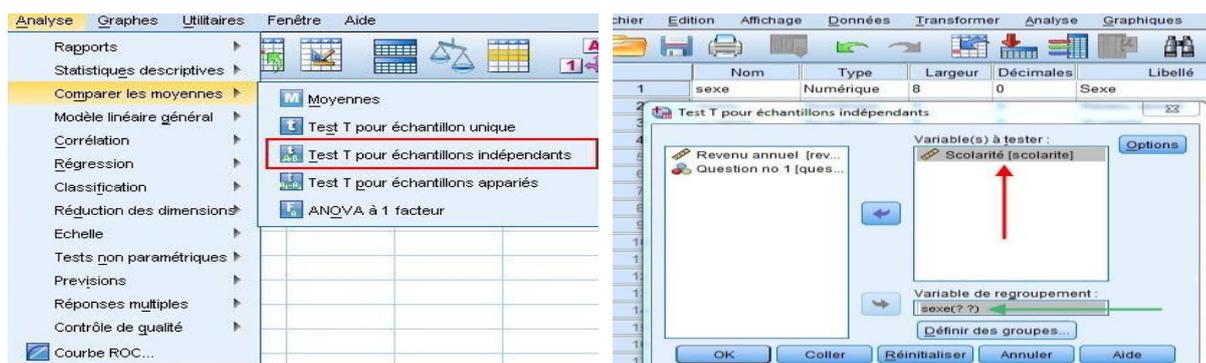
Dans la logique d'un test d'hypothèses - test t, khi-carré, etc. - il y a toujours 2 hypothèses statistiques.

- L'hypothèse nulle ou  $H_0$  - est une hypothèse qui postule qu'il n'y a pas de différence significative entre les moyennes des deux groupes (ou mesures); notez : Groupe 1 = Groupe 2 ou Mesure1 = Mesure2.
- L'hypothèse alternative ou  $H_1$  - correspond habituellement à l'hypothèse de votre recherche; notez : Groupe 1  $\neq$  Groupe 2 ou Mesure1  $\neq$  Mesure2. Ici  $\neq$  signifie n'égale pas.

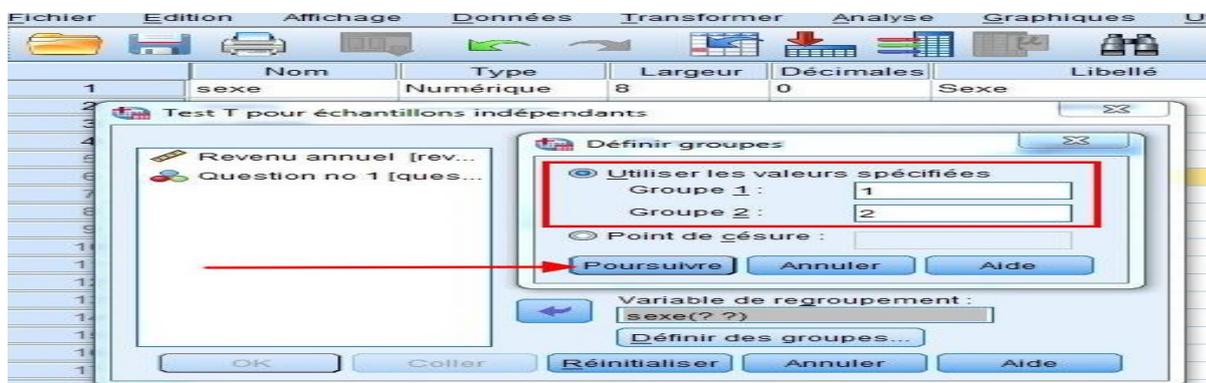
**Les étapes pour faire un test t à groupes indépendants :**

Si vos **groupes/échantillons** sont **indépendants**, choisir le menu **ANALYSE + COMPARER LES MOYENNES + TEST T POUR ÉCHANTILLONS INDÉPENDANTS**.

1. Au moyen des flèches  choisissez la Variable à tester (= la ou les VD/Indicateurs de votre recherche). Dans cet exemple, [Scolarité].
2. Choisir ensuite votre Variable de regroupement. Dans cet exemple ci-dessus, la variable [Sexe] est divisée en deux groupes : les hommes et les femmes. Le but de ce test est de comparer la moyenne des deux groupes - HOMMES et FEMMES - afin de vérifier l'hypothèse de recherche. Cette hypothèse avance que les niveaux de scolarité des hommes et des femmes sont différents au sein de la population.



3. Cliquez ensuite sur Définir les groupes pour assigner une valeur à vos deux groupes.
4. Une fenêtre s'ouvre...



1. Cliquez ensuite sur **Poursuivre**.

2. Notez que les points d'interrogation de la variable *Sexe* se sont transformés en 1 et 2, vos deux groupes.



**Comment analyser le résultat d'un TEST T :**

**H0 :** il n'y a pas de différence significative entre la scolarité des hommes et des femmes.

**H1 :** il y a une différence significative entre la scolarité des hommes et des femmes.

**Dans l'analyse** d'un test T à groupes indépendants, il y a 2 tableaux importants :

1. Le tableau des moyennes qui décrit vos groupes.
2. Le tableau du Test-t qui permet de comparer vos groupes.

Statistiques de groupe

		sexe	N	Moyenne	Ecart-type	Erreur standard moyenne
1	Scolarité	homme	15	17,20	1,568	,405
		femme	15	16,87	,834	,215

Test d'échantillons indépendants

	Test de Levene sur l'égalité des variances		Test-t pour égalité des moyennes					Intervalle de confiance 95% de la différence	
	F	Sig.	1	2	3	Différence moyenne	Différence écart-type	Inférieure	Supérieure
			t	ddl	Sig. (bilatérale)				
→ Hypothèse de variances égales	5,958	,021	7,27	28	,473	,333	,458	-,606	1,272
Hypothèse de variances inégales			,727	21,335	,475	,333	,458	-,619	1,286

Ici le **t** et **ddl** des tableaux ci-haut permettent à **SPSS** de calculer la valeur de **p** ou **Sig.(bilatérale)** de votre Test t. (Sig. veut ici dire **signification**).

Dans ce second tableau, voir ci-dessous, il y a 3 résultats importants :

1. Le résultat du test **t** (,727 dans le tableau ci-haut).
2. Le **ddl** ou degré de liberté, ici 28 (On écrit habituellement  $dl = 28$ ).
3. La valeur de **p** du test ou **Sig.(bilatérale)**, dans ce cas-ci 0,473 lorsque les variances sont égales donc accepter l'hypothèse nulle.

## 2-4- TEST KHI-DEUX :

- **Quand ?** Si votre recherche comporte deux groupes (ou deux mesures) et que votre variable dépendante est qualitative.
- **Pourquoi ?** Faire un test khi-deux ? Pour comparer les fréquences de ces deux groupes afin d'inférer une relation entre X (ex: sexe) et Y (Ex : les réponses - oui ou non - à la question no 1 de l'exercice **SPSS**).
- Les tests statistiques comme le khi-deux permettent au chercheur de rejeter ou non l'hypothèse nulle, donc de prendre une décision.
- Avant de procéder à ce test, il faut formuler vos hypothèses statistiques (Ho et H1).
- Attention : On dit aussi khi-carré ou Chi-deux; ce sont des synonymes.

## Les étapes pour faire un TEST KHI-DEUX :

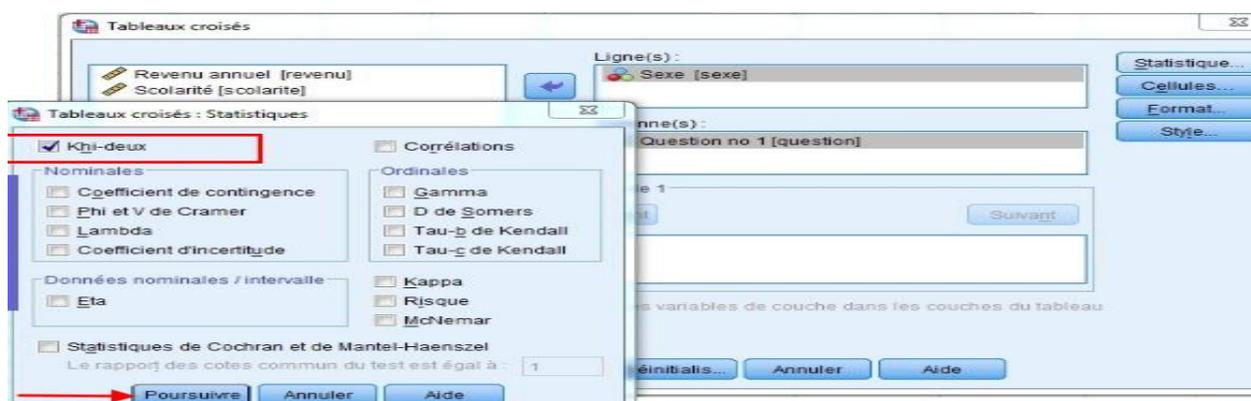
1. Choisir le menu ➡ **ANALYSE + STATISTIQUES DESCRIPTIVES + TABLEAUX CROISÉS.**



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



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



### Comment analyser les résultats d'un KHI-2 :

**H0** : il n'y a pas de différence significative entre la fréquence des hommes est différente de la fréquence des femmes.

**H1** : il y a une différence significative entre la fréquence des hommes est différente de la fréquence des femmes.

**Dans l'analyse** d'un Khi-deux, il y a 2 tableaux importants :

1. Le tableau 1 des effectifs ou des fréquences qui décrit vos groupes.
2. Le tableau 2 du Khi-deux qui permet de comparer ces groupes.

**Tableau croisé Sexe \* Question 1**

Effectif

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

**Tests du Khi-deux**

	Valeur	ddl	Signification asymptotique (bilatérale)	Signification exacte (bilatérale)	Signification exacte (unilatérale)
Khi-deux de Pearson	,159 <sup>a</sup>	2 1	,690		
Correction pour la continuité <sup>b</sup>	,000	1	1,000		
Rapport de vraisemblance	,159	1	,690		
Test exact de Fisher				1,000	,500
Association linéaire par linéaire	,153	1	,695		
Nombre d'observations valides	30				

Dans ce second tableau, il y a 3 résultats importants :

- Le résultat du test ou **Valeur** (0,159 dans le tableau ci-haut).
- Le **ddl** ou degré de liberté, ici ddl =1. On écrit habituellement dl.
- **Signification asymptotique (bilatérale)** ou valeur de p, dans ce cas-ci 0,690.donc refuser l'hypothèse nulle et accepter l'hypothèse alternative.

# Bibliographies

- Alboukadel Kassambara ; 2017; Practical Guide to Principal Component Methods in R.
- Biskin, B. H ,Comment on significance testing. Measurement and evaluation in counselling and development,
- Blais, J.-G, Statistique, méthodes quantitatives et analyse des données. Repères, essais en éducation, 1991.
- Cohen, J, The statistical power of abnormal-social psychological research : a review. Journal of abnormal and social psychology, 1962.
- IBM SPSS Statistics Base 24, Copyright IBM Corporation 1989, 2017
- IBM SPSS Statistics Core System Users Guide 24, Copyright IBM Corporation 1989, 2017
- Jérôme Pagés; 2013 ; Analyse factorielle multiple avec R ; EDP sciences.
- Maddock, J. E. et Rossi, J. S, Statistical power of articles published in three health psychology-related journals. Health Psychology,
- Marie Chavent ; 2021 ; Analyse en Composantes Principales(ACP) ; Université de Bordeaux.
- Mohamed DENNAI ;2016;Analyse en composantes principales et analyse factorielle, ACP & AFP ; AGIC - Edition Maïa .

## **Description du cours :**

L'analyse de données est un domaine issu du monde des statistiques qui vise à faire le lien entre les différentes données statistiques pour les classer, les décrire et les analyser de manière succincte.

Ce polycopié est destiné aux étudiants de Licence, Master ou Doctorat en économies et en sciences sociales et humaines en générale. Il s'articule autour des étapes clés d'une analyse des données.

## **Objectifs de cours :**

- Traiter et décrire l'information contenue dans des grands ensembles de données.
- Reconnaître les moyens théoriques et pratiques pour exploiter les informations issues de base de données statistiques multidimensionnelles grâce aux méthodes d'analyse statistique multivariées.
- Maîtriser les techniques qui ont pour but de décrire, de réduire, de classer et de clarifier les données, en dégagant, les liaisons, les ressemblances ou les différences entre les variables ou groupes de variables.
- Synthétiser l'information contenue dans la série statistique Interpréter correctement les graphiques et résultats fournis par les logiciels.

## **Mots mots-clés :**

Analyse en composantes principales ; teste d'hypothèse ; SPSS ; ANOVA ; Corrélation.

## **Course Description:**

Data analysis is a field from the world of statistics that aims to make the connection between different statistical data in order to classify, describe and analyze them succinctly.

This handout is intended for Bachelor's, Master's or Doctorate students in economics and social sciences and humanities in general. It revolves around the key steps of a data analysis.

## **Course Objectives:**

- Process and describe information contained in large data sets.
- Recognize the theoretical and practical means of exploiting information from multivariate statistical databases using multivariate statistical analysis methods.
- Master techniques for describing, reducing, classifying and clarifying data, identifying linkages, similarities or differences between variables or groups of variables.
- Synthesize the information contained in the statistical series Interpret cor

## **Keywords:**

Principal component analysis; hypothesis testing; SPSS; ANOVA; Correlation.

## وصف المحاضرة:

تحليل البيانات هو مجال من عالم الإحصاء يهدف إلى الربط بين البيانات الإحصائية المختلفة من أجل تصنيفها ووصفها وتحليلها بإيجاز. هذه المحاضرة مخصصة لطلاب الليسانس أو الماجستير أو الدكتوراه في الاقتصاد والعلوم الاجتماعية والإنسانية بشكل عام. تتمحور حول الخطوات الرئيسية لتحليل البيانات.

## الاهداف:

- معالجة ووصف المعلومات الواردة في مجموعات البيانات الكبيرة.
- التعرف على الوسائل النظرية والعملية لاستغلال المعلومات من قواعد البيانات الإحصائية متعددة المتغيرات باستخدام أساليب التحليل الإحصائي متعدد المتغيرات.
- إتقان تقنيات وصف البيانات واختزلها وتصنيفها وتوضيحها ، وتحديد الروابط أو أوجه التشابه أو الاختلاف بين المتغيرات أو مجموعات المتغيرات.
- تجميع المعلومات الواردة في السلسلة الإحصائية. وتفسير الرسوم البيانية والنتائج بشكل صحيح.

## الكلمات المفتاحية:

تحليل المكون الرئيسي؛ اختبار الفرضيات، برنامج SPSS، تحليل التباين، الارتباط.