

Contingence / Correspondance

7.	Contingence / Correspondance	2
7.1.	Table de contingence	2
7.2.	Distance et Test d'indépendance du Chi-2	5
7.3.	Ajustement d'une distribution empirique	7
7.4.	Analyse factorielle des correspondances (AFC)	8
7.4.1.	La matrice des données :	8
7.4.2.	Représentation dans N(I)	8
7.4.3.	Centrage	8
7.4.4.	Inertie et facteurs de N(I)	9
7.4.5.	Inertie et facteurs de N(J)	10
7.4.6.	Dualité et Formules de transition :	10
7.4.7.	Cartes factorielles et Interprétations	10
7.4.8.	Exemple	11

7. Contingence / Correspondance

7.1. Table de contingence

Etant donné un fichier de données individuelles (individus i en lignes et variables j en colonnes), on s'intéresse à deux variables-colonnes X et Y qui sont soit qualitatives (nominales ou ordinales) soit quantitatives (cardinales) discrètes avec un petit nombre de modalités (sinon il faut faire des regroupements), soit encore quantitatives recodées en classes. On suppose que le nombre de modalités de X et Y est respectivement p et q . Si les variables ont un trop grand nombre de modalités, et que les effectifs sont assez faibles, il peut être nécessaire de faire un recodage des variables en un plus petit nombre de modalités, pour que les fréquences conjointes (cf. infra) soient significatives et que l'on puisse interpréter leur tableau.

Un tri de profondeur 2 est un tri sur les deux colonnes à la fois. Il produit une table à double entrée de p lignes et q colonnes que l'on appelle « table de *contingence* » et qui donne les fréquences dites *conjointes* de chaque couple possible de modalités pour X, Y .

Ind\Var		X	Y	
i		x_i	y_i	
n				

Tableau de données individuelles

⇒

X \ Y		k		Total
j		n_{jk}		$n_{j.}$
Total		$n_{.k}$		$n = n_{..}$

Tableau croisé de X et Y

Les fréquences conjointes n_{jk} s'interprètent comme le nombre des individus qui ont la modalité j pour X et la modalité k pour Y . Les fréquences marginales sont celles que l'on obtient dans les marges par sommation : le point sert à positionner l'indice et distinguer par exemple $n_{.3}$ la fréquence marginale de la troisième modalité de X de $n_{3.}$ la fréquence marginale de la troisième modalité de Y .

$$\text{En ligne : } n_{j.} = \sum_k n_{jk} \quad \text{En colonne : } n_{.k} = \sum_j n_{jk}$$

A partir de ce tableau de fréquences *absolues*, on peut produire trois tableaux de fréquences relatives, selon que l'on rapporte la fréquence conjointe de la case (j,k) au total -ligne, au total-colonne, ou au total général. Dans le dernier cas on obtiendra les fréquences relatives conjointes : $f_{jk} = n_{jk} / n$. On en déduit les fréquences relatives marginales :

$$f_{j.} = f_j = \sum_k f_{jk} = \sum_k n_{jk} / n \text{ qui donne la fréquence relative de la modalité } j \text{ de } X.$$

$$f_{.k} = f_k = \sum_j f_{jk} = \sum_j n_{jk} / n \text{ qui donne la fréquence relative de la modalité } k \text{ de } Y.$$

Dans les deux premiers cas on obtient des fréquences relatives *conditionnelles* :

$f_j^k = f(j/k) = f_{jk} / f_{.k}$ qui donne la fréquences relatives de $X=j$ sachant que $Y=k$ et que l'on appelle plus communément pourcentage ligne.

$f_k^j = f(k/j) = f_{jk} / f_{j.}$ qui donne la fréquence relative de $Y=k$ sachant que $X=j$ que l'on appelle plus communément pourcentage colonne.

X \ Y		k		Total
j		f_{jk}		$f_{j.}$
Total		$f_{.k}$		1

Fréquences relatives conjointes

X \ Y		k		Total
j		f_k^j		1
Total		$f_{.k}$		1

Fréquences relatives conditionnelles en ligne

X \ Y		k		Total
j		f_i^k		$f_{i.}$
Total	1	1	1	1

Fréquences relatives conditionnelles en colonne

Dans **Excel** le tableau des fréquences conjointes peut être obtenu avec l'utilitaire d'analyse « tableaux dynamiques » en choisissant les deux variables à croiser dans la liste des champs et en les transportant dans « champ ligne » et « champ de colonne », puis en glissant déposant de même la seconde dans « Données » au centre du tableau. Il reste à choisir dans « champ pivot table » l'option « nombre » et dans « option », « afficher les données par » Normal (on aura les effectifs conjoints n_{jk}), ou %ligne, ou %colonne.

La question à résoudre est alors la mesure de la *liaison statistique* ou de l'*association* ou de la *contingence* entre les variables X et Y . Rappelons que l'on appelle contingence ce qui s'oppose à la nécessité, c'est-à-dire ce qui se produit effectivement sans que l'on puisse attribuer des causes ou des raisons.

Supposons pour commencer que X et Y ont 2 modalités chacune et que les marges sont fixées aux valeurs 70/30. On peut remplir un tel tableau à 2x2 cases de plusieurs façons dont les 4 cas suivants :

	Y1	Y2	
X1	70	0	70
X2	0	70	30
	70	30	100

 $X1 \Rightarrow Y2$

	Y1	Y2	
X1	0	70	70
X2	70	0	30
	70	30	100

Indépendance

	Y1	Y2	
X1	49	21	70
X2	21	9	30
	70	30	100

Liaison stat.

	Y1	Y2	
X1	44	26	70
X2	26	4	30
	70	30	100

 $\begin{matrix} 1 \\ \Rightarrow \\ Y1 \end{matrix}$

Dans le premier cas, si un individu est $X1$ il est forcément $Y1$: $X1$ est logiquement lié à $Y1$ et de même $X2$ est logiquement lié à $Y2$. Dans le second cas c'est l'inverse : $X1$ est logiquement lié à $Y2$, et $X2$ est logiquement lié à $Y1$. Dans le troisième cas il y a 70% de $Y1$ et 30% de $Y2$ dans chacune des deux catégories $X1$ et $X2$: être $X1$ ou $X2$ ne change rien quant à Y . C'est l'indépendance. Enfin le quatrième cas illustre une situation statistique intermédiaire entre la seconde liaison logique et la situation d'indépendance.

La mesure de liaison va s'effectuer par rapport à un tableau de référence qui est la situation d'*indépendance* ou absence de liaison. On dira que X et Y sont indépendantes statistiquement, si les profils des fréquences conditionnelles sont égaux, c'est à dire si quelle que soit la valeur j de X , pour les différentes sous-populations correspondant à $X=x_1, X=x_2, \dots, X=x_n$, la distribution conditionnelle de Y sachant $X=j$ est la même, et la même que la distribution marginale de Y . On dira symétriquement que X et Y sont indépendants si, quelle que soit la valeur k de Y , la distribution conditionnelle de X sachant $Y=k$ est la même, et la même que la distribution marginale de X . Dans ce cas ce sont les profils-colonnes qui seraient les mêmes. Donc formellement l'indépendance se traduit par les relations :

$$\forall j, f_j^k = f_{jk} / f_{.k} = f_j \Leftrightarrow \forall k, f_k^j = f_{jk} / f_j = f_{.k} \Leftrightarrow f_{jk} = f_j \cdot f_{.k}$$

Soit plus simplement sur les effectifs : $n_{jk}^* = (n_{j.}) (n_{.k}) / n_{..}$: l'effectif théorique dans une case serait le produit des effectifs marginaux correspondants divisés par n .

Une première analyse de la liaison entre X et Y peut donc se fonder sur la façon dont le profil des %ligne de la marge inférieure du tableau se déforme dans chacune des lignes du tableau, c'est-à-dire pour chacune des modalités de X . Ou de manière symétrique sur la façon dont le profil des %colonne de la marge droite du tableau se déforme dans chacune des colonnes du tableau, c'est-à-dire pour chacune des modalités de Y . Ou d'une autre manière encore, on étudiera le tableau des *écarts* entre les fréquences conjointes *observées* n_{jk} et les fréquences conjointes *théoriques* n_{jk}^* , celles que l'on aurait dans le cas de l'indépendance, pour des distributions marginales fixées. Le signe et l'importance et ces écarts, c'est à dire les sous effectifs et sureffectifs que l'on observe par rapport à ceux de l'indépendance renseignent sur les liaisons entre modalités.

Mais quelle confiance faut-il accorder à ces écarts ? S'ils sont assez grands on sera tenté de les attribuer à des phénomènes réels d'association entre les deux variables qui doivent s'interpréter en terme de phrases du type « ce n'est pas par hasard qu'il y a tel sureffectif mais bien à cause de, ou en raison de telle relation ». Par contre si ces écarts sont assez petits on les attribuera au hasard et on dira qu'ils sont non significatifs : en d'autres termes il serait dangereux d'en inférer une liaison due à un phénomène réel. Nous sommes donc en face d'une situation de test d'hypothèses, H_0 : Indépendance contre H_1 : liaison significative, un test pour lequel il faut définir une statistique et une région critique.

7.2. Distance et Test d'indépendance du Chi-2

D'un point de vue inférentiel, les fréquences absolues observées sont des réalisations de variables aléatoires suivant des lois multinomiales (généralisation de la loi binomiale). Sous l'hypothèse d'indépendance, l'effectif N_{jk} de chaque case suit une loi binomiale d'espérance $n.f_{j.}.f_{.k}$ et de variance $n.f_{j.}.f_{.k}(1-f_{j.})$. Si les effectifs sont suffisants - on retiendra la condition : effectif théorique $n.f_{j.}.f_{.k} > 4$ - Les variables N_{jk} centrées réduites suivent des lois asymptotiquement normales centrées-réduites, et la somme de leurs carrés qui prend la forme d'une distance entre tableau des fréquences observées et des fréquences théoriques :

$$\chi^2 = \sum_j \sum_k \frac{(n_{jk} - n f_{j.} f_{.k})^2}{n f_{j.} f_{.k}} = \sum_j \sum_k \frac{(n_{jk} - n_{jk}^*)^2}{n_{jk}^*}$$

Cette distance du χ^2 s'interprète comme une moyenne, sur toutes les cases du tableau, des écarts (*effectifs observés-effectifs théoriques*) pondérés par leurs poids relatifs calculés par rapport aux effectifs attendus. Le statisticien Karl Pearson a montré en 1900 que, sous l'hypothèse d'indépendance (H_0), cette distance dite du Khi-2 suivait une loi du même nom dont le seul paramètre s'appelle *degré de liberté du tableau* (ddl). Ce degré de liberté est le nombre de cases du tableau que l'on peut faire librement varier tout en respectant des marges fixées. Ddl n'est pas pq mais $(p-1)(q-1)$ car les n_{jk} sont des variables liées par les équations exprimant que leurs sommes sont égales aux marges fixées.

Si cette distance du χ^2 calculée est faible on peut conclure que les écarts sont dus au hasard et que la situation observée est compatible avec l'hypothèse d'indépendance. Si cette distance calculée dépasse la valeur critique donnée par la table on peut conclure à une liaison significative.

Attention : la distance du χ^2 n'est pas un bon indicateur de l'intensité de la liaison car elle est proportionnelle aux effectifs du tableau. Pour une *mesure* de la liaison (et non pas un test) on lui préférera l'indicateur :

$$\Phi^2 = \chi^2 / n = \sum_j \sum_k \frac{(f_{jk} - f_{j.} f_{.k})^2}{f_{j.} f_{.k}}$$

Si le test conclut à une liaison significative, il reste alors à l'interpréter, en partant des sur-effectifs ou sous-effectifs observés dans chaque case du tableau.

Exemple :

Le tableau ci-dessous croise le type de bac et le nombre de livres lus par ans dans un échantillon de 746 élèves des IFSI enquêtés en 1993. On donne dans chaque case le chiffre des effectifs observés, les fréquences en ligne et en colonne, les écarts à l'indépendance, et la contribution au χ^2 . Par exemple dans la case de ceux qui ont un bac général (ABCD) et ne lisent aucun livre, l'effectif observé est 8, ce qui représente 3% des 271 bacs généraux et 25,8% des 31 qui ne lisent aucun livre. L'effectif théorique dans le cas de l'indépendance serait $271 \times 31 / 746 = 11,3$. Donc il y a un sous effectif (écart à l'indépendance) de 3 personnes dans cette case. La contribution au χ^2 est de $(8 - 11,3)^2 / 11,3 = 0,96$.

BAC	NR	NB LIVRES				TOTAL	
		0	<10	<20	>20		
NON	23	10	97	46	22	198	
%lign	11.6	5.1	49.0	23.2	11.1	100%	
%col	39.0	32.3	28.6	22.1	20.2	26.5	
Ecart	7	2	7	-9	-7	0.0	
khi2	3.4	0.4	0.5	1.5	1.7	8	
ABCD	16	8	107	81	59	271	
%lign	5.9	3.0	39.5	29.9	21.8	100%	
%col	27.1	25.8	31.6	38.9	54.1	36.3	
Ecart	-5	-3	-16	5	19	-0.0	
khi2	1.4	0.9	2.1	0.4	9.5	14	
FGT	20	13	135	81	28	277	
%lign	7.2	4.7	48.7	29.2	10.1	100%	
%col	33.9	41.9	39.8	38.9	25.7	37.1	
Ecart	-2	1	9	4	-12	0.0	
khi2	0.2	0.2	0.7	0.2	3.8	5	
TOTAL	59	31	339	208	109	746	
%LIGNE	7.9	4.2	45.4	27.9	14.6	100	
%COL	100	100	100	100	100	100	
ECARTS	0.0	0.0	0.0	0.0	0.0	0.0	
KHI-2	5	2	3	2	15	27	

Le repérage des écarts *positifs* à l'indépendance (en gras) permet d'opposer ceux qui n'ont pas le bac ou un bac technique à ceux qui ont un bac général : les seconds ont une certaine propension à lire plus (sureffectifs). Ces écarts qui témoignent d'une liaison statistique (qui n'est pas systématique puisque nous avons trouvé par exemple 8 bacs ABCD qui ne lisent pas) sont-ils dus au hasard ou à une liaison significative ? La valeur du χ^2 pour tout le tableau est égale à 27 et supérieure à la valeur critique 15,5 trouvée dans la table pour un degré de liberté égal à $4 \times 2 = 8$ et un risque de 5%. Donc nous concluons à une liaison significative dans le sens déjà indiqué, et due principalement à des contributions importantes des forts lecteurs.

7.3. Ajustement d'une distribution empirique

Un cas particulier de la situation précédente est celui dans lequel nous n'avons qu'une seule variable. Le tableau n'a plus qu'une ligne qui représente la distribution de fréquence empirique de n observations sur les p modalités de la variable X . La question n'est plus de tester une liaison mais de tester l'adéquation de cette distribution statistique simple à une loi de probabilité hypothétique L , ou plus précisément l'hypothèse que les fréquences observées peuvent provenir de la répétition n fois d'un tirage de valeurs dans cette loi hypothétique L .

Valeurs observées	x_1	x_2	...	x_i	...	x_p	Total
Fréquences observées	n_1	n_2		n_i		n_p	$N = \sum n_i$
Probabilités / H_0	p_1	p_2		p_i		p_p	1
Fréquences théoriques	n^*_1	n^*_2		$n^*_i = np_i$		n^*_p	N
Distance du Khi-2	k_1	k_2		$k_i = (n_i - n^*_i)^2 / n^*_i$		k_p	$d\chi^2 = \sum k_i$

Le tableau des fréquences observées n_i est alors comparé case par case avec celui des effectifs théoriques $n^*_i = np_i$, c'est-à-dire les valeurs espérées (au sens de l'espérance mathématique) dans la case i que l'on aurait si les valeurs étaient produites par une loi Binomiale de paramètre (n, p_i) , p_i étant la loi probabilité fournie par la loi de l'hypothèse H_0 . La distance calculée entre effectif observé et effectif théorique est encore celle du Khi-2 :

$$d\chi^2 = \sum_i \frac{(n_i - np_i)^2}{np_i}$$

Cette distance suit sous l'hypothèse H_0 une loi du Khi-2 de degré de liberté $(p-k-1)$:

p est le nombre de comparaisons, k est le nombre de paramètres de la loi L que l'on a éventuellement estimés sur les mêmes données.

Si cette distance est supérieure à la valeur critique obtenue pour un risque α choisi, alors on rejeter l'hypothèse H_0 que les données proviennent d'une loi L .

Exemple :

On a observé les données suivantes de nombre de fautes par page dans un texte de 50 pages et l'on veut tester qu'elles proviennent d'une loi de Poisson.

Valeurs observées	0	1	2	3	4	5	6	7	>7	Total
Fréquences Observées	3	6	11	12	7	5	3	2	1	50
Probabilités théoriques	0,0498	0,1494	0,224	0,224	0,168	0,1008	0,0504	0,0216	0,012	1
Fréquences théoriques	2,49	7,47	11,2	11,2	8,4	5,04	2,52	1,08	0,6	50
Khi-2	0,10	0,29	0,00	0,06	0,23	0,00	0,77			1,46

On estime le paramètre λ de la loi de poisson par la moyenne empirique = 3,14 arrondi à 3, en prenant 8 pour la dernière classe. Pour obtenir un effectif théorique >5 on a regroupé les 3 dernières cases pour le calcul du χ^2 . Pour un degré de liberté de $7-1-1 = 5$ la valeur critique correspondant à un risque de 5% est 11,07. Notre distance calculée est 1,46. On accepte donc l'hypothèse d'une loi de Poisson de paramètre 3.

7.4. Analyse factorielle des correspondances (AFC)

L'analyse factorielle en général est une méthode de représentation d'un tableau de données par deux nuages de points dans des espaces de grande dimension. Pour rendre visibles ces nuages on les réduit par projection sur des sous-espaces de dimension plus petite (par exemple un plan) déterminés par de nouveaux axes dits factoriels, et dans lesquels ils conservent un maximum de leurs propriétés.

7.4.1. La matrice des données :

Considérons un tableau de données K dit « tableau de correspondances » entre deux ensembles $I=\{i\}$ à n modalités et $J = \{j\}$ à p modalités, et dont les éléments $\{k_{ij}, i=1,n \text{ et } j=1,p\}$ sont des « nombres » dont la somme en ligne et la somme en colonne prend un sens. On voit que le tableau de contingence qui croise 2 variables X et Y d'une enquête est un cas particulier d'un tel tableau (en remplaçant X et Y par I et J , j et k par i et j). Mais un tableau en euros qui distribue des consommations par pays et par poste de consommation, un tableau en minutes qui distribue des budgets-temps par activité et par classe sociale, un tableau qui recense des entreprises par secteur et par nombre de salariés sont d'autres exemples de tableau de correspondance.

On peut transformer ce tableau d'effectifs K en un tableau de fréquences $F=\{f_{ij}=k_{ij}/k\}$.

Chaque ligne de ce tableau a une somme $f_{i.} = \sum_j f_{ij}$ qui représente la fréquence associée à la modalité i de I . Chaque colonne a une somme $f_{.j} = \sum_i f_{ij}$ qui représente la fréquence associée à la modalité j de J .

7.4.2. Représentation dans $N(I)$

Si on s'intéresse au tableau des fréquences conditionnelles que l'on a appelé en 7.1 « pourcentages en ligne » et que l'on note $X = \{x_{ij} = f_{ij}/f_{i.}\}$. Chaque modalité i de I est représentée par un « profil ligne » de ce tableau, c'est-à-dire la suite $(x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{ip})$ des % correspondant à chaque modalité de J pour les individus ayant la modalité i de I . Cette suite est un vecteur à p composantes que l'on peut représenter par un point (i) dans un espace de dimension p . Cette notion d'espace de dimension p généralise celle d'espace à 2 ou 3 dimensions que vous connaissez : si $p = 2$ le vecteur à deux coordonnées (x_{i1}, x_{i2}) est représentable par un point dans un espace cartésien à 2 dimensions dont x_{i1} est l'abscisse et x_{i2} l'ordonnée ; et les n points représentant les profils-lignes formeront un nuage de points dans ce plan qui est un espace à deux dimensions R^2 . Si p est plus grand on dit encore que les n points $(1), (2), \dots, (i), \dots, (n)$ forment le nuage $N(I)$ des n points lignes dans un espace à p dimensions R^p . En fait c'est dans ce cas un sous espace de dimension $p-1$ parce que $\sum_j x_{ij} = \sum_i f_{ij} / f_{i.} = 1$

7.4.3. Centrage

La marge inférieure du tableau représente le profil-ligne moyen $(f_{.1}, f_{.2}, \dots, f_{.j}, \dots, f_{.p})$ de répartition des individus sur les modalités de J . Ce profil-ligne moyen sera représenté dans le nuage $N(I)$ par un point G , au centre de gravité du nuage. Si les deux variables étaient indépendantes, le profil de toutes les lignes serait égal à celui-ci, et tous les points de $N(I)$ seraient confondus en G . Mais les profils-lignes s'écartent plus ou moins de ce profil moyen et ce sont ces écarts que nous souhaitons analyser. Nous nous intéressons donc maintenant au tableau des « écarts » entre chaque profil-ligne (i) et celui de la marge ou variables « centrées » c'est-à-dire les lignes du tableau centré $Y = \{y_{ij} = (f_{ij}/f_{i.}) - f_{.j}\}$. Chacune de ces lignes peut encore être représentée par un point d'un nuage $N'(I)$ dans un espace de dimension $p-1$.

L'ennui c'est tout de même que nous ne pouvons plus représenter facilement cet espace.

Mais comme dans l'espace usuel à 2 ou 3 dimensions, nous pouvons toujours calculer des distances entre 2 points i et i' en écrivant que le carré de la distance est la somme des carrés des composantes chaque composante étant ici pondérée par $1/f_{.j}$:

$$d^2(i, i') = \frac{1}{n} \sum_j (x_{ij} - x_{i'j})^2 = \frac{1}{n} \sum_j (y_{ij} - y_{i'j})^2 = \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

$$d^2(i, G) = \frac{1}{n} \sum_j (x_{ij} - x_{.j})^2 = \frac{1}{n} \sum_j (y_{ij})^2 = \frac{1}{n} \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - f_{.j} \right)^2 = \frac{1}{n} \sum_j \frac{1}{f_{.j}} \left(\frac{f_{ij} - f_{i.} f_{.j}}{f_{i.}} \right)^2$$

On peut montrer que cette distance est invariante si on modifie le découpage en classes dans J , et en particulier si on fusionne 2 colonnes du tableau de données.

7.4.4. Inertie et facteurs de N(I)

En mécanique, on appelle inertie d'un point i de masse m à la distance d de l'origine la quantité $I(i) = md^2$.

L'inertie d'un point (i) de notre nuage représentatif des profils ligne centrés sera égale à

$$\text{Inertie}(i) = f_{i.} d^2(i, G) .$$

$$\text{Et Inertie totale de } N(I) = I_n = \sum_i f_{i.} d^2(i, G) = \frac{1}{n} \sum_i \sum_j \frac{(f_{ij} - f_{i.} f_{.j})^2}{f_{i.} f_{.j}} = \frac{1}{n} d\chi^2$$

Où $d\chi^2$ est la distance du Khi-2 sur tout le tableau.

Le principe de la recherche de facteurs est la suivante. Nous commençons par chercher une direction repérée par un vecteur u de longueur unitaire telle que l'inertie projetée sur cette direction soit maximale. Cette direction que l'on appelle *premier axe factoriel* est en quelque sorte la direction principale d'étirement du nuage, celle selon laquelle, si on devait résumer par une seule variable l'ensemble des modalités de I , ce serait par les coordonnées des projections du nuage $N(I)$ sur cette direction, coordonnées que l'on appelle *premier facteur*. Puis on recherche un second axe factoriel perpendiculaire au précédent qui maximise l'inertie projetée sur lui, et qui fournit un second facteur. Etc...jusqu'au $(p-1)$ ème facteur.

En fait on peut montrer que la recherche de ces facteurs revient à rechercher, pour $\alpha = 1, 2, 3, \dots, (p-1)$ les valeurs propres λ_α et les vecteurs propres u_α de la matrice

$$[p, p] : S = F'D_n^{-1} F D_p^{-1} = \left\{ s_{jj'} = \sum_i \frac{f_{ij} f_{ij'}}{f_{i.} f_{.j'}} \right\} \text{ ou encore la matrice symétrique } A =$$

$$\left\{ a_{jj'} = \sum_i \frac{f_{ij} f_{ij'}}{f_{i.} \sqrt{f_{.j} f_{.j'}}} \right\} = D_p^{-1/2} S D_p^{1/2} , \text{ c'est-à-dire tels que } A u_\alpha = \lambda_\alpha . \text{ On supposera dans}$$

ce cours que vous ne savez pas ce que cela veut dire et que vous faites confiance sur ce point dans le programme d'ordinateur qui vous fournira ces valeurs :

- les valeurs propres λ_α représentent l'inertie projetée sur l'axe α .
- les vecteurs propres u_α définissent la direction de l'axe factoriel α .
- Le facteur d'ordre α , est $F_\alpha = Y M u_\alpha =$ de coordonnées $f_{\alpha j} = \sum_j \frac{f_{ij}}{f_{i.} f_{.j}} u_{\alpha j} .$

7.4.5. Inertie et facteurs de N(J)

De façon tout à fait symétrique (en échangeant les rôles de I et J), on représente les p profils-colonnes (les % en col. de la matrice $X' = \{f_{ij}/f_{.j}\}$ non centrée ou de la matrice $Y' = \{f_{ij}/f_{.j} - f_{i.}\}$ centrée par rapport à la marge de droite) dans un espace R^n par un nuage N(J) de p points-colonnes dans l'espace R^n (en fait un sous espace de dimension $(n-1)$) munis chacun d'un poids $f_{.j}$. On définit de manière analogue le carré de la distance euclidienne entre deux points j et j' par :

$$d^2(j, j') = \sum_i \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}} \right)^2$$

On cherche de la même façon la succession des axes factoriels qui maximisent l'inertie de ce nuage de points projetée sur ces axes. Ce qui revient à chercher les valeurs propres μ_α et les vecteurs propres v_α d'une certaine matrice d'inertie T de dimension $[n, n]$.

Le facteur d'ordre α est $G_\alpha = \text{YM } v_\alpha =$ de coordonnées $g_{\alpha j} = \sum_i \frac{f_{ij}}{f_{i.} f_{.j}} v_{\alpha i}$.

Mais de fait il y a des relations simples (dites de dualité) entre les résultats des analyses dans N(I) et dans N(J).

7.4.6. Dualité et Formules de transition :

L'inertie totale est $\sum_i f_{i.} F_{\alpha i}^2 = \sum_j f_{.j} G_{\alpha j}^2 = \lambda_\alpha = \mu_\alpha$ (les valeurs propres sont les mêmes).

Les vecteurs propres sont liés par les relations : $u_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F' D_n^{-1} v_\alpha$; $v_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} F D_p^{-1} u_\alpha$

Les coordonnées factorielles sont centrées et déduites des vecteurs propres par les formules :

$$F_\alpha = \frac{\sqrt{\lambda_\alpha}}{f_{i.}} v_{\alpha i} \text{ avec } \sum_i f_{i.} F_{\alpha i} = 0 \quad ; \quad G_\alpha = \frac{\sqrt{\lambda_\alpha}}{f_{.j}} u_{\alpha j} \text{ avec } \sum_j f_{.j} G_{\alpha j} = 0.$$

Formules de transition : elles permettent de passer d'un nuage à l'autre :

$$F_{\alpha i} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \frac{f_{ij}}{f_{i.}} G_{\alpha j} \quad \text{et} \quad G_{\alpha j} = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{f_{ij}}{f_{.j}} F_{\alpha i}.$$

C'est à dire que, au coefficient $1/\sqrt{\lambda_\alpha}$ près les projections des points d'un nuage sont, sur un axe, les barycentres des projections des points de l'autre nuage. C'est ce qui permet une représentation simultanée de I et de J.

7.4.7. Cartes factorielles et Interprétations

Puisque les axes factoriels sont déterminés dans un ordre décroissant de part d'inertie, la projection des nuages de points N(I) et N(J) sur le sous espace des premiers facteurs assurera le meilleur compromis entre lisibilité et qualité de la représentation. Selon le même principe exactement du dessin industriel qui privilégie la projection sur un plan d'un objet à trois dimensions on se contentera en général de travailler sur une projection plane définie par les deux premiers axes.

L'interprétation du graphique (on dit aussi *carte* ou *mapping*) se fait en général en trois temps et relève d'un certain art :

1. On interprète chaque axe comme une nouvelle variable qui oppose telles caractéristiques à telles autres. On s'aide pour cela des contributions (absolues) de chaque point à l'inertie de l'axe, et des contributions relatives de l'axe à chaque point (ou cosinus carrés).

2. On interprète les proximités entre les points de la façon suivante :

- deux points de $N(I)$ sont proches si et seulement si ils ont même profil-ligne.
- deux points de $N(J)$ sont proches si et seulement si ils ont même profil-colonne.
- un point de $N(I)$ et un point de $N(J)$ sont proches soit parce qu'ils ont fortement associés soit par simple effet de barycentre (Voir l'exemple).

3. On tente parfois de faire des regroupements de modalités associées de I et de J qui sont dans la même zone de la carte et que l'on nomme un peu hardiment des classes. Mais il faut se souvenir que a) la carte n'est qu'une projection b) les points-modalités ne sont que des barycentres d'individus qui ont cette modalité.

7.4.8. Exemple

Le tableau de données ci-dessous donne pour un ménage type d'un pays de l'OCDE (i) les dépenses k_{ij} en dollars (ou écus ? c'était avant l'euro) pour le poste de consommation (j). On fournit les principaux résultats d'une AFC (Logiciel Statbox).

	ABT	HAC	LEC	MME	MED	TCO	LSC	AUT	TOTAL	
spn	1045	277	584	272	132	533	258	736	3837	Espagne
gre	1074	231	313	222	79	404	121	251	2695	Grèce
ita	1589	591	970	558	359	819	546	1056	6488	Italie
por	753	187	99	173	92	296	116	266	1982	Portugal
irl	1722	255	495	260	102	518	365	272	3989	Irlande
jap	2056	594	1761	505	990	889	913	1672	9380	Japon
rfa	1556	738	1743	767	1303	1332	816	899	9154	RFA
bel	1548	587	1332	757	772	885	472	1027	7380	Belgique
fce	1643	580	1503	678	711	1314	577	1011	8017	France
ned	1392	533	1418	550	923	795	693	992	7296	Pays-Bas
gb	1153	440	1237	413	81	992	589	1191	6096	Royaume-Uni
can	1452	516	1923	766	347	1286	891	1213	8394	Canada
usa	1589	767	2307	672	1642	1789	1079	1818	11663	Etats-Unis
dan	2093	531	2235	608	158	1551	856	870	8902	Danemark
swd	1921	614	2085	530	204	1318	806	510	7988	Suède
TOTAL	22586	7441	20005	7731	7895	14721	9098	13784	103261	
	Alimentation		Logement		Dep.Santé		Loisir/Ens			
	Habillement		Meubles		Transp/Comm		Autres			

1. On peut utiliser une AFC parce que les k_{ij} sont sommables en ligne et en colonne.

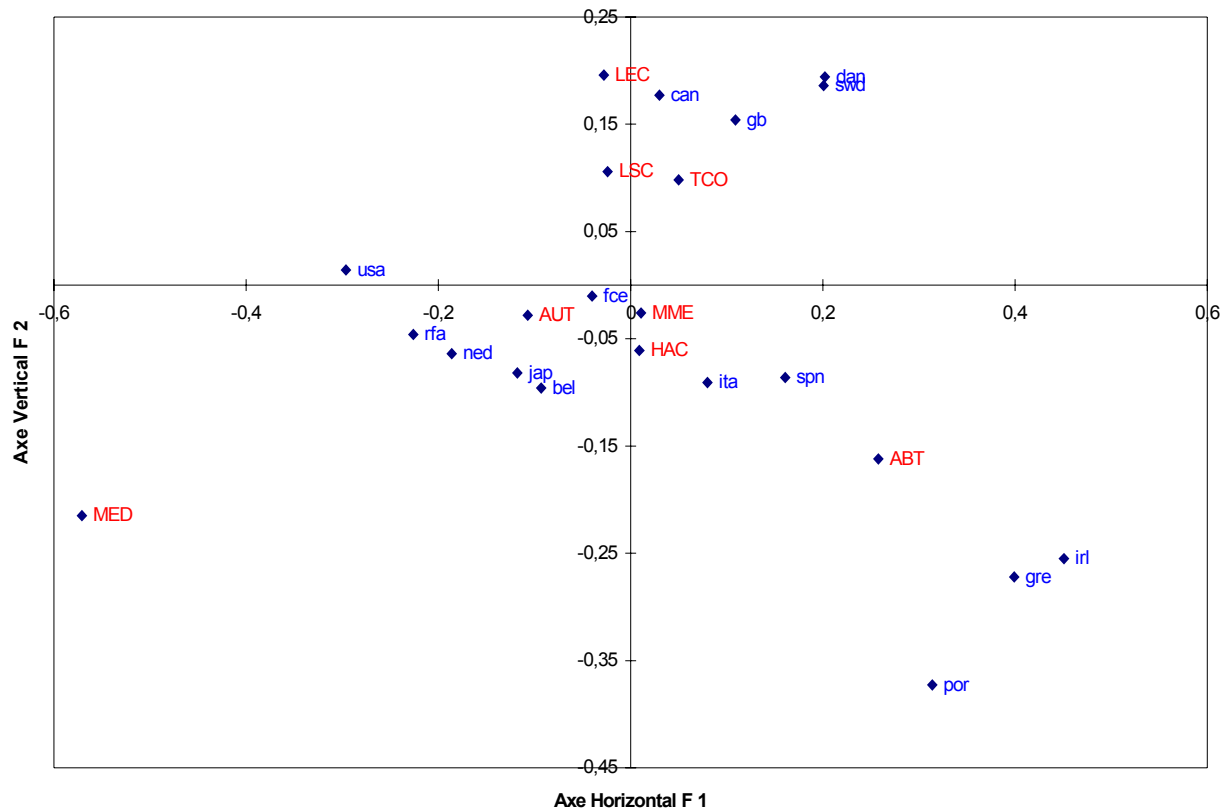
2. Les fréquences relatives sont les pourcentages en ligne suivants, et révèlent par exemple une structure de dépense de la Grèce très différente de la structure moyenne :

Grèce	0.399	0,086	0,116	0,082	0,029	0,150	0,045	0,093
Total	0,219	0,072	0,194	0,075	0,076	0,143	0,088	0,133

3. Le nombre des facteurs est de $8-1 = 7$. La somme des valeurs propres est 0,08046. Les 4 premières valeurs propres sont fournies :

	F1	F2	F3	F4
Valeurs propres	0,042	0,020	0,012	0,004

Avec les 2 premiers axes on cumule 0,062 d'inertie soit 77% de l'inertie.



4. L'axe factoriel F1 oppose l'alimentation (ABT) à la santé (MED) et la Grèce+Irlande+Portugal à USA (comme le confirment les valeurs des contributions). En gros la richesse au sens du PIB et une extension macro de la loi de Engel.

L'axe factoriel F2 oppose plus subtilement Alimentation+Santé (MED+ABT) à Logement+Loisirs+Transport, et pays du sud Europe et Irlande à Pays du Nord de l'Europe et Canada. En gros une autre dimension plus socio-culturelle du développement.

L'axe factoriel F3 isole le poste Autre (et Espagne + GB) à tout le reste. Mais on ne sait pas ce qu'est autre...et on n'a pas les coordonnées. Difficile d'en dire plus.

5. Japon et Belgique ont des structures de dépenses très semblables.

LSC et TCD représentent à peu près la même part de dépenses pour les 15 pays.

France et MME (Meubles). On ne peut rien dire de cette proximité entre points du centre des 2 nuages : effet de barycentre vraisemblable et pas de surdéveloppement français de la part dépensée en meubles.

Par contre, sur le bord du nuage, on peut dire que la Grèce a une forte part de ses dépenses en ABT (ce qui est confirmé par les % de la question 2).

Le tableau ci-dessous des contributions permet de confirmer et d'affiner l'interprétation.

Contribution absolue	F1	F2	F3	F4
ABT	35,1630	29,1827	2,2392	8,1754
HAC	0,0145	1,3590	0,0602	8,8666
LEC	0,3468	38,4380	4,1562	5,9158
MME	0,0219	0,2480	0,1748	40,4051
MED	59,8071	18,0179	11,4293	1,0811
TCO	0,8703	7,1341	1,9619	17,3883
LSC	0,1181	5,1163	0,0078	17,9777
AUT	3,6582	0,5040	79,9706	0,1899
spn	2,3442	1,3767	12,6019	0,2174
gre	10,0037	9,8796	0,9915	2,0775
ita	0,9858	2,6464	6,7995	3,3603
por	4,5537	13,6439	0,3784	4,1479
irl	18,9384	12,8109	5,3161	12,1657
jap	3,0328	3,0904	7,4983	40,9005
rfa	10,8632	0,9253	17,8097	3,9019
bel	1,4638	3,3281	0,0329	11,8790
fce	0,2886	0,0345	0,7997	11,1053
ned	5,8689	1,4474	0,7167	3,7050
gb	1,6921	7,1813	26,1553	0,7469
can	0,1839	13,0973	2,6935	2,0537
usa	23,7792	0,1268	0,0001	1,0805
dan	8,4630	16,6327	2,4354	0,8947
swd	7,5391	13,7785	15,7716	1,7626

Contribution Relative				
ABT	0,6979	0,2719	0,0125	0,0165
HAC	0,0051	0,2225	0,0059	0,3136
LEC	0,0165	0,8597	0,0555	0,0286
MME	0,0035	0,0187	0,0079	0,6584
MED	0,8347	0,1181	0,0447	0,0015
TCO	0,0925	0,3559	0,0585	0,1873
LSC	0,0209	0,4241	0,0004	0,3219
AUT	0,1386	0,0090	0,8496	0,0007
spn	0,3454	0,0952	0,5208	0,0032
gre	0,6502	0,3014	0,0181	0,0137
ita	0,1906	0,2402	0,3687	0,0659
por	0,3844	0,5407	0,0090	0,0355
irl	0,6756	0,2145	0,0532	0,0440
jap	0,2764	0,1322	0,1916	0,3779
rfa	0,6433	0,0257	0,2958	0,0234
bel	0,2570	0,2743	0,0016	0,2115
fce	0,1425	0,0080	0,1107	0,5559
ned	0,7698	0,0891	0,0264	0,0493
gb	0,1329	0,2647	0,5760	0,0059
can	0,0220	0,7364	0,0905	0,0249
usa	0,9247	0,0023	0,0000	0,0043
dan	0,4896	0,4517	0,0395	0,0052
swd	0,3954	0,3392	0,2320	0,0094