

Analyse des données avancée

Chapitre 9 : L'AFC et la métrique du khi-deux



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Un tableau d'une autre nature

Les profils

La métrique du khi-deux

Du khi-deux à l'inertie

Le calcul factoriel

Ce qu'il faut retenir

Un tableau d'une autre nature

Ce qui change

Deux variables **qualitatives** croisées. Chaque case porte un **effectif**, non une mesure.

Il n'y a ni moyenne, ni variance, ni corrélation à calculer sur des catégories : rien de l'ACP ne s'applique.

Le tableau du chapitre

Le même fichier de **300** clients, mais deux de ses variables qualitatives : la **catégorie socioprofessionnelle** et le **motif de fréquentation**.

Ces deux variables étaient illustratives en ACP. Elles deviennent ici les variables actives d'une analyse à part entière.

Le tableau croisé

observés	Prix	Choix	Proximité	Service	total
Étudiant	14	2	5	1	22
Employé	45	20	26	10	101
Cadre	8	40	8	16	72
Indépendant	12	19	12	32	75
Retraité	5	1	17	7	30
total	84	82	68	66	300

théoriques	Prix	Choix	Proximité	Service
Étudiant	6,16	6,01	4,99	4,84
Employé	28,28	27,61	22,89	22,22
Cadre	20,16	19,68	16,32	15,84
Indépendant	21,00	20,50	17,00	16,50
Retraité	8,40	8,20	6,80	6,60

Le khi-deux du chapitre 19 de l'inférence.

$$\chi^2 = \sum \frac{(n_{ij} - t_{ij})^2}{t_{ij}} = 110,41$$

Avec 12 degrés de liberté, le seuil à 5 % vaut 21,03.

L'indépendance est rejetée sans hésitation.

Et ensuite ? Le test dit qu'il y a une liaison. **Il ne dit pas laquelle.**

Cinq lignes, quatre colonnes, vingt écarts à commenter — l'AFC va les mettre sur un plan.

Fréquences et marges

$$f_{ij} = \frac{n_{ij}}{n}, \quad f_{i\cdot} = \sum_j f_{ij}, \quad f_{\cdot j} = \sum_i f_{ij}$$

Marges-lignes : 22, 101, 72, 75, 30. Marges-colonnes : 84, 82, 68, 66. Total $n = 300$.

Les profils

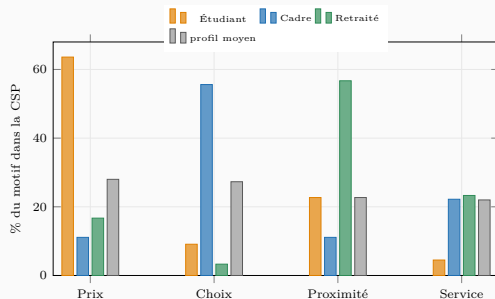
Le profil-ligne

$$\text{profil}(i) = \left(\frac{f_{i1}}{f_{i\cdot}}, \dots, \frac{f_{iJ}}{f_{i\cdot}} \right), \quad \text{de somme } 1$$

Trois profils, en pourcentages de ligne

Exemple					
	Prix	Choix	Proximité	Service	
Étudiant	63,6	9,1	22,7	4,5	
Cadre	11,1	55,6	11,1	22,2	
Retraité	16,7	3,3	56,7	23,3	
Profil moyen	28,0	27,3	22,7	22,0	

Diviser par la marge, en image



L'AFC compare des profils, pas des effectifs.

Il y a 22 étudiants et 101 employés : les comparer en effectifs n'aurait aucun sens.

En **pourcentages de ligne**, la comparaison redevient possible — et l'écart au **profil moyen** devient l'objet de l'analyse.

Un profil ligne égal au profil moyen n'apporte rien : il ira au centre du plan factoriel. C'est exactement ce que dit l'indépendance.

Intuition

Il y a **22** étudiants et **101** employés. Comparer leurs effectifs ferait apparaître une différence qui n'est que de **taille**, non de **comportement**.

Le profil élimine la taille et ne laisse que la structure. En AFC, on ne compare jamais des lignes : on compare des profils.

Une identité utile

$$\sum_i f_{i\cdot} \frac{f_{ij}}{f_{i\cdot}} = \sum_i f_{ij} = f_{\cdot j}$$

Le barycentre des profils-lignes, pondérés par leurs marges, **est le profil marginal des colonnes**.

Ce que cela impose

Le centre du nuage n'est donc pas choisi : c'est la marge, et rien d'autre.

Et il porte le sens le plus fort du tableau : le profil moyen **est le profil de l'indépendance**. Un profil qui s'en écarte est un profil qui a quelque chose à dire.

La métrique du khi-deux

La distance entre deux profils-lignes

$$d_{\chi^2}^2(i, \ell) = \sum_{j=1}^J \frac{1}{f_{\cdot j}} \left(\frac{f_{ij}}{f_{i\cdot}} - \frac{f_{\ell j}}{f_{\ell\cdot}} \right)^2$$

C'est une distance euclidienne **pondérée par l'inverse des marges-colonnes** : c'est le triplet $\mathbf{M} = \mathbf{D}_c^{-1}$ du chapitre 2.

Ce que le poids $1/f_{\cdot j}$ corrige

Un écart de 5 points sur une colonne qui pèse 50 % est un écart de un dixième : peu de chose.

Le même écart sur une colonne qui pèse 5 % **double** cette colonne. La division par la marge rend les deux comparables — sans elle, les modalités rares seraient invisibles.

L'énoncé

Si deux colonnes ont **le même profil**, les regrouper en une seule ne change **aucune distance entre lignes**.

Pourquoi c'est la propriété la plus importante de l'AFC

Le découpage des modalités est presque toujours arbitraire. Faut-il distinguer *cadre* et *profession libérale*? Découper l'âge en quatre tranches ou en six?

L'équivalence distributionnelle garantit que, si deux modalités se comportent pareil, **le résultat ne dépend pas du découpage**. Aucune autre métrique ne l'assure — et cet argument justifie le facteur $1/f_{.j}$ mieux que tout raisonnement d'échelle.

Le calcul, sur un couple

$$d^2(\text{Étudiant, Cadre}) = \frac{(0,636 - 0,111)^2}{0,280} + \frac{(0,091 - 0,556)^2}{0,273} + \frac{(0,227 - 0,111)^2}{0,227} + \frac{(0,045 - 0,222)^2}{0,220}$$
$$= 0,984 + 0,792 + 0,059 + 0,142 = 1,977$$

Les couples extrêmes

Couple	$d^2_{\chi^2}$
Étudiant – Employé	0,189
Cadre – Indépendant	0,543
Employé – Indépendant	0,832
Étudiant – Cadre	1,977
Cadre – Retraité	1,925

Du khi-deux à l'inertie

Le khi-deux et le phi-deux

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \quad e_{ij} = \frac{n_{i.} n_{.j}}{n}$$

$$\Phi^2 = \frac{\chi^2}{n} = \text{inertie totale du nuage des profils}$$

Exemple

Case	Observé	Attendu	Écart au carré rapporté
Étudiant – Prix	14	6,16	9,98
Cadre – Choix	40	19,68	20,98
Retraité – Proximité	17	6,80	15,30
Indépendant – Service	32	16,50	14,56

À elles seules : 60,8 sur $\chi^2 = 110,41$, soit 55 %.

Exemple

$$\Phi^2 = \frac{110,41}{300} = 0,3680$$

L'identité qui relie les deux mondes

L'inertie est le phi-deux

$$I = \sum_i f_i \cdot d_{\chi^2}^2(\text{profil}(i), \text{profil moyen}) = \Phi^2$$

La vérification, ligne par ligne

Ligne	Poids	Inertie	Part
Étudiant	0,073	0,0523	14,2 %
Employé	0,337	0,0637	17,3 %
Cadre	0,240	0,1085	29,5 %
Indépendant	0,250	0,0667	18,1 %
Retraité	0,100	0,0767	20,9 %
Total	1	0,3680	100 %

Application économique

Les **cadres** portent 29,5 % de l'information du tableau alors qu'ils ne sont que 24 % de l'effectif : c'est la catégorie la plus atypique.

Les **retraités** en portent 20,9 % en n'étant que 10 % — deux fois leur poids. Les **employés**, à l'inverse, en portent 17,3 % pour 34 % de l'effectif : ils sont proches du profil moyen.

Poids et inertie sont deux choses distinctes, et c'est leur écart qui est instructif.

Deux usages du même nombre

En statistique inférentielle, $\chi^2 = 110,4$ pour 12 degrés de liberté se compare à 21,0 : on rejette l'indépendance. La conclusion est **oui ou non**.

En analyse des données, le même 110,4 devient une inertie à **décomposer**. La question n'est plus « y a-t-il une liaison » mais « **laquelle** ».

Le test ferme la question ; l'analyse l'ouvre.

Le calcul factoriel

La matrice des écarts standardisés

$$\mathbf{S} = \mathbf{D}_r^{-1/2} (\mathbf{P} - \mathbf{r}\mathbf{c}') \mathbf{D}_c^{-1/2}$$

où $\mathbf{P} = (f_{ij})$, $\mathbf{r}$ et $\mathbf{c}$ sont les vecteurs de marges, $\mathbf{D}_r$ et $\mathbf{D}_c$ leurs matrices diagonales.

On lui applique la décomposition en valeurs singulières du chapitre 4 : $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$.

Trois remarques sur cette matrice

$\mathbf{P} - \mathbf{r}\mathbf{c}'$ est l'écart à l'indépendance : c'est le centrage propre à l'AFC.

Les deux facteurs $\mathbf{D}_r^{-1/2}$ et $\mathbf{D}_c^{-1/2}$ sont la métrique du khi-deux et la pondération par les marges.

Et $\sum_{i,j} s_{ij}^2 = \Phi^2$: la somme des carrés des éléments de $\mathbf{S}$ est l'inertie totale.

La décomposition de nos 0,3680

Axe	λ_k	Part	Cumul
1	0,2267	61,6 %	61,6 %
2	0,1030	28,0 %	89,6 %
3	0,0384	10,4 %	100 %

Le rang vaut $\min(5, 4) - 1 = 3$: trois axes au plus, quelle que soit la méthode.

Une valeur propre d'AFC est bornée par 1

$\sqrt{\lambda_k}$ est une **corrélation canonique** entre les deux variables qualitatives. Elle est donc comprise entre 0 et 1, et $\lambda_k \leq 1$ toujours.

Ici $\sqrt{\lambda_1} = 0,476$: la liaison est nette sans être écrasante. Une valeur propre proche de 1 signalerait deux variables presque confondues — donc redondantes.

Application économique

Deux axes portent 89,6 % de l'inertie. Le troisième, à 10,4 %, est le dernier possible : il ramasse ce qui reste, sans nécessairement porter du sens.

Le chapitre 10 lit le plan.

Ce qu'il faut retenir

Six points

1. En AFC, on ne compare pas des lignes mais des **profils** : la marge est éliminée.
2. Le **profil moyen est la marge**, et c'est le profil de l'indépendance.
3. La **distance du khi-deux** pondère chaque colonne par $1/f_{.j}$ — c'est le triplet $\mathbf{M} = \mathbf{D}_c^{-1}$.
4. Ce choix seul garantit l'**équivalence distributionnelle** : le résultat ne dépend pas du découpage des modalités.
5. $I = \Phi^2 = \chi^2/n$: l'inertie du nuage **est** le khi-deux rapporté à l'effectif. Ici 0,368.
6. Poids et inertie diffèrent : les cadres pèsent 24 % de l'effectif et portent 29,5 % de l'information.

La suite

Les axes sont calculés. Le chapitre 10 établit les relations barycentriques — celles qui autorisent, en AFC seulement, à lire lignes et colonnes **sur un même graphique**.