Analyse Factorielle des Correspondances (AFC)

Dr OM

LMI-SFA-UNA

2 février 2018

Structure de base des données

Objective mesurer des liaisons entre deux variables qualitatives :Khi-deux
exemple : Il s'agit de tester l'indépendance de deux variables qualitatives. Y a-t-il indépendance entre :

- la catégorie socioprofessionnelle et le vote à l'élection présidentielle ?
- le niveau d'études et les journaux lus ?

Êtes-vous	« pas du tout d'accord »	(1)
	« pas tellement d'accord »	(2)
	« peut-être d'accord »	(3)
	« bien d'accord »	(4)
	« entièrement d'accord »	(5)

avec cette phrase ? :

« On en a assez de ceux qui bloquent la vie du pays par leurs revendications ».

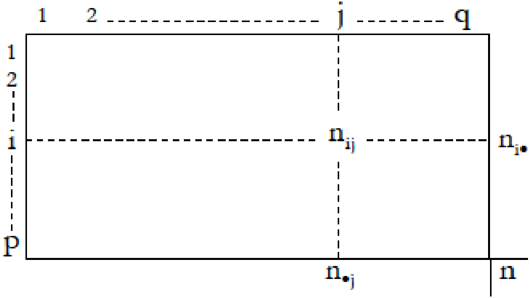
Existe-t-il un lien entre les réponses et la tendance politique ?

Tableau des profils lignes

Tendance politique	1	2	3	4	5	TOTAL
Extrême gauche	714	71	0	143	71	1 000
Gauche	284	216	199	174	127	1 000
Centre	87	106	228	335	244	1 000
Droite	16	86	156	271	471	1 000
Extrême droite	71	71	0	214	643	1 000
Indifférent	82	120	244	301	263	1 000
Non-réponse	88	124	269	285	233	1 000

Le tableau de contingence

Croisement de deux variables qualitatives I et J à p et q modalités.



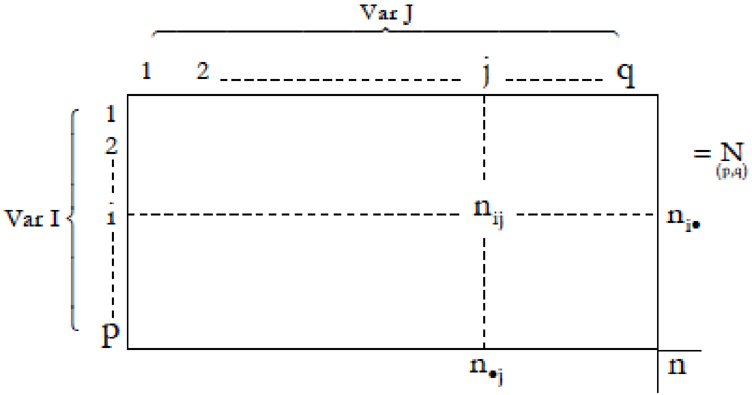
$$n_{i\bullet} = \sum_{j=1}^{q} n_{ij} \quad \text{(total ligne)}$$

$$n_{\bullet j} = \sum_{i=1}^{p} n_{ij} \quad \text{(total colonne)}$$

$$n = \sum_{i=1}^{p} \sum_{j=1}^{q} n_{ij} \quad \text{(total)}$$

Notations : tableau de contingence : N

Croisement de deux variables qualitatives à p et q modalités



Profils lignes - profils-colonnes - profils marginaux

- **p Profils des lignes** $\dfrac{n_{ij}}{n_{i\bullet}} = \dfrac{f_{ij}}{f_{i\bullet}}$

profil de la ligne i noté ℓ_i

$$\begin{pmatrix} \dfrac{n_{i1}}{n_{i\bullet}} & \dfrac{n_{i2}}{n_{i\bullet}} & \cdots\cdots & \dfrac{n_{iq}}{n_{i\bullet}} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \dfrac{f_{i1}}{f_{i\bullet}} & \dfrac{f_{i2}}{f_{i\bullet}} & \cdots\cdots & \dfrac{f_{iq}}{f_{i\bullet}} \end{pmatrix}$$

- **q Profils des colonnes**

profil de la colonne j

noté c_j

$$\begin{pmatrix} \dfrac{n_{1j}}{n_{\bullet j}} \\ \dfrac{n_{2j}}{n_{\bullet j}} \\ \dfrac{n_{pj}}{n_{\bullet j}} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \dfrac{f_{1j}}{f_{\bullet j}} \\ \dfrac{f_{2j}}{f_{\bullet j}} \\ \dfrac{f_{pj}}{f_{\bullet j}} \end{pmatrix}$$

Profils lignes - profils-colonnes - profils marginaux

Si les deux variables qualitatives I et J étaient indépendantes, les profils lignes seraient tous identiques, et donc identiques au profil marginal correspondant.

$$Independance \Rightarrow \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n} \Rightarrow n_{ij} = \frac{n_{\bullet j} * n_{i\bullet}}{n}$$

Profils lignes - profils-colonnes - profils marginaux

- On pouvait établir la relation précédente en raisonnant sur les profils colonnes.

$$\boxed{f_{ij} = f_{i\bullet}\, f_{\bullet j}}$$

$$\text{avec } f_{ij} = \frac{n_{ij}}{n} \text{ et } f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

$$f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

- Elle exprime clairement que dans le cas de l'indépendance le tableau de contingence est entièrement déterminé par ses marges

Définition du Khi-deux

- Pour chaque case, on peut donc calculer le nombre de cas attendus (sous hypothèse d'indépendance) $n_{ij} = \frac{n_{\bullet j} * n_{i\bullet}}{n}$
- On peut comparer les nombres de cas attendus E_{ij} aux nombres observés.

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - \frac{n_{i\bullet}\, n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet}\, n_{\bullet j}}{n}}$$

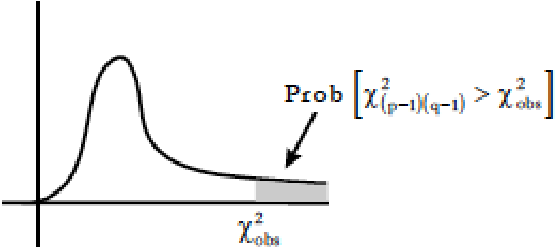
$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(n_{ij} - E_{ij}\right)^2}{E_{ij}}$$

Si les deux variables sont réellement indépendantes, cette expression suit une distribution du Khi-deux avec un nombre de degrés de liberté égal à : $(p-1)(q-1)$
Dans une table on lit $\chi^2_{\alpha,k}$ valeur ayant une probabilité α d'être dépassée pour une distribution du khi-deux avec $k=(p-1)(q-1)$ degrés de liberté.

1. Si $\chi^2 \leq \chi^2_{\alpha,k}$ On accepte H_o : independance
2. Si $\chi^2 > \chi^2_{\alpha,k}$ On rejette H_o : independance

Pratique sous logiciel statistique

- Calcul du χ^2 associé au tableau de contingence noté χ^2_{obs}.
- Probabilité pour une v.a. suivant une loi du khi-deux à $(p-1)(q-1)$ d.d.l. de dépasser χ^2_{obs}.



Si cette probabilité est faible (en général < 5 %), on rejette l'hypothèse d'indépendance entre les deux variables qualitatives.

AFC

- Répartition des habitants d'Abidjan selon leur lieu d'habitation et leur C.S.P.
  - Certains quartiers sont-ils proches ?
  - au sens même répartition des C.S.P. ?
  - Certaines C.S.P. sont-telles proches ?
  - Certaines C.S.P. sont-elles plus souvent associées à certains quartiers ?

L'analyse des correspondances traite des tableaux de contingence.

Représentation des profils lignes

- Les profils lignes sont considérés comme des individus.
- Les p profils-lignes forment un nuage de p points dans R^q
- A chaque profil-ligne est associé un poids égal à sa fréquence marginale profil ligne poids $f_{i\bullet}$.

On note $N(I)$ le nuage de points formé des profils-lignes pondérés : $(l_i; f_{i\bullet})$
Le centre de gravité g est défini par :

$$g_I = \sum_{i=1}^{p} f_{i\bullet} l_i$$

La jième coordonnée de g_I vaut $f_{\bullet j}$
Donc g_I = profil marginal de la variable J (à q modalités) $g_I = f_J$

Représentation des profils lignes

$N(J) =$ nuage de points formé des q profils - colonnes pondérés $\left(c_j, f_{\bullet j}\right)$

Le centre de gravité g_c est le profil marginal de la variable I à p modalités.

$$g_c = f_I$$

Le problème qui se pose est **l'étude de la dépendance** entre les deux variables qualitatives.

Dans le cas où les deux variables sont **indépendantes**, on a identité des profils :

$$(1) \qquad \frac{f_{ij}}{f_{i\bullet}} = f_{\bullet j} \qquad \text{profil - ligne}$$

$$(2) \qquad \frac{f_{ij}}{f_{\bullet j}} = f_{i\bullet} \qquad \text{profil - colonne}$$

$$\boxed{f_{ij} = f_{i\bullet}\, f_{\bullet j}}$$

Dans le cas de l'indépendance, le nuage des profils-lignes se réduit à un point g_ℓ

De même, le nuage des profils-colonnes se réduit à un point g_c.

$\Longrightarrow$ L'étude de la dépendance consiste à étudier la **forme des nuages**.

$\Longrightarrow$ Problème d'analyse en composantes principales.

Quelle métrique ?

Métrique du $\chi 2$

- Pour les profils lignes :

$$d^2_{\chi^2}(l_i, l_i') = \sum_{j=1}^{q} \frac{n}{n_{\bullet j}} (\frac{n_{ij}}{n_{i\bullet}} - \frac{n_{i'j}}{n_{i'\bullet}})^2$$

  - Donne un poids important aux différences portant sur les petits pourcentages.
  - Vérifie le principe d'équivalence distributionnelle : si deux colonnes ont le même profil, on les réunit en une seule d'effectif somme sans modifier les distances entre profilslignes.

- Pour les profils-colonnes :

$$d^2_{\chi^2}(c_j, c_j') = \sum_{j=1}^{q} \frac{n}{n_{i\bullet}} (\frac{n_{ij}}{n_{\bullet j}} - \frac{n_{ij'}}{n_{\bullet j'}})^2$$

Inertie du nuage N(I)

$I_{N(I)}$ l'inertie du nuage$N(I)$ calculée par rapport au centre de gravité f_J vaut

$$\frac{\chi^2}{N}$$

où $\chi^2 =$ Khi-deux associé au tableau de contingence étudié.
On obtient le même résultat pour l'inertie du nuage N(J).

l'A.C.P. du nuage des profils-lignes :

- Les profils-lignes jouent le rôle d'individus ; ils sont affectés des poids $f_{i\bullet}$
- La métrique utilisée pour le calcul des distances entre individus est la métrique du khi-deux
- Le premier axe principal du nuage des profils-lignes est la droite passant le plus près possible de l'ensemble des points de N(I).

l'A.C.P. du nuage des profils-lignes

Notons $\underline{a}^1$ la première composante principale

$$\underline{a}^1 = \begin{pmatrix} \vdots \\ \vdots \end{pmatrix} \leftarrow$$ coordonnées des p profils-lignes sur l'axe 1

Notons λ_1 la variance de $\underline{a}^1$ (égale à l'inertie portée par l'axe qui lui est associé).

$\underline{a}^2$ = deuxième composante principale de variance λ_2

$\underline{a}^3$ = troisième composante principale de variance λ_3

'A.C.P. du nuage des profils-colonnes

Notons $\underline{b}^1$ la première composante principale

$\underline{b}^1 = \begin{pmatrix} \vdots \end{pmatrix} \leftarrow$ coordonnées des q profils-colonnes sur l'axe 1

$\underline{b}^2 =$ deuxième composante principale

Les composantes principales de l'A.C.P. des profils-colonnes sont associées aux mêmes valeurs propres que les composantes principales de l'A.C.P. des profils-lignes.

$\underline{b}^1$ a pour variance λ_1

$\underline{b}^2$ a pour variance λ_2

Formules de transition

En notant b_j et a_i les j^{me} et i^{me} coordonnées des composantes principales b et a associées à la même valeur propre λ :

$$\sqrt{\lambda} b_i = \sum_{i=1}^{p} \frac{n_{ij}}{n_{\bullet j}}$$

$$\sqrt{\lambda} a_i = \sum_{j=1}^{q} \frac{n_{ij}}{n_{i\bullet}}$$

À λ près, la coordonnée d'une modalité i d'une variable est la moyenne des coordonnées des catégories de l'autre variable pondérées par les fréquences conditionnelles du profil de i.

Les modalités de la variable I sont représentées en tant qu'individus (profils-lignes) de l'A.C.P. des profils-lignes.

La modalité i de la variable I a pour coordonnées dans un espace de dimension k :

$\left(a_i^1 , a_i^2 , ... , a_i^k\right)$

avec $a_i^1 = i^{\text{ème}}$ coordonnée du vecteur $\underline{a}^1$

$a_i^2 = i^{\text{ème}}$ coordonnée du vecteur $\underline{a}^2$

.....

Les modalités de la variable I sont représentées en tant qu'individus (profils-lignes) de l'A.C.P. des profils-lignes.

Pour les modalités de la variable J, la modalité j a pour coordonnées :

$$\left(\sqrt{\lambda_1}\, b_j^1 \,,\, \sqrt{\lambda_2}\, b_j^2 \,,\, \ldots.. \sqrt{\lambda_k}\, b_j^k\right)$$

$$b_j^1 = j^{\text{ème}} \text{ coordonnée du vecteur } \underline{b}^1$$

$$b_j^2 = j^{\text{ème}} \text{ coordonnée du vecteur } \underline{b}^2$$

Les modalités du deuxième groupe (J) sont les barycentres des modalités du premier groupe (variable I).

(voir formules de transition)

Abandon du principe barycentrique

Les modalités de chaque ensemble sont représentées par les :

$$a_i^k, i = 1, \ldots, p$$

$$b_j^k, j = 1, \ldots, q$$

Cette représentation permet de déterminer les proximités entre certains éléments de I et certains éléments de J (compte tenu de la qualité de la représentation).

Contributions

de la ligne i à l'axe k

$$\frac{f_{i\bullet}\left(a_i^k\right)^2}{\lambda_k} \qquad \text{avec } f_{i\bullet} = \frac{n_{i\bullet}}{n}$$

de la colonne j à l'axe k

$$\frac{f_{\bullet j}\left(a_j^k\right)^2}{\lambda_k} \qquad \text{avec } f_{\bullet j} = \frac{n_{\bullet j}}{n}$$

Cosinus carrés

Modalité i représentée sur l'axe k

$$\frac{\left(a_i^k\right)^2}{d^2(i,G)}$$

Modalité j représentée sur l'axe k

$$\frac{\left(b_j^k\right)^2}{d^2(j,G)}$$

Aspects pratiques de l'interprétation

- L'interprétation peut se faire à partir des représentations graphiques (en s'assurant de la qualité de représentation de chaque modalité à l'aide des cos2).
- Quand le nombre de modalités est élevé, il est conseillé d'éditer d'abord le graphique des profils-lignes, puis celui des profilscolonnes, enfin la représentation simultanée.
- Les profils ayant des poids différents la lecture de leurs contributions à l'inertie de chaque axe s'avère très utile.
- On peut repérer les profils dont la contribution est supérieure au poids