Chapitre I : Analyse Descriptive : (Apprentissage Non Supervisé :Réduction de Dimensionnalité)

ANALAYSE FACTORIELLE DES CORRESPONDANCES: AFC

Introduction :

L'Analyse Factorielle des correspondances (AFC) introduite par J.P. Benzécri en 1962, est une forme particulière de l'ACP, adaptée au traitement de certain types de tableaux rectangulaires de données: les ***tableaux de contingence***.

- Elle permet d'étudier d'éventuelles relations existantes entre les variables qualitatives de n et m modalités:
- Si 2 variables qualitatives => *AFC Binaire*

  Sinon *AFC Multiple*
- Correspondance? Vars numériques=>Corrélation

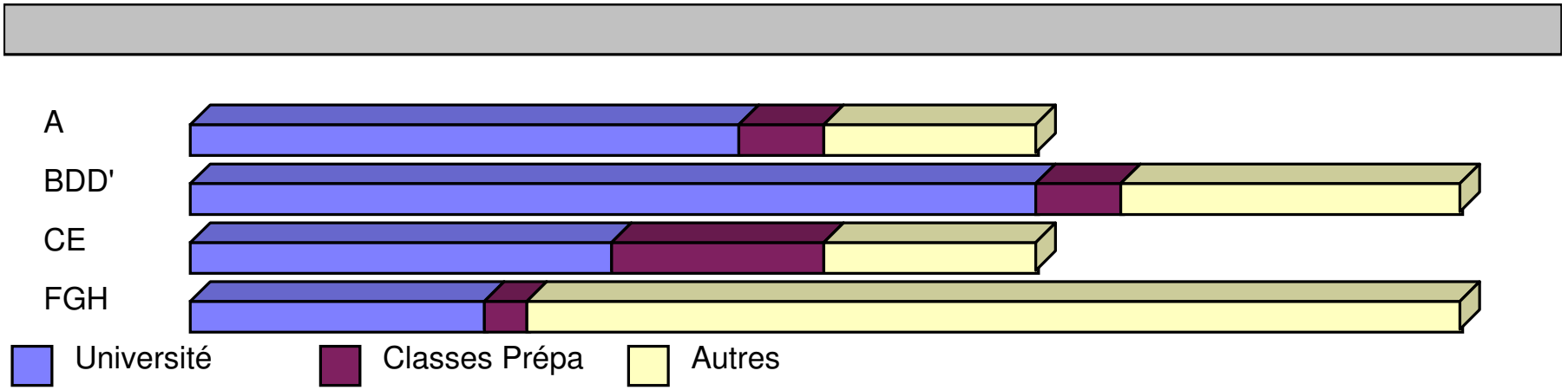
  Vars Nominales=> correspondance

Exemple : que deviennent les bacheliers ?

	destination			
	université	***classes prépa***	***autres***	***total***
A	13	2	5	**20**
BDD'	20	2	8	**30**
CE	10	5	5	**20**
FGH	7	1	22	**30**
total	**50**	**10**	**40**	**100**

Stats MEN 1975 - 1975 204 489 lycéens

Une représentation graphique intuitive .. Pas toujours suffisante …



Comment donner du sens à ces données?

Idée : ce qui est intéressant, c'est de mettre en évidence ce qui est inattendu dans ces répartitions

Inattendu = en quoi on dévie d'une répartition uniforme

On va donc

1. Évaluer ce que serait une situation d'uniformité
2. Calculer en quoi la situation constatée en diffère
3. Exprimer cette différence graphiquement pour pouvoir l'analyser, si possible en …
4. …. trouvant la meilleure manière de le faire

AFC: Appellation:

Pourquoi « des correspondances » ?

variables numériques $\Rightarrow$ Corrélation

variables nominales $\Rightarrow$ Correspondance

Pourquoi « factorielle » ?

Il s'agit de décomposer le tableau original en une somme de vecteurs/matrices qui sont chacun le produit de facteurs simples

Objectif de l'AFC Binaire:

- L'AFC consiste à rechercher la meilleure représentation **simultanée** de 2 ensembles constituants les lignes et les colonnes d'un tableau de contingence.

=>Dans ce cas la statistique classique nous donne par le test de **chi-deux** *le moyen de savoir si il existe une liaison entre les caractères étudiés mais ne permet pas de décrire cette liaison ce qui est précisément l'objet de l'AFC*

Objectif de l'AFC :

- Elle permet dans le plan des 2 premiers axes factoriels une représentation simultanée des ressemblances entre les lignes et colonnes et de la proximité entre eux.
- Elle permet également la visualisation et l'interprétation de la proximité entre modalités appartenant à la même variable

Principe de l'AFC:

- On observe 1 Population répartie sur 2 vars Qualit. Dans 1 tableau de contingence X de dimension np
- Et de terme général Kij.
- Pour chacune des 2vars on connait l'effectif de la population Kij : Ayant la modalité i de la var en ligne et la modalité j de la var en colonne.
- K est le total de la population.

	1	j	p	Total
1				
i		k_{ij}		Ki.
n				
Total		k.j		k

$$K_{i.} = \sum_{j=1}^{p} K_{ij} \qquad K_{.j} = \sum_{i=1}^{n} K_{ij} \qquad K = \sum_{j=1}^{p} K_{.j} = \sum_{i=1}^{n} K_{i.}$$

Principe de l'AFC: Tableau de Fréquence

- On préfère ramener le raisonnement en fréquence et donc l'effectif de la population en 1
- On construit le tableau de fréquence comme suit:

	1	j	p	Total
1				
i		Pij		Pi.
n				
Total		P.j		1

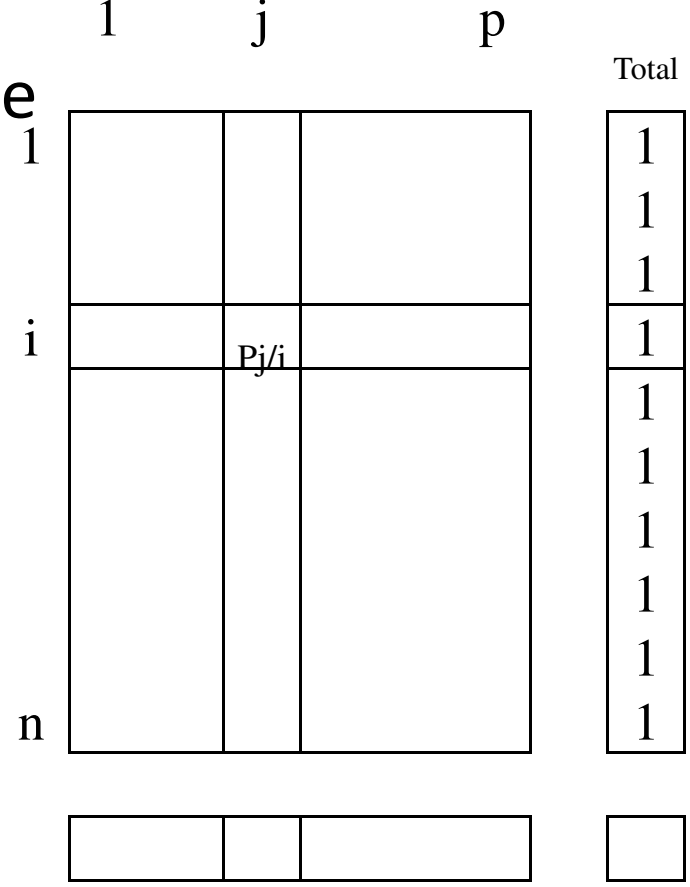
$$P_{ij} = \frac{K_{ij}}{K}$$

$$P_{i.} = \sum_{j=1}^{p} P_{ij} = \sum_{j=1}^{p} \frac{K_{ij}}{K} \qquad P_{.j} = \sum_{i=1}^{n} P_{ij} = \sum_{i=1}^{n} \frac{K_{ij}}{K}$$

Principe de l'AFC:
Analyse du nuage des profils en ligne

- On passe maintenant au tableau de Probabilité Conditionnelle ligne.

$$P_{j/i} = \frac{P_{ij}}{P_{i.}}$$



- L'AFC du tableau de départ sera donc 1 ACP de ce nouveau tableau
- Avec la distance de Chi-Deux.

Principe de l'AFC:

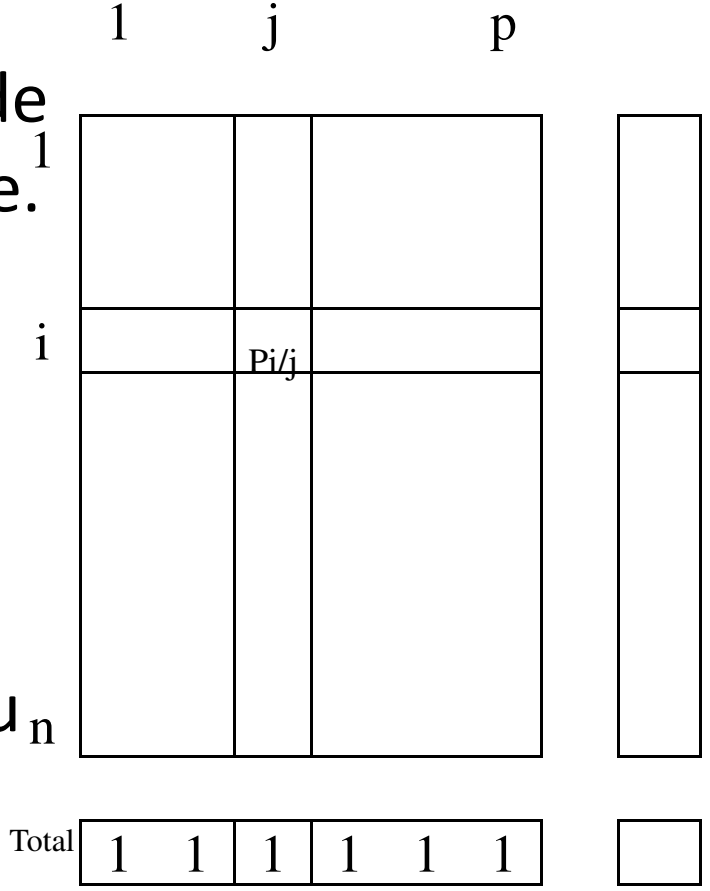
Analyse du nuage des profils en Colonne

- On passe maintenant au tableau de Probabilité Conditionnelle colonne.

$$P_{i/j} = \frac{P_{ij}}{P_{.j}}$$

- L'AFC du tableau de départ sera donc 1 ACP de ce nouveau tableau

Avec la distance de Chi-Deux.



➔ On va donc trouver 1 structure de dépendance entre ligne et colonne.

➔ On peut d abord tester l'hypothèse de non dépendance

➔ La méthode utilisée est Chi-Deux

Test de CHI-Deux

- Il s'applique sur le tableaux des effectifs.
- On doit poser une hypothèse:
  - H0 : V1 Indépendante de V2
  - H1 : V1 Dépendante de V2
  - ➢ Soit Oij = KxPij = Kij effectif observé(tableau initial)
  - ➢ Soit Eij = KxPi. x P.j est l'effectif théorique
  - ➢ I.E est l'indice de l'écart entre effectif Observé et effectif théorique:

$$\chi^2 \text{ cal} = I.E = \sum_{i}^{n}\sum_{j}^{p} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Test de CHI-Deux

- **Décision**:

➢ On regarde la table de Pearson pour(P-1;n-1) degré de liberté (ddl)et erreur α

  ➢ Si : **χ^2 cal (I.E) < χ^2 lu** ==> On accepte H0

  ➢ Si **χ^2 cal (I.E) >= χ^2 lu** ==> On accepte H1

  C.A.d on rejette l'indépendance ce qui justifie l'AFC qui permet de quantifier et visualiser la structure

Note sur le χ^2 : ses degrés de liberté

Définition:

- On appelle degré de liberté par ligne (ddll) le nombre de colonnes (de modalités) diminué de 1.
- On appelle degré de liberté par colonne (ddlc) le nombre de lignes (de modalités) diminué de 1.
- Le degré de liberté du khi-deux de la matrice est le produit ddll x ddlc = ddl.
- Pour une matrice donnée, le χ^2 à prendre en compte est en fait χ^2 / ddl

AFC: Distance entre 2 ligne dans R^p

- Il s'agit de calculer une matrice X de terme général : $$x_{ij} = \frac{P_{j/i}}{\sqrt{P_{.j}}}$$

❖ Ceci est pour avoir une mesure qui ne dépend pas du nombre des éléments d'une modalité. On pondère les éléments de matrice conditionnelle ligne par les poids relatif des modalités de l'ensemble

- En calcul ensuite :

$$d^2(i,i') = \sum_j^p (x_{ij} - x_{i'j})^2 = \sum_j^p \left[\frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}} - \frac{P_{ij}}{P_{i'.}\sqrt{P_{.j}}}\right]^2$$

AFC: Distance entre 2 colonnes dans R^n

- De la même façon en calcul:

$$d^2(j,j') = \sum_i^n \left(\frac{P_{i/j}}{\sqrt{P_{i.}}} - \frac{P_{i/j'}}{\sqrt{P_{i.}}} \right)^2 = \sum_i^n \frac{1}{P_{i.}} \left[\frac{P_{ij}}{P_{.j}} - \frac{P_{ij}}{P_{.j'}} \right]^2$$

Principe AFC:
Analyse du Nuage de Points dans R^p

- On retrouve la matrice X ou : $x_{ij} = \frac{P_{j/i}}{\sqrt{P_{.j}}}$
- On calcul la matrice de covariance XtX (pxp) de terme générique:

$$V_{jj'} = \sum_i P_{i.}(x_{ij} - \mu_{xj})(x_{ij'} - \mu_{xj'})$$

$$V_{jj'} = \sum_i^n r_{ij} r_{ij'} \qquad V_{jj'} = \sum_i^n \left[\frac{P_{ij}\, P_{ij'}}{P_{i.}\sqrt{P}_{.j}\sqrt{P_{.j'}}}\right] - \sqrt{P_{.j} P_{.j'}}$$

- Sachant que :

$$\mu_{xj} = \sqrt{P_{.j}}$$

AFC: Remarque 1:

$$Trace\,V = \sum_{j}^{p} V_{jj} = \sum_{j}^{p}\sum_{i}^{n} {r_{ij}}^2 = \sum_{j}^{p}\sum_{i}^{n} \left[\frac{P_{ij} - P_{i.}P_{.j}}{\sqrt{P_{i.}}\sqrt{P_{.j}}}\right]^2$$

- ==>

$$K.Trace\,V = \sum_{j}^{p}\sum_{i}^{n} \frac{K^2}{K}\left[\frac{P_{ij} - P_{i.}P_{.j}}{\sqrt{P_{i.}}\sqrt{P_{.j}}}\right]^2$$

$$K.Trace\,V = \sum_{j}^{p}\sum_{i}^{n} \left[\frac{\left(\boldsymbol{K}P_{ij} - \boldsymbol{K}P_{i.}P_{.j}\right)^2}{KP_{i.}P_{.j}}\right] = \sum_{j}^{p}\sum_{i}^{n} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} = I.E$$

- C'est l'indicateur d'écart du test χ^2 sous l'hypothèse d'indépendance.

AFC Remarque2:

- L’analyse du nuage des points dans R^n se fait de la même façon.
- les deux analyses peuvent être menées conjointement. Il est possible de représenter les deux ensembles: I et J, simultanément sur les plans factoriels, de telle sorte que la position d'un point de l'ensemble I (resp. J) y est interprétable par rapport à l'ensemble de tous les points de l'ensemble J (resp. I).

AFC Remarque 3:

- Les différentes notions présentées en ACP:
  - les coordonnées et plans factoriels (ou principaux)
  - les contributions
  - les éléments supplémentaires
- se retrouvent ici; elles conservent la même signification et s'utilisent de la même manière qu'en ACP.

Démonstration Vjj’:

$$V_{jj'} = \sum_{i}^{n} P_{i.}\left[\frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}} - \sqrt{P_{.j}}\right]\left[\frac{P_{ij'}}{P_{i.}\sqrt{P_{.j'}}} - \sqrt{P_{.j'}}\right]$$

$$V_{jj'} = \sum_{i}^{n} \frac{P_{i.}}{P_{i.}}\left[\frac{P_{ij} - P_{i.}P_{.j}}{\sqrt{P_{i.}}\sqrt{P_{.j}}}\right]\left[\frac{P_{ij'} - P_{i.}P_{.j'}}{\sqrt{P_{i.}}\sqrt{P_{.j'}}}\right]$$

$$V_{jj'} = \sum_{i} r_{ij} r_{ij'}$$

$$V_{jj'} = \sum_{i}^{n}\left[\frac{P_{ij}\,P_{ij'}}{P_{i.}\sqrt{P_{.j}}\sqrt{P_{.j'}}}\right] - \sqrt{P_{.j}P_{.j'}}$$

Démonstration $\mu_{xj} = \sqrt{P_{.j}}$

$$\mu_{xj} = \sum_{i}^{n} P_{i.} x_{ij} = \sum_{i}^{n} P_{i.} \frac{P_{ij}}{P_{i.}\sqrt{P_{.j}}} = \frac{1}{\sqrt{P_{.j}}} \sum_{i}^{n} P_{ij} = \frac{1}{\sqrt{P_{.j}}} P_{.j} = \sqrt{P_{.j}}$$