

Analyse Factorielle des Correspondances (AFC)

Jérôme Pagès (& François Husson)

Laboratoire de mathématiques appliquées - Agrocampus Rennes

husson@agrocampus-ouest.fr

Analyse Factorielle des Correspondances (AFC)

- 1 Données
- 2 Modèle d'indépendance
- 3 Les nuages et leur ajustement
- 4 Pourcentages d'inertie et inertie en AFC
- 5 Représentation simultanée des lignes et des colonnes
- 6 Aides à l'interprétation

Analyse Factorielle des Correspondances (AFC)

- 1 Données
- 2 Modèle d'indépendance
- 3 Les nuages et leur ajustement
- 4 Pourcentages d'inertie et inertie en AFC
- 5 Représentation simultanée des lignes et des colonnes
- 6 Aides à l'interprétation

Tableau de correspondances

		Ensemble J		
		1	j	J
Ensemble I	1			
	i		x_{ij}	
	I			

x_{ij} : nombre d'individus appartenant à l'élément i de l'ensemble I à l'élément j de l'ensemble J

Personnages de Mots

Phèdre (Racine)

Parfums

Milieux

Descripteur

Espèces

Nombre de fois que le personnage i a utilisé le mot j

Nombre de fois où le parfum i a été décrit par le mot j

Abondance de l'espèce j dans le milieu i

⇒ Exemples où le test d'indépendance du χ^2 peut être appliqué

Données historiques

Enquête du CREDOC (N. Tabard, 1974)

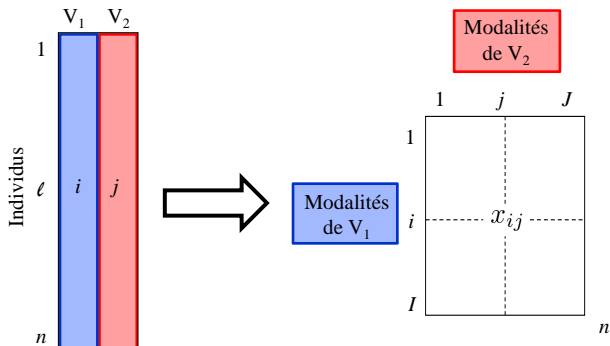
TABLEAU 37
REPONSES SIMULTANÉES A DES QUESTIONS D'OPINION

La famille idéale est celle où :	Activité convenant le mieux à une mère de famille quand les enfants vont à l'école :			
	rester au foyer	travailler à mi-temps	travailler à plein-temps	
les deux conjoints travaillent également	13	142	106	261
le mari a un métier plus absorbant que celui de sa femme	30	408	117	555
seul le mari travaille	241	573	94	908
	284	1 123	317	1 724

⇒ Etude de la liaison entre deux variables qualitatives

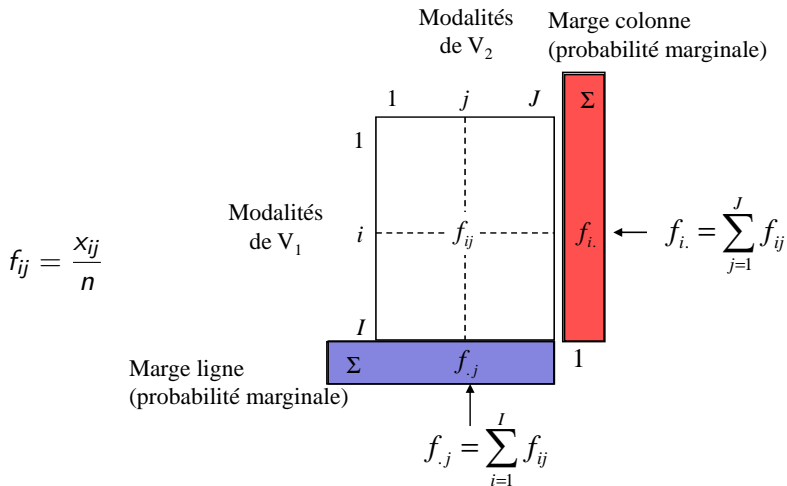
Données

n individus et 2 variables qualitatives



Distribution des n individus dans les $I \times J$ cases du tableau

Du tableau de contingences au tableau de probabilités



Liaison entre V_1 et V_2 : écart entre les données observées et le modèle d'indépendance

Analyse Factorielle des Correspondances (AFC)

- 1 Données
- 2 Modèle d'indépendance**
- 3 Les nuages et leur ajustement
- 4 Pourcentages d'inertie et inertie en AFC
- 5 Représentation simultanée des lignes et des colonnes
- 6 Aides à l'interprétation

Liaisons et indépendance entre deux variables qualitatives

Modèle d'indépendance :

Evènements indépendants : $P(A \text{ et } B) = P(A) \times P(B)$

Variables qualitatives indépendantes : $\forall i, \forall j, f_{ij} = f_{i.} \times f_{.j}$

\Rightarrow Probabilité conjointe = produit des probabilités marginales

Autres écritures : $\frac{f_{ij}}{f_{i.}} = f_{.j}$ $\frac{f_{ij}}{f_{.j}} = f_{i.}$

\Rightarrow Probabilité conditionnelle = probabilité marginale

Liaisons entre deux variables qualitatives

Ecart entre données obs (f_{ij}) et modèle d'indépendance (f_i, f_j)

- ① Significativité de la liaison (de l'écart) : test du χ^2

$$\chi_{obs}^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(\text{eff. observé} - \text{eff. théorique})^2}{\text{effectif théorique}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(n f_{ij} - n f_i \cdot f_j)^2}{n f_i \cdot f_j}$$

$$\chi_{obs}^2 = \sum_{i=1}^I \sum_{j=1}^J n \frac{(\text{probabilité observée} - \text{probabilité théorique})^2}{\text{probabilité théorique}} = n \Phi^2$$

- ② Intensité de la liaison = Φ^2 = écart entre probabilités théoriques et observées
- ③ Nature de la liaison = association entre modalités

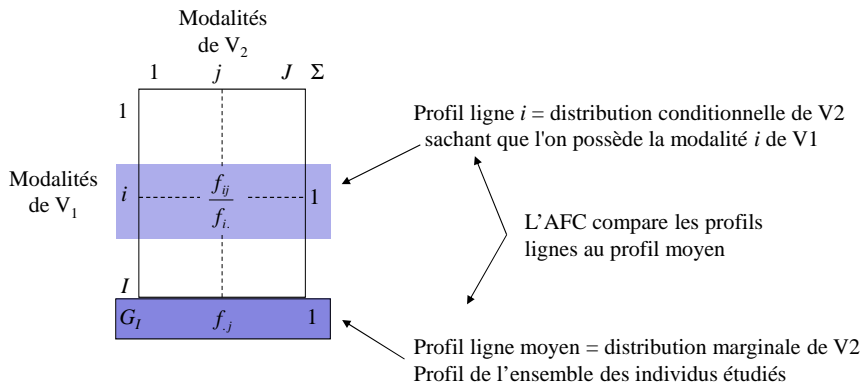
L'AFC travaille sur le tableau des probabilités

ne dit rien sur la significativité

visualise la nature de la liaison entre les deux variables

Comment l'AFC appréhende l'écart à l'indépendance ?

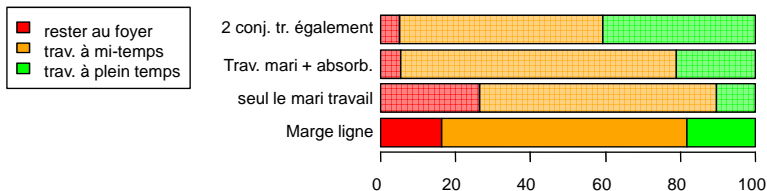
$$\text{Analyse par lignes : } \frac{f_{ij}}{f_{i.}} = f_{.j}$$



Approche multidimensionnelle de l'écart à l'indépendance

Comparaison du profil ligne au profil moyen

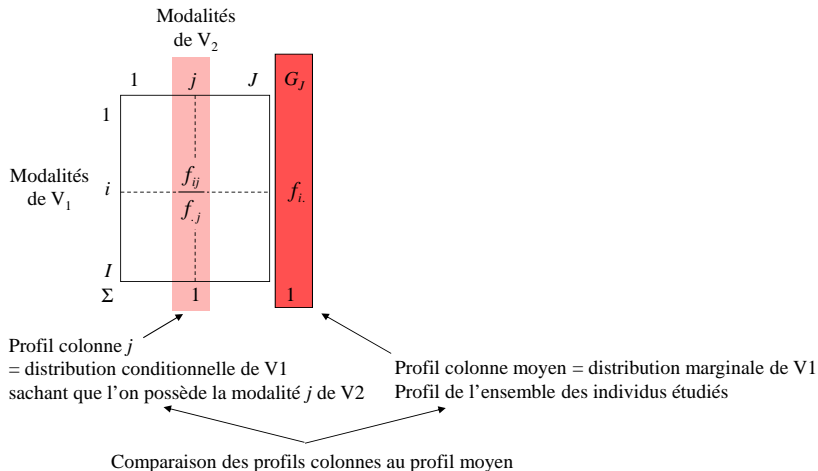
	rester au foyer	trav. à mi-temps	trav. à plein temps
2 conj. tr. également	4.98	54.41	40.61
trav. mari + absorbant	5.41	73.51	21.08
seul le mari travaille	26.54	63.11	10.35
marge ligne	16.47	65.14	18.39



Les femmes qui répondent *seul le mari travaille* répondent-elles de façon particulière à la question sur l'activité d'une mère de famille ?

Comment l'AFC appréhende l'écart à l'indépendance ?

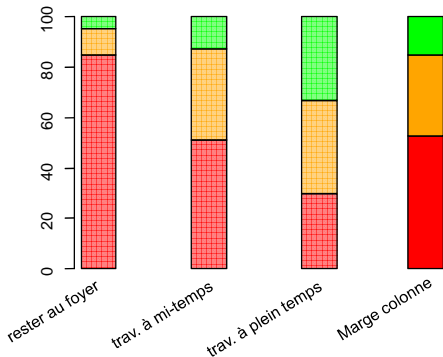
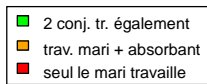
Analyse par colonnes : $\frac{f_{ij}}{f_{.j}} = f_{i.}$



Approche multidimensionnelle de l'écart à l'indépendance

Comparaison du profil colonne au profil moyen

	rester au foyer	trav. à mi-temps	trav. à plein temps	marge colonne
2 conj. tr. également	4.58	12.64	33.44	15.14
trav. mari + absorbant	10.56	36.33	36.91	32.19
seul le mari travaille	84.86	51.02	29.65	52.67

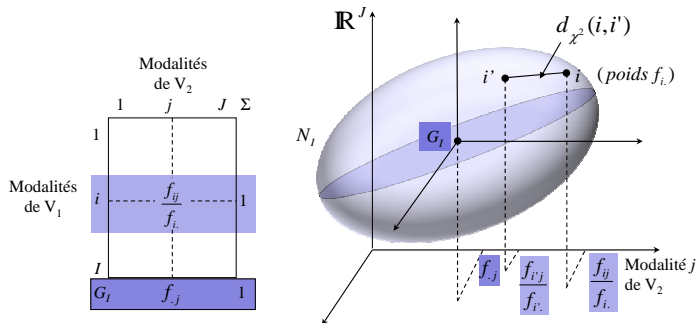


Les femmes qui répondent *travailler à mi-temps* répondent-elles de façon particulière à la question sur la famille idéale ?

Analyse Factorielle des Correspondances (AFC)

- 1 Données
- 2 Modèle d'indépendance
- 3 Les nuages et leur ajustement**
- 4 Pourcentages d'inertie et inertie en AFC
- 5 Représentation simultanée des lignes et des colonnes
- 6 Aides à l'interprétation

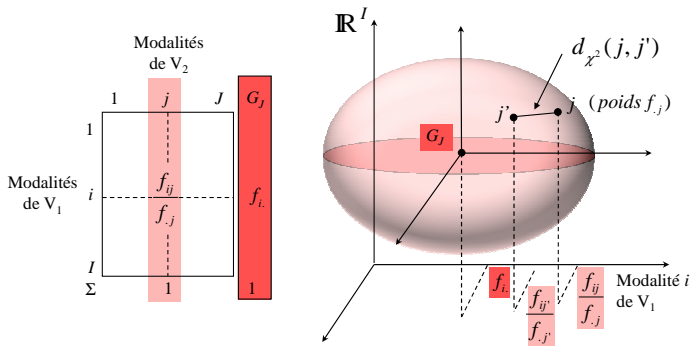
Le nuage des (profils) lignes



$$\text{Distance entre deux profils : } d_{\chi^2}^2(i, i') = \sum_{j=1}^J \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2$$

$$\text{Distance au profil moyen } G_I : d_{\chi^2}^2(i, G_I) = \sum_{j=1}^J \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - f_j \right)^2$$

Le nuage des (profils) colonnes



$$\text{Distance entre deux profils : } d_{\chi^2}^2(j, j') = \sum_{i=1}^I \frac{1}{f_i} \left(\frac{f_{ij}}{f_j} - \frac{f_{ij'}}{f_{j'}} \right)^2$$

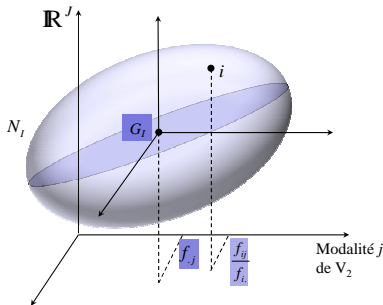
$$\text{Distance au profil moyen } G_J : d_{\chi^2}^2(j, G_J) = \sum_{i=1}^I \frac{1}{f_i} \left(\frac{f_{ij}}{f_j} - f_i \right)^2$$

Que se passe-t-il s'il y a indépendance ?

$$\text{Pour tout } i, \frac{f_{ij}}{f_i} = f_j$$

⇒ les profils sont confondus avec le profil moyen ⇒ N_i réduit à G_I

⇒ L'inertie du nuage est nulle



Idem pour les colonnes : pour tout j , $\frac{f_{ij}}{f_j} = f_i$.

Ecart à l'indépendance et inertie

Plus les données s'écartent de l'indépendance et plus les profils s'écartent de l'origine

$$\begin{aligned}
 \text{Inertie}(N_I/G_I) &= \sum_{i=1}^I \text{Inertie}(i/G_I) = \sum_{i=1}^I f_i \cdot d_{\chi^2}^2(i, G_I) \\
 &= \sum_{i=1}^I f_i \cdot \left(\sum_{j=1}^J \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - f_j \right)^2 \right) \\
 &= \sum_{i=1}^I \sum_{j=1}^J \frac{(f_{ij} - f_i \cdot f_j)^2}{f_i \cdot f_j} = \frac{\chi^2}{n} = \phi^2
 \end{aligned}$$

ϕ^2 mesure l'intensité de la liaison

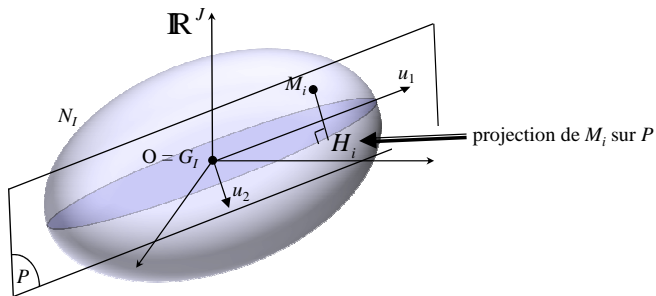
Etudier l'inertie de N_I revient à étudier l'écart à l'indépendance

Idem pour N_J : $\text{Inertie}(N_J/G_J) = \text{Inertie}(N_I/G_I)$ (dualité)

Représentation du nuage des lignes (ou des colonnes)

Décomposition de l'inertie de N_I par analyse factorielle

Projection de N_I sur une suite d'axes orthogonaux d'inertie maximum



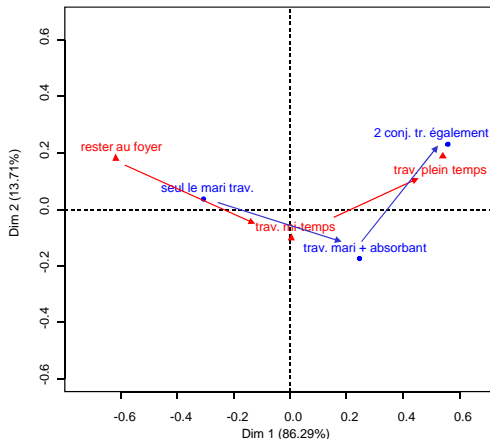
Trouver P tel que $\sum_{i=1}^I f_i \cdot (OH_i)^2$ est maximum

u_1 axe d'inertie maximum

u_2 axe d'inertie maximum avec $u_2 \perp u_1$

Inertie associée à l'axe s : $\sum_{i=1}^I f_i \cdot (OH_i^s)^2 = \lambda_s$

Règles d'interprétation sur l'exemple



1er axe :
attitude à l'égard
du travail féminin
(lignes et colonnes)
⇒ Dualité

	rester au foyer	trav. à mi-temps	trav. à plein temps	$O = G_J$
2 conj. tr. également	4.58	12.64	33.44	15.14
trav. mari + absorbant	10.56	36.33	36.91	32.19
seul le mari travaille	84.86	51.02	29.65	52.67

Analyse Factorielle des Correspondances (AFC)

- 1 Données
- 2 Modèle d'indépendance
- 3 Les nuages et leur ajustement
- 4 Pourcentages d'inertie et inertie en AFC**
- 5 Représentation simultanée des lignes et des colonnes
- 6 Aides à l'interprétation

Pourcentages d'inertie

- 1 Qualité de représentation de N_I par l'axe de rang s

$$\frac{\text{inertie projetée de } N_I \text{ sur } u_s}{\text{inertie totale de } N_I} = \frac{\sum_{i=1}^I f_i \cdot (OH_i^s)^2}{\sum_{i=1}^I f_i \cdot (OM_i)^2} = \frac{\lambda_s}{\sum_{i=k}^K \lambda_k}$$

	Inertie	Inertie (%)
F1	0.117	86.292
F2	0.019	13.708
Somme	0.135	100

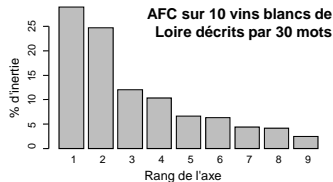
⇒ Ecart à l'indépendance bien résumé par le premier axe (86 %)

- 2 Inerties projetées s'additionnent d'un axe à l'autre (axes orthogonaux)

$$\sum_{k=1}^K \lambda_k = \text{Inertie } (N_I) = \Phi^2$$

Ici $n\Phi^2 = 1724 \times 0.135 = \chi^2 = 233.43 \Rightarrow \text{Proba. critique} = 10^{-49}$

- 3 La décroissance des inerties suggère le nombre d'axes à conserver

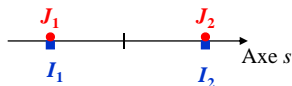
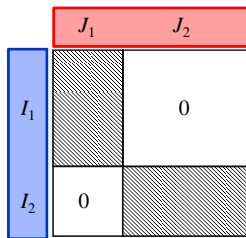


Inerties (= valeurs propres)

En AFC : $0 \leq \lambda_s \leq 1$

En ACP (normée) : $1 \leq \lambda_s$

A quelle structure correspond une valeur propre de 1 ?



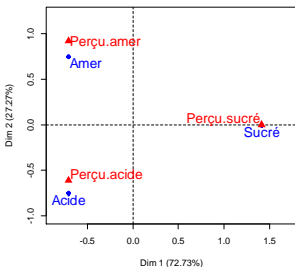
⇒ Partition en deux classes des lignes et des colonnes
Association exclusive des classes

Inerties (= valeurs propres)

Données : reconnaissance de trois saveurs (sucré, acide, amer)
 Pour chaque saveur, on a demandé à dix personnes de reconnaître la saveur d'une solution qui leur était présentée

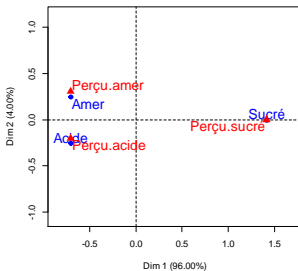
	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	9	1
Amer	0	3	7

AFC	V. Propre	%
Axe 1	1	72,727
Axe 2	0,375	27,273
Somme	1,375	100



	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100



Inerties (= valeurs propres)

TABLEAU 37
REPNSES SIMULTANÉES A DES QUESTIONS D'OPINION

La famille idéale est celle où :	Activité convenant le mieux à une mère de famille quand les enfants vont à l'école :			
	rester au foyer	travailler à mi-temps	travailler à plein-temps	
les deux conjoints travaillent également	13	142	106	261
le mari a un métier plus absorbant que celui de sa femme	30	408	117	555
seul le mari travaille	241	573	94	908
	284	1 123	317	1 724

	Inertie	Inertie (%)
F1	0.117	86.292
F2	0.019	13.708
Somme	0.135	100

$\lambda_1 = 0.117 \ll 1 \Rightarrow$ on est loin d'une association exclusive entre une ligne et une colonne

$\phi^2 = 0.135 \ll 2 \Rightarrow$ on est loin d'une liaison parfaite, i.e. d'une association exclusive entre les modalités des deux variables

Analyse Factorielle des Correspondances (AFC)

- 1 Données
- 2 Modèle d'indépendance
- 3 Les nuages et leur ajustement
- 4 Pourcentages d'inertie et inertie en AFC
- 5 Représentation simultanée des lignes et des colonnes**
- 6 Aides à l'interprétation

Représentation simultanée des lignes et colonnes

Relation de transition = propriétés barycentriques

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \underbrace{\sum_{j=1}^J \frac{f_{ij}}{f_i} G_s(j)}_{}$$

$F_s(i)$: coord. de la ligne i sur l'axe de rang s
 $\frac{f_{ij}}{f_i}$: jème élément du profil i
 $G_s(j)$: coord. de la colonne j sur l'axe de rang s
 λ_s : inertie associée à l'axe s (en AFC $\lambda_s \leq 1$)

Le long de l'axe de rang s , on calcule le barycentre de toutes les colonnes, chaque colonne j étant affectée du poids f_{ij}/f_i .

Le barycentre est ensuite d'autant plus écarté de l'origine que λ_s est petit : $1/\sqrt{\lambda_s} \geq 1$

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^I \frac{f_{ij}}{f_j} F_s(i)$$

Représentation simultanée et inertie

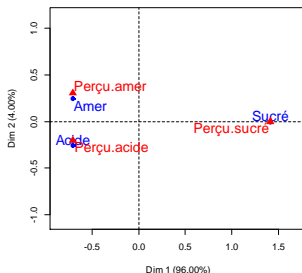
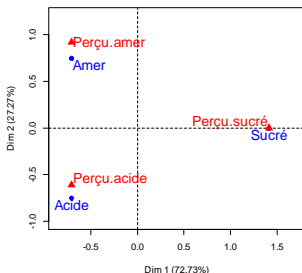
$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^I \frac{f_{ij}}{f_{.j}} F_s(i)$$

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	9	1
Amer	0	3	7

AFC	V. Propre	%
Axe 1	1	72,727
Axe 2	0,375	27,273
Somme	1,375	100

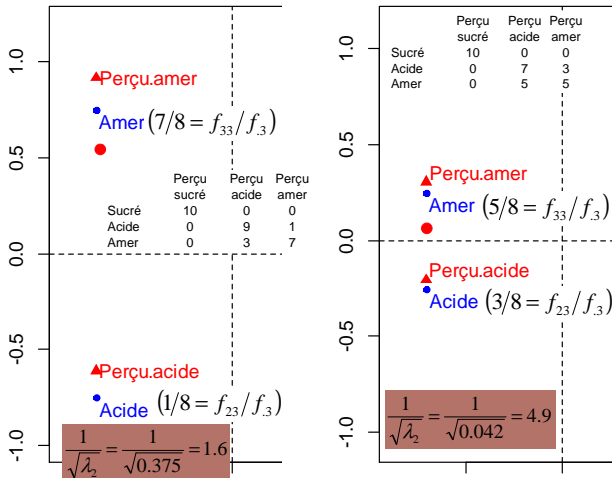
	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100

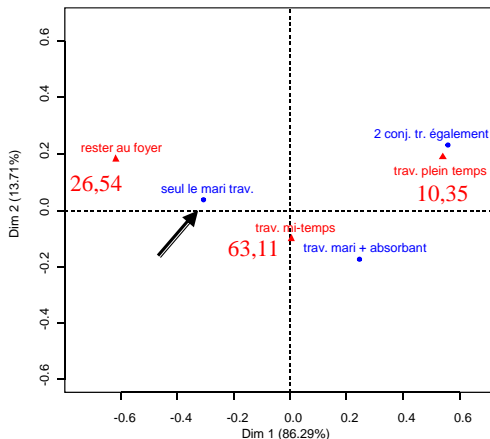


Représentation simultanée et inertie

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^I \frac{f_{ij}}{f_{.j}} F_s(i)$$

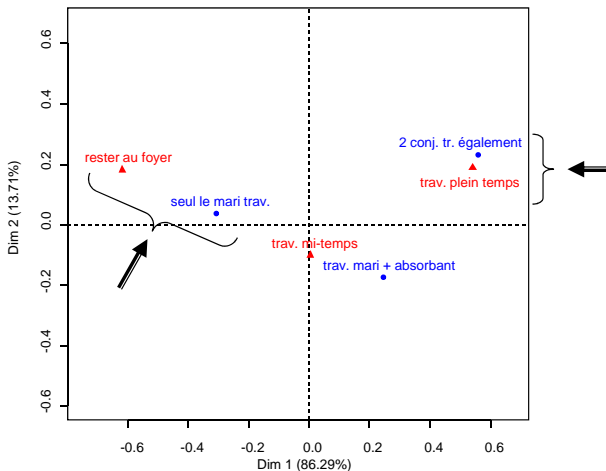


Propriété barycentrique



	rester au foyer	trav. à mi-temps	trav. à plein temps	Σ
2 conj. tr. également	4.98	54.41	40.61	100
trav. mari + absorbant	5.41	73.51	21.08	100
seul le mari travaille	26.54	63.11	10.35	100
$O = G_I$	16.47	65.14	18.39	100

Propriété barycentrique



Le 1er axe classe les modalités (des deux questions) depuis la plus défavorable au travail féminin jusqu'à la plus favorable

Retour aux données

Pourquoi *rester au foyer* est plus défavorable à l'égard du travail féminin que *seul le mari travaille* ?

TABLEAU 37
REPNSES SIMULTANÉES A DES QUESTIONS D'OPINION

La famille idéale est celle où :	Activité convenant le mieux à une mère de famille quand les enfants vont à l'école :			
	rester au foyer	travailler à mi-temps	travailler à plein-temps	
les deux conjoints travaillent également	13	142	106	261
le mari a un métier plus absorbant que celui de sa femme	30	408	117	555
seul le mari travaille	241	573	94	908
	284	1 123	317	1 724

Dans \mathbb{R}^I , $d_{\chi^2}^2(\text{rester au foyer}, G_J) = 0.416$

Dans \mathbb{R}^J , $d_{\chi^2}^2(\text{seul le mari travaille}, G_I) = 0.097$

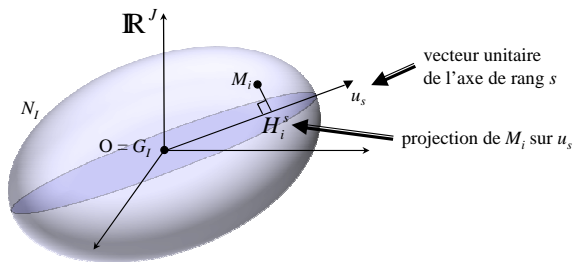
Analyse Factorielle des Correspondances (AFC)

- 1 Données
- 2 Modèle d'indépendance
- 3 Les nuages et leur ajustement
- 4 Pourcentages d'inertie et inertie en AFC
- 5 Représentation simultanée des lignes et des colonnes
- 6 Aides à l'interprétation**

Aides à l'interprétation : qualité de représentation

Indicateur de qualité de représentation d'un point (idem nuage) :

$$\frac{\text{inertie projetée de } M_i \text{ sur } u_s}{\text{inertie totale de } M_i} = \frac{f_i \cdot (OH_i^s)^2}{f_i \cdot (OM_i)^2} = \cos^2(\overrightarrow{OM_i}, u_s)$$

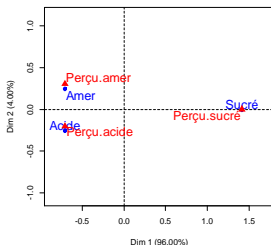


Indicateur montre dans quelle mesure l'écart d'un profil au profil moyen est complètement représenté par l'axe (ou par un plan)

Qualité de représentation : exemple

	Perçu sucré	Perçu acide	Perçu amer
Sucré	10	0	0
Acide	0	7	3
Amer	0	5	5

AFC	V. Propre	%
Axe 1	1	96
Axe 2	0,042	4
Somme	1,042	100



Qualité de représentation
(\cos^2)

	Axe1	Axe2
Sucré	1.000	0.000
Acide	0.889	0.111
Amer	0.889	0.111
Perçu.sucré	1.000	0.000
Perçu.acide	0.923	0.077
Perçu.amer	0.842	0.152

⇒ Interprétation des graphes basée sur points remarquables ayant une bonne qualité de représentation

Aides à l'interprétation : contribution

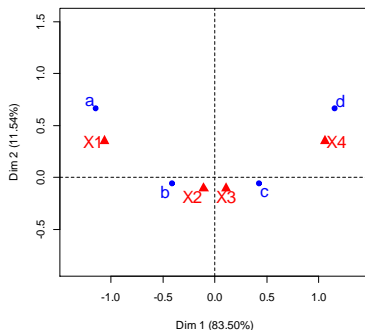
Indicateur brut : inertie projetée de M_i sur $u_s = f_i.(OH_i^s)^2$

Indicateur relatif : $\frac{\text{inertie proj. de } M_i \text{ sur } u_s}{\text{inertie de l'axe } s} = \frac{f_i.(OH_i^s)^2}{\lambda_s}$

- On peut additionner les contributions de plusieurs éléments
- Elles indiquent dans quelle mesure on peut considérer qu'un axe est dû à un élément ou à quelques éléments
- Compromis opérationnel entre distance à l'origine et poids
- Utiles pour les grands tableaux pour sélectionner un sous-ensemble d'éléments au début de l'interprétation (conjointement à la qualité de représentation)

Contribution : exemple

	X1	X2	X3	X4
a	1	1	0	0
b	5	10	10	0
c	0	10	10	5
d	0	0	1	1



	Inertie	%
Axe 1	0.258	83.501
Axe 2	0.036	11.538
Axe 3	0.015	4.96

	Axe1	Axe2
a	18.879	46.296
b	31.121	3.704
c	31.121	3.704
d	18.879	46.296
Σ	100	100

⇒ Les points extrêmes ne sont pas nécessairement ceux qui contribuent le plus à la construction des axes

Equivalence distributionnelle

Equivalence distributionnelle : si plusieurs lignes ayant le même profil sont regroupées en une seule, les résultats de l'AFC sont strictement équivalents (idem pour le regroupement de colonnes)

Application en analyse textuelle :

Grâce à l'équivalence distributionnelle, si 2 mots (ou plus) sont employés dans les mêmes circonstances, leurs coordonnées sont proches et faire l'analyse avec les deux termes ou avec un terme unique qui regroupe ces deux notions est strictement équivalent
⇒ notion très utile (regroupement des singuliers et pluriels, des conjugaisons des verbes, etc.)

Nombre maximum d'axes et V de Cramer

Nuage des lignes : I points dans un espace à J dimensions

$$\left. \begin{array}{l} J \text{ dim. mais 1 contrainte (profils)} \Rightarrow S \leq J - 1 \\ I \text{ points évoluent dans au plus } I - 1 \text{ dim.} \Rightarrow S \leq I - 1 \end{array} \right\} S \leq \min(I-1, J-1)$$

$$\Rightarrow \Phi^2 = \sum_{k=1}^{\min(I-1, J-1)} \lambda_k \leq \min(I-1, J-1)$$

d'où l'idée d'un indicateur borné de la liaison entre 2 variables :

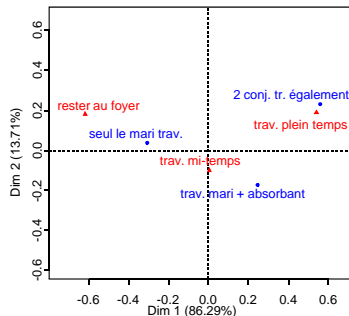
$$V \text{ de Cramer} = \frac{\Phi^2}{\min(I-1, J-1)} \in [0; 1]$$

Travail féminin			Trois saveurs			Trois saveurs		
$V = 0.135/2 = 0.0675$			$V = 1.375/2 = 0.6875$			$V = 1.042/2 = 0.521$		
AFC	V. Propre	%	AFC	V. Propre	%	AFC	V. Propre	%
Axe 1	0,117	86,292	Axe 1	1	72,727	Axe 1	1	96
Axe 2	0,019	13,708	Axe 2	0,375	27,273	Axe 2	0,042	4
Somme	0,135	100	Somme	1,375	100	Somme	1,042	100

Bilan sur l'exemple

TABLEAU 37
REPNSES SIMULTANÉES A DES QUESTIONS D'OPINION

La famille idéale est celle où :	Activité convenant le mieux à une mère de famille quand les enfants vont à l'école :			
	rester au foyer	travailler à mi-temps	travailler à plein-temps	
les deux conjoints travaillent également	13	142	106	261
le mari a un métier plus absorbant que celui de sa femme	30	408	117	555
seul le mari travaille	241	573	94	908
	284	1 123	317	1 724



L'AFC apporte une visualisation synthétique de l'écart à l'indépendance qui aide la compréhension du tableau (a fortiori avec de grands tableaux)

Sur ces données

- L'essentiel de l'écart à l'indépendance est structuré par l'attitude à l'égard du travail féminin
- La position des modalités le long de l'échelle d'attitude éclaire leur signification
Exemple : la proximité de *travailler à mi-temps* avec le profil moyen suggère que cette modalité est « neutre », à la différence de l'autre modalité moyenne (plutôt favorable au travail féminin)

Conclusion

Pour étudier la liaison entre deux variables qualitatives, on construit un tableau de contingence

Cette liaison réside dans l'écart entre le tableau de contingence et le modèle d'indépendance

L'analyse des correspondances :

- construit un nuage des lignes (et un nuage des colonnes) dont l'inertie totale mesure l'intensité de l'écart à l'indépendance
- décompose cette inertie totale sur une suite d'axes d'importance décroissante représentant chacun un aspect synthétique de la liaison entre les deux variables
- fournit une représentation des lignes et des colonnes dans laquelle la position d'un point reflète sa participation à l'écart à l'indépendance

Bibliographie

Pour approfondir l'analyse des correspondances dans le même esprit que cette vidéo :



Escofier B. & Pagès J. (2008)
Analyses factorielles simples et multiples ; objectifs, méthodes et interprétation.
4e édition. 318 p., Dunod, Paris.



Husson F., Lê S. & Pagès J. (2016)
Analyse des données avec R, 2e édition
240 p., Presses Universitaires de Rennes.