

Analyse des données

Filière: IAGI

Année universitaire: 2021-2022

AZMI Mohamed

Mohamed.azmi@ensam-casa.ma

Chapitre II : AFC

Analyse Factorielle des Correspondances

Introduction:

De l'ACP à l'AFC

ACP : Méthode factorielle de réduction de dimension pour l'exploration statistique des tables de données où les lignes représentent des individus et **les colonnes représentent des variables quantitatives**.

AFC : Méthode factorielle de réduction de dimension pour l'exploration statistique **d'une table de contingence définie par deux variables qualitatives**.

| X / Y | y_1 | | y_j | | y_J | Total |
|-------|----------|--|----------|--|----------|----------|
| x_1 | n_{11} | | n_{1j} | | n_{1J} | $n_{1.}$ |
| | | | | | | |
| x_i | n_{i1} | | n_{ij} | | n_{iJ} | $n_{i.}$ |
| | | | | | | |
| x_I | n_{I1} | | n_{Ij} | | n_{IJ} | $n_{I.}$ |
| Total | $n_{.1}$ | | $n_{.j}$ | | $n_{.J}$ | n |

n_{ij} = nombre d'observations ayant les modalités x_i et y_j

$n_{i.} = \sum_{j=1}^J n_{ij}$ Effectif marginal de la modalité x_i
= nombre d'observations ayant la modalité x_i

$n_{.j} = \sum_{i=1}^I n_{ij}$ Effectif marginal de la modalité y_j
= nombre d'observations ayant la modalité y_j

Introduction:

Données-Notation-Exemples

n_{ij} = nombre d'observations ayant les modalités x_i et y_j

$n_{i.} = \sum_{j=1}^J n_{ij}$ Effectif marginal;
= nombre d'observations ayant la modalité x_i

$n_{.j} = \sum_{i=1}^I n_{ij}$ Effectif marginal;
= nombre d'observations ayant la modalité y_j

$$n = \sum_{i,j}^{I,J} n_{ij}$$

$f_{ij} = \frac{n_{ij}}{n}$: probabilité d'avoir les modalités x_i et y_j

$f_{i.} = \sum_{j=1}^J \frac{n_{ij}}{n} = \frac{n_{i.}}{n}$: probabilité marginal *de* la modalité x_i

$f_{.j} = \sum_{i=1}^I \frac{n_{ij}}{n} = \frac{n_{.j}}{n}$: probabilité marginal *de* la modalité y_j

$$f_{..} = \sum_{i,j}^{I,J} \frac{n_{ij}}{n} = 1$$

Introduction:

Données-Notation-Exemples

| | | L'activité la plus appropriée pour une mère quand les enfants vont à l'école | | | |
|-------------------|--|--|-------------------------|-----------------------|-------------|
| | | Rester à la maison | Travail à temps partiel | Travail à temps plein | Total |
| la famille idéale | Les parents travaillent de façon égale | 13 | 142 | 106 | 261 |
| | Mari travaille plus | 30 | 408 | 117 | 555 |
| | Seul le mari travaille | 241 | 573 | 94 | 908 |
| Total | | 284 | 1123 | 317 | 1724 |

Enquête menée dans un contexte marqué par plusieurs luttes féministes, notamment en ce qui concerne l'accès des femmes au travail rémunéré (en France, les femmes ne pouvaient pas travailler sans le consentement de leur mari avant 1965).

Le tableau croise les réponses à deux questions :

- Quelle est votre perception de la famille idéale?
- Quelle est l'activité la plus appropriée pour une mère quand les enfants vont à l'école?

Introduction:

Données-Notation-Exemples

| | | L'activité la plus appropriée pour une mère quand les enfants vont à l'école | | | |
|-------------------|--|--|-------------------------|-----------------------|-------|
| | | Rester à la maison | Travail à temps partiel | Travail à temps plein | Total |
| la famille idéale | Les parents travaillent de façon égale | 13 | 142 | 106 | 261 |
| | Mari travaille plus | 30 | 408 | 117 | 555 |
| | Seul le mari travaille | 241 | 573 | 94 | 908 |
| Total | | 284 | 1123 | 317 | 1724 |

Quelle est la fréquence la plus importante?
Qu'est ce que vous en déduisez?

Introduction:

Le modèle d'indépendance et le test χ^2

L'étude du lien entre deux variables nécessite de positionner les données en fonction d'un point de départ donné ; dans ce cas, **l'absence de relation de dépendance**.

Le modèle d'indépendance précise ce critère comme point de départ. La relation standard d'indépendance entre deux événements $P[A \cap B] = P[A] \times P[B]$ est applicable à deux variables qualitatives.

Deux variables qualitatives sont indépendantes si : $\forall i; j \ f_{ij} = f_{i.} \times f_{.j} \Leftrightarrow \frac{n_{ij}}{n} = \frac{n_{i.}}{n} \times \frac{n_{.j}}{n}$

Ce qui veut dire que la probabilité conjointe f_{ij} **dépend uniquement des probabilités marginales** $f_{i.}$ et $f_{.j}$

Ce qui revient à comparer les **effectifs réels** n_{ij} aux **effectifs théoriques** du modèle de l'indépendance $\frac{n_{i.} \times n_{.j}}{n}$

Introduction:

Le modèle d'indépendance et le test χ^2

Table des effectifs réels $n_{i.}$

| | Rester à la maison | Travail à temps partiel | Travail à temps plein | Total |
|--|--------------------|-------------------------|-----------------------|-------------|
| Les parents travaillent de façon égale | 13 | 142 | 106 | 261 |
| Mari travaille plus | 30 | 408 | 117 | 555 |
| Seul le mari travaille | 241 | 573 | 94 | 908 |
| Total | 284 | 1123 | 317 | 1724 |

Table des effectifs théoriques $\frac{n_{i.} \times n_{.j}}{n}$

| | Rester à la maison | Travail à temps partiel | Travail à temps plein | Total |
|--|--------------------|-------------------------|-----------------------|-------------|
| Les parents travaillent de façon égale | 43,0 | 170,0 | 48,0 | 261 |
| Mari travaille plus | 91,4 | 361,5 | 102,1 | 555 |
| Seul le mari travaille | 149,6 | 591,5 | 167,0 | 908 |
| Total | 284 | 1123 | 317 | 1724 |

$13 < 43 \Rightarrow$ « Les parents travaillent de façon égale » et « Rester à la maison » se repoussent

$106 >> 48 \Rightarrow$ « Les parents travaillent de façon égale » et « Travail à temps plein » s'attirent

$573 \lesssim 591,5 \Rightarrow$ « Seul le mari travaille » et « Travail à temps partiel » se repoussent légèrement (ce qui est à l'encontre de ce qu'aurait pensé quelqu'un inhabitué à l'analyse des données. En effet, la grande valeur 573 est due aux fréquences marginales qui sont élevées)

Introduction:

Le modèle d'indépendance et le test χ^2

le test χ^2 est utilisé pour évaluer la signification de l'écart global de l'échantillon réel par rapport au modèle d'indépendance. Il est exprimé comme suit :

$$\chi^2 = \sum_{i,j} \frac{(\text{Actual Sample Size} - \text{Theoretical Sample Size})^2}{\text{Theoretical Sample Size}},$$
$$\chi^2 = \sum_{i,j} \frac{(nf_{ij} - nf_{i\bullet}f_{\bullet j})^2}{nf_{i\bullet}f_{\bullet j}} = n \sum_{i,j} \frac{(f_{ij} - f_{i\bullet}f_{\bullet j})^2}{f_{i\bullet}f_{\bullet j}} = n\Phi^2,$$

$f_{i\bullet}$: la fréquence des individus ayant choisi la catégorie i

$f_{\bullet j}$: la fréquence des individus ayant choisi la catégorie j

Φ^2 : mesure de la relation d'indépendance

Pour les grandes valeurs de n, et si l'hypothèse « $H_0 : X \text{ et } Y \text{ sont indépendantes}$ » est vraie, la statistique ci-dessus suit une loi de χ^2 (*khi-2*) à $(I-1)(J-1)$ degrés de liberté.

On rejette donc H_0 (et l'on conclut au caractère significatif de la liaison) si χ^2 dépasse une valeur particulière (valeur ayant une probabilité faible et fixée a priori – en général 0,05).

Introduction:

Le modèle d'indépendance et le test χ^2

| | Table des effectifs réels $n_{i.}$ | | | |
|-----------|------------------------------------|---------------|-------------|-------------|
| | Rester à la maison | temps partiel | temps plein | Total |
| Les 2 = | 13 | 142 | 106 | 261 |
| Mari + | 30 | 408 | 117 | 555 |
| Seul mari | 241 | 573 | 94 | 908 |
| Total | 284 | 1123 | 317 | 1724 |

| | Table des effectifs théoriques $\frac{n_{i.} \times n_{.j}}{n}$ | | | |
|-----------|---|---------------|-------------|-------------|
| | Rester à la maison | temps partiel | temps plein | Total |
| Les 2 = | 43,0 | 170,0 | 48,0 | 261 |
| Mari + | 91,4 | 361,5 | 102,1 | 555 |
| Seul mari | 149,6 | 591,5 | 167,0 | 908 |
| Total | 284 | 1123 | 317 | 1724 |

| | Distances de χ^2 | | | |
|-----------|-----------------------|---------------|-------------|--------------|
| | Rester à la maison | temps partiel | temps plein | Total |
| Les 2 = | 20,9 | 4,6 | 70,1 | 95,7 |
| Mari + | 41,3 | 6,0 | 2,2 | 49,4 |
| Seul mari | 55,9 | 0,6 | 31,9 | 88,3 |
| Total | 118,1 | 11,2 | 104,2 | 233,4 |

La valeur de χ^2 est 233,4 qui est largement significative
(p-value: $2,4 * 10^{(-49)}$)

Introduction:

Le modèle d'indépendance et le test χ^2

| | Distances de χ^2 | | | |
|-----------|-----------------------|---------------|-------------|--------------|
| | Rester à la maison | temps partiel | temps plein | Total |
| Les 2 = | 20,9 | 4,6 | 70,1 | 95,7 |
| Mari + | 41,3 | 6,0 | 2,2 | 49,4 |
| Seul mari | 55,9 | 0,6 | 31,9 | 88,3 |
| Total | 118,1 | 11,2 | 104,2 | 233,4 |

| | Distances de χ^2 en % | | | |
|-----------|----------------------------|---------------|-------------|--------|
| | Rester à la maison | temps partiel | temps plein | Total |
| Les 2 = | 8,96 | 1,98 | 30,04 | 40,98 |
| Mari + | 17,68 | 2,56 | 0,94 | 21,18 |
| Seul mari | 23,94 | 0,25 | 13,66 | 37,84 |
| Total | 50,58 | 4,78 | 44,63 | 100,00 |

- Le tableau distances de χ^2 en % illustre **la contribution de chaque cellule à l'écart de l'indépendance**
- L'association entre «Les parents travaillent de façon égale» et « le travail à temps plein » exprime la plus grande déviation de l'indépendance (30,04% du total)
- Notez bien aussi la très faible contribution du travail à temps partiel; 4,78 %

Introduction:

Exercice

- Le tableau ci-dessous croise 7 catégories socioprofessionnelles avec modes d'hébergements en vacances.

- Déterminer le tableau des effectifs théoriques en cas d'indépendance.
- Calculer la distance du χ^2 entre le tableau réel et le tableau en cas d'indépendance.
- Effectuer le test d'indépendance du χ^2 .

| CSP\Mode d'hébergement | Campi ng | Hôtel | Famille Amis | Locatio n gîte | Total 1 |
|-----------------------------|-------------|-------|-----------------|-------------------|---------|
| Agriculteur | 2 | | 8 | 2 | 12 |
| Cadre moyen | 4 | 2 | 1 | 5 | 12 |
| Chef d'entreprise | 1 | 5 | 1 | 3 | 10 |
| Employé | 8 | 1 | 3 | 3 | 15 |
| Ouvrier | 9 | | 3 | 2 | 14 |
| Profession intermédiaire | 3 | 1 | 2 | 13 | 19 |
| Retraité | 5 | 2 | 9 | 2 | 18 |
| Total (2) | 32 | 11 | 27 | 30 | 100 |

Introduction:

Le modèle d'indépendance: l'analyse de correspondance

- Les tableaux de contingence doivent être analysés en termes d'indépendance.
- L'AFC exprime le modèle d'indépendance comme suit : $\forall i,j$

$$f_{ij}= f_{i.} \times f_{.j} \Leftrightarrow \frac{f_{ij}}{f_{i.}}= f_{.j} \Leftrightarrow \frac{f_{ij}}{f_{.j}}= f_{i.}$$

| f_{ij} | | | | | Profils ligne $f_{ij} / f_{i.}$ | | | | | Profils colonne $f_{ij} / f_{.j}$ | | | | | $f_{i.}$ Probabilité marginale V1 Profil colonne moyen |
|-----------|--------------------|---------------|-------------|------------------|---------------------------------|--------------------|---------------|-------------|-------|-----------------------------------|--------------------|---------------|-------------|------------------|--|
| | Rester à la maison | temps partiel | temps plein | Total = $f_{i.}$ | | Rester à la maison | temps partiel | temps plein | Total | | Rester à la maison | temps partiel | temps plein | Total = $f_{.j}$ | |
| Les 2 = | 0,01 | 0,08 | 0,06 | 0,15 | Les 2 = | 0,05 | 0,54 | 0,41 | 1,00 | Les 2 = | 0,05 | 0,13 | 0,33 | 0,15 | $f_{.j}$ Probabilité marginale V2 Profil ligne moyen |
| Mari + | 0,02 | 0,24 | 0,07 | 0,32 | Mari + | 0,05 | 0,74 | 0,21 | 1,00 | Mari + | 0,11 | 0,36 | 0,37 | 0,32 | |
| Seul mari | 0,14 | 0,33 | 0,05 | 0,53 | Seul mari | 0,27 | 0,63 | 0,10 | 1,00 | Seul mari | 0,85 | 0,51 | 0,30 | 0,53 | |
| Total | 0,16 | 0,65 | 0,18 | 1,00 | Total = $f_{.j}$ | 0,16 | 0,65 | 0,18 | 1,00 | Total | 1,00 | 1,00 | 1,00 | 1,00 | |

Introduction:

Le modèle d'indépendance l'analyse de correspondance

- L'AFC exprime le modèle d'indépendance comme suit : $\forall i,j$

$$f_{ij} = f_{i.} \times f_{.j} \Leftrightarrow \frac{f_{ij}}{f_{i.}} = f_{.j} \Leftrightarrow \frac{f_{ij}}{f_{.j}} = f_{i.}$$

- $\frac{f_{ij}}{f_{i.}}$ est la **probabilité conditionnelle** de la catégorie j (pour la var 2) sachant la catégorie i (pour la var 1).
- L'indépendance survient lorsque la **probabilité conditionnelle est égale à la probabilité marginale**.
- Cette perception de l'indépendance est similaire à ce que l'on pourrait s'attendre à trouver : **l'indépendance** survient si la probabilité de porter j (de V2) ne dépend pas de la catégorie portée par i V1, et vice versa.

Introduction:

Le nuage des profils ligne N_I

Profils ligne $f_{ij} / f_{i\bullet}$

| | Rester à la maison | temps partiel | temps plein | Total |
|---|--------------------|---------------|-------------|-------|
| Les 2 = | 0,05 | 0,54 | 0,41 | 1,00 |
| Mari + | 0,05 | 0,74 | 0,21 | 1,00 |
| Seul mari | 0,27 | 0,63 | 0,10 | 1,00 |
| Total=$f_{\bullet j}$ | 0,16 | 0,65 | 0,18 | 1,00 |

- Les profils lignes sont :
 - $i=1 \Rightarrow (0.05; 0.54; 0.41)$
 - $i=2 \Rightarrow (0.05; 0.74; 0.21)$
 - $i=3 \Rightarrow (0.27; 0.63; 0.10)$
- Les coordonnées du point moyen $G_I = (0.16; 0.65; 0.18)$

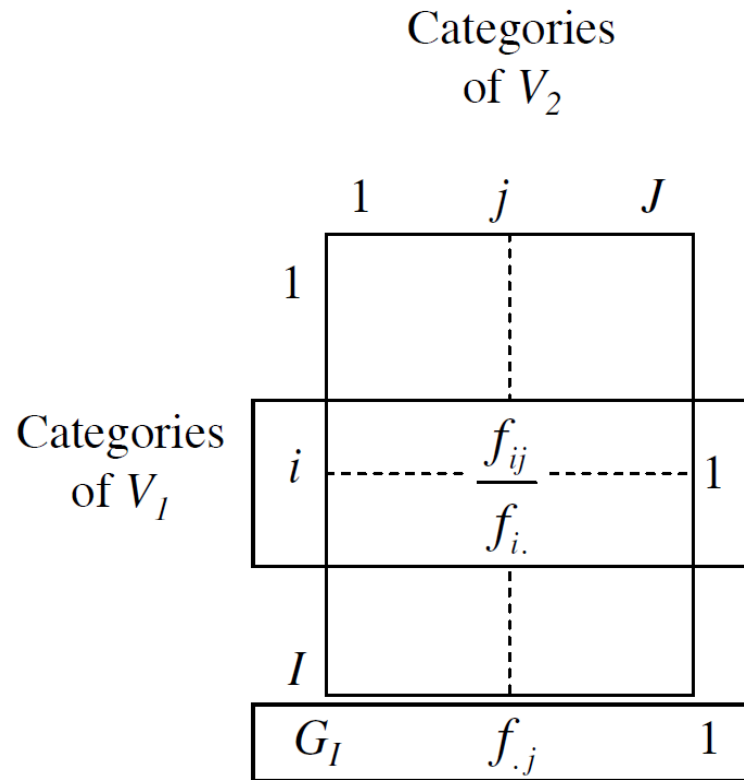
Par rapport au nuage d'individus en ACP, le nuage des lignes en AFC présente deux caractéristiques essentielles :

- Chaque point ligne i est caractérisé par un poids $f_{i\bullet}$ qui augmente avec la fréquence
- La distance entre individus consiste à attribuer à chaque dimension j le poids $1/f_{\bullet j}$

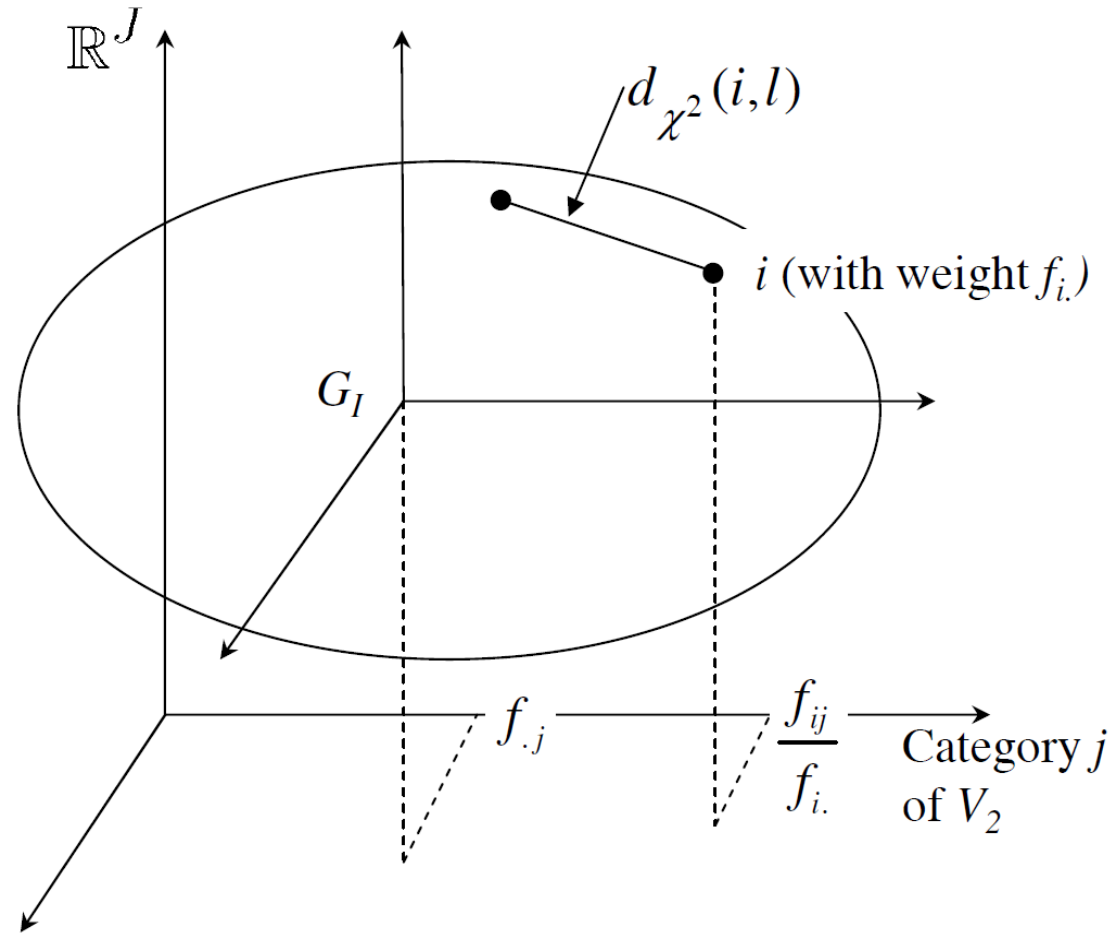
$$d_{\chi^2}^2(i, l) = \sum_{j=1}^J \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{lj}}{f_{l\bullet}} \right)^2$$

Introduction:

Le nuage des profils ligne N_I



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



Introduction:

Le nuage des profils ligne N_I

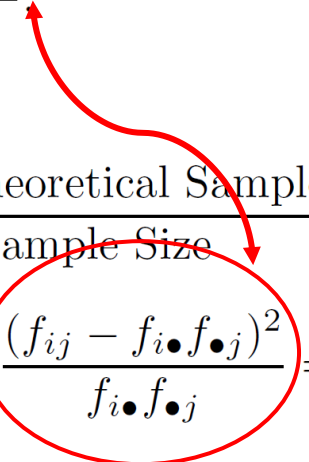
$$d_{\chi^2}^2(i, l) = \sum_{j=1}^J \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - \frac{f_{lj}}{f_{l\bullet}} \right)^2$$

L'inertie du point i par rapport au centre G_I est :

$$\text{Inertia}(i/G_I) = f_{i\bullet} d_{\chi^2}^2(i, G_I) = f_{i\bullet} \sum_{j=1}^J \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2$$

$$\text{Inertia}(i/G_I) = \sum_{j=1}^J \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}}$$

$$\chi^2 = \sum_{i,j} \frac{(\text{Actual Sample Size} - \text{Theoretical Sample Size})^2}{\text{Theoretical Sample Size}},$$

$$\chi^2 = \sum_{i,j} \frac{(nf_{ij} - nf_{i\bullet} f_{\bullet j})^2}{nf_{i\bullet} f_{\bullet j}} = n \sum_{i,j} \frac{(f_{ij} - f_{i\bullet} f_{\bullet j})^2}{f_{i\bullet} f_{\bullet j}} = n\Phi^2,$$


- L'inertie totale du nuage de points lignes par rapport à G_I est égale Φ^2 .
- Φ^2 mesure l'intensité de la relation entre les deux variables du tableau de contingence.
- Examiner la distribution des profils ligne en termes de G_I signifie examiner la différence entre les données et le modèle d'indépendance.
- C'est ce que fait l'AFC en soulignant les directions de la plus grande inertie pour le nuage des profils ligne.

Introduction:

Le nuage des profils colonne N_J

Profils ligne $f_{ij} / f_{.j}$

| | Rester à la maison | temps partiel | temps plein | Total |
|----------------------------------|--------------------|---------------|-------------|-------|
| Les 2 = | 0,05 | 0,13 | 0,33 | 0,15 |
| Mari + | 0,11 | 0,36 | 0,37 | 0,32 |
| Seul mari | 0,85 | 0,51 | 0,30 | 0,53 |
| Total=$f_{.j}$ | 1,00 | 1,00 | 1,00 | 1,00 |

Dans les tableaux de contingence, les lignes et les colonnes jouent des rôles symétriques : on peut étudier soit $V1 \times V2$, soit $V2 \times V1$. C'est l'une des principales différences entre l'AFC et l'ACP

- Les profils lignes sont :

- $i=1 \Rightarrow (0.05; 0.11; 0.85)$
- $i=2 \Rightarrow (0.13; 0.36; 0.51)$
- $i=3 \Rightarrow (0.33; 0.37; 0.30)$

- Les coordonnées du point moyen: $G_J = (0.15; 0.32; 0.58)$

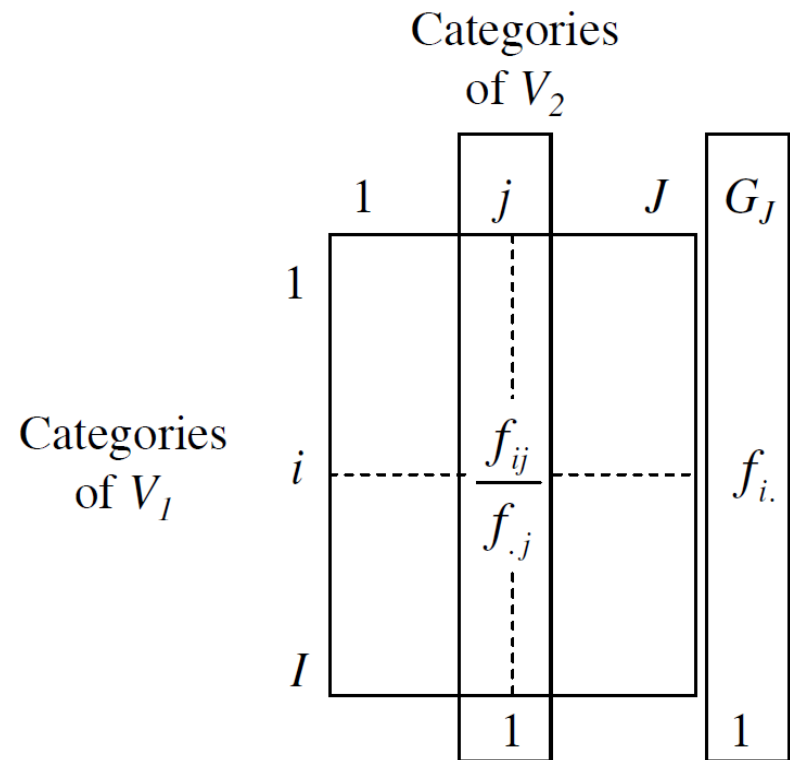
- Chaque point colonne j est caractérisé par un poids $f_{.j}$ qui augmente avec la fréquence
- La distance au carré entre deux colonnes consiste à attribuer à chaque dimension i le poids $1/f_i$.

$$d_{\chi^2}^2(j, k) = \sum_{i=1}^I \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ik}}{f_{\bullet k}} \right)^2$$

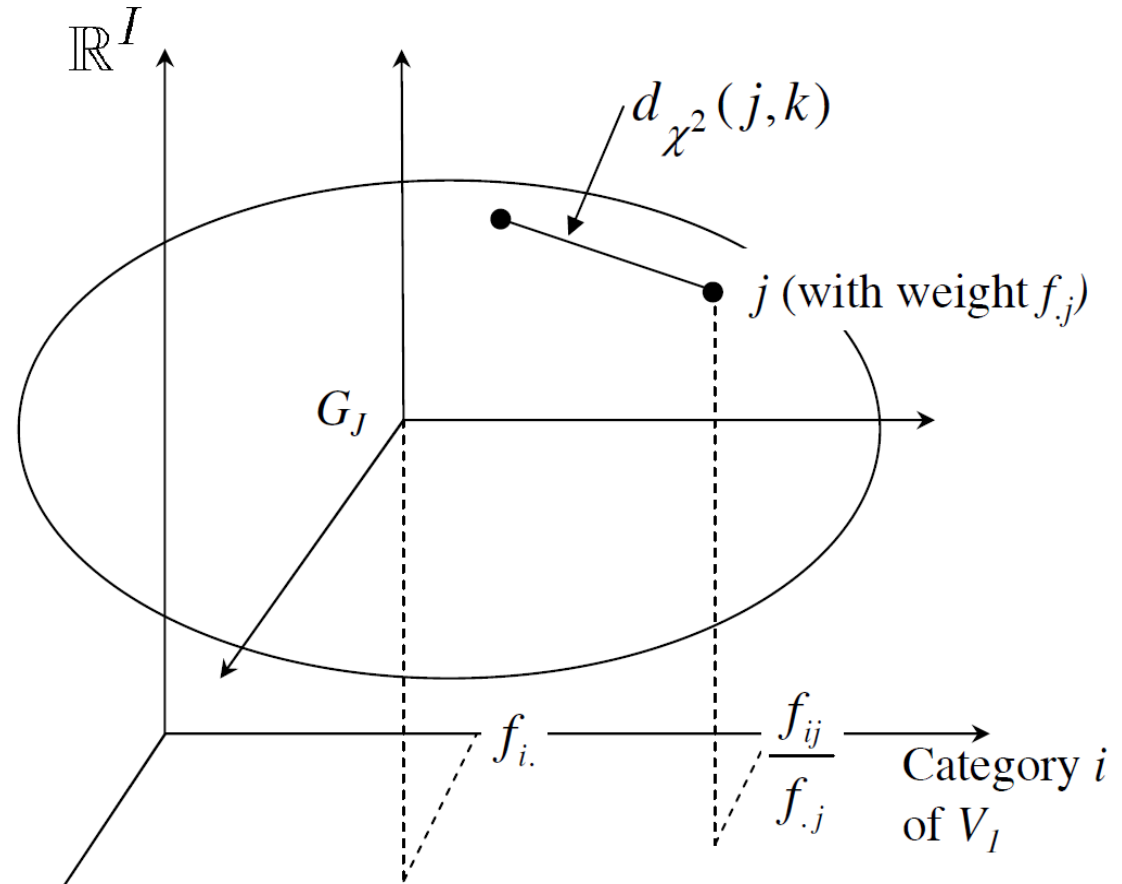
- L'inertie du nuage des points colonne = L'inertie du nuage des points ligne = Φ^2

Introduction:

Le nuage des profils colonne N_J



$$d_{\chi^2}^2(j, k) = \sum_{i=1}^I \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - \frac{f_{ik}}{f_{\bullet k}} \right)^2$$



Introduction:

Les facteurs principaux

Nous travaillons à peu près de la même manière que lors de l'ajustement du nuage des individus dans ACP. ie; nous visons une séquence d'axes orthogonaux avec une inertie maximale.

- **Meilleur axe de projection** : On cherche \vec{u}_s tel que $\sum_{i=1}^I f_i (OH_i^s)^2$ soit maximum; les H_i^s désignant les projetées orthogonales du profil i sur l'axe \vec{u}_s .
- **Meilleure représentation plane P** : On construit ainsi de manière itérative une suite d'axes de directions $\vec{u}_1, \vec{u}_2, \dots$ telle que:
 - \vec{u}_1 donne la direction qui maximise l'inertie projetée.
 - \vec{u}_2 donne la direction du reste de l'espace qui maximise l'inertie projetée.
 - ...
- le nombre maximal de dimensions nécessaires pour représenter parfaitement les nuages des profils ligne et colonne est $\min\{(I-1), (J-1)\}$
- l'inertie λ_s associée au facteur défini par \vec{u}_s mesure la partie de Φ^2 exprimée par ce facteur.

Introduction:

Interprétation des résultats

Valeurs propres (Inertie projetée) de l'AFC

| | Eigenvalue | Percentage of variance | Cumulative percentage |
|-------|------------|---------------------------|--------------------------|
| Dim 1 | 0.117 | 86.29 | 86.29 |
| Dim 2 | 0.019 | 13.71 | 100.00 |

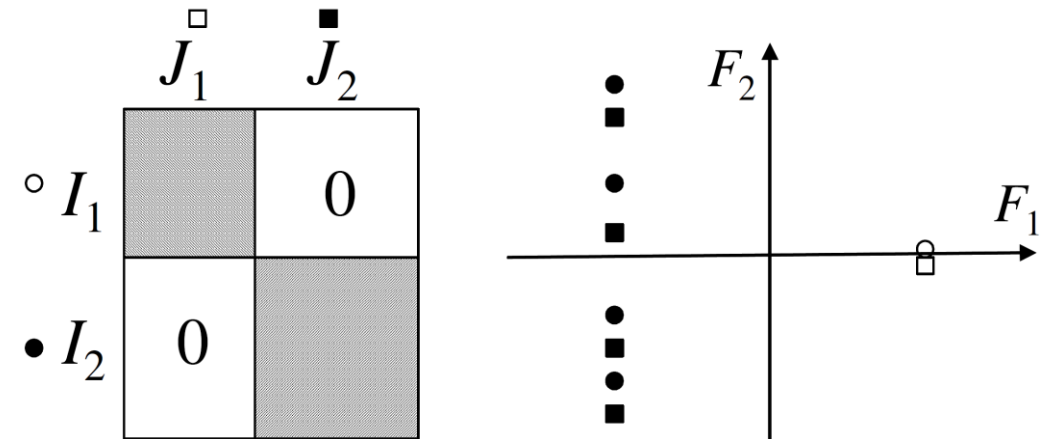
- Φ^2 quantifie la relation entre V1 et V2 ($\Phi^2 = \frac{\chi^2}{n} = \frac{233,4}{1724} \simeq 0.136$ et aussi $\Phi^2 = \sum \text{valeurs propres} = 0.117 + 0.019 \simeq 0.136$)
- Vu que l'inertie associée à un facteur (ou dimension) est une partie de la relation globale entre les deux variables V1 et V2, il semble naturel d'exprimer cette dimension en pourcentage ($\frac{\text{valeur propre}}{\Phi^2} \times 100$) :
- Dans l'exemple, il apparaît donc que la première dimension représente $\simeq 86,29\%$, et donc presque toute la différence entre l'échantillon réel (le tableau de données) et l'échantillon théorique selon l'hypothèse d'indépendance.

Introduction:

Interprétation des résultats

$$0 \leq \lambda_s \leq 1 \quad \text{et} \quad \Phi^2 = \sum_{s=1}^{\min\{I-1, J-1\}} \lambda_s \Rightarrow \Phi^2 \leq \min\{I-1, J-1\}$$

- Dans le cas particulier où $\lambda_s = 1$, l'ensemble des I lignes (et des J colonnes) peut être divisé en deux sous-ensembles I_1 et I_2 (resp J_1 et J_2) ; I_1 (resp I_2) est exclusivement associé à J_1 (resp J_2). Cette structure de données indique une forte relation entre les deux variables V1 et V2 (la fig. ci-contre).
- Dans la pratique, les valeurs propres d'une AFC ne sont presque jamais exactement égale à 1.
- Dans l'exemple, les valeurs propres sont plutôt petites. Bien qu'elle soit associée à une forte dépendance (test de χ^2), **par conséquent, ce qui est illustré dans l'exemple n'est qu'une tendance, même si elle est très significative.**



| | Eigenvalue | Percentage of variance | Cumulative percentage |
|-------|------------|------------------------|-----------------------|
| Dim 1 | 0.117 | 86.29 | 86.29 |
| Dim 2 | 0.019 | 13.71 | 100.00 |

Valeurs propres (Inertie projetée) de l'AFC

Introduction:

Interprétation des résultats

- Le nombre maximal de dimensions nécessaires pour représenter les nuages des profils ligne et colonne est $\min\{(I-1), (J-1)\}$.

$$0 \leq \lambda_s \leq 1 \text{ et } \Phi^2 = \sum_{s=1}^{\min\{(I-1), (J-1)\}} \lambda_s \Rightarrow \Phi^2 \leq \min\{(I-1), (J-1)\}$$

- En rapportant la valeur observée de Φ^2 à son maximum théorique, nous sommes conduits à l'indicateur statistique connu sous le nom du V de Cramer, défini comme:

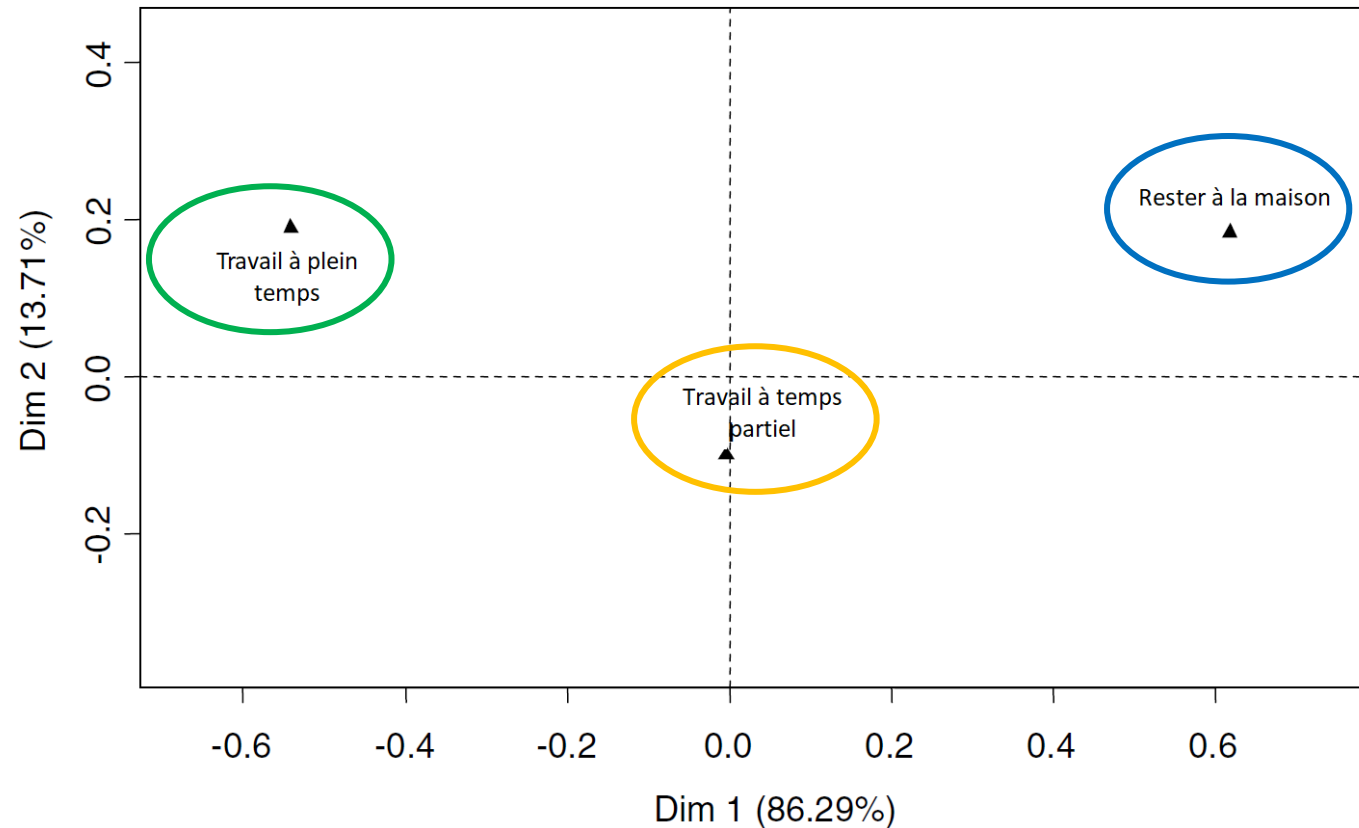
$$V = \left(\frac{\Phi^2}{\inf(I-1, J-1)} \right)^{1/2}$$

- $V = 0 \Rightarrow$ indépendance
- $V = 1 \Rightarrow$ relation maximale; chaque catégorie de la variable ayant le plus grand nombre de catégories est exclusivement associée à une seule catégorie de l'autre variable.
- Ainsi, lorsqu'on présente plusieurs variables qualitatives (pour les mêmes individus), on peut produire une matrice V de la même manière qu'on pourrait construire une matrice de corrélation.

- L'inertie λ_s dépend des coordonnées des profils lignes sur le facteur s :
 - la distance entre un profil et l'origine peut être interprétée comme un écart par rapport au profil moyen (Indépendance) et contribue donc à la relation entre V1 et V2.
 - La proximité de deux profils ligne i et j exprime également un écart similaire par rapport au profil moyen. Ces catégories, i et j , (de V1) ont une tendance d'association (c'est-à-dire plus que si les variables étaient indépendantes) aux mêmes catégories de V2.
- Le fait que les deux profils i et j soient opposés par rapport à l'origine exprime deux manières opposées de s'éloigner du profil moyen : les catégories de V2 auxquelles i est plutôt associé sont aussi celles auxquelles j est moins associé.

Introduction:

Exemple: Profils colonne



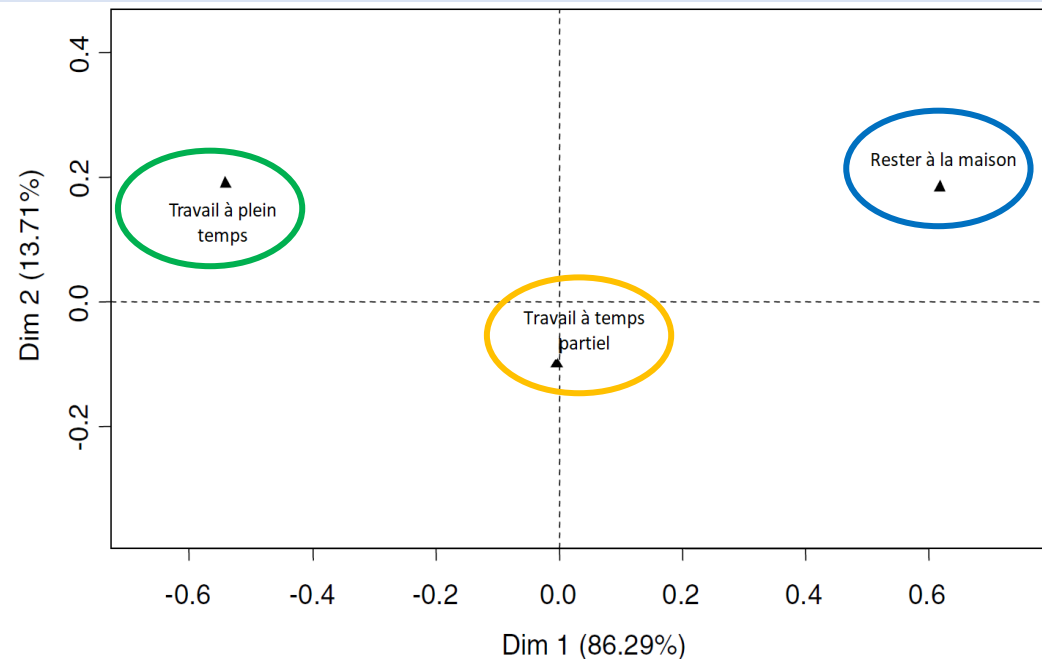
Profils colonne $f_{ij} / f_{.j}$

| | Rester à la maison | temps partiel | temps plein | Total= $f_{.i}$ |
|-----------|--------------------|---------------|-------------|-----------------|
| Les 2 = | 0,05 | 0,13 | 0,33 | 0,15 |
| Mari + | 0,11 | 0,36 | 0,37 | 0,32 |
| Seul mari | 0,85 | 0,51 | 0,30 | 0,53 |
| Total | 1,00 | 1,00 | 1,00 | 1,00 |

L'opposition des catégories sur le graphique représente inévitablement une opposition en termes de profil

Introduction:

Exemple: Profils colonne



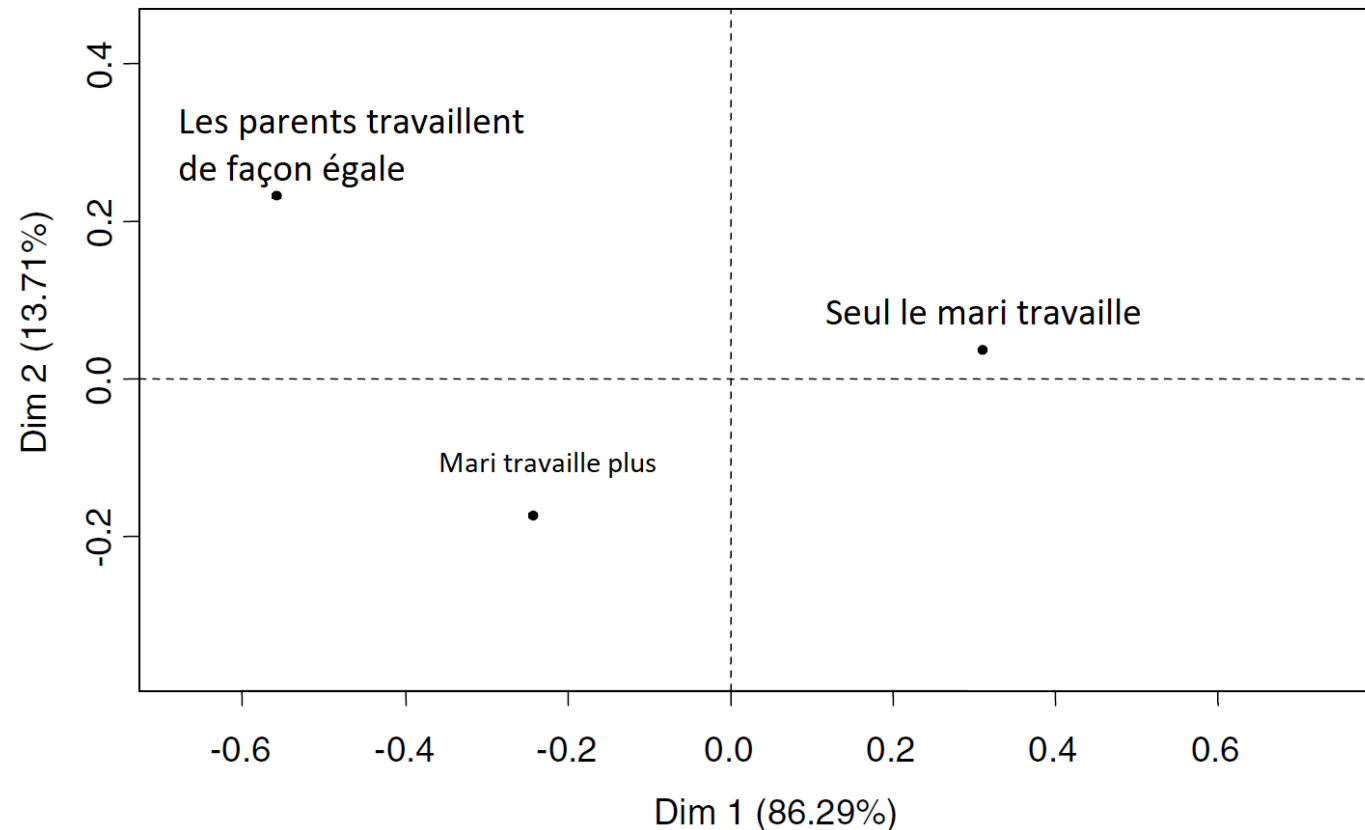
| | Profils colonne $f_{ij} / f_{.j}$ | | | |
|-----------|-----------------------------------|---------------|-------------|------------------|
| | Rester à la maison | temps partiel | temps plein | Total = $f_{.i}$ |
| Les 2 = | 0,05 | 0,13 | 0,33 | 0,15 |
| Mari + | 0,11 | 0,36 | 0,37 | 0,32 |
| Seul mari | 0,85 | 0,51 | 0,30 | 0,53 |
| Total | 1,00 | 1,00 | 1,00 | 1,00 |

- Les femmes qui ont répondu **rester à la maison** ont répondu :
 - Seul le mari travaille plus souvent que la population moyenne : 85 % contre 53 %.
 - Les deux parents travaillent moins souvent que la population moyenne : 5 % contre 15 %.
- Les femmes qui ont répondu **temps plein** ont répondu :
 - Seul le mari travaille moins souvent que la population moyenne 30 % contre 53 %.
 - Les deux parents travaillent plus souvent que la population moyenne, soit 33 % contre 15 %.
- La catégorie travail à **temps partiel** est extrêmement proche du centre de gravité (profil moyen). Cette réponse ne semble pas particulièrement informative : **lorsqu'une femme donne cette réponse, cela ne nous donne pas d'indications sur ce que pourrait être sa réponse à la question 1**

On peut dire que le 1er facteur organise les catégories de la seconde variable, des moins favorables au travail des femmes aux plus favorables.

Introduction:

Exemple: Profils ligne



| Profils ligne $f_{ij} / f_{i.}$ | | | | |
|---------------------------------|--------------------|---------------|-------------|-------|
| | Rester à la maison | temps partiel | temps plein | Total |
| Les 2 = | 0,05 | 0,54 | 0,41 | 1,00 |
| Mari + | 0,05 | 0,74 | 0,21 | 1,00 |
| Seul mari | 0,27 | 0,63 | 0,10 | 1,00 |
| Total= $f_{.j}$ | 0,16 | 0,65 | 0,18 | 1,00 |

La première dimension organise les catégories allant des moins favorables au travail des femmes (seul le mari travaille) aux plus favorables (les deux parents travaillent également).

Là encore, on peut appeler cette dimension « attitude à l'égard du travail des femmes »

- Jusqu'à présent, nous avons considéré séparément le nuage de lignes dans \mathbb{R}^J et le nuage de colonnes dans \mathbb{R}^I . Chacun de ces nuages a été projeté dans ses directions d'inertie maximale; projections qui ont été commentées séparément, chacune avec sa propre optimalité (avec chacune maximisant l'inertie projetée).
- Cependant en AFC, comme en ACP, les analyses du nuage des lignes et du nuage des colonnes sont étroitement liées en raison de leurs relations de dualité.
- La dualité découle du fait que nous analysons le même tableau de données, mais de différents points de vue (lignes ou colonnes). Cette dualité est particulièrement apparente et fructueuse dans l'AFC, car les lignes et les colonnes des tableaux de contingence sont fondamentalement les mêmes, c'est-à-dire les catégories des variables catégorielles.

Introduction:

Représentation superposée

Les 3 relations de la projection superposée des nuages des profils ligne N_I et colonne N_J

1. Les nuages N_I et N_J (Profils ligne et profils colonne) ont la même inertie totale. Dans l'AFC, la nature claire et même cruciale de cette inertie totale illustre bien que nous étudions la même chose par N_I et N_J .

$$Inertie (N_I) = \sum_{i=1}^I \sum_{j=1}^J \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \Phi^2 \text{ et } Inertie (N_J) = \sum_{j=1}^J \sum_{i=1}^I \frac{(f_{ij} - f_{i.}f_{.j})^2}{f_{i.}f_{.j}} = \Phi^2$$

2. L'inertie du nuage N_I et l'inertie du nuage N_J projetées sur l'axe du rang s (u_s et v_s respectivement) sont égales

$$\sum_{i=1}^I f_{i.} (OH_i^s)^2 = \sum_{j=1}^J f_{.j} (OH_j^s)^2 = \lambda_s$$

3. La troisième relation, et celle qui est essentielle à l'interprétation, rassemble les coordonnées des lignes et celles des colonnes sur les axes du même rang.

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

- $F_s(i)$ est la coordonnée du profil de ligne i sur la dimension de rang s
- $G_s(j)$ est la coordonnée du profil de colonne j sur la dimension de rang s
- λ_s est l'inertie de N_I (et de N_J respectivement) projetée sur la dimension de rang s en R^I (et en R^J respectivement).

Introduction:

Représentation superposée

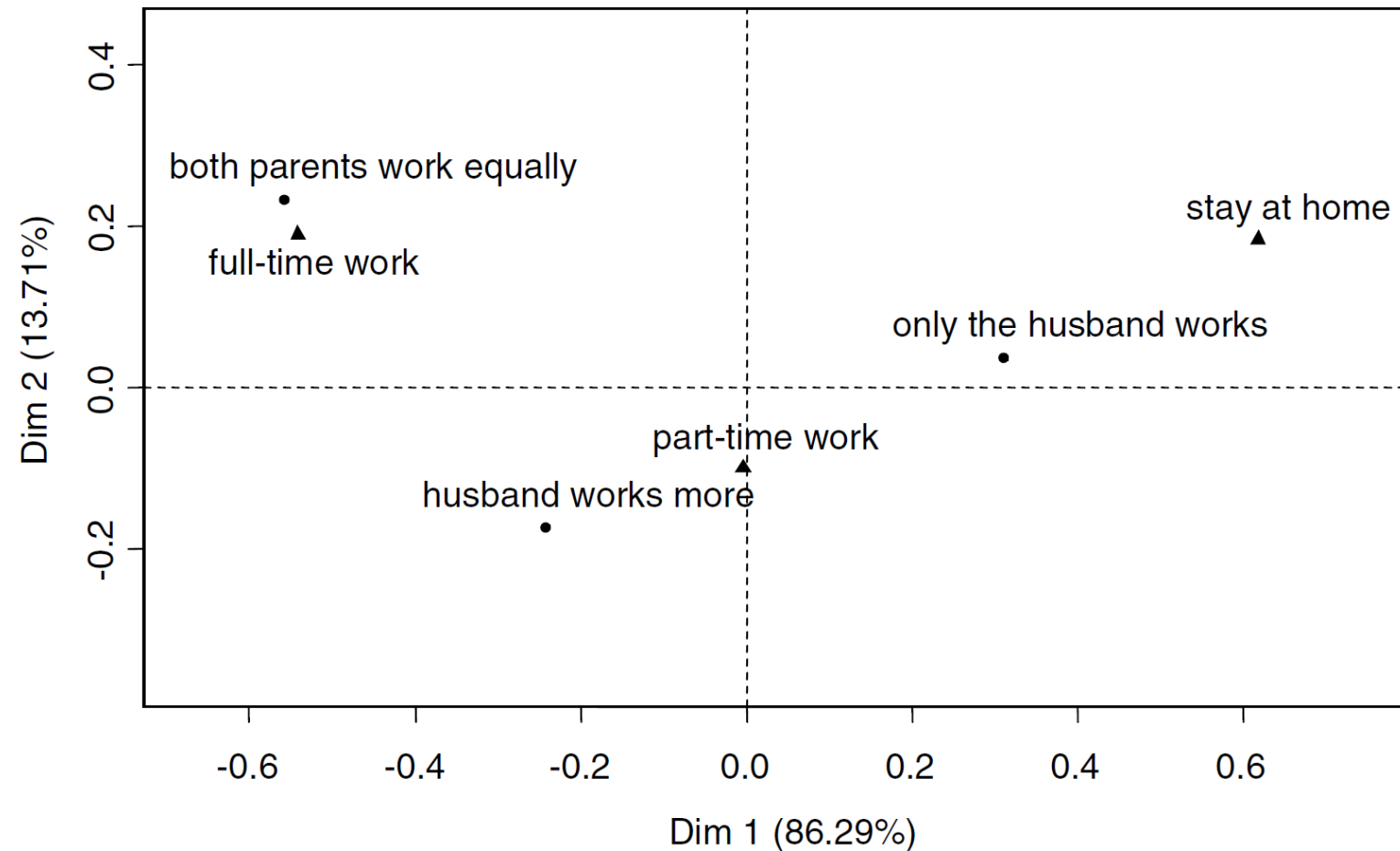
- Ainsi pour les dimensions s de cette représentation superposée, à hauteur du facteur multiplicatif $1/\sqrt{\lambda_s}$
 - Une ligne i se trouve au barycentre des colonnes, chaque colonne j ayant un poids $f_{ij}/f_{i\bullet}$ (ces poids sont positifs et totalisent 1).
 - Une colonne j se trouve au niveau du barycentre des lignes, chaque ligne i ayant un poids $f_{ij}/f_{\bullet j}$ (ces poids sont également positifs et totalisent 1).
- Cette propriété, qu'on appelle barycentrique (ou parfois pseudo-barycentrique) est aussi connue sous le nom de relations de transition car elle est utilisée pour transiter d'un espace \mathbb{R}^I ou \mathbb{R}^J à l'autre.

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

Introduction:

Représentation superposée

- La propriété barycentrique est utilisée pour interpréter la position d'une ligne par rapport à toutes les colonnes et la position d'une colonne par rapport à toutes les lignes;
- Chaque ligne (ou colonne) est proche des colonnes (et lignes) avec lesquelles elle est la plus étroitement liée, et est loin des colonnes (et lignes) avec lesquelles elle est la moins étroitement liée.

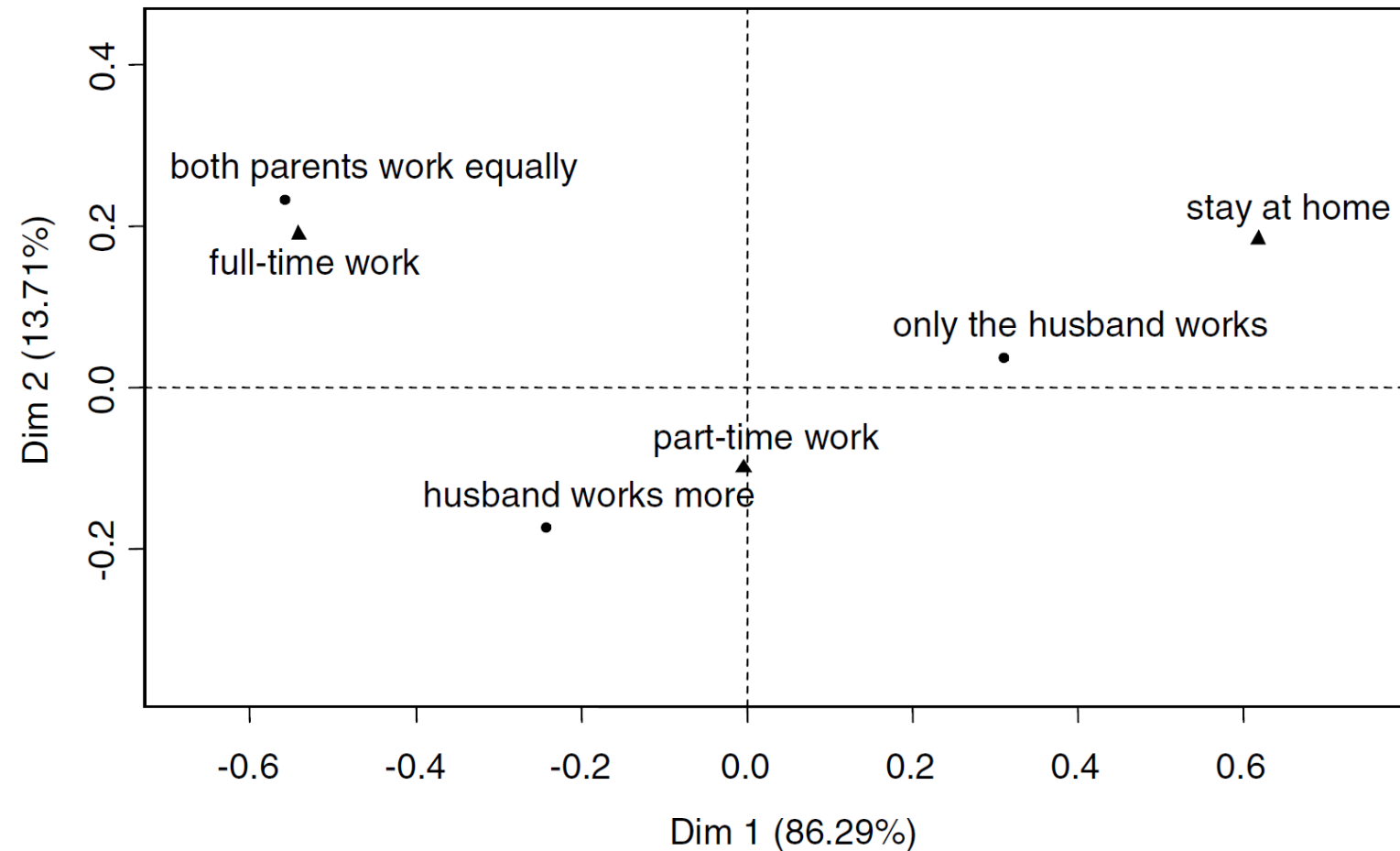


Introduction:

Représentation superposée

Ainsi, dans l'exemple :

- *Rester à la maison* est du même côté que *seul le mari travaille*, une catégorie avec qui elle est fortement liée, et est de l'autre côté des deux autres catégories, auxquelles elle est faiblement associée.
- *Les deux parents travaillent de façon égale* est du même côté que *travail à temps plein*, et les deux sont du côté opposé de la modalité *rester à la maison*.



- Il faut noter que l'origine des axes coïncide avec le profil moyen (barycentre) de chacun des deux nuages. Ainsi, lorsqu'un profil ligne a une coordonnée positive :
 - Il est généralement plus associé aux catégories j qui ont des coordonnées positives qu'il ne l'aurait été dans le modèle d'indépendance.
 - Il est généralement moins associé aux catégories j qui ont des coordonnées négatives qu'il ne l'aurait été dans le modèle d'indépendance.
- Il faut préciser que l'on peut commenter la position d'une ligne par rapport aux positions de **toutes** les colonnes mais il est **impossible** de tirer des conclusions sur la distance entre une ligne **spécifique** et une colonne **spécifique**.
- Les associations entre lignes et colonnes que nous souhaitons commenter doivent être vérifiées directement à partir des données (la table de contingence).

- La représentation simultanée produite par l'AFC est conçue pour visualiser la nature de la relation entre les variables (c'est-à-dire les relations entre lignes et colonnes) et ne nous dit rien de son intensité. Cette intensité est calculée à partir des valeurs propres qui sont des composantes de Φ^2 .
- Dans la pratique de l'AFC, les deux aspects de la relation entre deux variables (nature et intensité) sont identifiés par des outils distincts (graphiques et valeurs propres).

Introduction:

Représentation superposée

- Dans l'exemple traité, la première dimension oppose les catégories favorables au travail des femmes à celles qui ne le sont pas.
- Plus précisément, la première dimension organise les catégories des deux variables, des moins favorables au travail des femmes « *rester à la maison* » aux plus favorables « *les deux parents travaillent également* ».
- Dans cette optique, l'AFC laisse entendre que « *rester à la maison* » est une réponse beaucoup moins favorable au travail des femmes que « *seul le mari travail* ». Ce résultat nous indique comment les participants ont interprété les choix multiples associés aux deux questions.
- Il convient donc de concentrer la recherche sur les données afin d'identifier l'origine de la différence perçue par l'AFC entre ces deux catégories. La catégorie la plus éloignée de l'origine, « *rester à la maison* », correspond à la plus grande différence du profil moyen, comme en témoigne sa contribution à Φ^2 (118,07 pour « *rester à la maison* »; 88,34 pour le « *seul le mari travail* »).

Introduction:

Représentation superposée

- En termes plus tangibles, on peut observer que 85% des répondants « *rester à la maison* » ont également répondu « *seul le mari travail* », ces données regroupent ainsi les deux réponses les moins favorables au travail des femmes.
- D'autre part, dans seulement 27 % des cas, les femmes qui ont répondu « *seul le mari travail* » ont donné ces deux réponses défavorables.
- Dans ce cas, on pourrait dire que « *rester à la maison* » prédispose les femmes à donner une deuxième réponse défavorable au travail des femmes. « *rester à la maison* ». Elle est donc moins favorable au travail des femmes que « *seul le mari travail* ».

Profils colonne $f_{ij} / f_{.j}$

| | Rester à la maison | temps partiel | temps plein | Total= $f_{.j}$ |
|-----------|--------------------|---------------|-------------|-----------------|
| Les 2 = | 0,05 | 0,13 | 0,33 | 0,15 |
| Mari + | 0,11 | 0,36 | 0,37 | 0,32 |
| Seul mari | 0,85 | 0,51 | 0,30 | 0,53 |
| Total | 1,00 | 1,00 | 1,00 | 1,00 |

Profils ligne $f_{ij} / f_{i.}$

| | Rester à la maison | temps partiel | temps plein | Total |
|-----------------|--------------------|---------------|-------------|-------|
| Les 2 = | 0,05 | 0,54 | 0,41 | 1,00 |
| Mari + | 0,05 | 0,74 | 0,21 | 1,00 |
| Seul mari | 0,27 | 0,63 | 0,10 | 1,00 |
| Total= $f_{i.}$ | 0,16 | 0,65 | 0,18 | 1,00 |

Introduction:

Contribution des points à l'inertie d'une dimension

- L'inertie associée à une dimension peut être décomposée par des points. La contribution du point i à l'inertie de la dimension du rang s est exprimée par :

$$Ctr_s(i) = \frac{\text{l'inertie de } i \text{ projetée sur la dimension du rang } s}{\text{l'inertie de } N_I \text{ projetée sur la dimension du rang } s} = \frac{f_i \cdot (OH_i^s)^2}{\sum_{i=1}^I f_i \cdot (OH_i^s)^2} = \frac{f_i \cdot (OH_i^s)^2}{\lambda_s}$$

- L'identification des points qui contribuent le plus à la dimension facilite l'interprétation.
- Par exemple, le cas extrême d'une dimension construite seulement par un ou deux points peut donc être détecté immédiatement : l'interprétation peut donc se focaliser sur ce(s) point(s) et éviter des généralisations erronées.

Introduction:

Contribution des points à l'inertie d'une dimension

- Par exemple, les contributions de « *seul le mari travail* » et « *Les parents travaillent de façon égale* » sur la première dimension montrent les rôles respectifs des **poids** et des **distances** en comparant deux contributions proches.

$$Ctr_1(\textit{seul le mari travail}) = \frac{0.5267 \times 0.3096^2}{0.1168} = \frac{0.5267 \times 0.0958}{0.1168} = 0.432$$

$$Ctr_1(\textit{Les parents travaillent de façon égale}) = \frac{0.1514 \times 0.5586^2}{0.1168} = \frac{0.1514 \times 0.312}{0.1168} = 0.404$$

- Graphiquement, le point « *Les parents travaillent de façon égale* » est (approximativement) deux fois plus éloigné de l'origine que « *seul le mari travail* », suggérant ainsi une plus grande influence. Cependant, le poids de « *Les parents travaillent de façon égale* » est (environ) trois fois moins, ce qui suggère simultanément une influence moindre.

Introduction:

Qualité de représentation des points

- la qualité de représentation d'un point i par la dimension du rang " s "

$$q_{lt_s}(i) = \frac{\text{l'inertie de } i \text{ projetée sur la dimension du rang } s}{\text{l'inertie totale de } i} = \frac{(OH_i^s)^2}{(Oi)^2} = \cos^2(\overrightarrow{Oi}, \overrightarrow{OH_i^s})$$

Dans la pratique, les qualités de représentation sont principalement utilisées :

- Lors de l'examen d'une catégorie spécifique; la qualité de la représentation nous permet de sélectionner le plan sur lequel cette catégorie est exprimée avec le plus de précision.
- Lorsque l'on recherche un petit nombre de catégories pour illustrer la signification d'une dimension à l'aide de données brutes. Ceci est extrêmement utile pour communiquer les résultats. D'abord, les catégories avec les coordonnées les plus extrêmes sont sélectionnées, puis celles d'entre elles qui sont les mieux représentées sont sélectionnées.

Introduction:

Distance et inertie dans l'espace initial

Deux perspectives peuvent être envisagées pour identifier les catégories les plus ou les moins « responsables » de l'écart par rapport à l'indépendance.

- 1. La distance par rapport au profil moyen, qui ne tient pas compte de la taille de la catégorie.

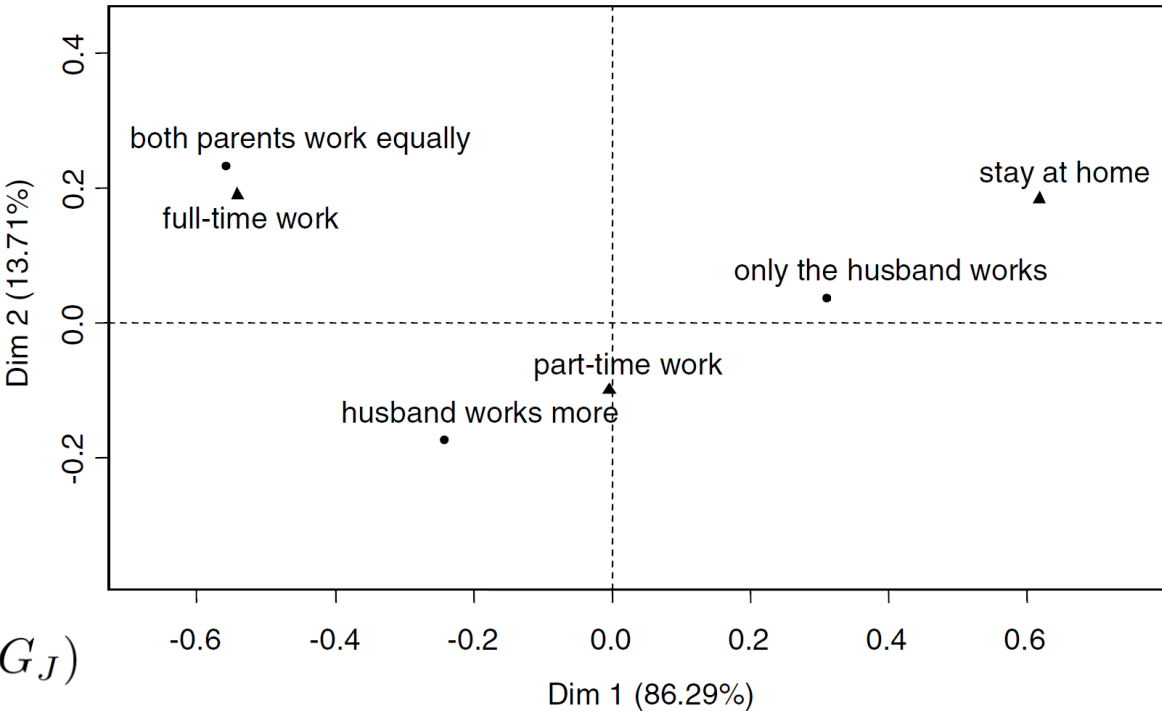
$$d^2_{\chi^2}(i, G_I) = \sum_{j=1}^J \frac{1}{f_{\bullet j}} \left(\frac{f_{ij}}{f_{i\bullet}} - f_{\bullet j} \right)^2 \qquad d^2_{\chi^2}(j, G_J) = \sum_{i=1}^I \frac{1}{f_{i\bullet}} \left(\frac{f_{ij}}{f_{\bullet j}} - f_{i\bullet} \right)^2$$

- 2. L'inertie qui a déjà été utilisée dans la décomposition χ^2 , c'est le produit de la distance et de la fréquence relative à la catégorie en question.

$$\text{Inertia}(i/G_I) = f_{i\bullet} d^2_{\chi^2}(i, G_I) \qquad \text{Inertia}(j/G_J) = f_{\bullet j} d^2_{\chi^2}(j, G_J)$$

| | Both work equally | Husband works more | Only husband works |
|----------|-------------------|--------------------|--------------------|
| Distance | 0.3665 | 0.0891 | 0.0973 |
| Inertia | 0.0555 | 0.0287 | 0.0512 |

| | Stay at home | Part-time work | Full-time work |
|----------|--------------|----------------|----------------|
| Distance | 0.4158 | 0.0099 | 0.3287 |
| Inertia | 0.0685 | 0.0065 | 0.0604 |



Introduction:

Exercice

- Le tableau ci-dessous croise 7 catégories socioprofessionnelles avec modes d'hébergements en vacances.

Copier le tableau sur Excel et faites les calculs suivant:

- Calculer les profils ligne et les profils colonne
- Calculer les distances par des profils ligne rapport à G_I ainsi que leurs inerties
- Calculer les distances par des colonne ligne rapport à G_J ainsi que leurs inerties
- Interpréter les résultats

| CSP\Mode d'hébergement | Camping | Hôtel | Famille Amis | Location gîte | Total 1 |
|--------------------------|---------|-------|--------------|---------------|---------|
| Agriculteur | 2 | | 8 | 2 | 12 |
| Cadre moyen | 4 | 2 | 1 | 5 | 12 |
| Chef d'entreprise | 1 | 5 | 1 | 3 | 10 |
| Employé | 8 | 1 | 3 | 3 | 15 |
| Ouvrier | 9 | | 3 | 2 | 14 |
| Profession intermédiaire | 3 | 1 | 2 | 13 | 19 |
| Retraité | 5 | 2 | 9 | 2 | 18 |
| Total (2) | 32 | 11 | 27 | 30 | 100 |

TP AFC sur R