Département Informatique
Analyse des données
Travaux pratiques
Séance 4

L'Analyse Factorielle des Correspondances

Le script a été réalisé à partir de la version 3.5.3 du logiciel R!

Préambule

Rappel des modalités d'évaluation :

- Ce TP peut être fait en binôme ou seul étant donné les circonstances
- Le rapport de ce TP sera soumis à évaluation. La date de remise de vos rapports (1 pour les questions de cours en rouge et 1 pour l'analyse de l'enquête) sur Celene au plus tard pour le 5 juin.

Cette séance est consacrée à la compréhension de l'Analyse Factorielle des Correspondances. Pour cela on propose d'étudier deux réponses parmi bien d'autres extraites d'une enquête du CREDOC publiée en 1974 par Nicole Tabard, intitulée "*Besoins et aspirations des familles et des jeunes*" travaux repris dans un cadre pédagogique [1] . Les réponses fournies correspondent aux deux questions suivantes :

Question 1 : Parmi les trois modèles suivants, quel est celui qui se rapproche le plus de l'image idéale que vous vous faites d'une famille ? Dit autrement, quelle est la famille idéale pour vous ?

a- Une famille où les deux conjoints ont un métier qui les absorbe autant l'un que l'autre
b- Une famille où la femme a une profession moins absorbante que celle de l'homme
c- Une famille où l'homme seul exerce une profession

Question 2- En distinguant la période où les enfants sont petits et celle où tous vont à l'école, quel est selon vous le type d'activité qui conviennent le mieux à une mère de famille ? dit autrement, quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ?

a- Au foyer
b- Travail extérieur à mi-temps
c- Travail extérieur à plein temps

Précisons ici, que nous sommes en possession de deux variables qualitatives ayant chacune 3 modalités possibles. Les réponses sont synthétisées dans le tableau suivant :

Question 1 : Quel est le nom usuel donné à ce tableau?

[1] Jean-Baptiste Pressac, Laurent Mell. Analyse factorielle des correspondances sous R - Partie I. Traitements et analyses de données quantitatives en SHS, Mar 2017, Brest, France. Hal-01516697

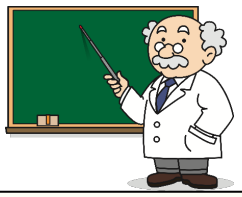
TABLEAU 37
REPONSES SIMULTANEES A DES QUESTIONS D'OPINION

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

Une interrogation : Pourquoi utiliser l'AFC pour analyser ce type de réponse?

Comme mentionné par les auteurs, il est difficile de savoir à partir de ce tableau si les femmes sont favorables ou non au travail féminin. En effet, l'observation des chiffres indique que 908 femmes sur 1 724 (soit 52 %) ont répondu que la famille idéale est celle où *"seul le mari travaille"*. Elles sont néanmoins 1 123 sur 1 724 (soit 65 %) à avoir répondu que l'activité convenant le mieux à une mère de famille, *quand ses enfants vont à l'école, est de travailler à mi-temps*. On relève une apparente contradiction ! On est en droit donc de se poser la question de l'existence d'un lien (ou encore liaison ou dépendance) entre ces deux variables. L'AFC va nous permettre de lever l'interrogation. Grâce à l'AFC et l'application R, nous allons chercher à visualiser la nature de la liaison entre les deux questions.

Question 2 : Qu'appelle-t-on une liaison ?



Travail demandé

Le script suivant va vous permettre de découvrir la démarche adoptée pour arriver à conclure sur la nature de la liaison entre ces deux variables. L'objectif de ce travail est avant tout de vous aider à comprendre et concrétiser le concept d'AFC. Votre investissement se "résume" à dérouler pas à pas les instructions, de faire l'effort d'en comprendre la syntaxe et surtout d'en observer les résultats fournis. **Ici où là dans le texte, des questions de cours apparaissent auxquelles il vous est demander de répondre sur une feuille séparée.**

En guise de conclusion à ce travail, vous prendrez le temps de résumer en un petit paragraphe, les principales étapes nécessaires pour connaitre la nature de la liaison entre deux variables. Vous résumerez en quelques lignes, la philosophie générale (l'esprit) de la méthode.

Quelques commentaires pour une bonne compréhension du script

Création du projet RStudio

Rappel de la procédure : Depuis le menu *File > New Project* choisir *New Directory* et puis *New Project*. Saisir le nom du répertoire contenant votre futur script ainsi que le chemin de sauvegarde. Puis retourner dans le menu "*File > New Fil*e et sélectionner *R script*. Pour finir copier le fichier de données "*fichier_data*" fourni dans le répertoire contenant le script.

Définir la localisation du répertoire de travail

Il est important d'indiquer à R le choix de votre répertoire de travail. Il existe pour cela une commande *setwd* ("chemin absolu"). La définition du répertoire de travail peut aussi être faite par le biais de la commande "*Set As Working Directory*" via l'onglet *More*. Cet onglet correspond au symbole de la roue crantée qui apparaît dans la fenêtre en bas à droite de Studio en ayant choisi le volet *Files* !

Par ailleurs, la commande "*getwd ()*" permet d'afficher la localisation du répertoire de travail sous la forme d'un chemin absolu :

Lire le jeu de données

La lecture est possible à partir de la commande "*read.table()*". Cette fonction permet de lire un fichier dans un format tabulaire et de créer une dataframe à partir de ce dernier. Les paramètres inclus indiquent :

1. L'argument *header=TRUE* permet de spécifier que la première ligne du jeu de données correspond aux intitulés des colonnes.
2. L'argument *row.names* =1permet de spécifier que le nom des lignes est contenu dans la première colonne.
3. L'argument *sep=*";" indique que le séparateur de colonne est le point-virgule.
4. L'argument *check.names=FALSE* permet d'empêcher R de modifier la syntaxe du nom des modalités de variable.
5. L'argument *fileEncoding="latin1*" permet de spécifier l'encodage du fichier.

Afin de faciliter par la suite la lecture et l'emploi du jeu de données, le tableau est stocké dans la variable *enquete*.

Dans la fenêtre se trouvant en haut et à droite, il est possible, très facilement et à n'importe quel moment, d'afficher le tableau de contingence "enquete" ou autres variables créées tout au long du script, à partir du volet Environnement (fenêtre Data) et le symbole présentant un petit tableau, sur l'extrême droite de chaque ligne présente.

Étant donné que seules les trois premières colonnes du tableau sont utilisées ici, les suivantes doivent être supprimées. La commande "*enquete [1:3]* " est employée pour cette raison. Le crochet [1,3] indique que l'on souhaite utiliser que les trois premières colonnes. Dans la fenêtre "Data", on observe bien que

le nombre de variables affichées a été réduit.

	au foyer	mi-temps	plein temps
2 conj. tr. également	13	142	106
trav. mari + absorbant	30	408	117
seul le mari trav.	241	573	94

Tableau de contingence réduit (deux variables et 3 modalités chacune)

Il est possible de renommer les modalités des deux variables en utilisant la fonction "*dimnames*" selon l'exemple donné dans le script suivant :

```
dimnames(enquete)[[1]][1]<-"Les 2 conjoints travaillent"
dimnames(enquete)[[1]][2]<-"Travail du mari plus absorbant"
dimnames(enquete)[[1]][3]<-"Seul le mari travaille"
dimnames(enquete)[[2]][1]<-"Rester au foyer"
dimnames(enquete)[[2]][2]<-"Travail à mi-temps"
dimnames(enquete)[[2]][3]<-"Travail à plein-temps"
```

Commentaires :

[[i]], l'indice entre doubles crochets indique le souhait de renommer la modalité ou les modalités en ligne (i=1) ou en colonne (i=2).

[i], l'indice entre crochets indique la $i^{ième}$ modalité de la variable selon la ligne ou la colonne.

En suivant ce codage, écrire `dimnames(enquete)[[1]][1] <-"Les 2 conjoints travaillent"` veut dire que l'on rebaptise la première modalité de la ligne!

De même, écrire, `dimnames(enquete)[[2]][3] <-"Travail à plein-temps"` veut dire que l'on rebaptise la troisième modalité de la colonne!

Dans l'étude présente, il est utile de calculer les marges lignes et colonnes du tableau de contingence.

Question 3 : Pourquoi est-il utile de connaitre les marges ligne et colonne ?

Question 4 : donner l'expression en fonction de l'indice ligne i et colonne j, de chacune d'elle ?

Pour cela, il est proposé ce qui suit :

```
enquete_marges <- enquete
enquete_marges$Total <-rowSums(enquete_marges)
enquete_marges[nrow(enquete_marges)+1,] <- colSums(enquete_marges)
row.names(enquete_marges)[nrow(enquete_marges)] <-"Total"
```

Le tableau 1, fourni par la commande "enquete_marges" une fois le script précédent exécuté, est donné ci-dessous :

	Rester au foyer	Travailler a mi-temps	travailler a plein-temps	Total
les deux conjoints travaillent	13	142	106	261
le travail du mari plus absorbant	30	408	117	555
Seul le mari travaille	241	573	94	908
Total	284	1123	317	1724

Tableau 1

Il est intéressant de calculer les pourcentages des réponses en ligne et en colonne. Pour cela on calcule les profils ligne et colonne ainsi que les profils moyen ligne et colonne.

Question5 : donner l'expression, en fonction de l'indice ligne i et colonne j, du profil ligne d'un tableau de contingence ?

Question6 : donner l'expression, en fonction de l'indice ligne i et colonne j, du profil colonne d'un tableau de contingence ?

Script pour le calcul de l'ensemble des profils ligne :

```
enquete_pourcentage_ligne <- enquete
enquete_pourcentage_ligne[nrow(enquete_pourcentage_ligne) +1,]<- colSums(enquete_pourcenta ge_ ligne)
row.names(enquete_pourcentage_ligne)[nrow(enquete_pourcentage_ligne)] <-"profil ligne moyen"
enquete_pourcentage_ligne<-rowPercents(enquete_pourcentage_ligne)
```

Tableau fourni par la commande "*enquete_pourcentage_ligne*"

	Rester au foyer	Travailler a mi-temps	travailler a plein-temps	Total	Count
les deux conjoints travaillent	5.0	54.4	40.6	100	261
le travail du mari plus absorbant	5.4	73.5	21.1	100	555
Seul le mari travaille	26.5	63.1	10.4	100	908
profil ligne moyen	16.5	65.1	18.4	100	1724

Commentaire :

Pour rappel, la ligne “Profil ligne moyen” correspond à la répartition en pourcentage des modalités à la question sur “l’activité qui convient le mieux à une mère de famille quand les enfants vont à l’école”, quelque-soit la réponse à la question sur la famille idéale. Le profil

ligne moyen peut être comparé aux profils lignes (la répartition en pourcentages ou la distribution de probabilité d'une modalité en ligne). Ici, aucun des trois profils lignes n'est proche du profil ligne moyen.

Script pour le calcul de l'ensemble des profils colonne:

```
enquete_pourcentage_colonne <- enquete
enquete_pourcentage_colonne$Total<-rowSums(enquete_pourcentage_colonne)
enquete_pourcentage_colonne<-colPercents(enquete_pourcentage_colonne)
dimnames(enquete_pourcentage_colonne)[[2]][4]<-"Profil colonne moyen"
```

Tableau 2 fourni par la commande "enquete_pourcentage_colonne"

	Rester au foyer	Travailler a mi-temps	travailler a plein-temps	Profil colonne moyen
les deux conjoints travaillent	4.6	12.6	33.4	15.1
le travail du mari plus absorbant	10.6	36.3	36.9	32.2
Seul le mari travaille	84.9	51.0	29.7	52.7
Total	100.1	99.9	100.0	100.0
Count	284.0	1123.0	317.0	1724.0

Tableau 2

Commentaire :
Ce tableau permet de constater que la répartition des réponses sur la famille idéale pour la modalité "Travail à mi-temps" est le plus proche de la répartition des réponses à la question sur la famille idéale. Autrement dit, le profil colonne "Travail à mi-temps" est le profil colonne le plus proche du profil colonne moyen. Cette similitude se traduira sur le graphe de l'AFC comme nous le verrons plus loin.

Étant donné que l'AFC travaille en réalité sur le tableau de probabilités, il sera nécessaire de passer en paramètre de la fonction R à venir le tableau de contingence modifié. Celui-ci est calculé en divisant les valeurs du tableau de contingence par le nombre d'individus (on effectue le calcul sur le tableau de contingence avec marge pour mieux constater que l'effectif total du tableau de probabilité est bien égal à 1, ce qui est la marque d'une distribution de probabilités). Le script suivant illustre ce traitement :

```
enquete_tableau_probabilite<-enquete_marges / 1724
enquete_tableau_probabilite
```

Tableau3 fourni par la commande " enquete_tableau_probabilite"

	Rester au foyer	Travailler a mi-temps	travailler a plein-temps	Total
les deux conjoints travaillent	0.007540603	0.08236659	0.06148492	0.1513921
le travail du mari plus absorbant	0.017401392	0.23665893	0.06786543	0.3219258
Seul le mari travaille	0.139791183	0.33236659	0.05452436	0.5266821
Total	0.164733179	0.65139211	0.18387471	1.0000000

Tableau 3

Question 7 : Montrez, en les exprimant sous forme littérale (2 phrases) que les valeurs contenues dans le tableau des profils lignes et des profils colonnes correspondent à des probabilités conditionnelles ?

L'objectif de ce travail est de visualiser la nature de la liaison entre les deux variables qualitatives : ***quelle est la famille idéale pour vous ?*** d'une part et d'autre part ***quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ?***

Mais faut-il encore que cette liaison soit significative. Pour le savoir, il est utile de réaliser le test du Khi2 !

Question 8 : Pourquoi cette notion est-elle importante ?

Étape du test du Khi2

Le test du khi2 est un outil qui va permettre de savoir si la liaison est s**ignificative.** Notion qui se rapproche de celle de l'indépendance ou pas entre les deux variables. Précisons en passant que ce test n'indique pas l'intensité de cette liaison. Information donnée par le phi2 (pas vue en cours). Afin de réaliser ce test, il "suffit" d'appeler une fonction fournie de base avec le logiciel R. Il s'agit de la fonction *chisq.test* que l'on retrouve dans le script donné ci-contre :

Question 9 : Donner l'expression du khi2

```
khi2enquete<- chisq.test(enquete)
```

La variable *khi2enquete* fournie les informations suivantes : X-squared = 233.43, df = 4, p-value < $2.2e^{-16}$

Commentaire :

Le terme X-squared est à lire Khi2 soit Khi au carré. La fonction chisq.test donne, entre autres, la valeur du Khi2 qui est un **indicateur de la significativité de la liaison**. Mais ce qui est intéressant ici est la p-value. Nous voyons ici que la p-value est inférieure à $2{,}2 \times 10^{-16}$. Cela signifie que la probabilité que les variables soient indépendantes est très très faible ! Il est donc possible de rejeter l'hypothèse d'indépendance entre les deux

variables. Pour autant, cela ne veut pas dire que les variables soient dépendantes. Les réponses à la question sur la famille idéale sont probablement liées aux réponses concernant l'activité convenant le mieux à une mère de famille dont les enfants vont à l'école.

Quelques rappels sur le rôle et l'intérêt du test du Khi2

Le test du khi2 permet de **déterminer la probabilité que les deux variables d'un tableau de contingence soient indépendantes**, c'est-à-dire qu'il n'existe pas de relation entre les modalités en ligne et les modalités en colonne (les unes ne conditionnent pas les autres, et réciproquement). Dit autrement cela veut dire que le "fait d'appartenir à une modalité de la première variable n'a pas d'influence sur la modalité d'appartenance de la deuxième variable". Dans ce test, l'hypothèse de départ suppose qu'il y a indépendance entre les deux variables. Si nous acceptons l'hypothèse d'indépendance, nous n'aurons pas d'utilité à réaliser une AFC car les points projetés seront extrêmement proches ou confondus avec le centre de gravité. Si nous rejetons l'hypothèse d'indépendance (critère p-value $< 0,05$), l'hypothèse alternative suppose que la liaison entre les deux variables est significative sans que nous puissions définir l'intensité de la liaison.

Quelques précisions supplémentaires : pour que le test du khi2 soit opératoire, il doit respecter un certain nombre de conditions:

1. L'**effectif total** du tableau de contingence doit être supérieur ou égal à 20.
2. L'**effectif marginal** du tableau de contingence doit toujours être supérieur ou égal à 5.
3. L'**effectif théorique** des cases du tableau de contingence doit être supérieur à 5 dans 80% des cases du tableau de contingence

Comme dit, du fait que nous ayons obtenu une p-value inférieure à $2,2x10^{-16}$ et, par extension, inférieure au seuil de 0,05, nous **rejetons l'hypothèse d'indépendance entre les deux variables**.

Comme le résultat est significatif, tout en respectant les conditions de validité du test du khi2, la valeur du test est sauvegardée dans la variable khi2enquete.

Interprétation du test du Khi2

Le test du khi2 est symétrique. Les lignes et les colonnes du tableau croisé (tableau de contingence) sont interchangeables. Le résultat du test sera exactement le même. Il n'y a pas de "sens de lecture" du tableau.

Il est possible d'afficher le tableau de contingence d'origine (**tableau des données observées**) en sélectionnant la valeur observed.

Tableau 4 fourni par la commande " khi2enquete$observed

	Rester au foyer	Travailler a mi-temps	travailler a plein-temps
les deux conjoints travaillent	13	142	106
le travail du mari plus absorbant	30	408	117
Seul le mari travaille	241	573	94

Tableau 4

De la même manière, il est possible d'afficher le tableau d'indépendance (**tableau des effectifs théoriques**) en sélectionnant la valeur expected. Dans ce cas, on calcule le tableau des pourcentages théoriques, en multipliant pour chaque case la proportion observée dans la population des deux modalités correspondantes. Puis, le tableau des effectifs théoriques se calcule en multipliant le tableau des pourcentages théoriques par l'effectif total.

Tableau5 fourni par la commande " khi2enquete$expected"

	Rester au foyer	Travailler a mi-temps	travailler a plein-temps
les deux conjoints travaillent	42.99536	170.0133	47.9913
le travail du mari plus absorbant	91.42691	361.5226	102.0505
Seul le mari travaille	149.57773	591.4640	166.9582

Tableau5

Le tableau des effectifs théoriques n'a que peu d'intérêt en lui-même mais en a davantage comparativement au tableau des données observées.

Il est utile aussi d'afficher le tableau des résidus (**tableau des écarts à l'indépendance**) en sélectionnant la valeur "residuals". Un résidu positif signifie que les effectifs dans la case sont supérieurs à ceux attendus sous l'hypothèse d'indépendance. Et l'inverse pour un résidu négatif.

donnees_residus<-khi2enquete$residuals

Tableau6 fourni par la commande " donnees_residus"

	Rester au foyer	Travailler a mi-temps	travailler a plein-temps
les deux conjoints travaillent	-4.574496	-2.148441	8.373594
le travail du mari plus absorbant	-6.424239	2.444409	1.479859
Seul le mari travaille	7.475127	-0.759211	-5.646384

Tableau6

Exprimé d'une autre manière, l'écart à l'indépendance représente l'**écart entre l'effectif observé et l'effectif théorique**, et ceci pour chacune des cases du tableau de contingence. L'écart à l'indépendance "est un effectif et c'est un invariant, indépendant du choix des lignes et des colonnes (c'est la différence entre l'effectif observé et l'effectif théorique : le résultat est donc un effectif)."

On peut noter ce qui suit :

1. Un **écart à l'indépendance positif** correspond à une **attraction** entre les deux modalités pour la case observée.
2. À l'inverse, un **écart à l'indépendance négatif** correspond à une **opposition** entre les deux modalités pour la case observée.

Plus la valeur de l'écart à l'indépendance est importante, plus l'attraction/opposition entre les modalités est forte.

Rappel de l'objectif

Notre objectif est de **visualiser la nature de la liaison entre les deux variables qualitatives**. Sachant qu'une liaison correspond à **l'écart entre les données observées et le modèle d'indépendance**, nous souhaitons donc **visualiser la nature de l'écart à l'indépendance entre ces variables**.

Par ailleurs, il y a **trois façons de caractériser la liaison** entre les deux variables qualitatives.

1. La **significativité** de la liaison (qui se mesure avec le test du khi2).
2. L'**intensité** de la liaison (qui se mesure, avec le Phi2). Pas vu cette année.
3. La **nature** de la liaison (qui correspond à l'association entre les modalités et qui est représentée par le biais de l'AFC.

Chargement des packages

Le test du Khi2 a permis d'écarter l'hypothèse d'indépendance. Il y a donc une liaison entre les modalités des deux variables. De fait, nous pouvons faire une AFC pour visualiser la nature de la liaison. L'outil est disponible dans le **package FactoMineR** (dédié à l'analyse multidimensionnelle de données), c'est un choix parmi d'autres possibles offrant ce type de méthode statistique.

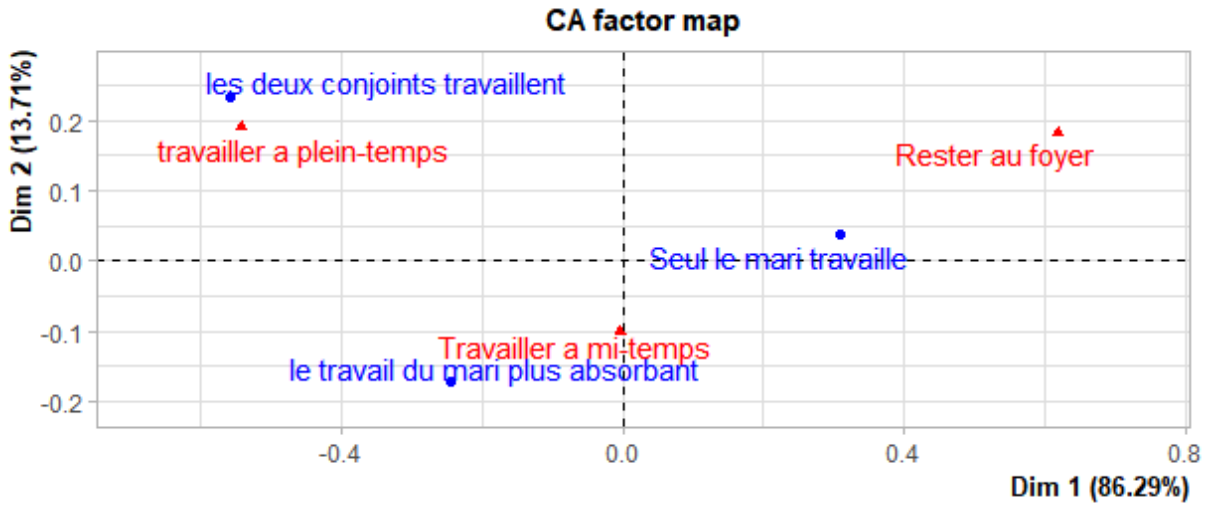
library(FactoMineR) est déjà mentionné en début de script

Les résultats obtenus par l'AFC

Lors du précédent test du khi2, une p-value inférieure à $2{,}2\text{x}10^{-16}$ avait été obtenue permettant de rejeter l'hypothèse d'indépendance entre les deux variables et par conséquent définissant la liaison entre ces deux variables comme significative.

De ce constat il est opportun de réaliser une AFC afin de visualiser la nature de la liaison. La fonction *CA()* sur le fichier de données enquete est donc utilisée comme suit:

Le script *afc_enquete<-CA(enquete)* permet de fournir le graphique1ci-dessous:



Graphique1

Question 9 : Comment sont pondérés les variables et les individus lors d'une AFC ?

Question 10 : Dans la commande *'afc_enquete<-CA(enquete)'*, c'est le tableau *enquete* qui est passé en paramètre mais selon vos connaissances comment et pourquoi sera introduite la métrique du Khi2 ?

Rappel des deux questions originelles :

1. Quelle est la famille idéale pour vous ? en bleu
2. Quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ? en rouge

Lors d'une première lecture de ce tableau de contingence, il était soulevé une **apparente contradiction**. À la question "Quelle est la famille idéale pour vous ?", il apparaissait que 908 femmes sur 1 724 (visible dans la **marge colonne**), soit environ 53 % des répondantes, déclarent "Seul le mari travaille" et seulement 261 femmes sur 1 724 (environ 15 %) déclarent "Les deux conjoints travaillent également". Sur la base de ces premières réponses, il est possible d'émettre l'hypothèse, qu'à cette époque, une majorité était en faveur d'un modèle familial où seul le mari travaille.

À côté de ça, à la question "Quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ?", elles sont 1 440 sur 1 724 (visible dans la **marge ligne**), soit environ 84 %, à être en faveur du travail à mi-temps ou à plein-temps. Les réponses à cette question semblent indiquer que les femmes sont moins hostiles au travail féminin (bien au contraire).

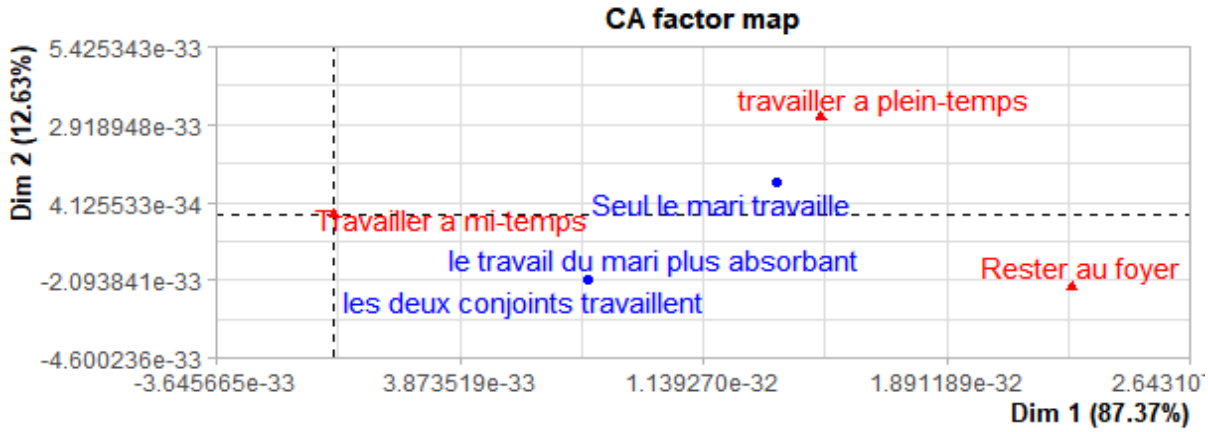
Par conséquent, à ce stade de l'interprétation, l'analyse fait face a priori à une **contradiction**.

De cela, il est possible de dire que le tableau de contingence ne permet pas de savoir si les femmes des années 70 sont favorables ou non à l'activité féminine. Par contre, une **première lecture du graphe de l'AFC** nous permet de dire que les modalités des réponses s'associent entre elles des plus favorables au travail féminin aux plus défavorables au travail féminin.

Avant d'approfondir, plus en détail, l'interprétation de cette AFC, nous allons faire un pas de côté et voir ce qui se passe dans le cas où il y aurait **indépendance entre les deux variables**.

Pour cela, il est nécessaire d'exécuter une AFC avec les données du modèle d'indépendance :

Le script *afc_enquete_independance<-**CA**(khi2enquete$expected)* fournit alors le graphique2 suivant:



Graphique2

La lecture de ce graphique nous permet de voir que les points sont quasiment tous **confondus avec le centre de gravité**, correspondant au **profil moyen**. La représentation graphique est trompeuse mais l'échelle des axes va dans le sens de notre interprétation. Simplement, ce qu'il y a à retenir de ce graphe, c'est que, lorsqu'il y a indépendance entre les deux variables, tous les points sont confondus avec l'origine. Du fait qu'il n'y ait pas d'écarts à l'indépendance, il n'y a graphiquement rien à exploiter, rien à interpréter, rien à analyser. Ce graphe donne à voir ce que nous avons précédemment énoncé, à savoir que :

1. Si nous acceptons l'hypothèse d'indépendance (p-value > 0,05 dans le cas d'un test du khi2), nous n'aurons pas d'utilité à réaliser une AFC car les points projetés seront extrêmement proches ou confondus avec le centre de gravité donc confondus avec le centre du graphe.
2. La réalisation d'un test du khi2 est donc fortement conseillée avant la réalisation d'une AFC.
3. Plus précisément, le test du khi2 conditionne l'éventuelle réalisation d'une AFC.

Les résultats qui suivent vont aussi dans le sens du propos :

Le script *summary(afc_enquete_independance)* permet de faire les commentaires suivants :

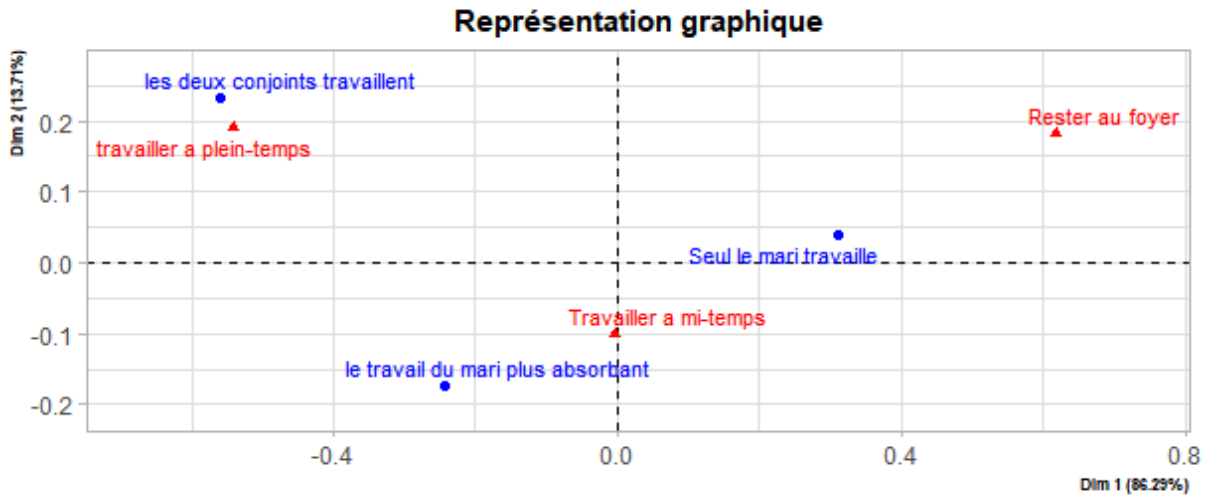
1. Les coordonnées (dim.n) des modalités en ligne et en colonne, sur chacune des dimensions, sont égales à 0!
2. Il en va de même pour leur contribution à la construction (ctr) de chaque dimension et leur qualité de représentation (cos^2) sur chaque dimension.

Quittons ce cas d'indépendance et revenons maintenant à notre premier jeu de données où la liaison entre les variables est significative.

Les différentes fonctions proposées par le logiciel R associées aux multiples packages disponibles, nous permettent l'affichage du graphe

La commande *plot(afc_enquete, cex = 0.7, cex.axis = 0.6, cex.lab = 0.8, title = Représentation*

graphique", selectRow = "cos2 0.7", selectCol = "cos2 0.7") fournit le graphisme suivant:



Question 11 : À quoi correspond la figure ci-dessus ?

Question 12 : Que représentent les valeurs de pourcentage associées aux 2 axes (dim1 et dim2) ?

Question 13 : Que peut-on dire sur la qualité de la représentation obtenue ?

Question 14 : Peut-on donner un sens (une interprétation) à chacun des axes ?

Descriptif des arguments utilisés dans la fonction *plot()* :

1. L'argument cex= permet de modifier la taille des points et des textes liés à ces points.
2. L'argument cex.axis permet de modifier la taille des annotations d'axe
3. L'argument tittle= permet de donner un titre au graphique
4. L'argument selectRow= permet de sélectionner les modalités en ligne à afficher en fonction de critères particuliers comme le cos^2= 0,7 par exemple.

Ce ne sont que quelques arguments, parmi d'autres, en vue d'améliorer la représentation graphique. Déjà vus dans les séances de TP précédentes. Dans le jeu de données qui nous concerne, ces améliorations restent mineures et n'ont que peu d'impact sur la représentation graphique. Par contre, dans le cas où nous aurions à travailler sur un tableau de contingence, dont chacune des variables comporte au moins une vingtaine de modalités, ces améliorations graphiques prendraient d'autant plus d'importance.

La commande *summary(afc_enquete, nbelements = Inf)* donne un ensemble d'informations permettant certains commentaires:

1. Le résultat du test du khi2 (uniquement sur les lignes et les colonnes actives) avec la p-value.
2. Un tableau avec les valeurs propres, les pourcentages d'inertie associés à chaque dimension.
3. Un tableau avec les résultats sur les lignes actives avec leur coordonnée sur chaque dimension (Dim.n), leur contribution à la construction (ctr) de chaque dimension et leur qualité de représentation (cos2) sur chaque dimension.

4. Un tableau avec les résultats sur les colonnes actives (dim.n, ctr, cos2).

Noter que la fonction *summary* n'affiche, par défaut, que les 10 réponses les plus significatives.

5. L'argument *nbelements=Inf* permet de retirer cette limite. Dans l'exemple présent cela n'est pas utile puisque chacune des deux variables ne dispose que de trois modalités.
6. Mais, dans un certain nombre de cas, cet argument est précieux.

On rappelle que la variance est aussi appelée dans la littérature "**valeur propre**" ou "**inertie**". La variance mesure l'intensité de la liaison entre les deux variables expliquées par cet axe. **La variance est comprise entre 0 et 1**. Le total des variances (le **Phi2**) est la mesure de l'intensité de la liaison (liaison = l'écart à l'indépendance).

Ici, le Phi2 = 0.117 + 0.019 = 0.136

Question : à quoi comparer cette valeur ?

Les variances étant *au maximum* égales à 1, le Phi2 peut avoir *au maximum* une valeur de 2. 0.136 << 2 : L'intensité de la liaison est faible. Les réponses s'associent globalement peu entre elles. Mais rares sont les cas où l'intensité de la liaison entre les variables est forte.

Question15 : Pourquoi le maximum de phi2 est-il de 2 ?

Interprétation des points sur le graphe

Le centre de l'axe correspond au profil ligne moyen et au profil colonne moyen. Nous avions précédemment calculé les profils moyens et nous avions effectivement constaté que le profil de la modalité "Travail à mi-temps" était proche du profil colonne moyen.

	Rester au foyer	Travailler à mi-temps	Travailler à plein-temps	Total	Count
les deux conjoints travaillent	5.0	54,4	40,6	100	261
le travail du mari plus absorbant	5.4	73,5	21,1	100	555
Seul le mari travaille	26,5	63,1	10,4	100	908
Profil ligne moyen	16,5	65,1	18,4	100	1724

Tableau des pourcentages en ligne

	Rester au foyer	Travailler à mi-temps	Travailler à plein-temps	Profil colonne moyen
Les deux conjoints travaillent	4.6	12.6	33.4	15.1
Le travail du mari plus absorbant	10.6	36.3	36.9	32.2
Seul le mari travaille	84.9	51	29.7	52.7
Total	100	99.9	100	100
Count		1123	317	1724

Tableau des pourcentages en colonne

Peut-on interpréter la position entre deux point lignes ou deux points colonnes ?

La réponse est “oui” ! Les points lignes (resp. les points colonnes) qui ont des profils lignes (resp. des profils colonnes) similaires sont proches sur le graphe. Ce n’est le cas d’aucun des points du graphe issu des données de Nicole Tabard.

Par contre, on ne peut pas interpréter la proximité entre un point ligne et un point colonne. On peut seulement dire que les lignes sont “du côté” des colonnes auxquelles elles s’associent le plus, dans la mesure où ces colonnes sont éloignées du centre de gravité. Et respectivement, les colonnes sont “du côté” des lignes avec lesquelles elles s’associent le plus, dans la mesure où ces lignes sont éloignées du centre de gravité. On en veut pour preuve que sur le graphe “seul le mari travaille” est du côté de “rester au foyer”.

Sur le graphe, “Seul le mari travaille” et “du côté” de “rester au foyer”. Si on revient sur le tableau de contingence (affiché ci-dessous), cela peut sembler contradictoire puisque 573 personnes ont répondu à “seul le mari travaille” et “travailler à mi-temps” alors que seulement 241 personnes ont répondu à “seul le mari travaille” et “rester au foyer”. “Seul le mari travaille” ne devrait-il pas être du côté de “Travailler à mi-temps”

	Rester au foyer	Travailler à mi-temps	Travailler à plein-temps	Total
Les deux conjoints travaillent	13	142	106	261
Le travail du mari plus absorbant	30	408	117	555
Seul le mari travaille	241	573	94	908
Total	284	1123	317	1724

C’est tout simplement parce qu’il ne faut pas regarder le tableau de contingence mais le tableau des pourcentages en colonnes car on veut comparer une modalité en ligne avec les trois autres modalités en colonne. Or dans ce tableau, la ligne “Seul le mari travaille” représente 84,9% des réponses à la colonne “Rester au foyer”. C’est son meilleur score comparé aux autres colonnes. C’est pourquoi on peut dire que “Seul le mari travaille” est “du côté” de “Rester au foyer”.

Question 16 : Au final, et en utilisant vos connaissances générales sur la notion d’analyse factorielle (réduction de dimensions), quelle particularité de cette analyse serait-il bon de rappeler ici en conclusion ?

Conclusion

Comme le sujet l'annoncé en introduction, l'objectif était de nous concentrer sur **l’analyse factorielle des correspondances (AFC)** et d’étudier les **éventuelles** liaisons entre les modalités de deux variables qualitatives. Ce travail vous à parmi de découvrir un certain nombre d’éléments de lecture, d’analyse et d’interprétation sur comment aborder la problématique de l'analyse multivariée à deux variables ici.