Analyse factorielle des correspondances sous R - Partie I

Jean-Baptiste Pressac, Laurent Mell

Analyse factorielle des correspondances sous R - Partie I

Jean-Baptiste Pressac, Laurent Mell

14 Avril 2017

Auteurs

Jean-Baptiste PRESSAC (CRBC, CNRS) (Jean-Baptiste.Pressac@univ-brest.fr)
Laurent MELL (LABERS, Université de Bretagne Occidentale) (laurent.mell@univ-brest.fr)

Si vous avez des commentaires à formuler ou des remarques à effectuer, vous pouvez nous contacter par mail (Jean-Baptiste.Pressac@univ-brest.fr, laurent.mell@univ-brest.fr). Par la même occasion, si vous souhaitez échanger plus en détail et/ou contribuer à cette démarche, nous sommes aussi disponibles.

Introduction

Ce premier travail s'inscrit dans une démarche de réflexion, plus large, que nous souhaitons entreprendre autour **l'analyse de données multidimensionnelles**. La spécificité de ce premier travail réside dans le fait que nous allons nous concentrer sur **l'analyse factorielle des correspondances (AFC)**. Notre objectif sera d'étudier les **éventuelles** liaisons entre les modalités de deux variables qualitatives. Par ailleurs, nous avons fait le choix d'utiliser le logiciel R ainsi que l'environnement de développement RStudio.

Les données

Les données sur lesquelles nous allons travailler proviennent, en partie, du MOOC Analyse des données multidimensionnelles sur la plateforme FUN. Ce MOOC est proposé par François Husson, Jérôme Pagès et Magalie Houée-Bigot.

Les données sont issues d'un enquête du CREDOC publiée en 1974 par Nicole Tabard, intitulée Besoins et aspirations des familles et des jeunes. Le MOOC ne nous donne pas d'informations sur les circonstances dans lesquelles le questionnaire a été établi ni sur le nombre total de questions. Nous savons uniquement qu'il a porté sur 1724 femmes. L'AFC ne nous permettra pas d'analyser l'intégralité du questionnaire. De toute façon, nous allons nous focaliser sur une relation spécifique. Nous allons étudier l'articulation des réponses qualitatives à deux questions :

- Quelle est la famille idéale pour vous ?
- Quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ?

Le point de départ de l'analyse est le tableau de contingence reproduit ci-dessous. C'est ce type de données (les marges des totaux mis à part) que nous fournirons à la fonction de calcul de l'AFC.

Comme le souligne François Husson dans le MOOC, il est difficile de savoir à partir de ce tableau si les femmes sont favorables ou non au travail féminin. En effet, 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 (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. L'AFC va nous permettre d'étudier le lien entre ces deux questions et de lever cette apparente contradiction. Elle va notamment nous permettre de visualiser la nature de la liaison entre les deux questions. Mais qu'est ce qu'une liaison ?

Une liaison entre deux variables est l'écart entre les données observées et le modèle d'indépendance. Mettons pour l'instant de côté cette notion, nous y reviendrons plus tard.

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

Figure 1:

Création du projet RStudio

Commençons par ouvrir le logiciel RStudio et par créer un nouveau projet depuis le menu File > New Project. Choisissons ensuite "New directory" > "Empty project" puis saisissons *AFC-sous-R* dans Directory name. Téléchargeons le fichier AnaDo_JeuDonnees_TravailFemme.csv du tableau de contingence et plaçons-le dans le répertoire "AFC-sous-R". Créons ensuite notre script R depuis File > New file > R script.

Définir la localisation du répertoire de travail

Il est très important, d'emblée, de définir le répertoire de travail. Il existe une commande `getwd()` qui permet d'afficher la localisation du répertoire de travail sous la forme d'un chemin absolu :

```
getwd()
```

```
## [1] "/Users/ubo/Documents/Recherche/R/AFC-sous-R"
```

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 dans la fenêtre en bas à droite de RStudio.

Importation du jeu de données

Nous lisons ensuite le fichier CSV à partir de la fonction `read.table()` :

```
wfemmes <- read.table("AnaDo_JeuDonnees_TravailFemme.csv", header=TRUE,
                      row.names=1, sep=";", check.names=FALSE, fileEncoding="latin1")
```

- La fonction `read.table()` permet de lire un fichier dans un format tabulaire et de créer une dataframe à partir de ce dernier.

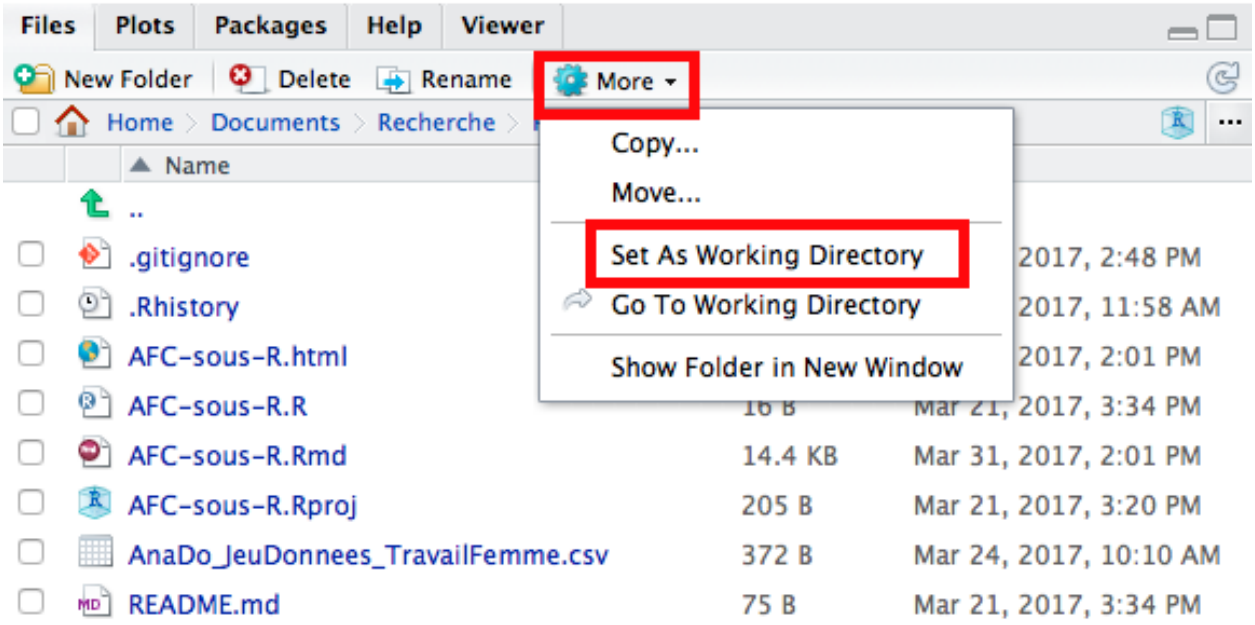


Figure 2:

- L'argument `header=TRUE`permet de spécifier que la première ligne du jeu de données correspond aux intitulés des colonnes.
- L'argument`row.names=1`permet de spécifier que le nom des lignes est contenu dans la première colonne.
- L'argument `sep=";"`indique que le séparateur de colonne est le point-virgule.
- L'argument `check.names=FALSE` permet d'empêcher R de modifier la syntaxe du nom des modalités de variable.
- 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, nous **stockons le résultat dans la variable `wfemmes`**.

Nous pouvons, très facilement et à n'importe quel moment, afficher le tableau de contingence `wfemmes` à partir du volet Data dans la fenêtre en haut à droite de RStudio.

Seules les trois premières colonnes du jeu de données sont utilisées ici. Nous avons donc besoin de suppprimer les colonnes dont nous n'aurons pas l'utilité. Plus spécifiquement, nous n'allons **sélectionner que les trois premières colonnes** avec la commande suivante :

```
wfemmes <- wfemmes[1:3]
```

Dans cette commande :

- En spécifiant `[1:3]` (entre crochets), nous indiquons que nous ne sélectionnons que les 3 premières colonnes.

Nous pouvons voir que, dans le volet Data dans la fenêtre en haut à droite de RStudio, le nombre de variables affichées a été réduit.

Afin d'amliorer la lecture du tableau et, par la suite, du graphique, nous allons renommer les modalités des

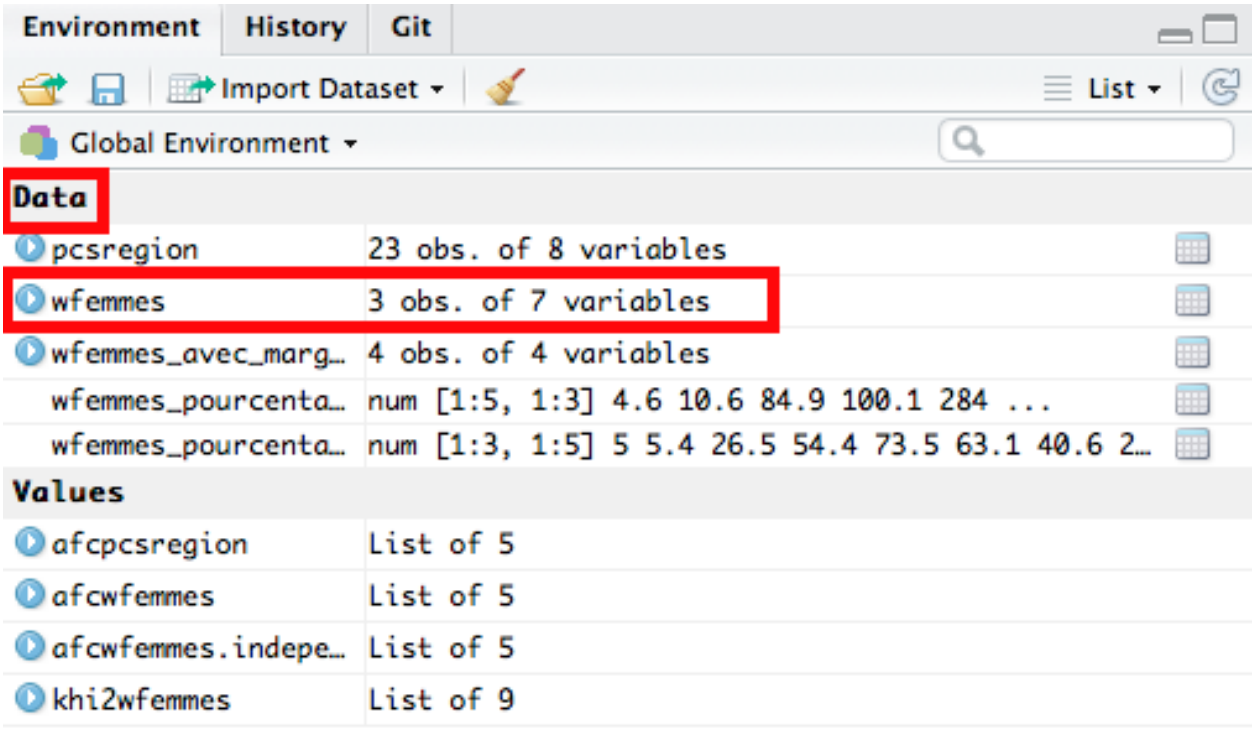


Figure 3:

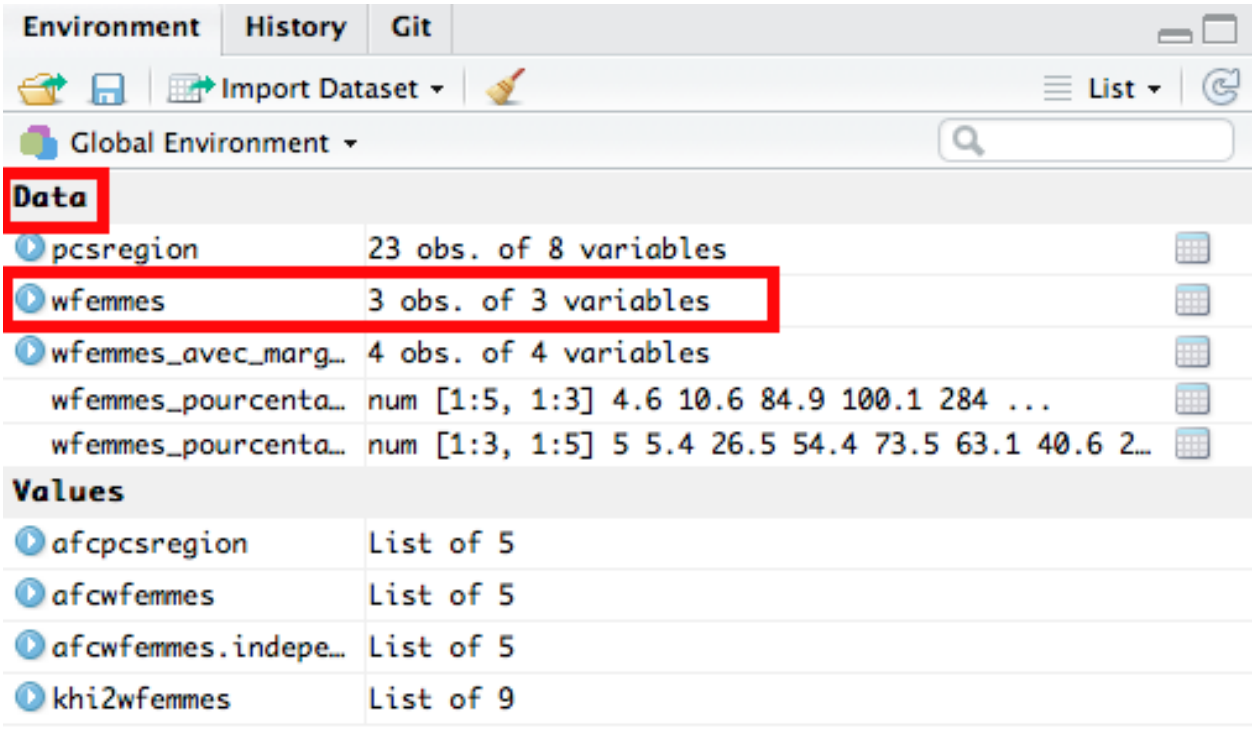


Figure 4:

deux variables grâce à la fonction `dimnames()` :

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

Pour la fonction `dimnames()` :

- `[[1]]` (entre deux doubles crochets) permet de spécifier que nous allons renommer une ou plusieurs modalités en ligne.
- `[[2]]` (entre deux doubles crochets) permet de spécifier que nous allons renommer une ou plusieurs modalités en colonne.
- `[1]` (entre crochets) permet de spécifier que nous allons renommer la première modalité d'une variable.
- `[2]` (enter crochets) permet de spécifier que nous allons renommer la deuxième modalité d'une variable.

Suivant cette logique, lorsque nous écrivons, par exemple, `(dimnames(wfemmes)[[2]][3]<-"...")`, cela signifie que nous allons renommer la troisième modalité en colonne.

Notez que nous pouvons calculer sous R les marges lignes et les marges colonnes du tableau de contingence de la manière suivante :

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

```
##                               Rester au foyer Travail à mi-temps
## Les 2 conjoints travaillent                13                142
## Travail du mari plus absorbant             30                408
## Seul le mari travaille                    241                573
## Total                                     284               1123
##                               Travail à plein-temps Total
## Les 2 conjoints travaillent                     106   261
## Travail du mari plus absorbant                  117   555
## Seul le mari travaille                           94   908
## Total                                           317  1724
```

Affichons ce tableau de manière plus agréable :

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Total
Les 2 conjoints travaillent	13	142	106	261
Travail du mari plus absorbant	30	408	117	555
Seul le mari travaille	241	573	94	908
Total	284	1123	317	1724

Il est aussi intéressant de calculer les pourcentages en ligne et les pourcentages en colonne avec la librairie `Rcmdr`.

```
library (Rcmdr)
```

```
## Loading required package: splines
```

```
## Loading required package: RcmdrMisc
```

```
## Loading required package: car
```

```
## Loading required package: sandwich
```

```
## L'interface graphique de R Commander n'est utilisable que dans des sessions interactives
```

```
wfemmes_pourcentage_en_ligne <- wfemmes
wfemmes_pourcentage_en_ligne[nrow(wfemmes_pourcentage_en_ligne)+1, ] <- colSums(wfemmes_pourcentage_en_l
row.names(wfemmes_pourcentage_en_ligne)[nrow(wfemmes_pourcentage_en_ligne)] <- "Profil ligne moyen"
wfemmes_pourcentage_en_ligne <- rowPercents(wfemmes_pourcentage_en_ligne)
```

Table 2: Tableau des pourcentages en ligne

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Total	Count
Les 2 conjoints travaillent	5.0	54.4	40.6	100	261
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

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.

Calculons maintenant le tableau des pourcentages en colonne.

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

Table 3: Tableau des pourcentages en colonne

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Profil colonne moyen
Les 2 conjoints travaillent	4.6	12.6	33.4	15.1
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

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 répartion 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.

Nous verrons également que l'on passera en paramètre à la fonction R de calcul de l'AFC, le tableau de contingence. Mais l'AFC travaille en réalité sur le tableau de probabilités que l'on peut calculer 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) :

```
wfemmes_tableau_de_probabilite <- wfemmes_avec_marges / 1724
```

Table 4: Tableau de probabilités

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Total
Les 2 conjoints travaillent	0.0075406	0.0823666	0.0614849	0.1513921
Travail du mari plus absorbant	0.0174014	0.2366589	0.0678654	0.3219258
Seul le mari travaille	0.1397912	0.3323666	0.0545244	0.5266821
Total	0.1647332	0.6513921	0.1838747	1.0000000

Rappellons que notre objectif est de **visualiser la nature de la liaison entre deux variables qualitatives**. Mais faut-il encore que cette liaison soit **significative**.

Pour ce faire, nous réalisons un **test du Khi2**.

Test du Khi2

Le test du khi2 mesure la **significativité d'une liaison mais pas son intensité**. Afin de réaliser ce test du khi2, nous utilisons une fonction fournie de base avec le logiciel R. Il s'agit de la fonction `chisq.test`. Il n'est pas nécessaire d'installer une librairie supplémentaire afin de réaliser ce test.

```
chisq.test(wfemmes)
```

```
##
##  Pearson's Chi-squared test
##
## data:  wfemmes
## X-squared = 233.43, df = 4, p-value < 2.2e-16
```

Le terme `X-squared` est à lire `Khi2` soit Khi au carré. La fonction `chisq.test` nous donne, entre autres, la valeur du Khi2 qui est un **indicateur de la significativité de la liaison**. Mais ce qui nous interesse ici est la `p-value`. Nous voyons ici que la `p-value` est inférieure à 2,2x10^{-16}. Cela signifie que la **probabilité que les variables soient indépendantes** est inférieure à 2,2x10^{-16}. Ce qui nous permet 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.

Test du Khi2 - Explications

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 et comme le rappelle très clairement Julien Barnier, 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 nulle (H0) suppose qu'il y a indépendance entre les deux variables. Si nous acceptons l'hypothèse d'indépendance (H0), nous n'aurons pas d'utilité à réaliser une AFC car les points projetés seront extrêmement proches ou confondus avec le centre de gravité, confondus avec le centre du graphe. Si nous rejetons l'hypothèse d'indépendance (p-value $< 0,05$), l'hypothèse alternative (H1) suppose que la liaison entre les deux variables est significative sans que nous puissions définir l'intensité de la liaison.

Rappelons que pour que le test du khi2 soit opératoire, il doit respecter un certain nombre de conditions (pour reprendre les propos de Claude Grasland) :

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

Du fait que nous ayons obtenu une p-value inférieure à $2{,}2 \times 10^{-16}$ 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 conditons de validité du test du khi2, nous stockons le résultat de la fonction dans la variable `khi2wfemmes`.

```
chisq.test(wfemmes) -> khi2wfemmes
```

Test du Khi2 - Aide à l'interprétation

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

Nous pouvons afficher le tableau de contingence d'origine (**tableau des données observées**) en sélectionnant la valeur `observed`.

```
khi2wfemmes$observed
```

```
##                                Rester au foyer Travail à mi-temps
## Les 2 conjoints travaillent                 13                142
## Travail du mari plus absorbant              30                408
## Seul le mari travaille                     241                573
##                                Travail à plein-temps
## Les 2 conjoints travaillent                      106
## Travail du mari plus absorbant                   117
## Seul le mari travaille                            94
```

De la même manière, nous pouvons afficher le tableau d'indépendance (**tableau des effectifs théoriques**) en sélectionnant la valeur `expected`. Dans ce contexte, nous calculons 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.

Pour plus de détails, nous vous recommandons la lecture de ce document de Julien Barnier : Tout ce que vous n'avez jamais voulu savoir sur le Khi2 sans jamais avoir eu envie de le demander.

```
khi2wfemmes$expected
```

```
##                                Rester au foyer Travail à mi-temps
## Les 2 conjoints travaillent           42.99536           170.0133
## Travail du mari plus absorbant        91.42691           361.5226
## Seul le mari travaille               149.57773           591.4640
##                                Travail à plein-temps
## Les 2 conjoints travaillent                 47.9913
## Travail du mari plus absorbant             102.0505
## Seul le mari travaille                     166.9582
```

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

Nous pouvons aussi 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 supérieur à ceux attendus sous l'hypothèse d'indépendance. Et l'inverse pour un résidu négatif.

```
khi2wfemmes$residuals
```

```
##                                Rester au foyer Travail à mi-temps
## Les 2 conjoints travaillent          -4.574496          -2.148441
## Travail du mari plus absorbant       -6.424239           2.444409
```

```
## Seul le mari travaille                7.475127         -0.759211
##                                Travail à plein-temps
## Les 2 conjoints travaillent                8.373594
## Travail du mari plus absorbant             1.479859
## Seul le mari travaille                    -5.646384
```

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. D'ailleurs, comme le note Philippe Cibois, 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)."

Par ailleurs,

- Un **écart à l'indépendance positif** correspond à une **attraction** entre les deux modalités pour la case observée.
- 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 bien 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 deux variables qualitatives**.

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

- La **significativité** de la liaison (qui se mesure avec le test du khi2).
- L'**intensité** de la liaison (qui se mesure, entre autre, avec le Phi2).
- 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. Pour notre part, nous avons choisi d'utiliser le **package FactoMineR** (dédié à l'analyse multidimensionnelle de données) mais il y en existe d'autres qui peuvent être utilisés pour réaliser ce type de méthode statistique.

Nous chargeons donc la librairie `FactoMineR` permettant de réaliser plusieurs analyses de données multi-dimentionnelles (AFC, ACP, ACM, etc.).

```
library (FactoMineR)
```

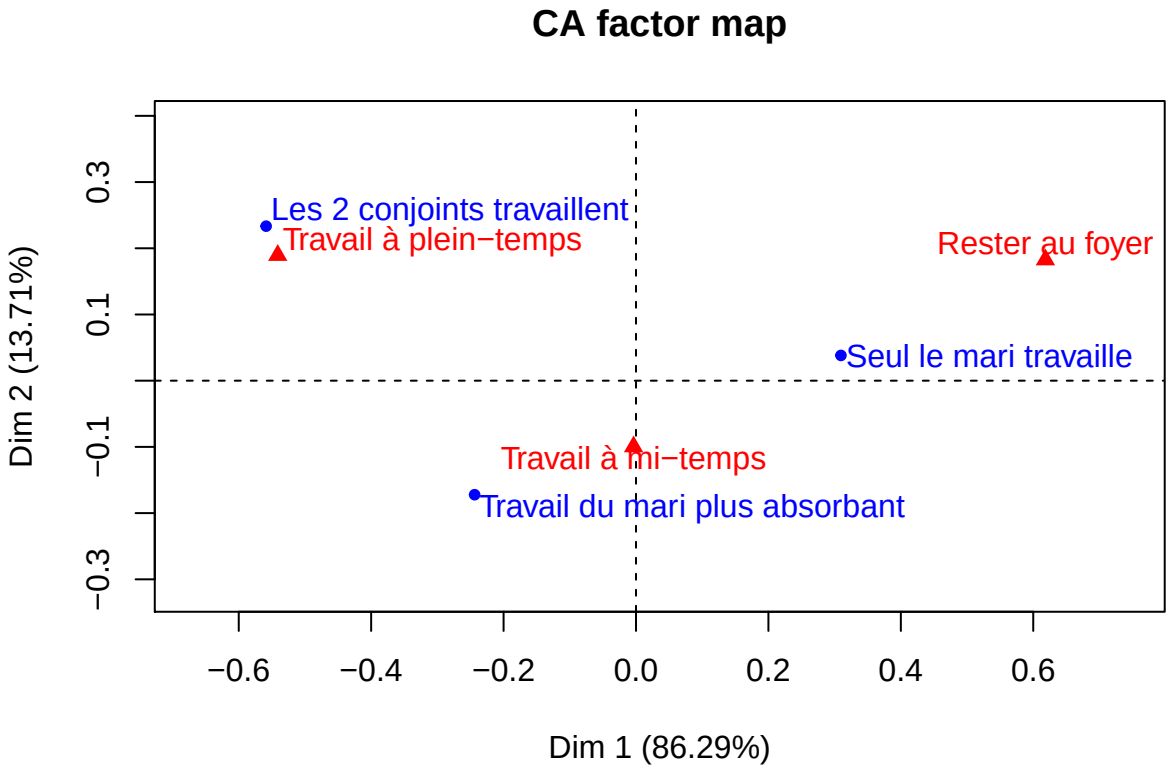
AFC - Les résultats

Lors du précédent test du khi2, nous avons obtenu une p-value inférieure à $2{,}2\text{x}10^{-16}$. Nous avons donc rejeté l'hypothèse d'indépendance entre les deux variables et admis que la liaison entre ces deux variables est signficative.

Nous sommes en droit de réaliser une AFC afin de visualiser la nature de la liaison. Pour ce faire, nous allons employer la fonction `CA()`, fournie par le package `FactoMineR` que nous venons de charger. Précédemment,

nous avons stocké notre jeu de données dans la variable `wfemmes`. Nous exécutons une analyse factorielle des correspondances (AFC) sur notre jeu de données de la manière qui suit :

```
afcwfemmes <- CA(wfemmes)
```



Afin de faciliter, par la suite, la lecture et l'emploi des résultats cette AFC, nous les **stockons dans la variable** `afcwfemmes`.

Revenons un instant sur ce fameux **tableau 37**, issu de l'enquête de Nicole Tabard, croisant les deux variables (questions) :

- Quelle est la famille idéale pour vous ?
- Quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ?

Il est important de rappeler que les résultats de cette enquête ont été publiés en 1974. Il est fort à parier que la répartition des réponses serait totalement, si ce n'est en grande partie, différente aujourd'hui.

Lors d'une première lecture de ce tableau de contingence, François Husson soulève une **apparente contradiction**. À la question "Quelle est la famille idéale pour vous ?", nous voyons 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, nous pouvons 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).

Du coup, à ce stade de l'interprétation, nous nous retrouvons *a priori* face une **contradiction**.

De cela, nous pouvons 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

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

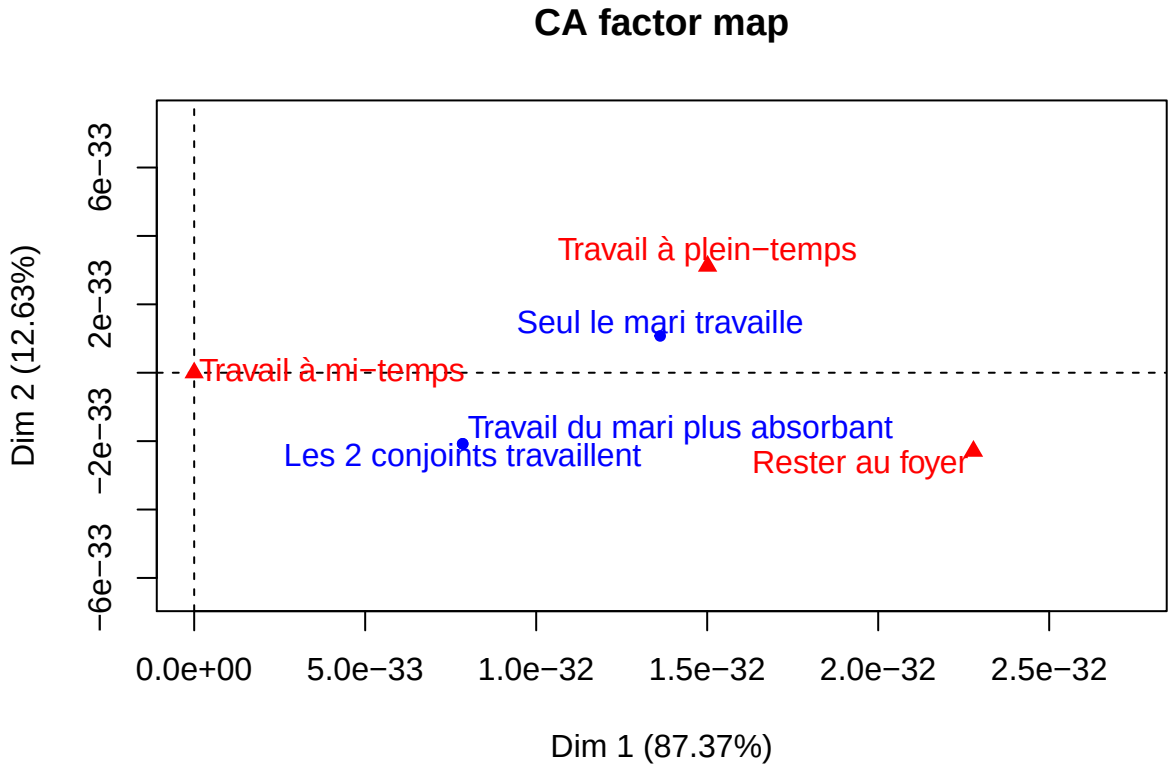
Figure 5:

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**.

Si nous réalisons une AFC avec les données du modèle d'indépendance :

```
afcwfemmes.independance <- CA(khi2wfemmes$expected)
```



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 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 :

- 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, confondus avec le centre du graphe.
- La réalisation d'un test du khi2 est donc fortement conseillée avant la réalisation d'une AFC.
- Plus précisément, le test du khi2 conditionne l'éventuelle réalisation d'une AFC.

Les résultats qui suivent vont aussi dans le sens de notre propos :

```
summary(afcwfemmes.independance)
```

```
##
## Call:
## CA(X = khi2wfemmes$expected)
##
## The chi square of independence between the two variables is equal to 2.223396e-29 (p-value =  1 ).
##
## Eigenvalues
##                        Dim.1   Dim.2
## Variance               0.000   0.000
## % of var.             87.369  12.631
## Cumulative % of var.  87.369 100.000
##
## Rows
##                                 Iner*1000   Dim.1 ctr cos2   Dim.2 ctr
```

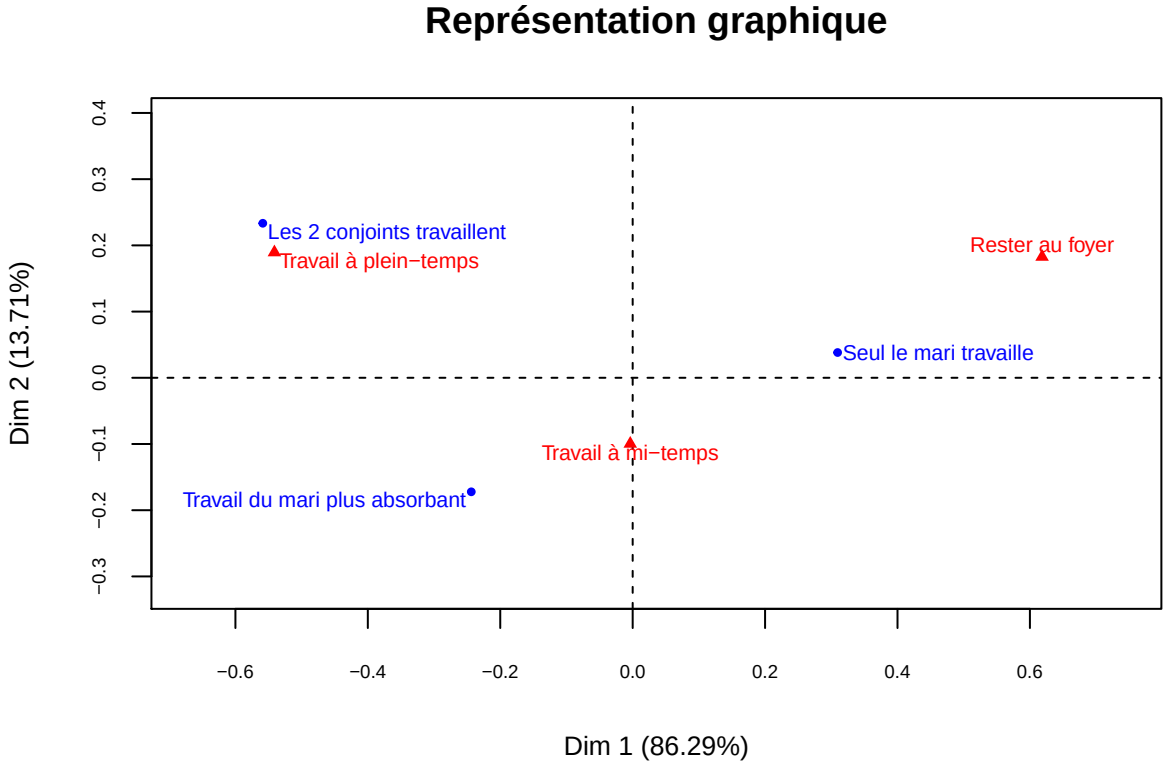
```
## Les 2 conjoints travaillent    |         0 |     0   0    0 |     0   0
## Travail du mari plus absorbant |         0 |     0   0    0 |     0   0
## Seul le mari travaille         |         0 |     0   0    0 |     0   0
##                                cos2
## Les 2 conjoints travaillent       0 |
## Travail du mari plus absorbant    0 |
## Seul le mari travaille            0 |
##
## Columns
##                                 Iner*1000   Dim.1 ctr cos2   Dim.2 ctr
## Rester au foyer                |         0 |     0   0    0 |     0   0
## Travail à mi-temps             |         0 |     0   0  NaN |     0   0
## Travail à plein-temps          |         0 |     0   0    0 |     0   0
##                                cos2
## Rester au foyer                   0 |
## Travail à mi-temps              NaN |
## Travail à plein-temps             0 |
```

Dans le cas où il y a indépendance entre les deux variables, le **test du khi2 nous donne une p-value = 1**. Les coordonnées (dim.n) des modalités en ligne et en colonne, sur chacune des dimensions, sont égales à 0. Il en va de même pour leur contribution à la construction (ctr) de chaque dimension et leur qualité de réprésentation (cos2) 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, ainsi que celles associées aux multiples packages disponibles, nous permettent l'affichage du graphe.

```
plot(afcwfemmes, cex = 0.7, cex.axis = 0.6, cex.lab = 0.8, title = "Représentation graphique",
     selectRow = "cos2 0.7", selectCol = "cos2 0.7")
```



Voici le détail des arguments que nous avons utilisés avec la fonction `plot()` :

- L'argument `cex=` permet de modifier la taille des points et des textes accompagnant ces points.
- L'argument `cex.axis` permet de modifier la taille des annotations d'axe.
- L'argument `cex.lab` permet de modifier la taille des intitulés d'axe.
- L'argument `title=` permet de donner un titre au graphique.
- L'argument `selectRow=` permet de sélectionner les modalités en ligne à afficher en fonction de critères particuliers.
- L'argument `selectCol=` permet de sélectionner les modalités en colonne à afficher en fonction de critères particulier.

Ce ne sont que quelques arguments, parmi d'autres, en vue d'améliorer la représentation graphique. 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.

Appuyons-nous sur un exemple pour illustrer notre propos. Nous allons utiliser des données de l'INSEE sur les spécificités socioprofessionnelles régionales.

Importons donc le jeu de données.

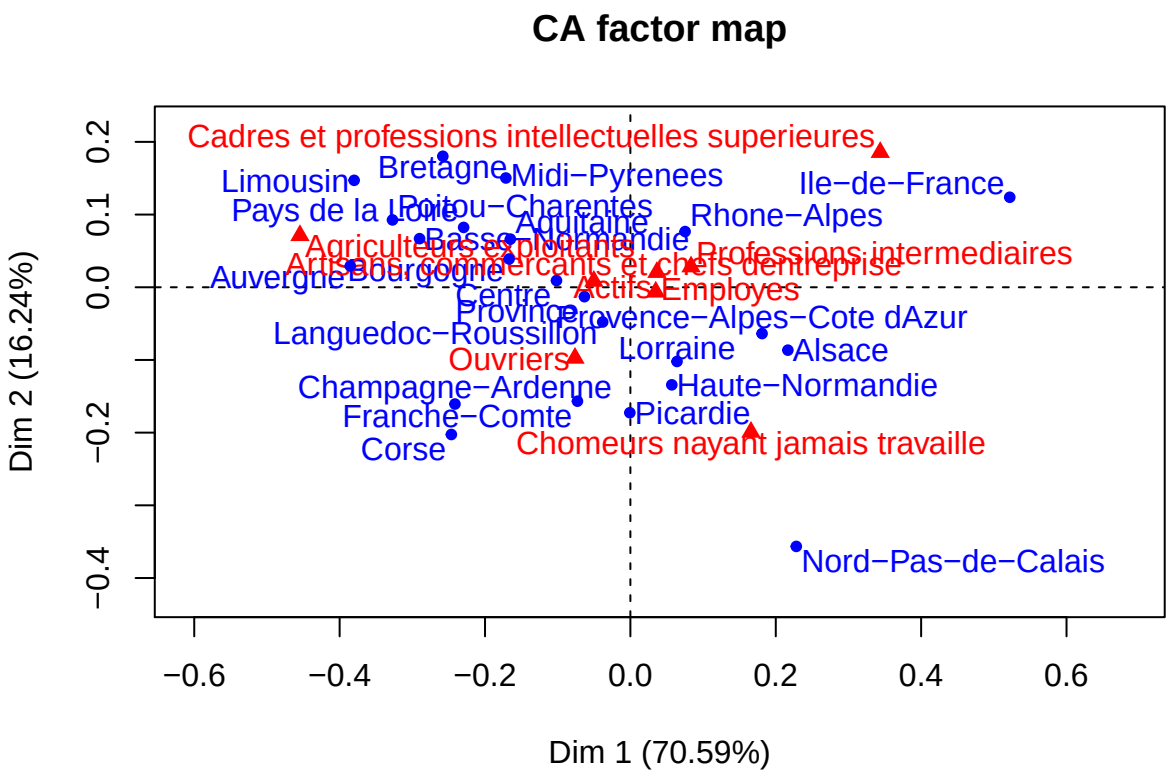
```
pcsregion <- read.table("pcsregion.csv", header=TRUE, row.names=1,
                        sep=";", check.names=FALSE, fileEncoding="latin1")
```

Lisons le jeu de données.

```
View(pcsregion)
```

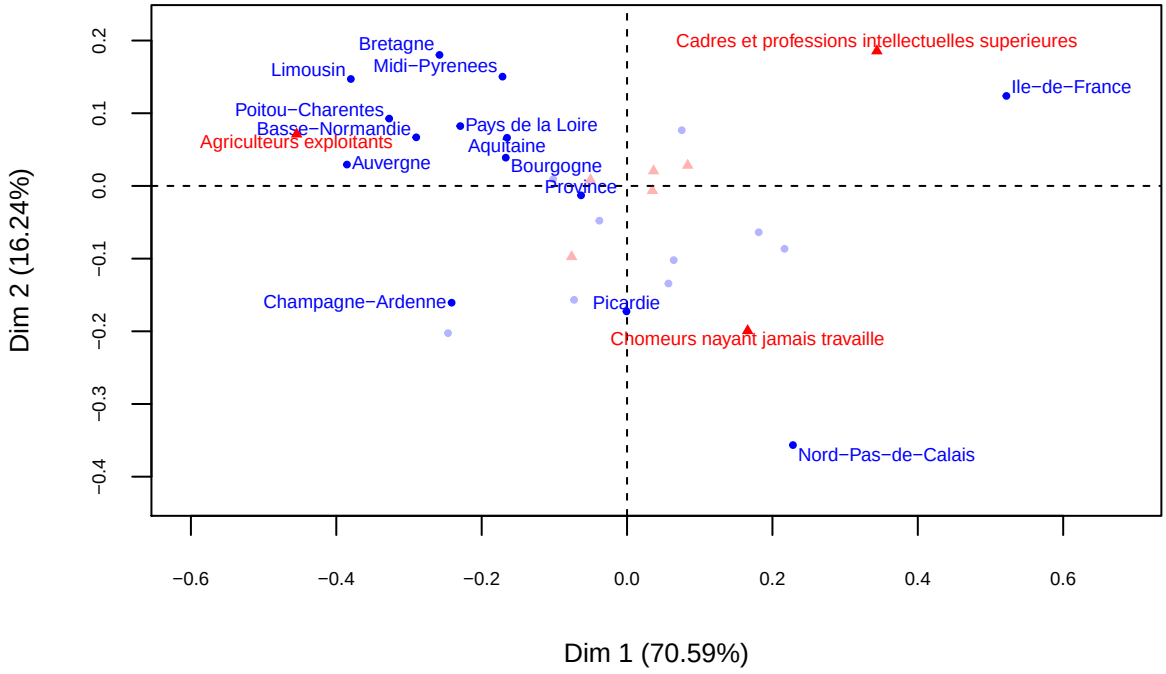
Lorsque nous réalisons une AFC avec ce jeu de données, nous obtenons ceci :

```
afcpcsregion <- CA(pcsregion)
```



Soyons honnêtes. Avec "seulement" 30 modalités (8 modalités en colonne et 23 modalités en ligne), nous voyons que **le graphe est (en grossissant légèrement le trait) "totalement" illisible**. C'est dans ce contexte que l'utilisation des différents arguments, que nous avons détaillés plus haut autour de la fonction `plot()`, prennent leur véritable importance.

```
plot(afcpcsregion, cex = 0.6, cex.axis = 0.6, cex.lab = 0.8,
     title = "Spécificités socioprofessionnelles régionales",
     selectRow = "cos2 0.7", selectCol = "cos2 0.7")
```



De cette manière, nous pouvons, par exemple, jouer sur la taille des noms de modalité et ne sélectionner que celles qui ont une qualité de représentation significative (arbitrairement cos2 $> 0{,}7$). Alors, bien évidemment, il ne s'agit que d'un exemple pour souligner l'éventail des possibilités mis à notre disposition en vue d'améliorer la représentation graphique, ainsi que la lecture et l'analyse des données.

Revenons maintenant à nos données initiales quant aux représentations sur le travail féminin.

Le nombre d'axes à interpréter

Au-delà de la simple lecture visuelle du graphe, il existe d'autres moyens d'interpréter nos données. À partir de là, il serait intéressant de déterminer le nombre d'axes à interpréter.

Les valeurs propres (i.e. variance) des axes :

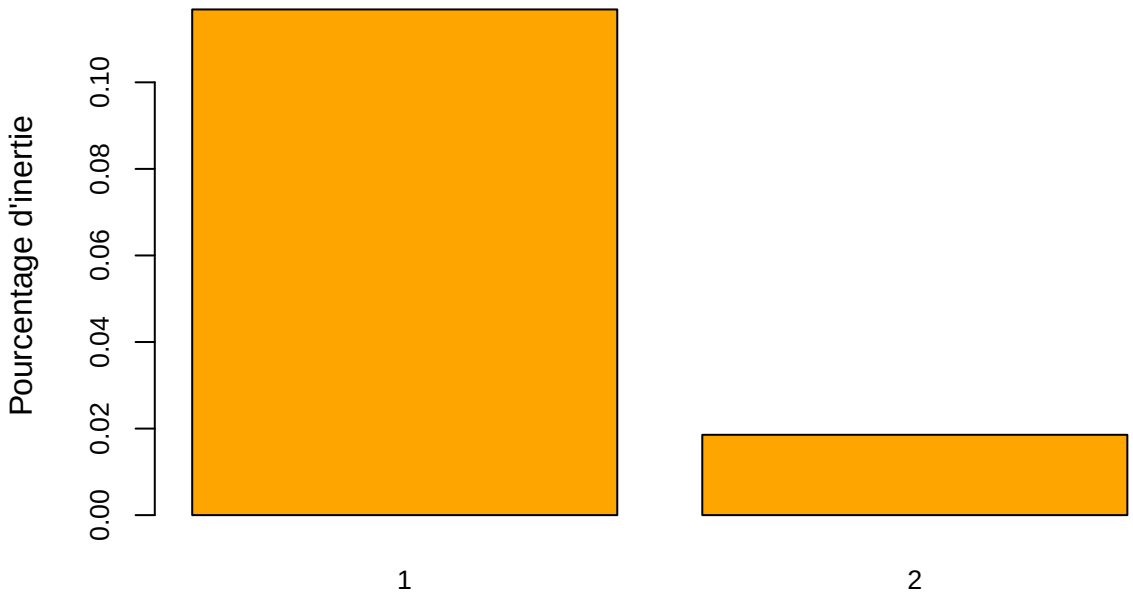
```
afcwfemmes$eig
```

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.11684002              86.29218                          86.29218
## dim 2 0.01856045              13.70782                         100.00000
```

Dans le résultat de `afcwfemmes$eig`, les axes (aussi appelés *facteurs*) sont appelés `dim`. La *valeur propre* ou `eigenvalue` représente la quantité d'information expliquée par l'axe. La deuxième colonne renvoie au *pourcentage d'inertie*, correspondant à la proportion des variations totales des variables originales expliquées par chaque axe.

Nous pouvons, à la suite, effectuer une représentation de l'histogramme des valeurs propres :

```
barplot(afcwfemmes$eig[,1], main = "Histogramme des valeurs propres",
        sub = "Représentations des femmes sur le travail féminin",
        xlab = "Axes", ylab = "Pourcentage d'inertie", cex.axis = 0.8,
        cex.names = 0.8, col ="orange", names.arg = 1:nrow(afcwfemmes$eig))
```



Détails des arguments utilisés dans la fonction `barplot()` :

- L'argument `main=` correspond au titre du graphique.
- L'argument `sub=` correspond au sous-titre du graphique.
- L'argument `xlab=` correspond au titre de l'axe x.
- L'argument `ylab=` correspond au titre de l'axe y.
- L'argument `cex.axis=` permet de modifier la taille des valeurs numériques du graphique.
- L'argument `cex.names=` permet de modifier la taille des valeurs lettrées du graphique.
- L'argument `col=` permet de modifier la couleur des barres.
- L'argument `names.arg=` permet de nommer chaque barre.

L'histogramme va servir visuellement à déterminer le nombre d'axes à interpréter à l'aide du critère du "coude" de Cattell. Lorsqu'on observe un décrochement entre deux axes, on détermine les axes avant ce décrochement comme étant les axes à interpréter. Ce n'est pas la seule règle possible. Un autre critère de choix, appelé « critère de Kaiser », suggère de garder tous les axes dont la valeur propre est supérieure à la valeur propre moyenne (qui vaut par définition 1 divisé par le nombre d'axes). Pour plus d'informations, nous vous invitons à vous reporter au lien suivant afin d'approfondir ce travail quant au nombre d'axes à retenir.

Ce jeu de données ne permet pas de réellement saisir cette démarche qui vise à déterminer le nombre d'axes à interpréter. Pour ce faire, reprenons nos données sur les spécificités socioprofessionnelles régionales.

Voici les valeurs propres des axes :

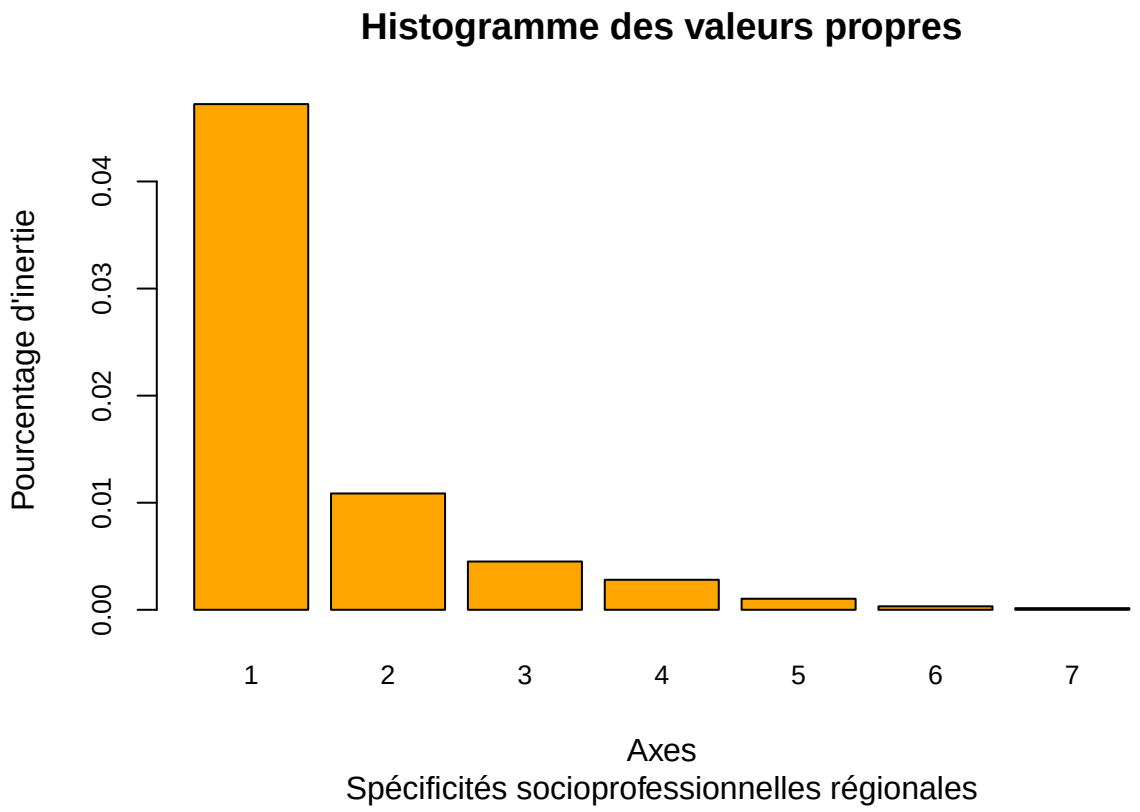
```
afcpcsregion$eig
```

```
##         eigenvalue percentage of variance
## dim 1 0.0472259511             70.5870208
## dim 2 0.0108625347             16.2358606
## dim 3 0.0045071587              6.7366965
## dim 4 0.0028026530              4.1890299
## dim 5 0.0010305300              1.5402981
```

```
## dim 6 0.0003226686              0.4822818
## dim 7 0.0001530858              0.2288122
##       cumulative percentage of variance
## dim 1                           70.58702
## dim 2                           86.82288
## dim 3                           93.55958
## dim 4                           97.74861
## dim 5                           99.28891
## dim 6                           99.77119
## dim 7                          100.00000
```

Dans ce nouvel exemple, nous abtenons sept facteurs (ou axes). Présentons, dans la foulée, l'histogramme des valeurs propres :

```
barplot(afcpcsregion$eig[,1], main = "Histogramme des valeurs propres",
        sub = "Spécificités socioprofessionnelles régionales",
        xlab = "Axes", ylab = "Pourcentage d'inertie", cex.axis = 0.8,
        cex.names = 0.8, col ="orange", names.arg = 1:nrow(afcpcsregion$eig))
```



À la vue de cet histogramme, nous pourrions ne retenir que les deux premières dimensions et réaliser une AFC qu'en fonction de ces deux dernières. Toutefois, rien ne vous empêche de "relancer" une AFC avec la dimension 3, voire la dimension 4. Ce n'est pas parce que les deux premières dimensions sont les plus explicatives que les suivantes n'ont rien à dire.

Les résultats

La fonction `summary` est extrêmement utile dans ce contexte.

```r
summary(afcwfemmes, nbelements = Inf)
```

```
##
## Call:
## CA(X = wfemmes)
##
## The chi square of independence between the two variables is equal to 233.4304 (p-value =  2.410248e-
##
## Eigenvalues
##                        Dim.1   Dim.2
## Variance               0.117   0.019
## % of var.             86.292  13.708
## Cumulative % of var.  86.292 100.000
##
## Rows
##                                  Iner*1000    Dim.1    ctr   cos2    Dim.2
## Les 2 conjoints travaillent    |    55.487 | -0.559 40.432  0.851 |  0.233
## Travail du mari plus absorbant |    28.675 | -0.244 16.371  0.667 | -0.172
## Seul le mari travaille         |    51.239 |  0.310 43.197  0.985 |  0.038
##                                    ctr   cos2
## Les 2 conjoints travaillent    44.429  0.149 |
## Travail du mari plus absorbant 51.436  0.333 |
## Seul le mari travaille          4.135  0.015 |
##
## Columns
##                                  Iner*1000    Dim.1    ctr   cos2    Dim.2
## Rester au foyer                |    68.489 |  0.618 53.913  0.920 |  0.183
## Travail à mi-temps             |     6.478 | -0.004  0.007  0.001 | -0.100
## Travail à plein-temps          |    60.434 | -0.541 46.079  0.891 |  0.189
##                                    ctr   cos2
## Rester au foyer                29.613  0.080 |
## Travail à mi-temps             34.853  0.999 |
## Travail à plein-temps          35.533  0.109 |
```

Cette fonction `summary` nous permet d'obtenir :

- Le résultat du test du khi2 (uniquement sur les lignes et les colonnes actives) avec la p-value.
- Un tableau avec les valeurs propres, les pourcentages d'inertie associés à chaque dimension.
- Un tableau avec les résultats sur les lignes actives avec leur coordonnée (dim.n) sur chaque dimension, leur contribution à la construction (ctr) de chaque dimension et leur qualité de réprésentation (cos2) sur chaque dimension.
- Un tableau avec les résultats sur les colonnes actives (dim.n, ctr, cos2).
- Un tableau (optionnel) avec les résultats sur les éléments supplémentaires en ligne avec la coordonnée (dim.n) et la qualité de représentation (cos2).
- Un tableau (optionnel) avec les résultats sur les éléments supplémentaires en colonne avec la coordonnée (dim.n) et la qualité de représentation (cos2).

Par ailleurs, la fonction `summary` n'affichera, par défaut, que les 10 réponses les plus significatives.

- L'argument `nbelements=Inf` permet de retirer cette limite.
- Dans l'exemple qui nous concerne ce n'est utile puisque chacune des deux variables ne dispose que de trois modalités.
- Mais, dans un certain nombre de cas, cette argument est précieux.

Ce qu'il appelle 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`

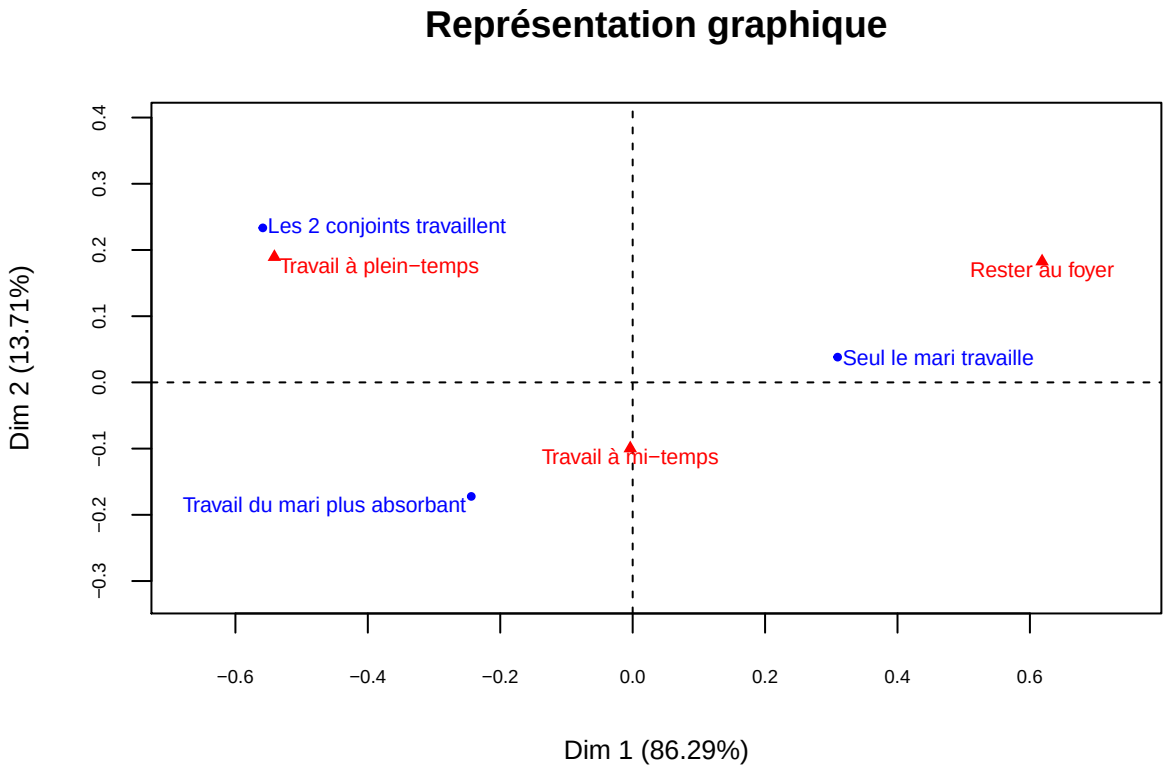
À quoi comparer cette valeur ?

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.

Le `% of var`, soit le pourcentage de variance, représente le **% de variance expliqué par l'axe**. Autrement dit, la dimension 1 (ou l'axe 1) explique 86.292 % de la variance entre les deux variables. Et le plan constitué des axes 1 et 2 explique 100% de la variance. Ce qui est normal puisque les pourcentages se répartissent sur tous les axes. S'il y avait eu 8 axes, le pourcentage d'inertie se serait réparti sur les 8 axes.

Ce qu'il faut savoir c'est que le nombre d'axes est égal au nombre minimum entre le nombre de colonnes - 1 et le nombre de lignes - 1 du tableau de contingence. Autrement, si vous avez un tableau de contingence avec en ligne 9 modalités et en colonnes 15 modalités. On fait 9 - 1 = 8 puis 15 - 1 = 14 et on garde 8.

Interpréter la position des points sur le graphe



Le centre de l'axe correspond au profil ligne moyen et au profil colonne moyen. Nous avions précedemment 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.

Table 5: Tableau des pourcentages en ligne

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Total	Count
Les 2 conjoints travaillent	5.0	54.4	40.6	100	261
Travail du mari plus absorbant	5.4	73.5	21.1	100	555

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Total	Count
Seul le mari travaille	26.5	63.1	10.4	100	908
Profil ligne moyen	16.5	65.1	18.4	100	1724

Table 6: Tableau des pourcentages en colonne

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Profil colonne moyen
Les 2 conjoints travaillent	4.6	12.6	33.4	15.1
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

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 peut ne peut pas interpréter la proximité entre un point ligne et un point colonne. On peut seulement dire que 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	Travail à mi-temps	Travail à plein-temps	Total
Les 2 conjoints travaillent	13	142	106	261
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 meileur score commparé aux autres colonnes. C’est pourqoui on peut dire que“Seul le mari travaille” est “du côté” de “Rester au foyer”.

Table 8: Tableau des pourcentages en colonne

	Rester au foyer	Travail à mi-temps	Travail à plein-temps	Profil colonne moyen
Les 2 conjoints travaillent	4.6	12.6	33.4	15.1
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

Conclusion

Comme nous l'avons annoncé en introduction, notre 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. Avec ce premier travail, nous avons apporté un certain nombre d'éléments de lecture, d'analyse et d'interprétation. Nous pousuivrons cet exercice **dans une seconde partie** prochainement. Nous chercherons, entre autres, à approfondir la question de l'interprétation de l'AFC (détermination du nombre d'axes à interpréter, interprétation des coordonnées des modalités actifs, interprétation de la contribution et de la qualtité de représentation des modalités, élaboration de la typologie, etc.).