Les méthodes d'analyse d'enquêtes

Philippe Cibois

Professeur émérite de sociologie.
Université de Versailles – St-Quentin en Yvelines

Introduction

La procédure d'enquête est utilisée quand, dans un domaine donné, on se trouve confronté à une situation d'incertitude quant aux causes d'un état de chose. De ce fait on est amené à poser des questions, souvent à des personnes, pour inventorier leurs opinions, leurs pratiques, leur situation, leur passé. De ce vaste coup de filet sans hypothèse préalable, on espère tirer des explications sur les phénomènes en cause.

Cette procédure d'enquête est souvent couteuse en temps et en moyens mobilisés mais le résultat est souvent décevant car ceux qui font des enquêtes ne disposent pas en général de méthodes pour les explorer en profondeur et se contentent donc de résultats superficiels. Des méthodes efficaces existent cependant, certaines datant des années soixante comme l'analyse factorielle des correspondances, d'autres plus récentes comme la régression logistique.

Le but du présent ouvrage est de donner au créateur d'enquête les moyens de comprendre les méthodes qui lui permettront, en utilisant les logiciels disponibles, de réaliser lui-même un dépouillement d'enquête efficace.

La stratégie utilisée pour mettre en œuvre les méthodes d'analyse est de respecter la situation d'incertitude de départ et de ne pas imposer une méthode qui force les résultats dans un sens ou un autre mais qui laisse émerger d'éventuelles surprises. A cette fin le processus d'analyse sera caractérisé par l'utilisation du concept de *variable d'intérêt* : tout enquête est faite quand on est face à un phénomène dont on veut rendre compte et cette focalisation détermine une ou plusieurs "variables d'intérêt" dont on veut rendre compte. On proposera donc une première méthode qui consiste à repérer quelles sont les questions de l'enquête qui sont le plus liées à cette variable d'intérêt. On utilisera à cette fin le PEM, *Pourcentage de l'Ecart Maximum* qui permet de faire ce travail. Comme cette méthode est très simple au point de vue théorique elle permettra de comprendre les notions *d'indépendance* dans un tableau et *d'écart à l'indépendance*, qui sont indispensables pour la bonne intelligence des méthodes suivantes.

Une fois repérées les variables qui sont liées au phénomène étudié, on utilisera une méthode, *l'analyse des correspondances*, qui permettra de faire une analyse globale du phénomène, c'est-à-dire qui positionnera les différentes modalités de la variable d'intérêt dans un univers de modalités suffisamment riche pour que des hypothèses de travail puissent en être issues, mais suffisamment limité pour que l'analyse ne soit pas submergée par trop de données.

Une fois cette vue d'ensemble établie, l'analyse se focalisera sur des points précis qui demandent une investigation complémentaire car l'analyse précédente, comme une carte qui englobe un vaste territoire, est peu précise et trop incertaine. De l'analyse globale, on passe à *l'analyse locale*, et de l'hypothèse de travail à sa vérification.

Dans ce but un utilisera également la *régression logistique* sur données d'enquêtes qui permet d'estimer l'effet d'une variable sur une autre "toutes choses égales par ailleurs", c'est-à-dire par exemple en neutralisant l'effet de l'âge quand on étudie l'effet de l'origine sociale.

On montrera enfin qu'il est possible de retrouver dans la population observée des types de répondants en classant les individus en fonction des résultats précédents. Ce retour aux données est une précaution qui permet de vérifier la résistance des résultats et d'éviter que les *types-idéaux* obtenus ne s'ancrent pas assez dans la réalité.

Pour chaque méthode, on partira d'exemples simples pour faire comprendre les concepts utilisés, sans en donner les justifications mathématiques qui n'ont pas leur place dans un ouvrage introductif. Ensuite un exemple en vraie grandeur suivi tout au long du livre permettra de juger des capacités de la démarche.

Vingt ans après

Ce livre se substitue au *Que sais-je ?* 2095 intitulé *l'Analyse factorielle* paru en 1983. En effet les attentes des lecteurs ne sont plus les mêmes : quand une nouvelle technique apparait, on cherche à comprendre comment elle fonctionne et on soulève volontiers le couvercle pour démonter l'intérieur. Dans les années 1980, je me souviens avoir dû expliquer comment fonctionnait un ordinateur mais ces temps sont révolus : on n'éprouve plus ce besoin aujourd'hui pas plus que pour le téléphone ou pour un moteur électrique. Pour utiliser le vocabulaire de la sociologie des sciences[1], l'ordinateur est utilisé aujourd'hui comme une *boite noire* : on veut n'en connaitre que ce qui est utile à un bon usage.

Il en est de même pour les techniques statistiques : en vingt ans d'enseignement régulier de ces techniques, j'ai vu la demande des utilisateurs évoluer, passant d'un désir très fort de savoir comment l'analyse factorielle produisait ses résultats à un objectif différent, comment bien utiliser la méthode. Le présent livre prend acte de cette évolution : la part du principe de la méthode y est réduite pour laisser plus de place à des exemples commentés d'utilisation et à des règles de bonne pratique. Cependant des annexes permettront à ceux qui le désirent de parvenir à une compréhension des algorithmes utilisés par l'analyse des correspondances et la régression logistique.

Comme dans le précédent ouvrage, je veux redire ici ma dette à l'enseignement de Georges Th. Guilbaud qui est à l'origine de la présentation de l'analyse des correspondances faite ici, ainsi qu'aux formalisations opérées par Henry Rouanet et son équipe. Pour la régression logistique, Louis-André Vallet reste celui vers lequel on peut toujours se tourner en cas de doute sur la méthode.

[1] Dominique Vinck, *Sociologie des sciences*, Paris, A. Colin, 1995

Chapitre 1 : repérer les questions pertinentes

On suppose donc une enquête déjà existante, soit issue d'une recherche, soit en vue de l'analyse secondaire d'une enquête disponible et rendue accessible aux chercheurs. On suppose que les données de cette enquête sont utilisables par le biais soit d'un logiciel international comme SAS ou SPSS, soit d'un logiciel libre comme Trideux, développé par l'auteur et dont les exemples de ce livre sont issus. Les méthodes de dépouillement présentées ici sont indépendantes des logiciels : les aspects pratiques en dépendent évidemment et il faudra s'y reporter pour plus de détails.

On suppose donc que l'on a un fichier d'individus dont le nombre est variable, qui peut aller de quelques dizaines à plusieurs centaines de milliers : il peut sembler paradoxal d'envisager un dépouillement d'enquête avec moins de cent individus mais regarder attentivement le contenu d'un fichier est un objectif valable même si la possibilité d'étendre les résultats obtenus à une population de référence est faible. Quand on fait une enquête, quelque soit le nombre d'individus, on veut légitimement avoir une description de la population enquêtée : si l'effectif en est faible, on ne pourra que constater l'état de la population ; si l'effectif est important on pourra généraliser les résultats à la population dont l'enquête est issue, sous réserve que l'échantillon a été prélevé de manière raisonnée, par exemple par la méthode des quotas ou en sélectionnant des populations spécifiques. Il faut bien distinguer la description des données d'une part, des résultats qui peuvent être généralisés à l'ensemble de la population étudiée d'autre part. Pour pouvoir généraliser on utilisera des tests statistiques, essentiellement celui du khi-deux que l'on supposera connu : on se souviendra que le khi-deux étant sensible aux effectifs, dès qu'une population d'enquêtés devient importante, il devient rare que le khi-deux d'un tableau croisé ne soit pas significatif.

Dans la suite, on utilisera comme exemple des données assez classiques en terme d'effectif, c'est-à-dire de plusieurs centaines d'individus. Il ne faut cependant pas avoir peur des faibles effectifs car on peut faire une bonne description d'une centaine d'individus. Même si on ne peut généraliser les résultats obtenus à la population de référence d'une manière certaine, si la cohérence des résultats est grande, leur valeur probatoire apparaitra aux lecteurs qui les considéreront comme des pistes à poursuivre, des tendances intéressantes à explorer par d'autres enquêtes.

Quand, dans une recherche de type ethnologique, on dispose de peu d'informateurs, on leur demande d'être de bonne qualité et personne ne se formalise de leur faible effectif. Quand dans une recherche historique on nous présente le cas particulier d'un petit gentilhomme du Cotentin qui a tenu tout au long de son existence un journal, les généralisations qui sont faites à partir de ce cas unique n'ont de valeur que dans la mesure où l'on s'assure que notre homme est représentatif de son corps social[2]. C'est d'ailleurs ce qui permet à la micro-histoire de porter tous ses fruits et c'est une attitude analogue qui doit conduire celui qui a peu de données à les utiliser au mieux, à en tirer toutes les informations, à être suffisamment motivé pour aller le plus loin possible (sans tomber dans l'acharnement

[2] Madeleine Foisil, *Le sire de Gouberville, un gentilhomme normand au XVI^e siècle*, Flammarion, 2001

de celui qui veut obtenir une certaine orientation des résultats, mais le débutant est plutôt trop modeste dans ses prétentions).

A l'inverse, celui qui a beaucoup d'individus et qui leur a posé beaucoup de questions risque d'être noyé par la masse d'informations dont il dispose potentiellement. Dans ce cas également, une bonne description des données s'impose : les méthodes descriptives sont les mêmes dans les deux cas, ce n'est affaire que de degrés, de nombre de questions que l'on peut prendre en compte dans une même analyse.

On suppose donc que l'on a une population d'individus à laquelle on a posé un certain nombre de questions. Ces questions seront de deux catégories : des questions d'opinion ou relatives aux pratiques de l'individu dans le domaine enquêté, ou pouvant l'éclairer ; des questions indépendantes de l'enquête mais relatives à la connaissance de l'enquêté en général telles que l'âge, le sexe, la catégorie socioprofessionnelle, le plus haut diplôme obtenu, ou d'autres encore comme le revenu, l'affiliation politique (ou religieuse) qui relèvent de l'opinion ou de la description mais qui ont en commun de ne pas être spécifiques d'une enquête donnée.

On supposera dans la suite que toutes ces questions seront utilisées sous formes de catégories, de modalités : la question "sexe" a deux modalités de réponse ; masculin ou féminin. La variable âge qui a pu être recueillie en codant l'âge en clair doit être recodée en tranches d'âges : le recueil des données doit se faire, non pas au niveau le plus fin possible (pour l'âge, la date de naissance précise avec le jour et le mois) mais au niveau le plus fin *pertinent* : pour des adultes ce sera l'année, pour des enfants du primaire, ce peut être le trimestre, pour des plus jeunes, le mois ou une durée moindre. Il faut que ce qui soit recueilli soit pertinent pour la finalité de l'enquête étant entendu que l'on pourra toujours facilement recoder les données d'une manière logicielle : par exemple, il est bon de recueillir les données en mettant des catégories d'accord ou de désaccord qui respectent les nuances d'une opinion (tout à fait d'accord, à peu près d'accord, etc.). Dans le courant du dépouillement, il sera souvent utile d'effacer ces nuances, de perdre de l'information pour en gagner par ailleurs par confrontation à de nombreuses autres modalités.

Si on doit éviter de proposer la non-réponse à un enquêté, on doit l'enregistrer : on proposera dans la suite des méthodes qui permettent de tirer de l'information de ce type de modalité par comparaison avec les réponses que l'enquêté a donné aux autres questions. On ne doit pas éliminer les non répondants : ils peuvent être porteurs d'une attitude qu'il faut éventuellement prendre en compte.

I Première étape : les préalables

On suppose donc que l'on dispose d'une population qui a répondu à de nombreuses questions. Les réponses sont enregistrées dans un logiciel et le travail minimum a été fait, c'est-à-dire que pour chaque question on dispose d'un identifiant alphanumérique de quelques caractères et que les modalités ne sont repérées que par leur numéro. Par exemple la question V02 qui se trouve être le sexe du répondant, a deux modalités de réponse dont on sait par le biais des guides de codage que la modalité notée 1 correspond au sexe masculin et la modalité 2 au sexe féminin. S'il y avait des non-réponses, elle seraient codées 0 (c'est rare pour cette question, quand cela arrive, on a vérifié que ce sont souvent des hommes pour

qui c'est "normal" d'être homme alors qu'une femme n'oublie jamais sa condition féminine).

La première opération à faire est de confectionner un instrument de travail que l'on imprimera immédiatement et qui est la distribution des réponses à toutes les questions, appelé souvent "tri à plat" des réponses, par opposition à "tri croisé" qui fait intervenir plusieurs questions en même temps. Les résultats peuvent être de cette forme :

```
Question V02 Code-max.   2
Tot.     1     2
512    244   268
100   47.7  52.3
```

Il s'agit de la question V02 (en fait le sexe de l'enfant car il s'agit d'une enquête de sociologie concernant une population d'enfant au collège). Le code maximum que peut prendre cette question est évidemment 2. Le total est de 512 individus qui correspondent à 100%. Il y a 244 individus codés 1 c'est-à-dire de sexe masculin, qui représentent 47,7% de la population, et 268 de sexe féminin représentant 52,3%. Il n'y a pas de non-réponses[3].

Rapidement, on éprouvera le besoin de rendre les résultats plus lisibles et de faire en sorte que la modalité 1 soit codée "masculin", la 2 "féminin" et que la question V02 s'appelle "sexe". Un tel travail ne doit pas être fait au départ pour deux raisons : il est rapidement décourageant par le temps de travail qu'il demande ; il ne doit être fait que sur des questions, des variables (les deux mots sont assez interchangeables) qui ont été travaillées, étudiées, recodées éventuellement, que l'on s'est appropriée par un examen attentif.

Une enquête se dépouille en prenant un certain temps, variable selon la dextérité informatique et selon les désirs, la motivation de celui qui fait le dépouillement : qui manie bien un logiciel peut espérer en une semaine complète de travail arriver à des résultats non négligeables. Il faut de ce fait conserver une trace écrite des opérations faites : ouvrir un fichier de traitement de texte et y reporter les résultats intermédiaires et les recodages faits est une bonne pratique. On se constitue ainsi un *journal de bord* des opérations faites qui constitue un grand secours quand on est obligé de reprendre un traitement après quelques jours.

Il faut garder trace de la manière dont on a construit des variables nouvelles à partir d'anciennes. Par exemple pour l'enquête scolaire qui nous sert d'exemple, on dispose évidemment de l'année de naissance de l'enfant et de la classe où il est arrivé. En tenant compte de ces deux indications, on peut construire une variable nouvelle, que l'on va appeler AGS en code simplifié, "Age scolaire" en clair et qui aura trois modalités : 1 = "en avance ", 2 = "à l'heure", 3="en retard"[4]. Ceux qui sont en avance représentent 18% de l'ensemble, ils sont 74% à être à l'heure et

[3] Sachant que 0,1% de la population correspond à une demi individu, une précision plus grande serait illusoire. On arrondi au plus près et l'on garde toujours un chiffre après la virgule, quelque soit la précision, afin de bien distinguer typographiquement les effectifs observés, qui sont toujours des entiers, de ce qui relève d'un calcul comme les pourcentages.

[4] autre recodage possible : "à l'heure" contre "en retard".

seulement 8% à être en retard. Ces chiffres manifestent une réussite qui ne se retrouve pas dans l'ensemble de la population des collèges et qui manifeste que notre échantillon est spécifique : il l'est par construction car son but est de comparer les motivations de parents d'élèves scolarisés soit dans des "écoles nouvelles", soit dans des collèges à recrutement social équivalent et étant perçus comme de "bons établissements". Les "écoles nouvelles" sont des écoles publiques (Decroly) ou privées non confessionnelles (La Source à Meudon, l'Ecole Alsacienne) qui se caractérisent par une pédagogie différente mise au point par des réformateurs comme Decroly ou Cousinet qui ont cherché à mieux partir des intérêts des enfants et à mettre au point des techniques pédagogiques spécifiques (qui se sont d'ailleurs répandues ensuite, ce qui fait qu'on peut se demander ce qu'il en reste aujourd'hui : c'est une des réponses attendue de cette enquête).

Comme cette enquête va opposer deux types d'élèves : ceux qui sont en école nouvelle et ceux qui sont dans des collèges à recrutement social analogue et de bon niveau, une *variable d'intérêt* privilégiée va être précisément cette question à deux modalités "Ecole Nouvelle", "Collège de bonne réputation" : la distribution de cette question n'est pas pertinente en soi dans la mesure où, par construction, chaque enquêteur devait interroger quatre élèves de collège à bonne réputation et un d'école nouvelle. Ce n'est qu'en la croisant avec d'autres questions que l'on verra l'effet de cette variable[5].

Un premier tableau croisé va nous permettre de voir de premiers résultats et de mettre au point un outil qui nous servira dans la suite : un indicateur de la force de liaison entre modalités (ou entre questions). Nous effectuons donc le tri croisé entre le sexe et la variable d'intérêt, le type de collège.

```
Croisement de la question 17A type d'écoles avec la question
V02 sexe
Le Khi-deux du tableau est de    3.8
Degré liberté =   1 Prob.= 0.047 **
```

Il s'agit d'un tableau à 2 lignes et 2 colonnes et donc à un degré de liberté puisqu'en fixant l'effectif d'une case, toutes les autres se déduisent des marges.

Le khi-deux est significatif au seuil de 5% ($p < 0,05$ codé souvent avec deux étoiles)

```
Le PEM du tableau est de    18.6%
```

Par PEM il faut entendre Pourcentage de l'Ecart Maximum : il s'agit d'un indicateur d'attraction qui vaut pour l'ensemble du tableau (PEM global) ou pour une case du tableau (PEM local). Nous allons expliciter en premier lieu le PEM local.

---

[5] "Les stratégies éducatives des classes moyennes et supérieures salariées", enquête dirigée par François de Singly et Philippe Cibois dans le cadre du Deug de sociologie de l'Université de Paris V en 1991-1992

Dans le tableau ci-dessous, on trouve 4 nombres dans chaque case (et leur somme en marge) :

- l'effectif (N=) : pour la case "Féminin en Ecole nouvelle, il est de 60 individus ;

- le pourcentage en ligne (%Ligne) : sur 268 élèves de sexe féminin, les 60 en école nouvelle représentent 22,4% du total (soit plus que 19,1%, le pourcentage toutes lignes confondues, ce qui indique une attraction)

- la contribution au khi-deux qui est égale à l'effectif en écart à l'indépendance au carré divisé par l'effectif théorique.

Ici l'effectif théorique (produit des marges par le total) est de 98 x 268 / 512 = 51,30. L'écart à l'indépendance est de (observé – théorique) 60 – 51,30 = 8,70. La contribution au khi-deux est de $8{,}70^2$ / 51,30 = 1,5

- le PEM, Pourcentage de l'Ecart Maximum (%Attrac). On a noté que pour cette case, l'écart à l'indépendance est 8,70 individus. Si la liaison entre sexe féminin et école nouvelle était à son maximum, les 268 filles ne pourraient pas être à l'école nouvelle (dont l'effectif n'est que de 98 individus) mais les 98 élèves de l'école nouvelle pourraient être de sexe féminin. Donc 98 est le maximum de la case et l'écart à l'indépendance dans le cas de ce maximum serait de (maximum – théorique) 98 – 51,30 = 46,70

Comparons l'écart observé 8,70 à l'écart dans le cas du maximum 46,70 ce qui nous donne une proportion de 8,70 / 46,70 = 0,186 ou 18,6% en pourcentage. Cette valeur est suivie d'une étoile sur le tableau pour signaler qu'elle est issue d'un tableau croisé significatif[6].

N= %Ligne / Khi2 %Attrac	Ecole Nouvelle	Collège BonneRép	Total en ligne
Masc	38 15.6	206 84.4	244 100
	1.6 -18.6*	0.4 18.6*	2.0 47.7
Fémi	60 22.4	208 77.6	268 100
	1.5 18.6*	0.3 -18.6*	1.8 52.3
Total	98 19.1	414 80.9	512 100
en colonne	3.1	0.7	3.8 100

Dans un tableau 2 x 2, tous les PEM sont symétriques, c'est-à-dire de même valeur absolue et de signes opposés, c'est-à-dire correspondant non à une attraction, mais à une répulsion dans le cas d'un PEM négatif. Le PEM global est pris en faisant la somme des écarts positifs observés à l'indépendance par rapport à la somme des écarts positifs dans le cas de la liaison maximum : on vérifie facilement qu'il est aussi égal à 18,6%. Ce résultat est général : dans le cas d'un tableau 2 x 2, le PEM global et le PEM local (positif) sont identiques. Le calcul du PEM peut être étendu à des tableaux plus grands ayant un ordre sur les marges (que l'on peut toujours établir par une méthode d'analyse factorielle).

---

[6] Philippe Cibois, "Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence", *Bulletin de méthodologie sociologique*, 1993, n°40, p.43-63.

Empiriquement, des cas de PEM très élevés (supérieurs à 50%) manifestent une liaison tellement forte qu'ils sont l'indice d'une redondance des indicateurs : par exemple, dans toute enquête, on vérifie que le PEM entre le fait d'être à la retraite et d'être dans une tranche d'âge supérieur à 60 ans est toujours très élevé. Inversement, quand la liaison est inférieure à 10%, elle peut être l'effet du hasard et c'est pour cette raison qu'on associe toujours au PEM le test du khi-deux. On constate empiriquement que les PEM intéressants se situent entre 10 et 50%.

Quand on dépouille une enquête, il faut immédiatement intégrer tout résultat obtenu, en étant bien conscient qu'il pourra être remis en cause dans la suite. Par exemple ici, on doit immédiatement prendre acte de la liaison entre sexe féminin et écoles nouvelles : c'était d'ailleurs l'une des hypothèses qui étaient proposées au moment de la construction de l'enquête de vérifier si les écoles nouvelles, en mettant l'accent sur les aspects relationnels, n'étaient pas en train de moderniser la définition traditionnelle du rôle féminin.

Quand on commence à dépouiller une enquête, il faut progressivement s'approprier les données, en faire l'expérience et c'est une bonne pratique de commencer par explorer quelques hypothèses simplement par le biais de tris croisés. Par exemple, une autre hypothèse de départ qu'il est facile de vérifier était que les écoles nouvelles étaient privilégiées par des parents de classe moyenne ou supérieure dont les enfants avaient des difficultés scolaires. Nous allons utiliser à cette fin, l'âge scolaire, variable que nous avons construite et qui est un indicateur "objectif" des difficultés du parcours scolaire.

On a le tableau croisé suivant :

Croisement de la question AGS Age scolaire avec la question 17A type d'écoles
Le Khi-deux du tableau est de 8.8
Degré liberté = 2 Prob.= 0.012 **

N= %Ligne / Khi2%Attrac	Ecole Nouvelle	Collège BonneRép	Total en ligne
En avance	16 17.4	76 82.6	92 100
	0.1 -9.1	0.0 9.1	0.2 18.0
A l'heure	67 17.7	312 82.3	379 100
	0.4 -7.6	0.1 7.6	0.5 74.0
En retard	15 36.6	26 63.4	41 100
	6.5 21.6*	1.5-21.6*	8.1 8.0
Total en colonne	98 19.1	414 80.9	512 100
	7.1	1.7	8.8 100

On voit que les contributions au khi-deux qui rendent le tableau significatif sont associées précisément au fait d'être en retard scolaire, qu'il y a une attraction (PEM de 21,6% significatif) entre ce retard scolaire et l'école nouvelle. On voit donc que

l'hypothèse qui avait été faite est d'une certaine manière confirmée, mais à la condition de bien voir que les élèves en retard sont très minoritaires, y compris dans l'école nouvelle (83 des 98 soit 85% des élèves d'école nouvelle sont à l'heure ou en avance).

A partir de ce deuxième tri croisé, on voit que chaque tableau croisé apporte une information utile, mais ponctuelle, il manque à la fois la vue d'ensemble et la prise en compte des nombreuses autres questions de l'enquête. Nous allons maintenant mettre au point une procédure qui permette une découverte systématique des éléments intéressants de l'ensemble des tris croisés possibles.

II Sélectionner les questions pertinentes

Ce que nous voulons repérer, ce sont les questions qui sont pertinentes par rapport à la variable d'intérêt de notre enquête, l'opposition entre écoles nouvelles et collèges de bonne réputation. Nous allons donc croiser systématiquement cette variable d'intérêt avec toutes les autres questions de l'enquête mais ne sélectionner que celles qui sont le plus en attraction avec elle.

Le questionnaire, qui comprenait plus de cent questions, non seulement testait la situation sociale de la mère avec une grande précision sur le plan du métier, de la formation (y compris celle des grands-parents), des gouts, des affiliations politiques, religieuses, associatives, du couple, mais envisageait aussi :

- une description fine de l'enfant : ses "qualités", ses "défauts", son attitude en famille ;

- son comportement et son niveau scolaire, les raisons du choix du collège, ses matières préférées, son avenir ;

- ses loisirs ;

- le style des relations que la mère avait avec l'enfant (complicité, fermeté, etc.).

L'indicateur que nous utiliserons sera le PEM global, identique au PEM local positif dans le cas d'un tableau 2 x 2 (comme plus haut, le croisement avec le sexe). Dans le cas d'un tableau qui a davantage de colonnes, comme dans le tableau précédent, on se référera à la publication de présentation de la méthode[7]. Si le PEM n'est pas disponible sur le logiciel que l'on utilise, on pourra prendre des indicateurs analogues comme le V de Cramèr (dont le PEM est une extension qui tient compte des possibilités actuelles de calcul). L'important est de disposer d'un indicateur qui, pour une question donnée, donne automatiquement la liste des autres questions de l'enquête avec lesquelles elle est en attraction.

Les résultats, pour la question « type d'école » (choix d'une école nouvelle contre choix d'un collège de bonne réputation) sont les suivants : on prend les questions par ordre d'attraction décroissante de façon à avoir un premier choix d'une vingtaine de questions. Ces questions peuvent être regroupées autour de plusieurs thèmes :

[7] Cibois 1993

- les raisons du choix du collège lui-même : s'il était proche ou non ; si l'on y cultivait l'autonomie ou la réussite scolaire et par qui le choix a été fait (un parent, les deux, l'enfant a-t-il été associé à ce choix ?)

- comment est envisagé la scolarité de l'enfant : est-ce que l'enfant est satisfait de l'enseignement qu'il reçoit ? Pour le futur, faut-il le pousser ou le laisser suivre son rythme ? Si l'on a prévu le lycée où il irait. Jusqu'où pense-t-on qu'il ira (université ou grandes écoles ?).

- un certain nombre de questions concernent le style éducatif des parents : ce que l'on souhaite obtenir comme résultat (respect des autres, savoir-vivre, sens des responsabilités, etc.) ; quel type de sanction on envisage éventuellement (privation, réprimande) ; si les parents ont le sentiment ou non de reproduire le style d'éducation qu'ils ont reçu eux-mêmes ; si l'enfant connait les opinions politiques de ses parents.

- questions portant sur les activités de l'enfant : ses activités préférées, s'il pratique la compétition sportive, ses jeux préférés, ce qu'il a reçu à Noël

- il y a peu de questions relatives aux parents sinon la catégorie socioprofessionnelle de la mère et ses loisirs favoris.

Cette vingtaine de questions, qui comportent à peu près 200 modalités de réponses (soit une dizaine de modalités par question en moyenne) n'est qu'un point de départ pour commence à se faire une opinion sur le contenu de l'enquête.

Cette procédure qui consiste à passer par la variable d'intérêt pour sélectionner les questions à l'avantage d'aider à commencer la recherche avec un nombre suffisant de modalités de 200, qui est un bon point de départ. Ce n'est qu'un point de départ qu'il faudra affiner dans la suite. Pour le traiter, nous allons utiliser l'analyse factorielle des correspondances.

Chapitre 2. L'analyse factorielle des correspondances

I Décomposition des écarts à l'indépendance

Avant de montrer comment utiliser cette technique, il faut en comprendre quels sont les concepts fondamentaux. A cette fin nous partirons du tableau suivant qui, issu toujours de la même enquête, croise l'intérêt vis-à-vis de la religion de la personne interrogée (la mère de l'enfant) avec sa position politique. Dans la catégorie marquée « ni gauche ni droite », on a regroupé les réponses faisant référence au mouvement écologique ou qui refusent de se positionner sur une échelle gauche/droite.

Position politique	Intérêt vis-à-vis de la religion			
	Fort	Moyen	Nul	Total
Droite	24	41	7	72
Centre	14	30	12	56
Gauche	28	89	74	191
Ni G ni D	46	83	64	193
Total	112	243	157	512

Tableau 1 : effectif observé

Faire l'analyse des correspondances de ce tableau conduit à construire un graphique où chaque point représente un intitulé de ligne ou de colonne. Un point ligne sera proche d'un point colonne quand on pourra repérer une attraction entre cette ligne et cette colonne, attraction repérée par un fort écart à l'indépendance.

La situation d'indépendance dans un tableau se définit de la façon suivante : en moyenne dans ce tableau, la proportion de fort intérêt est de 112 / 512 = 0,219 soit 21,9%. Si cette proportion s'appliquait au 72 personnes de droite, l'effectif qu'il y aurait serait de 0,219 x 72 = 15,8 personnes. Cet effectif correspondrait au cas fictif où il y aurait indépendance entre les lignes et les colonnes puisqu'il est calculé simplement par produit des marges divisé par le total.

Pour l'ensemble du tableau les résultats sont les suivants.

Position politique	Intérêt vis-à-vis de la religion			
	Fort	Moyen	Nul	Total
Droite	15,8	34,2	22,1	72
Centre	12,3	26,6	17,2	56
Gauche	41,8	90,7	58,6	191
Ni G ni D	42,2	91,6	59,2	193
Total	112	243	157	512

Tableau 2 : indépendance

Comme il s'agit d'un cas fictif, on l'appelle tableau des effectifs théoriques sous l'hypothèse d'indépendance.

Les observation sont soit au-dessus de l'indépendance comme pour la première case Droite et fort intérêt où l'on a un écart à l'indépendance de 24 – 15,75 = 8,25 personnes en écart positif. Par contre on a un écart négatif entre la gauche et le fort intérêt : 28 – 41,8 = -13,8 où le déficit manifeste une répulsion. Quand on est de gauche, on est moins que la moyenne à avoir un fort intérêt pour la religion.

Le tableau général est le suivant :

Position politique	Intérêt vis-à-vis de la religion			
	Fort	Moyen	Nul	Total
Droite	8,3	6,8	-15,1	72
Centre	1,8	3,4	-5,2	56
Gauche	-13,8	-1,7	15,4	191
Ni G ni D	3,8	-8,6	4,8	193
Total	112	243	157	512

Tableau 3 : écarts à l'indépendance

On voit qu'il y attraction entre la droite (et dans une mesure plus faible le centre) avec l'intérêt fort ou moyen ; une attraction entre la gauche et une absence d'intérêt. Pour ceux qui refusent le positionnement politique traditionnel, ils se retrouvent dans les extrêmes et fuient l'intérêt moyen.

Ces résultats sont tout à fait classiques en sociologie[8] : en France l'opposition gauche / droite se superpose souvent à l'opposition vis-à-vis de la religion (catholique souvent). Quant à la position moyenne, elle reflète souvent un attachement traditionnel en voie de se distendre : ceux qui ne se situent pas dans l'opposition politique classique font leur choix soit pour soit contre le domaine religieux.

Chaque écart à l'indépendance est le résultat de l'opération *effectif observé – effectif théorique*. On peut étendre cette opération au tableau en disant que ce qui est vrai au niveau de chaque case l'est aussi au niveau du tableau dans son ensemble. Le tableau observé est ainsi décomposé en une somme de deux tableaux : théorique + écarts à l'indépendance. En reprenant les intitulés des tableaux on a :
T observé = T théorique + T écarts

C'est cette décomposition qui va être poursuivie par l'analyse des correspondances où le tableau des écarts va être d'abord approximé par un tableau le plus proche de lui mais où, comme dans le tableau d'indépendance, chaque case sera connue par ses marges.

[8] René Rémond, *Les droites en France*, Paris, Aubier, 1982 ; Jean-François Sirinelli (dir.), *Les droites françaises*, Gallimard, 1995 ; Guy Michelat et Michel Simon, *Classe religion et comportement politique*, Paris, Presses de la FNSP et ed. sociales, 1977 ; Jean-Marie Donegani, *La liberté de choisir*, Paris, Presses de la FNSP, 1993.

Voici l'approximation du tableau des écarts avec les coefficients marginaux qui permettent de le construire[9] : les valeurs des cases du tableau sont très proches du tableau 3 des écarts. Pour s'en convaincre il suffit de faire la différence terme à terme dans un tableau du reste.

Position	Intérêt vis-à-vis de la religion			
politique	Fort	Moyen	Nul	Coeff.
Droite	9,0	5,7	-14,7	-3,147
Centre	2,8	1,8	-4,6	-0,993
Gauche	-10,4	-6,7	17,1	3,645
Ni G ni D	-1,4	-0,9	2,3	0,495
Coeff.	-2,854	-1,826	4,680	

Tableau 4 : approximation des écarts

Le tableau suivant est le reste : ce qu'il faut ajouter terme à terme pour retrouver les écarts.

Position	Intérêt vis-à-vis de la religion			
politique	Fort	Moyen	Nul	Coeff.
Droite	-0,7	1,1	-0,4	-0,394
Centre	-1,1	1,6	-0,5	-0,586
Gauche	-3,4	5,0	-1,6	-1,826
Ni G ni D	5,2	-7,7	2,5	2,807
Coeff.	1,851	-2,742	0,891	

Tableau 5 : reste

Dans l'approximation où se trouve la plus grande partie des écarts, les nombres vont en valeur absolue jusqu'à 17 (gauche – intérêt nul), tandis que dans le reste, la plus forte valeur est proche de 8 (en négatif : ni gauche ni droite – intérêt moyen).

Le tableau du reste peut également être obtenu par la multiplication terme à terme de coefficients marginaux.

Si on regarde comment se sont répartis les écarts, on voit que dans l'approximation, ce sont surtout les oppositions des trois premières lignes (l'échelle politique traditionnelle) qui ont été prises en compte tandis que dans le reste c'est plutôt la ligne d'opposition à la répartition gauche / droite habituelle qui est présente. On voit ainsi que la décomposition en tableaux séparés met en relief pour chacun un aspect des données, pour lequel on emploie le mot de *facteur*.

On peut utiliser les couples de coefficients marginaux de chaque « facteur », ceux de l'approximation et ceux du reste, comme abscisse et ordonnées des points dans un graphique. Le résultat est le suivant :

[9] Voir l'annexe pour les détails complémentaires.

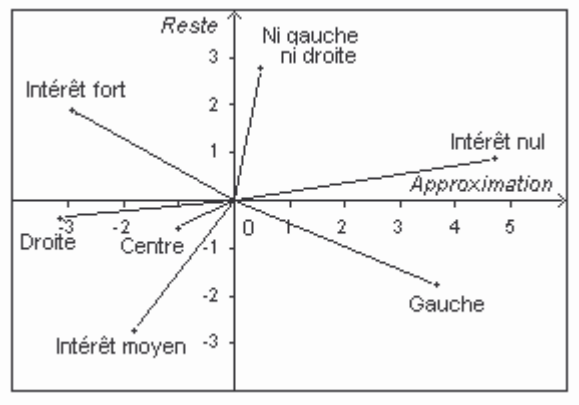


Figure 1 : Facteur *approximation* en abscisse, *reste* en ordonnée

Les règles de lecture pour ce *plan factoriel* permettent de retrouver l'information des écarts à l'indépendance : on doit regarder l'angle au centre entre point ligne et point colonne. Trois cas de figures sont possibles :

1) l'angle est inférieur à 90° : par exemple entre gauche et intérêt nul, ou droite et intérêt fort. Ceci signifie qu'il y a *attraction* entre ces modalités et que l'écart à l'indépendance est fort (les deux attractions citées sont les plus fortes avec des écarts de 15,4 et 8,3)

2) le cas opposé se présente quand l'angle est proche de 180° : par exemple entre droite et intérêt nul, gauche et intérêt fort et ni gauche ni droite et intérêt moyen. Ceci signifie qu'il y a *opposition* ou *répulsion* entre ces modalités et que l'écart à l'indépendance prend des valeurs négatives fortes (les cas cités correspondent aux trois plus bas niveaux d'écart : -15,1, -13,8 et -8,6).

3) le cas intermédiaire se situe quand l'angle est proche de 90° : par exemple entre gauche et intérêt moyen. Ceci signifie qu'il y a *indépendance* entre ces modalités : l'écart à l'indépendance est faible (ici c'est la plus faible valeur absolue des écarts de 1,7 : elle est négative car l'angle est légèrement supérieur à 90°)

Ces trois cas de figure d'attraction, d'indépendance ou d'opposition représentent toutes les éventualités possibles et toutes les situations que l'on observe sont intermédiaires entre ces cas types. Quand on a à traiter un grand nombre de modalités, et cela va être le cas pour dépouiller une enquête, on s'intéresse surtout aux attractions, c'est-à-dire aux proximités angulaires entre modalités qui déterminent des zones du graphique auxquelles il est parfois possible de donner un nom qui relève de l'interprétation.

Plus un point est proche du centre, et moins ses attractions ou oppositions sont fortes. Quand un point est strictement au centre, cela veut dire qu'il en situation d'indépendance avec toutes les autres modalités. Ici le point qui se rapproche le plus de cette situation est le point « centre » dont la ligne d'écarts à l'indépendance est la plus faible de tout le tableau.

Enfin, si l'on prend en compte les tableaux d'approximation et de reste individuellement (ou ce qui revient au même si on ne regarde que les abscisses des points ou les ordonnées), on peut donner un nom à chaque approximation, appelée aussi *facteur*. Le tableau d'approximation correspond à l'opposition politique traditionnelle : gauche areligieuse contre droite favorable (1er facteur), tandis que le tableau du reste est spécifique de la position ni gauche ni droite et de son refus à l'intérêt moyen (2^{e} facteur).

En résumé, l'analyse des correspondances d'un tableau croisé consiste à décomposer les écarts à l'indépendance de ce tableau en plusieurs tableaux connus par leurs coefficients marginaux qui sont appelés traditionnellement les vecteurs propres de chaque tableau.

Sur le graphique associé à ces vecteurs propres, les intitulés des lignes ou colonnes, par leurs positions angulaires réciproques permettent de retrouver les écarts à l'indépendance des données.

Le nombre total de tableaux nécessaires est égal au plus petit nombre de lignes et colonnes, ici trois lignes : le premier tableau, numéroté zéro correspond à l'indépendance, le tableau suivant est le premier facteur, le suivant le deuxième. Quand le nombre de facteurs nécessaire est plus grand que deux, les facteurs suivants, plus faibles sont soit négligés, soit interprétés séparément.

II Contributions des modalités, des tableaux.

A la décomposition des écarts à l'indépendance se juxtapose une décomposition des contributions au khi-deux de chaque case. Cette contribution, donnée traditionnellement par la formule :
(observé – théorique)² / théorique, peut être lue aussi comme écart² / théorique ou encore comme le produit de l'écart par le rapport écart / théorique. Cette dernière manière de voir nous signale que la contribution du khi-deux est homogène à l'unité de compte (l'individu dans une enquête), et que l'écart observé est pondéré par un rapport qui va dans le sens de l'augmentation quand l'écart (en valeur absolu) est plus grand que le théorique; et de la diminution dans le cas contraire.

Dans le cas ici traité, toutes les valeurs absolues des écarts sont inférieures au théorique et donc tous les rapports sont inférieurs à l'unité et réducteurs. Nous allons procéder ci-dessous à la décomposition additive du khi-deux global initial. Les contributions de chaque case sont sommées en ligne, en colonne et sur le total :

Position	Intérêt vis-à-vis de la religion			
politique	Fort	Moyen	Nul	Total
Droite	4,3	1,4	10,3	16,0
Centre	0,3	0,4	1,6	2,2
Gauche	4,5	0,0	4,1	8,6
Ni G ni D	0,3	0,8	0,4	1,5
Total	9,5	2,6	16,3	28,4

Tableau 6 : khi-deux du tableau initial

Le principe de la décomposition est de calculer la contribution au khi-deux de chaque case dans le tableau d'approximation et de reste. Nous avons donc les deux tableaux de khi-deux suivants :

Position politique	Intérêt vis-à-vis de la religion			
	Fort	Moyen	Nul	Total
Droite	5,1	1,0	9,8	15,9
Centre	0,7	0,1	1,3	2,0
Gauche	2,6	0,5	5,0	8,0
Ni G ni D	0,0	0,0	0,1	0,1
Total	8,4	1,6	16,1	26,1

Tableau 7 : khi-deux du tableau d'approximation

Position politique	Intérêt vis-à-vis de la religion			
	Fort	Moyen	Nul	Total
Droite	0,0	0,0	0,0	0,1
Centre	0,1	0,1	0,0	0,2
Gauche	0,3	0,3	0,0	0,6
Ni G ni D	0,6	0,6	0,1	1,4
Total	1,0	1,1	0,2	2,3

Tableau 8 : khi-deux du tableau du reste

On vérifie facilement que si cette fois la décomposition ne se fait pas de manière additive au niveau des cases, elle se fait au niveau des totaux de lignes, de colonne et du total général. Pour le total général de 28,4 il se répartit en 26,1 pour l'approximation + 2,3 pour le reste. Cette répartition est très inégalitaire et peut se mesurer par un pourcentage, dit traditionnellement « d'explication » qui est de 26,1 / 28,4 = 0,919 soit 91,9% pour le premier facteur et 8,1% pour le deuxième, ce qui justifie le vocabulaire employé d'*approximation* pour le premier facteur (puisque 9 sur 10 de l'information repérée par le khi-deux s'y trouve) et de *reste* pour le 2^{e}, éventuellement négligeable.

On se sert des totaux de khi-deux de chaque facteur pour évaluer la contribution de chaque ligne ou colonne dans un facteur. Par exemple, pour le premier on voit que la plus forte contribution des lignes est celle de la droite qui représente 15,9 sur un total de 26,1 soit 60,9%. Dans les programmes habituels ces contributions sont données en millièmes et non en pourcent (ici 609 pour mille). Quand on a beaucoup de modalités, on voit immédiatement celles qui ont le plus contribué à la fabrication d'un facteur, ce qui permettra de l'interpréter avec sécurité.

Le khi-deux de chaque tableau, qui lui est propre est appelé aussi *valeur propre* du facteur : cette valeur propre est exprimée par un dérivé du khi-deux, le khi-deux divisé par l'effectif total (ou phi-deux). Il est pour le premier facteur de 26,1 / 512 = 0,051.

III Procédure de codage en tableau de Burt

Dans le cas précédent, nous avions une question en colonne (l'intérêt vis-à-vis de la religion) et une question en ligne (l'opinion politique), or au précédent chapitre, nous avions repéré une vingtaine de questions. Pour pouvoir traiter plus de deux questions en même temps, on prend comme tableau à traiter, non le tableau croisé ordinaire, mais un tableau spécial, appelé *tableau de Burt*[10], qui consiste à faire un tableau entièrement symétrique pour les lignes et les colonnes et où, par exemple en ligne, se trouvent toutes les modalités de toutes les questions retenues. En croisant avec les mêmes modalités en colonne, on juxtapose les tris croisés précédents et un tableau diagonal où se trouvent les effectifs de chaque modalité. L'exemple précédent mis en tableau de Burt permettra d'en comprendre le principe.

	Droi	Cent	Gauc	NiNi	Fort	Moy	Nul	Tot.
Droit	72				24	41	7	144
Cent		56			14	30	12	112
Gauc			191		28	89	74	382
NiNi				193	46	83	64	386
Fort	24	14	28	46	112			224
Moy	41	30	89	83		243		486
Nul	7	12	74	64			157	314
Tot.	144	112	382	386	224	486	314	2048

Tableau 9 : tableau de Burt

Le tableau initial (politique en ligne, intérêt en colonne) se trouve en haut à droite. En bas à gauche, c'est le même tableau mais ce qui était en ligne se trouve en colonne et réciproquement. Les deux tableaux diagonaux n'ont d'effectif que sur la diagonale et cet effectif est le total marginal du tableau d'origine.

Avec une telle disposition, on peut mettre maintenant autant de questions que l'on veut. Ce tableau de Burt peut s'interpréter, pour une modalité donnée comme l'effectif correspondant à la population ayant en même temps les deux modalités. Par exemple pour la première modalité en ligne et toutes les autres en colonne sont "droite" et "droite", les 72 de droite, sont "droite" et "centre", évidemment personne (blanc correspondant à zéro), puis sont de droite et d'intérêt fort 24, etc. Le total de marge correspond ici, où il y a deux questions, à 2 fois l'effectif de marge (n fois s'il y a n questions). Le total général correspond à 2 x 2 = 4 fois l'effectif de l'enquête car l'effectif total se trouve dans chacun des 4 tableaux (n^2 pour n questions).

On vérifiera que dans un tableau de Burt, l'écart à l'indépendance d'une case correspond strictement à l'écart à l'indépendance du tableau d'origine, ce qui fait que la décomposition factorielle est analogue avec toutefois des différences au niveau de la décomposition du khi-deux car il y a beaucoup de khi-deux artificiel dans le

[10] En hommage au psychologue Cyril Burt (1883-1971)

tableau de Burt. En effet, comme le tableau d'origine est dupliqué, son khi-deux l'est aussi et le khi-deux lié aux tableaux diagonaux est complètement artificiel.

Les résultats sont présentés dans Trideux (d'une manière assez classique) de la façon suivante :

```
Facteur   1 Valeur propre =  0.375746
Pourcentage du total = 29.4
Facteur   2 Valeur propre =  0.284394
Pourcentage du total = 22.3
```

On voit déjà que le codage en tableau de Burt fait baisser énormément les pourcentages d'explication de chaque facteur (qui étaient de 92 et 8%). En effet, du fait du khi-deux artificiel, davantage que 2 facteurs sont nécessaires pour rendre compte de l'intégralité des données (5 ici mais seuls les 2 premiers ne sont pas artificiels). Il ne faut donc pas utiliser ces pourcentages pour interpréter l'analyse : l'indicateur pertinent devient maintenant la valeur propre elle-même en utilisant la règle empirique suivante :

- quand la valeur propre est supérieure à 0,1 (ce qui est le cas ici pour les deux facteurs utiles), cela indique une forte liaison entre les questions utilisées ;

- quand la valeur propre est inférieure à 0,1 mais supérieure à 0,01, on est dans le cas standard, habituel ;

- enfin quand la valeur propre est inférieure à 0,01, la liaison entre les questions est faible.

Qu'on se trouve souvent dans le cas standard vient du fait que les questions que l'on met dans une analyse n'indiquent ni de trop fortes liaisons (parce qu'on les connaitraient déjà) ni de trop faibles (car on ne veut pas croiser des réalités trop hétérogènes).

```
Coordonnees factorielles (F= ) et   contributions pour le
facteur (CPF)
Modalites en colonne
*---*------*----*------*----*
ACT.    F=1  CPF    F=2  CPF
*---*------*----*------*----*
V931    870  309    265   38   Nul
V932   -219   30   -528  232   Moyen
V933   -743  161    773  230   Fort
V951  -1275  304   -256   16   Droite
V952   -518   39   -490   46   Centre
V953    557  154   -447  131   Gauche
V954     75    3    680  307   NiGniD
*---*------*----*------*----*
*   *      *1000*      *1000*
*---*------*----*------*----*
```

Pour la liste des questions (on n'indique que les colonnes mais les lignes seraient strictement identiques), on a pour chaque facteur (F=), les coordonnées factorielles (ou vecteurs propres) et la contribution en millièmes. Les vecteurs propres donnent un graphique très proche du précédent où la croisée des axes est marquée par un angle droit et où la position de chaque point correspond à la première lettre de son intitulé :

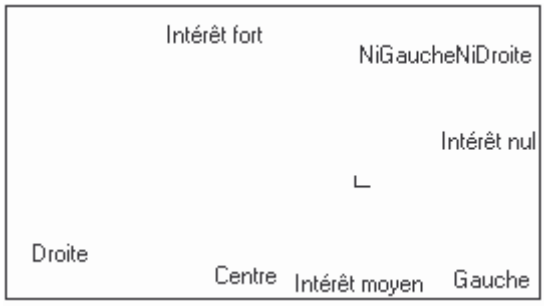


Figure 2 : plan factoriel du tableau de Burt, facteur 1 horizontal, facteur 2 vertical.

IV Modalités supplémentaires

Dans une enquête, certaines questions sont de nature différente des questions spécifiques de l'enquête : ce sont celles qui sont employées dans toutes les enquêtes comme le sexe, l'âge, le niveau d'étude, la catégorie socioprofessionnelle ou ces variables spécifiques que sont l'opinion politique ou religieuse. Ces variables sont souvent appelées explicatives, terme datant d'une époque où l'on croyait expliquer la superstructure par l'infrastructure. Si l'on ne prétend plus *expliquer* par ces variables, on pense toujours que ces variables vont au moins éclairer la situation, illustrer par leur présence un aspect important de la réalité ; on leur donne donc aussi le nom de variables *illustratives*.

Pour les rendre opérantes, on met ces modalités en variables *supplémentaires* dans l'enquête. Pour comprendre ce qu'est une variable supplémentaire, dans le tableau de Burt précédent où l'on a deux variables de type politique et religieux, si on veut savoir où se positionnent les personnes ayant mis leurs enfants dans une école nouvelle comparées à celles qui ont choisi un collège de bonne réputation, la stratégie de dépouillement consiste à ne pas mettre strictement ensemble ces trois questions mais à respecter la différence de nature du choix d'école en mettant cette question en variable supplémentaire[11].

Mettre une question en supplémentaire ne modifie en rien l'analyse des autres variables : c'est une fois l'analyse faite que l'on effectue des calculs supplémentaires pour mettre dans le graphique chaque modalité au plus près d'une modalité ordinaire (dite *active* par opposition à supplémentaire) qui serait identique à la modalité supplémentaire.

En ajoutant la question de l'école en variable supplémentaire, on n'a rien changé aux résultats indiqués plus haut. On a simplement les lignes suivantes ajoutées : la contribution qui est indiquée est hypothétique car c'est celle qu'aurait une modalité active identique.

```
*---*------*----*------*----*
SUP.    F=1  CPF    F=2  CPF
*---*------*----*------*----*
17A1    187    9    124    5 EcoleNouv
17A2    -44    2    -29    1 CollReput
*---*------*----*------*----*
```

[11] Quand il y a beaucoup de questions, ce sont bien les modalités illustratives qui sont mises en éléments supplémentaires.

Les contributions des deux modalités sont faibles, elles seront proches du centre. Toutefois on voit que les parents d'école nouvelle sont dans la direction du pôle d'intérêt nul et donc en faible conjonction avec lui, tandis que les parents ayant choisi un collège de bonne réputation sont en légère conjonction avec le pôle droite / intérêt moyen.

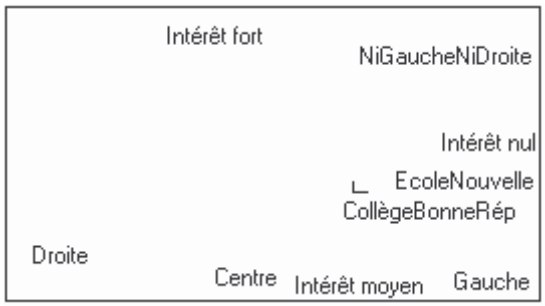


Figure 3 : type d'école en supplémentaire

Cette technique de l'élément supplémentaire est utilisée toutes les fois qu'une modalité est hétérogène par rapport aux modalités actives. C'est le cas des données incertaines, difficiles à interpréter comme les non-réponses à une question, c'est aussi le cas des modalités à très faible effectif qui peuvent perturber une analyse. Plutôt que de perdre complètement l'information, on met la modalité en élément supplémentaire.

V Résumé

Ce qui est visualisé par un graphique factoriel, ce sont les écarts à l'indépendance entre modalités, c'est-à-dire les attractions entre modalités. L'analyse va présenter une première approximation de l'ensemble des écarts par un premier axe ou premier facteur. Les écarts restants vont être approximés par un deuxième facteur et ainsi de suite.

Pour chaque facteur, les contributions des différentes modalités sont plus ou moins fortes : on se servira des plus fortes pour comprendre ce qui a été pris en compte par un facteur. Quand le nombre de modalités est grand, cette sélection est indispensable pour ne porter sur le graphe que les modalités les plus contributives.

Les facteurs sont faits avec les variables actives : comme les modalités par proximité de conjonction dessineront des types de répondants, les variables actives seront les questions qui sont spécifiques à l'enquête (opinions ou comportements). Une fois la typologie faite, on illustrera ces types en ajoutant des modalités supplémentaires qui seront les questions standards de toute enquête (sexe, âge, etc.).

On notera qu'on ne fera pas l'inverse qui consisterait à faire une typologie des variables de statut en actives sur laquelle on projetterait en supplémentaires les modalités de l'enquête. Avec une telle procédure, et si l'enquête était bien représentative de la population globale, on devrait toujours avoir la même typologie de statut social, puisque indépendante de l'enquête. On ne fait pas cela car ce sont des typologies de répondants spécifiques aux enquêtes que l'on veut obtenir. Cependant, comme on le verra plus loin, il pourra être utile de faire une telle analyse pour choisir les questions explicatives à prendre en compte dans une analyse "toutes choses égales par ailleurs".

Chapitre 3 : rechercher des types de répondants avec l'analyse des correspondances

Disposant maintenant de l'outil qu'est l'analyse des correspondances, nous allons préciser les règles qui permettent de l'utiliser efficacement. Attention, il s'agit d'un processus cumulatif qui suppose un certain nombre d'essais, de modifications dans le choix des questions, de recodages, de mises en supplémentaires. La technique, comme d'ailleurs toute technique statistique appliquée à des données réelles, suppose une expérience, des règles de l'art que nous allons essayer de communiquer à travers l'exemple que nous allons suivre. Avant d'arriver à un résultat final simple, convaincant, facile à exposer, il faut passer par des étapes où la mise au point relève plus des règles de la bonne cuisine que de l'interprétation des lois statistiques.

Rappelons où nous en sommes : par la technique de la variable d'intérêt (ici le type d'école), nous avons sélectionné une vingtaine de questions rassemblant près de 200 modalités de réponses, pour le moment toutes actives.

I Première analyse : la queue de comète

Cette première analyse porte précisément sur 192 modalités, toute active. Les premières valeurs propres sont standards puisque comprises entre 0,01 et 0,1 (les pourcentages d'explication, qui dépendent pour leur plus grande part du nombre de modalités, ne signifient rien : le plus fort ici n'est que de 4%).

Examinons avant toute chose le premier plan factoriel (premier facteur horizontal, deuxième facteur vertical). Le résultat est tout à fait décevant puisque seuls une douzaine de modalités apparaissent, les 180 autres étant superposées au centre.

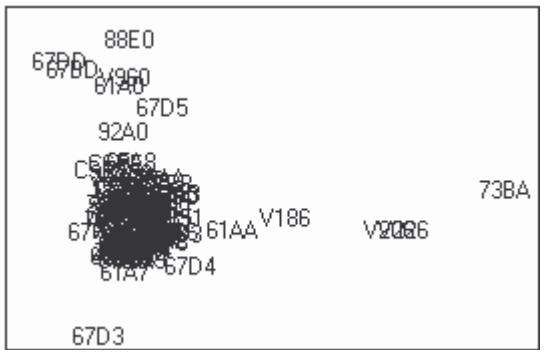


Figure 4 : la comète et ses queues

Quand sur un graphique comme celui-ci, les points sont superposés, il est inutile d'éditer le même graphique avec les noms longs qui donnent du sens aux numéros de modalités. Cependant, pour pouvoir ne serait-ce que lire les numéros des modalités, il faut désintriquer manuellement les points sur le graphique (en général en se servant de la souris). Ces légères modifications de position n'ont aucune conséquence sur l'interprétation. C'est ce qui est fait dans la figure 5 : on s'aperçoit que ces modalités sont de deux catégories :

- les non-réponses : ce sont toutes les modalités qui se terminent par zéro

- les modalités à faible effectif (de 1 individu à 12)

Comme il s'agit de modalités qui sont incertaines quant à leur interprétation, elles vont être mises en modalités supplémentaires. Dans le logiciel Trideux, les non-réponses sont par défaut mises en éléments supplémentaires.

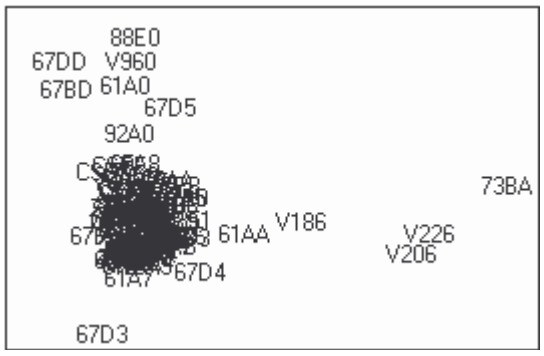


Figure 5 : graphique de la fig.4 rendu plus lisible

Attention, la même démarche, de mettre en supplémentaires les éléments isolés, doit être faite plusieurs fois avant de faire la « fission » de la comète. Quand on y arrive, la répartition est meilleure mais on en encore trop de points au centre. Pour rendre le graphique lisible, il faut n'afficher que les points les plus contributifs, en commençant par mettre le seuil à 1, puis 2, puis 5, 10, 20 : arrivé à ce seuil où le graphique devient lisible, on peut redescendre progressivement jusqu'à 15, ce qui donne la première analyse de la figure 3 où une trentaine de points sont représentés.

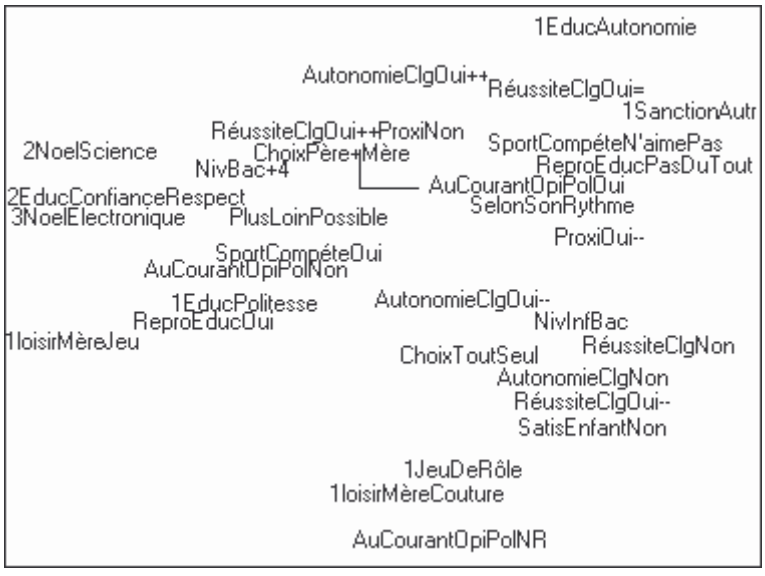


Figure 6 : première analyse, contribution minimum de 15 pour mille

Comme des proximités (angle au centre faible) indiquent des attractions entre modalités, commençons par le point le plus en haut à droite du graphique (1EducAutonomie) qui indique que la première priorité éducative donnée par la mère est l'autonomie, le fait de rendre responsable l'enfant. C'est un aspect qui a été jugé

très important dans le choix du collège (AutonomieClgOui++) alors que la proximité géographique n'a pas été déterminante (ProxiNon) et que le niveau scolaire ne l'a été que moyennement (RéussiteClGOui=). Dans ce type, on a l'impression de ne pas reproduire le modèle éducatif que l'on eu soi-même (ReproEducPasDu Tout), ce qui fait que la première réponse envisagée comme type de sanction, (1SanctionAutre) est que les réponses traditionnelles sont repoussées (privation, réprimande, etc.). La compétition sportive n'y est pas favorisée.

Les frontières du type ne sont pas nettes : on a indiqué les réponses en conjonction à 45° avec la première modalité choisie mais une réponse comme le fait que l'enfant soit au courant des opinions politiques de ses parents est partagée avec le type du répondant de la partie inférieure droite qui est d'accord aussi pour dire que, en ce qui concerne l'avenir de l'enfant, il faut qu'il aille à son rythme.

Pour le type de répondant en bas à droite, la satisfaction de l'enfant à l'école n'est pas nécessaire (SatisEnfantNon) et l'autonomie n'a pas fait partie du choix du collège qui d'ailleurs n'a pas été une décision délibérée mais qui s'est faite toute seule.

Le troisième type, à gauche, s'oppose au deux autres : on y met l'accent sur les valeurs éducatives traditionnelles que l'on veut reproduire, le respect, la confiance, la politesse. En matière d'avenir, on veut que l'enfant aille le plus loin possible et la compétition (sportive) y est favorisée. C'est un choix des parents et le collège a été l'objet d'un choix en fonction de son niveau.

Cette première analyse montre trois types : éducation mettant l'accent sur l'autonomie, éducation mettant l'accent sur les valeurs traditionnelles et un troisième type qui ne semble pas avoir fait de choix spécifique. On peut faire l'hypothèse raisonnable que le premier type doit être lié aux écoles nouvelles, mais lesquelles spécifiquement ?

Bien que l'on ait déjà traité près de 200 modalités, on voit qu'il manque beaucoup de choses pour répondre aux questions initiales de l'enquête, par exemple le sexe de l'enfant, le choix des différents collèges, le niveau de l'enfant, les caractéristiques sociales des parents.

II Analyse finale

Pour arriver à l'analyse finale, il va falloir introduire les questions qui vont permettre de répondre aux hypothèses de départ qui sont celles de cette enquête. Le questionnaire utilisé devait permettre de tester si le choix d'une école nouvelle pouvait être lié soit :

- à une stratégie de *rattrapage* : l'enfant a des difficultés dans le système scolaire standard et à défaut d'une bonne réussite scolaire, il cultive les nouvelles valeurs de notre époque que sont l'authenticité, la capacité relationnelle, l'autonomie ;

- à une stratégie de *reconversion* : même si l'enfant n'a pas de difficultés scolaire, certains parents pensent que ces mêmes nouvelles valeurs (autonomie, capacité relationnelle, authenticité) sont celles qui vont s'imposer dans la vie présente et qu'il faut en doter ses enfants.

Autre sous-hypothèse : est-ce qu'il n'y aurait pas une tendance à ce que l'on insiste davantage pour les garçons sur la réussite scolaire et pour les filles sur ces nouvelles valeurs, plus "douces", plus liées traditionnellement à l'insistance associée au modèle féminin du relationnel.

Pour répondre à la première hypothèse, il faut introduire des indicateurs de niveau scolaire, pour la 2^{e} il faut le sexe de l'enfant.

D'une manière générale, il faut introduire en modalités supplémentaires les variables de statut social : sexe, âge, etc. ainsi que toute question que l'on juge pertinente : le nombre n'est pas un obstacle car on peut sans difficulté traiter plusieurs centaines de modalités.

Pour l'analyse finale, on a donc ajouté des variables de statut ou apparentées : sexe, niveau scolaire de l'enfant, opinion politique de la personne interrogée. On en arrive au total de 273 modalités dont 152 supplémentaires. Comme le plan factoriel est évidemment trop chargé, on sélectionne les modalités en fonction de leur contribution.

On ne prend pas le même niveau de contribution pour les modalités actives (20) et pour les supplémentaires (6). En effet, les modalités supplémentaires, par construction ne peuvent être très en relation avec les actives car elles ne sont pas de même nature. Dans toute enquête, les questions d'opinion ou de comportement associées à un même domaine présentent entre elles des attractions fortes, par contre avec les variables de statut, les relations existent mais sont moins fortes. Comme on veut faire apparaitre des supplémentaires pour illustrer l'analyse, on utilise un niveau moins élevé de contribution.

Pour le choix de ces seuils, il faut procéder par essais et erreurs : si le seuil est trop faible, on a trop de points, s'il est trop fort, on n'en n'a plus assez. Il faut trouver un équilibre qui dépend aussi de la place dont on dispose. En cas de difficultés, il est possible de présenter un plan général schématique comme celui-ci où les ellipses et leur titre sont le fruit d'une interprétation mais qui aident la lecture.

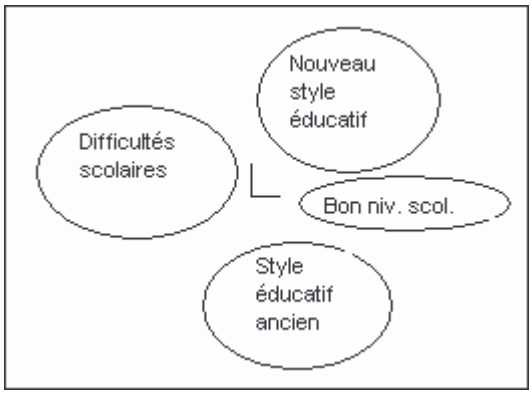


Figure 7 : analyse finale : schéma d'ensemble

Commençons par interpréter les points situés en proximité de l'axe horizontal du graphique (qui correspond à l'opposition la plus forte). Dans l'ellipse de droite on a regroupé des points qui sont autant de manifestations d'un *bon niveau scolaire* :

Figure 8 : Bon niveau scolaire

III Type : bon niveau scolaire

- "Niveau élève très bon[12]" : il s'agit de la question suivante "Comment percevez-vous votre enfant du point de vue scolaire", et c'est la meilleure perception.

- "Aide des parents non nécessaire" : réponse négative à la question "Le soir, pour son travail scolaire, vous ou son père intervenez-vous ?".

- "1ère Matière Scolaire NR", A la question de savoir s'il y a des matières scolaires dans lesquelles l'enfant se montre le meilleur, la mère répond que non, il est bon dans toutes les matières (NR signifie Non réponse).

A ces indicateurs d'un bon niveau scolaire sont associés des comportements comme :

- le fait de ne pas trouver de sujet de désaccord avec l'enfant quand on demande d'en lister plusieurs ("Désaccord aucun") où le fait que le premier "défaut" de l'enfant soit le fait d'être "timide" ce qui n'est pas un défaut très marqué (les "défauts" et "qualités" n'ont pas été proposés *a priori*, il s'agit des termes mêmes des répondants). On voit une atmosphère familiale apaisée où le "défaut" de l'enfant est cohérent avec le calme qui y règne.

- l'insistance sur la lecture : lire est la première activité préférée par l'enfant ("1Lire") et, à Noël, il a reçu des livres et d'autres cadeaux. D'une certaine façon, cette insistance sur la lecture appartient à la fois à l'excellence scolaire (car au collège l'incitation à la lecture est forte) et en même temps à un style cultivé de loisirs.

Un indicateur de statut est associé à ce pôle d'excellence scolaire : il s'agit d'une modalité supplémentaire qui est en attraction avec les modalités qui sont spécifiques des répondants de ce secteur. Il s'agit d'enfants de sexe féminin (à gauche de l'ellipse qui n'est qu'un repère visuel pour aider et n'est pas construite par l'analyse statistique mais par celui qui interprète le graphique). On retrouve là ce résultat bien connu que les filles réussissent mieux scolairement que les garçons[13].

Symétriquement au bon niveau scolaire, se trouve sur le côté gauche de l'axe un pôle de difficultés scolaires.

---

[12] Sur le graphique et pour gagner de la place, des abréviations ont été employées et cette modalité est marquée "NivElèveTrèsBon". Dans la suite du texte les modalités sont marquées sous leur forme explicitée.

[13] Plus le point est près du centre, *moins* il est lié aux points dans la même direction. Ici par exemple la modalité "Niveau de l'élève très bon" représente 20% de la population mais elle est plus forte chez les filles (24%) et symétriquement plus faible chez les garçons (15%) : l'attraction n'est pas très forte mais statistiquement significative. D'une manière générale cf. Christian Baudelot et Roger Establet, *Allez les filles !* Seuil, 1998

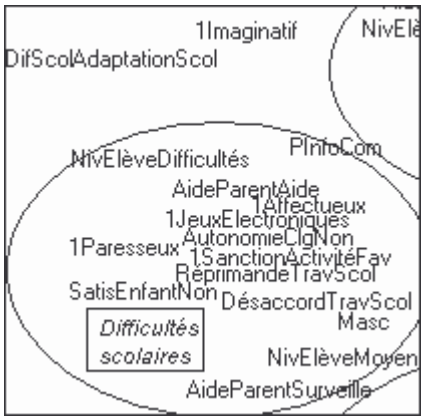


Figure 9: difficultés scolaires

IV Type : difficultés scolaires

La perception du niveau scolaire est soit "moyen" (en bas de l'ellipse) soit considéré comme "ayant des difficultés" (en haut) : ces deux appréciations sont les plus basses de l'échelle proposée qui allait de "très bon élève" à "élève ayant quelques difficultés".

De ce fait la question scolaire devient l'objet d'investissements constants et de conflits :

- A la question de savoir comment le soir, le père ou la mère interviennent pour le travail scolaire de l'enfant, une des réponse de ce pôle est que l'on "regarde chaque soir ce qu'il a à faire et ce qu'il a fait" ("Aide : parents surveillent") : c'est la réponse située en bas de l'ellipse précisément du côté du pôle de style éducatif traditionnel (que nous verrons ensuite) alors qu'en haut, la réponse est simplement qu'on l'aide ("Aide : parents aident") et cette réponse est plus proche du pôle de l'éducation nouvelle.

- le travail scolaire est cité comme la première occasion de réprimande ou de punition et également comme premier sujet de désaccords entre parents et enfant. L'enfant n'est d'ailleurs pas satisfait non plus de l'enseignement qu'il reçoit de ses professeurs ("Satisfaction Enfant : Non"). Comme le travail scolaire est le problème, le collège n'a de ce fait pas été choisi parce qu'on y développe l'autonomie ("Autonomie Collège : Non").

- si une sanction est envisagée, l'option choisie est "mon enfant a une de ses activités favorites qui est supprimée provisoirement".

- cette situation entraine une perception du premier défaut de l'enfant comme étant "paresseux" (intitulé qui regroupe aussi peu courageux, "flemmard", pas studieux) ou "étourdi" (ou distrait, "tête en l'air", manque d'attention). Par contre, la première qualité évoquée est le fait que l'enfant soit "affectueux", ce qui est évidemment indépendant des problèmes scolaires.

Comme indicateur de statut, on trouve, symétriquement au type *bon élève*, que les garçons sont plus nombreux que les filles à être en difficulté scolaire. Il y a également un autre statut du père, celui d'appartenir aux professions de l'information, des arts et des spectacles ("PèreInfoCom") qui est en haut du regroupement, proche du nouveau style éducatif : ceci signifie que ces parents ont à la fois des traits éducatifs nouveaux et des enfants en difficulté scolaire.

En conclusion de cette analyse de l'opposition horizontale du graphique, il apparait que ce qui distingue d'abord les perceptions, c'est le niveau scolaire : les styles éducatifs que nous allons voir maintenant, dans la mesure où ils s'opposent dans l'autre dimension du graphique (verticale), nous indiquent que l'on pourra trouver des styles éducatifs associés soit à de bons, soit à de mauvais résultats scolaires.

V Type : style éducatif ancien

Ce style éducatif, situé en bas du graphique, est qualifié *d'ancien* par rapport au style qui revendique la nouveauté, mais il est tout à fait classique dans les catégories sociales étudiées.

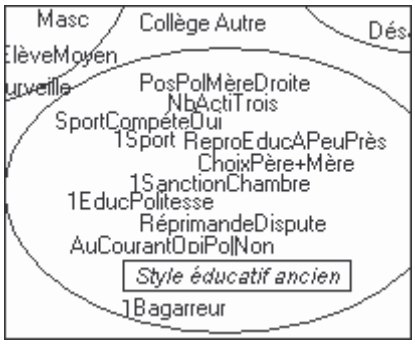


Figure 10 : style éducatif ancien

- La réponse donnée à la question sur ce quoi on insiste en priorité dans l'éducation, est la politesse ("1Educ Politesse"), réponse qui regroupe aussi, le fait de "savoir bien se tenir", le fait d'être "bien élevé", d'avoir du "savoir vivre".

- En insistant sur cette valeur, les parents ont d'ailleurs le sentiment de donner "à peu près" l'éducation qu'ils ont reçue de leurs propres parents et non "tout à fait", car l'éducation qu'ils donnent n'est pas le mode le plus "traditionnel". Par exemple, le choix du collège est une décision commune du père et de la mère[14] ("Choix Père + Mère"), cependant, parents et enfants ne vivent pas sur un pied d'égalité comme le montre le fait que les enfants ignorent les positions politiques du parent ("Au courant opinion politique : non").

- si sanction il doit y avoir, c'est le confinement dans la chambre qui est choisi ("1 Sanction Chambre")

- le sport et la compétition jouent ici un rôle important : le sport est la première activité préférée de l'enfant et il fait un sport de compétition. D'ailleurs, l'occasion de réprimande citée est le comportement violent de l'enfant, le fait qu'il se dispute ("Réprimande Dispute"), ce qui n'est pas sans lien avec les activités proposées (qui sont nombreuses "Nombre d'activités = 3"). Les enfants (en petit nombre), dont le défaut est d'être "bagarreurs", se trouvent dans cette zone.

[14] Le standard relationnel du milieu est une forme d'égalité dans les prises de décision du couple, ce qui n'était peut être pas le cas dans la génération des propres parents des personnes interrogées.

Comme indicateur de statut, l'opposition entre styles d'éducation se fait plus sur une base d'affiliation politique : la mère qui répond s'autopositionne plutôt à droite. On retrouve symétriquement le positionnement politique à gauche pour le nouveau style éducatif que nous allons maintenant étudier.

VI Type : nouveau style éducatif

Ce nouveau style éducatif se trouve symétriquement en haut : on y retrouve La Source au centre et deux autres collèges, Decroly un peu à gauche et l'Ecole Alsacienne, plus à droite, (c'est-à-dire participant en même temps au type "bon niveau scolaire"). Ce qui caractérise ce type, c'est précisément le refus de pratiques que le questionnaire, fait pour toucher tout le monde, présentait comme "normales" et qui sont précisément refusées ici.

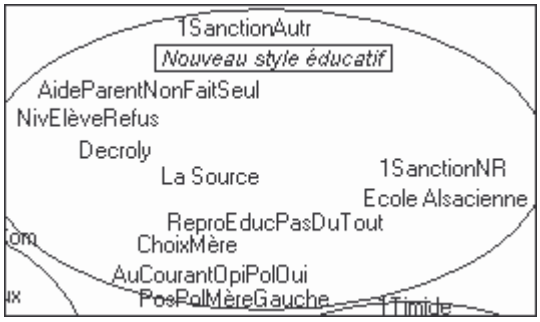


Figure 11 : un nouveau style éducatif

Par exemple, le questionnaire demandait de dire "à quelle occasion vous réprimandez ou vous punissez ?". On trouve ici d'abord le refus de répondre ("1SanctionNR") ou la codification "autre" ("1SanctionAutr") qui indique que le répondant n'a pas pris les réponses les plus fréquentes comme "recevoir une explication" ou "être privé de ses activité favorites" ou "devoir rester dans sa chambre" : les attitudes les plus classiques sont ici refusées. De même à la question de la perception du niveau scolaire de l'enfant (allant de *très bon* à *difficultés* avec les intermédiaires), ces parents choisissent la réponse "je refuse ce genre de classement" (codé "NivEleveRefus"). L'intervention des parents dans le travail scolaire est refusée ("il doit le faire tout seul" codé "Aide Parent Non Fait Seul").

Ces refus sont perçus par les parents de ce secteur comme une nouveauté éducative : par rapport à l'éducation qu'eux-mêmes ont reçu, ils ne pensent pas du tout reproduire la même ("ReproEducPasDuTout") : un autre élément de cette nouveauté est dans les rapports entre parents et enfants où l'on trouve plus de transparence des parents, dont les opinions politiques (d'ailleurs de gauche) sont connues des enfants.

Pour mieux comprendre ce nouveau style éducatif, nous allons procéder à un agrandissement de cette zone en faisant apparaitre un plus grand nombre de points qui contribuent moins que les précédents à la fabrication de ce portrait-robot

statistique, mais qui permettent d'en mieux préciser la nature[15]. Dans cet agrandissement, nous avons soulignés les points déjà vus dans le graphique général précédent.

Plusieurs modalités sont issues de la question sur la priorité en matière d'éducation : de plus cette question était une "question ouverte", c'est-à-dire que les réponses n'étaient pas proposées par le questionnaire mais que les mots mêmes des répondants étaient acceptés librement. Dans le pôle du nouveau style éducatif, on repère deux thèmes principaux :

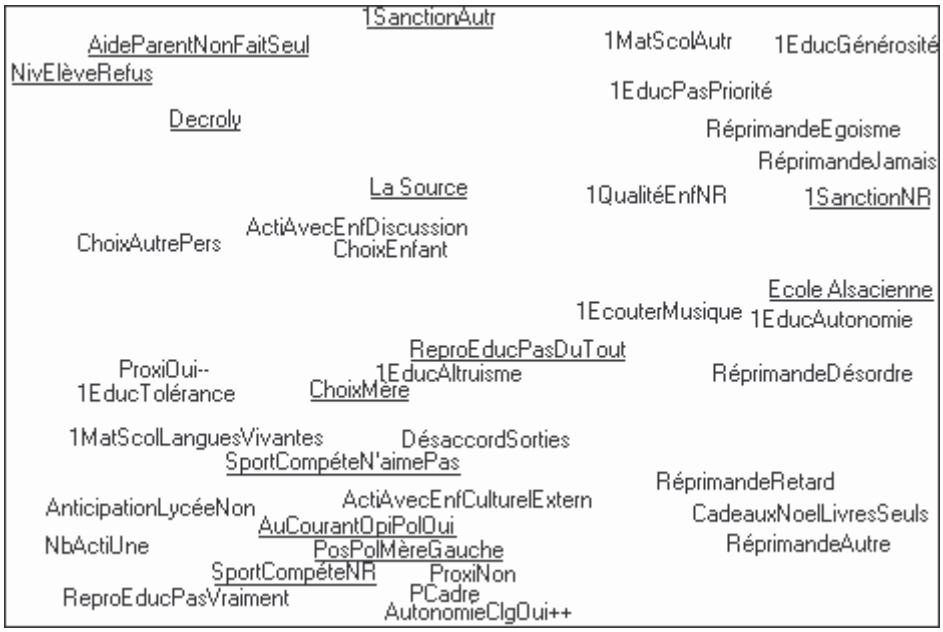


Figure 12 Nouveau style éducatif (agrandissement)

1) l'autonomie de l'enfant, son éveil ; dans la même catégorie on peut mettre le fait de dire que l'on n'a pas de priorité éducative, ce qui, d'une manière plus radicale, laisse entendre que l'enfant doit être laissé autonome.

2) l'attention aux autres, la tolérance, l'ouverture, le dialogue, la générosité.

C'est la présence simultanée de ces deux thèmes qui définit le style éducatif : il s'oppose à la "confiance mutuelle" du bon niveau scolaire qui peut être interprétée comme une exigence de transparence sans problème, il s'oppose encore plus radicalement à l'exigence de "franchise" du style traditionnel qui est une injonction à une transparence forcée. L'opposition est encore plus forte avec l'exigence d'obéissance qui se trouve dans le style traditionnel quand on a de mauvais résultats et le développement du sens de l'effort lié aux mauvais résultats[16].

En négatif, c'est-à-dire en ce qui concerne les occasions de réprimande, soit on refuse le terme, soit on stigmatise ce qui va contre les valeurs altruistes comme l'égoïsme, la vulgarité et la colère (ainsi que le désordre, signe d'une autonomie en

[15] Niveaux de CPF pris à 3 pour mille. Il faut bien préciser que c'est ce choix qui fait apparaitre petit à petit les points, en fonction de leur importance dans la création des types : la méthode est inductive, c'est-à-dire que ce n'est pas l'interprète qui fait la sélection des points à son gré.

[16] Modalités non présentées dans le graphique général et qui n'apparaissent que sur le graphique à grande échelle dans sa partie non montrée ici.

cours). Ceci n'empêche pas les occasions de désaccord, comme en ce qui concerne les sorties : l'autonomie est une conquête progressive et les parents s'estiment un droit de surveillance.

Nous somme là au cœur de la tension éducative : l'injonction d'autonomie peut sembler une contradiction[17] :

- comment être autonome et dépendant de sa famille ? Ce qui est défini par ce style, c'est une priorité donnée dans l'éducation, non une reconnaissance (illusoire) d'autonomie existante. L'enfant d'âge collège est toujours dépendant et l'éducation passe par des activités faite avec lui telles qu'on les trouve signalées dans le graphique : de discussion, d'activités culturelles faites à l'extérieur ou à la maison, mais aussi par le biais d'activités d'épanouissement personnel (écouter de la musique, programmer), plus que par le sport de compétition qui n'est pas apprécié.

- comment être autonome et ouvert aux autres ? Cette double exigence manifeste que l'autonomie n'est pas vécue comme un isolement mais comme une phase nécessaire d'intégration dans un groupe dont on accepte les règles. Etre autonome et ouvert aux autres, c'est entrer de plein pied dans la tradition démocratique où par la discussion, qui suppose des ressources personnelles et une conviction autonome, on est confronté aux autres que l'on respecte.

VII Retour aux hypothèses de départ

Cette enquête avait été faite pour tester plusieurs hypothèses dont certaines sont faciles à éliminer.

1) le nouveau style des écoles nouvelles serait l'amplification de valeurs féminines : on a vu que s'il y a une opposition de *genre*, elle concerne la traditionnelle meilleure réussite scolaire des filles, dans le cadre de la première opposition que met en avant l'enquête.

2) les écoles nouvelles seraient destinées à rattraper des élèves de classe moyenne supérieure en échec scolaire. Cette hypothèse tombe du fait que les deux axes du graphique sont indépendants (orthogonaux sur le graphique). Il y a une opposition en terme de niveau scolaire (axe horizontal) et une en terme de style éducatif (axe vertical). Il y a des familles de chaque style éducatif dans chaque niveau scolaire. Ce que le graphique nous suggère cependant, c'est que *l'Ecole Alsacienne*, située en haut (nouveau style éducatif) et à droite (bon niveau scolaire) correspond à l'alliance de ces deux qualités, *La Source* est à un niveau d'indépendance.

Une modalité semble cependant manifester que la difficulté scolaire, faite d'inadaptation au système traditionnel conduit à trouver dans l'Ecole nouvelle une solution. En effet, dans le tout premier graphique, le point le plus en haut à gauche (donc appartenant au mauvais niveau et au nouveau style éducatif) correspond à l'indication de difficultés scolaires au primaire en terme d'adaptation au système scolaire, ou de manque d'intérêt ("DifScolAdaptationScol"). On ne peut donc exclure

[17] comme l'injonction d'amour, le "double bind", la célèbre double contrainte impossible à réaliser car il est contradictoire de demander d'aimer si l'amour est un mouvement libre.

que des difficultés d'adaptation conduisent des parents vers des écoles nouvelles plus respectueuses des démarches de chacun[18].

3) quant à la troisième hypothèse qui assimile le choix d'un nouveau style éducatif à une reconversion à des valeurs nouvelles d'autonomie, de capacité relationnelle et d'authenticité, elle est vérifiée d'une certaine façon mais ce qui est mis en avant par le choix des parents, c'est la racine profonde de ces nouvelles valeurs, que révèle cette tension entre autonomie et ouverture aux autres comme nous allons le voir maintenant. Les écoles nouvelles ne sont pas caractérisées par le fait de mettre l'accent sur l'autonomie et sur les capacités relationnelles, elles sont caractérisées par un nouveau style de positionnement où l'individu, ayant confiance en lui-même, a aussi la volonté d'entrer en dialogue avec les autres.

Il nous faut cependant revenir un peu en arrière en notant que le premier clivage apparu dans notre population étudiée est relatif à la perception du niveau scolaire de l'enfant avec toutes les pratiques éducatives qui vont avec. C'est l'opposition majeure dans notre population enquêtée, comme c'est le souci fondamental de toute famille ayant un enfant d'âge scolaire. Quel que soit le style éducatif, il n'est pas possible de s'affranchir de la réussite scolaire. Ce qu'apporte l'enquête, c'est qu'à ce souci commun peuvent être associés des styles éducatifs différents : le style classique des milieux sociaux favorisés (ceux de notre enquête) fait de pression, de compétition, d'inculcation des valeurs traditionnelles de l'enfant "bien élevé" ; mais aussi ce style d'éducation nouvelle fait de cette tension entre construction d'une autonomie et souci du groupe.

VIII Education nouvelle et société

Pour rendre compte de cette tension, il faut examiner la dynamique qui est à l'origine du mouvement de l'Education nouvelle au 20^{e} siècle[19]. En effet, si les réformateurs pédagogiques des origines (Montaigne, Port-Royal, Comenius, Rousseau) ont toujours eu l'idée que la dynamique de l'apprentissage passait par l'intérêt propre de l'enfant (contre l'idée que l'enfant pouvait être instruit contre son gré), la spécificité des réformateurs contemporains a été de tenir compte aussi des avancées scientifiques d'une part et de la question *politique* d'autre part.

Je n'insisterai pas ici sur l'apport de la psychopédagogie, en particulier des apports de Piaget, qui a été personnellement partie prenante du mouvement de l'Education nouvelle, mais je voudrais souligner l'aspect *politique,* au sens large, du mouvement. Prenons le cas d'Henri Wallon (1879-1962) à la fois psychologue et homme politique, lui aussi partie prenante du mouvement de l'Education nouvelle : il insiste sur l'étude nécessaire de l'enfant[20], à la fois d'un point de vue individuel (qui relève de la psychologie) et d'un point de vue collectif (qui relève de l'étude du milieu de l'enfant).

Comme on le sait, le rapport Langevin-Wallon, issu d'une commission réunie dès la fin de la guerre et qui rendit son rapport en juin 1947, propose une réforme de l'enseignement qui, dans un but de démocratisation propose une unification des

---

[18] Cette modalité est significativement liée à des indicateurs de mauvais niveau scolaire d'une part et est prise par quelques enfants de *Decroly*, ce qui explique sa position sur le graphique.

[19] Annick Raymond, *L'éducation morale dans le mouvement de l'Education nouvelle*, L'Harmattan, 2002

[20] cf. Annick Raymond 2002 : p.140

réseaux scolaires et, dans un but d'efficacité pédagogique, reprend des acquis de l'éducation nouvelle.

Comme le dit dans ce sens l'introduction du rapport : "les études primaires, secondaires, supérieures sont trop souvent en marge du réel. L'école semble un milieu clos, imperméable aux expériences du monde. Le divorce entre l'enseignement scolaire et la vie s'accentue par la permanence de nos institutions scolaires au sein d'une société en voie d'évolution accélérée. Ce divorce dépouille l'enseignement de son caractère éducatif. Une réforme est urgente qui remédiera à cette carence de l'enseignement dans l'éducation du producteur et du citoyen et lui permettra de donner à tous une formation civique, sociale, humaine".[21]

Ce rapport restera lettre morte même si sous la 5e République, les réformes utiliseront certaines de ses propositions comme l'orientation, en en détournant l'objet. Alors que le projet visait la démocratisation de l'enseignement, on doit bien se rendre compte, comme le souligne l'historien Antoine Prost, que "la démocratisation a progressé jusqu'au début des années soixante dans une structure scolaire pensée par des conservateurs avec une volonté proprement réactionnaire de défense et illustration des humanités, alors qu'au contraire, les réformes de 1959, 1963 et 1965, qui voulaient assurer l'égalité des chances devant l'école et la démocratisation de l'enseignement ont, dans les faits, organisé le recrutement de l'élite scolaire au sein de l'élite sociale."[22]. En effet la procédure d'orientation a été détournée : la formation professionnelle a été utilisée comme une voie d'échec, ce qui en a fait un repoussoir. Les méthodes actives issues de l'école nouvelle ont été ignorées.

La visée *politique* est présente dans le rapport et, dans le paragraphe consacré à l'éducation morale et civique, on trouve une citation de Paul Langevin, autre acteur du mouvement de l'Education nouvelle : "l'école fait faire à l'enfant l'apprentissage de la vie sociale et, singulièrement, de la vie démocratique. Ainsi se dégage la notion du groupe scolaire à structure démocratique auquel l'enfant participe comme futur citoyen et où peuvent se former en lui, non par les cours et les discours, mais par la vie et l'expérience, les vertus civiques fondamentales : sens de la responsabilité, discipline consentie, sacrifice à l'intérêt général, activités concertées et où on utilisera les diverses expériences de *self-government* dans la vie scolaire".

Si le système scolaire dans son ensemble reste loin de cet idéal, nous devons noter qu'il reste au cœur des aspirations des parents qui choisissent l'école nouvelle. La vie démocratique ne s'apprend pas comme un concept mais comme une expérience dans laquelle chaque enfant doit à la fois cultiver son autonomie, pour exister lui-même et, en même temps, mettre en œuvre l'acceptation des autres, ce *vouloir vivre ensemble* qui est selon Renan[23] ce qui rend possible la vie d'une collectivité, d'une nation[24]. Cette pratique est une valeur, une éthique et comme telle

---

[21] Le rapport Langevin-Wallon a été rendu disponible récemment dans : Claude Allègre et Philippe Meirieu, *Pour l'école*, Mille une nuits, 2004, il est également disponible sur internet par exemple à l'adresse http://perso.wanadoo.fr/claude.rochet/ecole/docs/langevin.pdf

[22] Antoine Prost, *L'enseignement s'est-il démocratisé ?* Presses universitaires de France, 1992, p.201.

[23] Ernest Renan, *Qu'est-ce qu'une nation ?* Edité par Joël Roman, Presses Pocket, 1992

[24] Aujourd'hui, le penseur qui a le mieux théorisé cette pratique sociale faite d'autonomie et d'ouverture aux autres est Habermas qui a montré que la pratique démocratique consiste à accepter le meilleur argument proposé dans la discussion. Cf. entre autres ouvrages, Jürgen Habermas, *De l'éthique de la discussion*, Les éditions du Cerf, 1992, Champs/Flammarion.

doit être apprise et transmise. L'apprentissage de la vie démocratique, fait d'autonomie personnelle et de souci de la collectivité, est ce qui est recherché dans une Ecole nouvelle, comme l'a montré empiriquement cette enquête.

IX Retour à l'analyse locale

Une fois la vue d'ensemble établie, il est possible de revenir à l'analyse *locale*, c'est-à-dire de centrer sont attention sur une modalité particulière en repérant où cette modalité se situe et avec lesquelles elle est liée.

A titre d'exemple, nous centrerons notre analyse locale sur la modalité « élève moyen » qui se situe dans le pôle des difficultés scolaires. A cette fin, on recherche quelles sont toutes les modalités qui sont significativement en attraction avec le fait d'être élève moyen : on utilise à nouveau le PEM Pourcentage de l'écart maximum. Les résultats peuvent être manifestés de deux manières : d'abord en utilisant un morceau du graphe factoriel et en visualisant par un trait la présence (et si l'on veut l'intensité) d'une attraction.

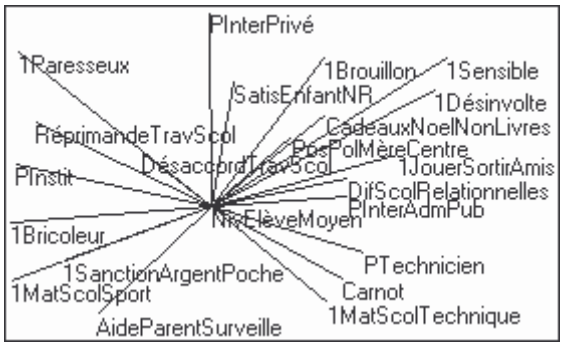


Figure 13 : Graphe des attractions de l'élève moyen

On peut aussi présenter le profil des PEM qui donne par attraction décroissante les autres modalités de l'analyse significativement[25] en attraction.

Profil de la modalité V333 NivElèveMoyen
N=126

Nom	PEM	Obs.	Test	Intitulé
V7A5	100	2	*	1Bricoleur
CSP7	100	2	*	PInstit
35A9	78	5	***	1MatScolTechnique
V8AB	62	5	**	1Désinvolte
35A8	56	8	***	1MatScolSport
CSP9	50	5	**	PInterAdmPub
V7AS	34	9	**	1Sensible
CSPA	34	6	*	PInterPrivé
V8AR	31	25	***	1Paresseux

[25] On utilise pour les test du khi-deux les repères suivants : 3 étoiles, significatif au seuil de 1%, 2 étoiles : 5%, une étoile : 10%

```
V851   30    8 *   1SanctionArgentPoche
CSPB   29    7 *   PTechnicien
17B4   28   10 **  Carnot
V320   26   18 *** SatisEnfantNR
67A3   24   88 **  CadeauxNoelNonLivres
84A1   23   41 *** RéprimandeTravScol
V392   22   32 *** DifScolRelationnelles
V8A7   15   14 *   1Brouillon
89A2   13   48 *** DésaccordTravScol
V361   13   32 **  AideParentSurveille
V952   12   19 *   PosPolMèreCentre
61A1   12   63 *   1JouerSortirAmis
```

Il y a un certain « bruit » dans ce genre de profil (mais le bruit est moins gênant que le « silence ») : par exemple les deux profils à 100% signifient que les 2 individus dont la première qualité est d'être bricoleur sont tous des élèves moyens. On retrouve dans le profil les doléances des parents sur le niveau scolaire, le fait de mettre en avant les matières où l'enfant est « bon » : le technique et le sport, des parents de classe moyenne au style éducatif traditionnel sanctionnant par le biais de l'argent de poche, et où la lecture n'est pas privilégiée.

On peut ainsi « expliquer » une modalité par les autres modalités qui lui sont liées comme on explique un mot du dictionnaire par d'autres mots en lien sémantique avec lui. Plus le nombre de modalités pris dans l'analyse sera vaste et plus le profil sera riche et l'explication intéressante.

X Retour à la méthode

Il n'est possible de montrer comment on dépouille une enquête qu'en le faisant, en interprétant un graphique où beaucoup de modalités sont présentes afin qu'il soit assez riche. On a vu aussi que cette interprétation revient aux hypothèses de départ pour montrer comment elles sont soit réfutées soit modifiées. Enfin, les résultats de l'analyse ont été replacés dans une problématique plus vaste qui est celle du domaine étudié, ici la sociologie de l'éducation.

Il y a là un effet qui peut être dangereux : dans la mesure où sont injectées ici beaucoup de connaissances extérieures à l'enquête, on peut légitimement se demander si les graphiques factoriels n'ont pas servi de test projectif des opinions de l'analyste.

Pour lever ce doute, nous allons maintenant procéder à des vérifications empiriques à propos des quatre types de répondants qui ont été isolés en revenant aux données elles-mêmes. Nous étudierons en détail le type que nous avons le plus développé, qui est au cœur de l'enquête, le nouveau style d'éducation en nous posant une question simple : quelle est l'importance numérique de ce type de répondant et comment pouvons nous le définir ? A cette fin nous allons montrer comment il est possible de constituer une nouvelle variable qui définirait l'appartenance au type.

XI Construire une nouvelle variable d'un type

La vérification le plus simple est le comptage : nous allons prendre les modalités qui sont apparues dans le nouveau style pédagogique et nous allons compter combien d'individus ont en commun ces modalités. Nous utiliserons donc les questions suivantes présentes dans le graphique correspondant (détaillé, non présenté ici) :

1) plusieurs qualités de l'enfant apparaissent ; prenons celles qui sont données en premier : attentionné, autonome, curieux, ouvert, rapide, serviable et la non-réponse qui refuse ce genre de jugement. Ces qualités se caractérisent par leur aspect non scolaire et par les qualités humaines d'ouverture et d'attention aux autres. Il va de soi qu'un tel test est projectif du projet éducatif des parents.

2) inversement, quand on évoque un cas éventuel de réprimande, le parent propose les cas suivants : désordre, vulgarité, égoïsme, colère ou encore le refus de cette éventualité. C'est l'aspect négatif du fait de se centre sur soi qui est sanctionné, non le manque d'ardeur au travail.

3) la question suivante évoque en conséquence une sanction possible : ici soit on refuse de répondre, soit une autre issue est envisagée.

4) les valeurs éducatives proposées ici sont : l'autonomie, la confiance, le respect, la générosité, la capacité d'éveil, la tolérance.

5) quand on demande le niveau scolaire de l'élève, la réponse est qu'on refuse ce genre de classement.

6) le parent, quand on lui demande s'il a l'impression de reproduire le modèle éducatif qu'il a reçu répond que ce n'est pas le cas du tout.

7) enfin, l'enfant est au courant des opinions politiques de ses parents.

On a donc 7 questions dont certaines ont plusieurs modalités dans le style repéré : nous allons simplement compter combien d'individus ont de modalités de ce type : le maximum est 7 car quand plusieurs modalités d'une question sont présente, il ne s'agit pas de réponses multiples mais de réponses proches et chaque répondant ne peut en prendre qu'une par question.

Nous sommes ainsi en mesure de construire un indicateur simple d'appartenance au style : ceux qui en auront 7 constitueront le type pur et ceux qui n'en n'auront aucune seront des opposants stricts.

Voici le comptage du nombre de modalités, de 0 au maximum observé.

Tot.	0	1	2	3	4	5
512	66	182	164	69	16	15
100	12.9	35.5	32.0	13.5	3.1	2.9

En examinant cette distribution, la première réaction peut être la déception : aucun individu n'atteint le maximum, ni même 6 modalités et ils sont peu nombreux à en avoir 4 ou 5. Les deux tiers de la population ont une ou deux modalités du type, ni refus strict, ni adhésion notable.

Comme ce phénomène est permanent, quelque soit l'enquête, il vaut mieux comprendre la situation de la façon suivante : ce que nous propose l'analyse des correspondances dans les regroupement que l'on observe ne sont pas des types à l'état pur, mais des types à l'état approché dont la présence simultanée de modalités forme un tout logique intelligible. C'est ce que Weber a appelé un « type-idéal » : c'est ce qui a fait à la fois la séduction de ce type d'analyse et sa difficulté quand on a pris pour des types réels ce qui n'était que type-idéal[26].

L'expérience montre qu'une bonne approximation du type est donnée quand on regroupe les individus qui ont au moins la moitié des cas observés, c'est-à-dire ici, puisque 5 est le maximum, ceux qui en ont 3, 4 ou 5, c'est-à-dire les 100 individus qui représentent 20% de la population.

Cette population ne se confond pas avec les enfants dont les parents ont choisi une école nouvelle : si l'on croise le type approché du nouveau style éducatif avec le type d'école, on a les résultats suivants :

	Nouveau style éducatif		
	Non	Oui	Total
Ecole nouvelle	65 66,3	33 33,7	98 100
Collège autre	347 83,8	67 16,2	414 100
Total	412 80,5	100 19,5	512 100

Tableau 10 : Ecole et style éducatif

Il y a bien attraction entre les parents d'Ecole nouvelle et le nouveau style éducatif puisque ce style représente 20% de la population et que, dans la population d'Ecole nouvelle, le pourcentage est supérieur (34%). Sur le graphique, les écoles nouvelles sont bien proches du nouveau style éducatif mais c'est bien une *attraction* qui est indiquée, non une exclusivité. Le nouveau style éducatif se retrouve aussi chez des parents qui n'ont pas fait le choix de l'Ecole nouvelle.

La nouvelle variable qui a été créée à partir d'un type-idéal (et que l'on peut appeler pour cette raison variable idéale-typique) permet de résumer l'information, de la synthétiser, et va nous permettre maintenant d'aller plus loin en tentant, sur cette variable, d'appliquer des techniques d'analyse « toutes choses égales par ailleurs » pour pouvoir discerner d'une manière fine ce qui peut rendre compte de cette attitude.

Avant d'étudier ces techniques dont la plus utilisée est la régression logistique, nous allons rester encore un chapitre avec l'analyse des correspondances pour en regarder quelques figures classiques.

[26] Cf. Chapitre suivant pour plus de précisions sur ce point.

Chapitre 4 : les figures de l'analyse des correspondances

I La forme en parabole : effet Guttman

Une configuration très classique fait que le nuage des modalités se présente, dans le premier plan factoriel (axes 1 et 2) sous la forme d'une parabole

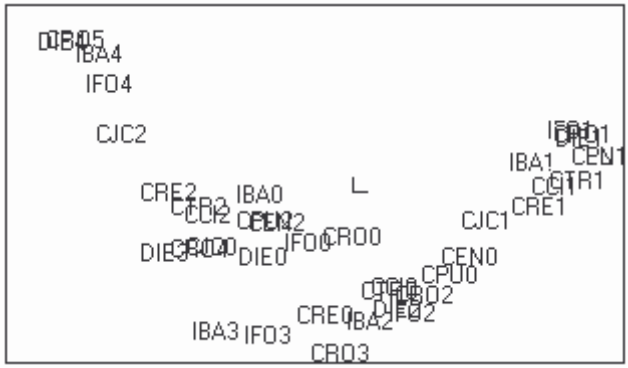


Figure 14 Parabole des modalités

Cette forme en parabole est appelée "Effet Guttman" du nom du sociologue Louis Guttman (1916-1987) connu pour ses recherches méthodologiques, en particulier sur les échelles de réponses. Ce genre de configuration se produit quand précisément il y a des liens multiples entre les réponses. Sur le même exemple, qui sera explicité ensuite, on projette le graphe des PEM entre modalités et l'on voit qu'elles sont liées deux à deux par proximité. Pour rendre compte de ces fortes liaisons, l'analyse des correspondances construit un premier axe d'opposition entre très en accord et très opposés à ce qui fait le premier axe et un deuxième axe qui, artificiellement, oppose les positions extrêmes aux positions moyennes. Il n'y a pas lieu de s'extasier sur cette configuration artificielle qui signale simplement le phénomène d'une forte liaison entre les diverses questions de l'enquête qui est également repérée par la force des premières valeurs propres (ici égale à 0,4 pour le premier facteur, c'est-à-dire largement au-dessus du seuil empirique de 0,1 que l'on a déjà donné comme valeur repère d'une forte liaison).

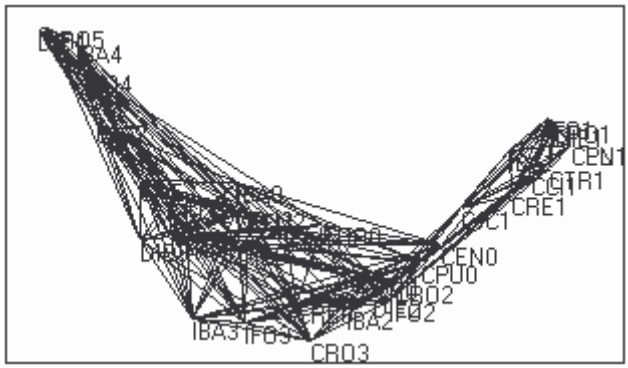


Figure 15 Parabole et PEM entre modalités

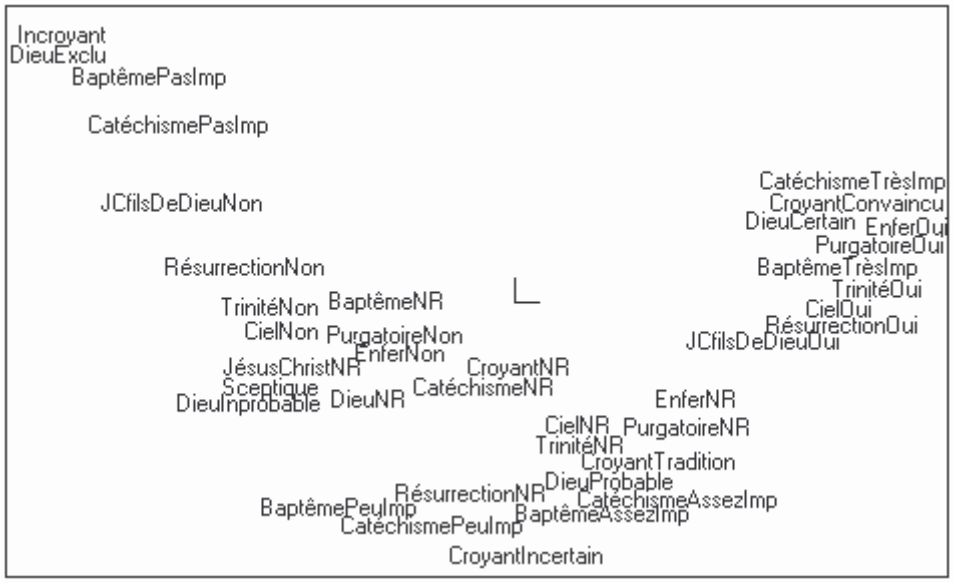


Figure 16 Croyances religieuses

La figure 16 est issue d'une enquête réalisée en 1986 sur les rapports entre les français et le catholicisme[27]. On voit que le côté droit de la parabole correspond au pôle des catholiques convaincus (l'existence de Dieu est certaine, le baptême des enfants et leur instruction religieuse sont très importants, l'interrogé se dit croyant convaincu, il croit en la trinité, en la résurrection du Christ, au Ciel, au Purgatoire et à l'Enfer). Inversement le pôle gauche correspond à l'incroyance, à l'exclusion de Dieu, au rejet de l'importance du baptême et du catéchisme. L'aspect plus intéressant de cet effet Guttman se situe dans le bas de la parabole, dans le passage de la croyance ferme à la croyance incertaine puis au scepticisme. Par exemple on voit que le croyant par tradition, s'il considère que Dieu est probable, se réfugie dans la non-réponse pour le dogme traditionnel (Enfer, Purgatoire, Ciel, Trinité) mais juge encore assez importante la formation religieuse des enfants. D'une manière symétrique, le sceptique, s'il dit non aux mêmes dogmes, se réfugie dans la non-réponse pour la formation religieuse des enfants. Il n'y a que l'incroyant affirmé qui la rejette comme pas importante du tout. L'intérêt de cette échelle de croyance (Guttmanienne) est l'éclairage qu'elle permet d'apporter sur la manière dont a été comprise la question d'autodéfinition de la croyance où l'on a demandé si l'intéressé se définissant comme : "un croyant convaincu, un croyant par tradition, un croyant incertain, un sceptique, un incroyant". *A priori*, entre "incertain" et "sceptique", l'écart de sens est faible et l'on pourrait avancer qu'il y a équivalence entre les deux termes. Au vu du graphique, avec le point *Croyant incertain* le plus en bas, intermédiaire entre la tradition et le scepticisme, on voit bien que les répondants se sont appuyés sur la gradation qui leur était proposée. Il ont bien vu qu'il s'agissait déjà d'une échelle en 5 points du plus croyant au moins croyant où ils pouvaient se positionner d'une manière fine. Ceci explique la cohérence des réponses.

On voit sur cet exemple ce que signifie de parler de questions liées entre elles : cela veut dire qu'ici toutes les modalités de réponses de type croyant convaincu sont prises à peu près par les mêmes individus. Des questions liées entre elles impliquent des répondants typiques : on rencontre fréquemment ce phénomène quand on met

[27] Guy Michelat, Julien Potel, Jacques Sutter, Jacques Maitre, *Les français sont-ils encore catholiques ?* Paris, éditions du Cerf, 1991

dans une même analyse des questions qui se présentent de la même façon dans leur réponses comme "tout à fait d'accord", "assez d'accord", "plutôt pas d'accord" et "pas d'accord du tout". La routine de la réponse l'emporte et le questionneur, qui n'a pas trop cherché à approfondir son problème, recueille, et c'est justice, une réponse stéréotypée de l'enquêté.

II Effets des faibles effectifs

Nous avons déjà repéré une figure fréquente de l'analyse des correspondances, celle qui correspond à un faible effectif : il s'agit de la "comète et de sa queue" (chapitre 3 figure 4) où le noyau central correspond à un bloc agglutiné de questions et la queue de la comète à une ou plusieurs modalités à faible effectif. Ce problème est plus général car plusieurs points à faible effectifs peuvent entrainer la création d'un plan factoriel qui est tout à fait interprétable mais qui peut être un piège[28]. On utilise maintenant des données issues de l'enquête sur les pratiques culturelles des français de 1989[29] dans laquelle on trouve un ensemble de questions portant sur les sorties suivantes effectuées ou non dans l'année précédente : les sorties sont classées par ordre d'importance décroissante ; la population est celle des 4722 adultes de l'enquête ; ce qui est pris en compte est le fait d'avoir effectué la sortie indiquée dans les 12 derniers mois.

	Effectif	%
Cinéma	2106	44,6
Bal	1131	24,0
Discothèque ou boite	1104	23,4
Match	1014	21,5
Théâtre	621	13,2
Concert musique classique	457	9,7
Concert de rock	427	9,0
Cirque	378	8,0
Spectacle de danse	294	6,2
Concert de jazz	281	6,0
Opéra	156	3,3

On a donc 11 activités (et les 11 non-activités aux effectifs complémentaires mises en supplémentaires). On éclairera la compréhension du graphique en mettant en variables supplémentaires le sexe, l'âge et le niveau de diplôme du répondant. (figure 17)

On a sur la gauche du graphique un regroupement de sorties à forte charge culturelle : opéra, musique classique, spectacle de danse, théâtre, qui sont pratiquées par un public diplômé de 2e ou 3e cycle universitaire (ou grandes écoles).

---

[28] Philippe Cibois, "Les pièges de l'analyse des correspondances", *Histoire & Mesure*, 12 (3/4), 1997, pp. 299-320.

[29] Olivier Donnat et Denis Cogneau, 1990, *Les pratiques culturelles des français 1973-1989*, La découverte / La documentation française.

On distingue à droite une culture de niveau de diplôme peu élevé associant sortie au bal et au match, une culture correspondant à des âges croissants n'excluant pas toute sortie (cirque comme accompagnateur d'enfants), une culture jeune de boite et de rock, avec le jazz qui sert d'intermédiaire avec les sorties "distinctives". En effet, si le facteur vertical est liés à l'âge (des plus jeunes aux plus âgés en allant du haut en bas), l'axe horizontal correspond à l'opposition de "distinction" au sens de Pierre Bourdieu[30], c'est-à-dire à des pratiques dont le prestige culturel entraine la rareté et la distinction.

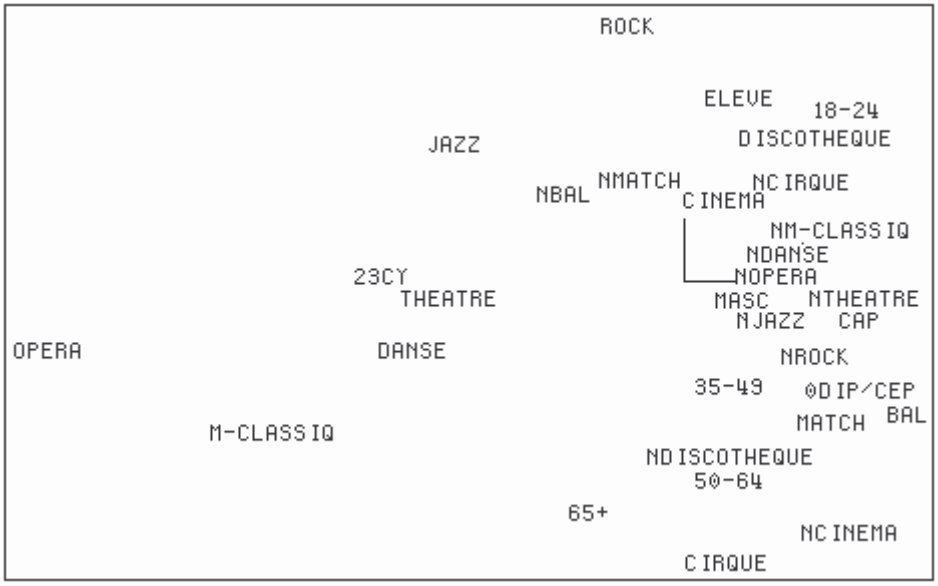


Figure 17 Sorties effectuées l'année précédente

Cependant cette rareté a un effet pervers : en effet si l'on procède au comptage des sorties multiples on voit dans le tableau ci-dessous que ceux qui ont fait 2 types de sorties ne représentent que 6% de l'ensemble et que pour 3 types de sorties et plus, on arrive à moins de 4% .

	Effectif	%
5 sorties	12	0,3
4 sorties	51	1,1
3 sorties	114	2,4
2 sorties	283	6,0
1 sortie	637	13,5
0 sortie	3625	76,8
Total	4722	100

Cet effet de distinction peut se présenter du fait de quelques conjonctions entre modalités rares : pour s'en préserver, quelques comptages sous forme de tris-croisés permettent de repérer l'importance numérique de ces co-occurrences. Le cas échéant, la mise en éléments supplémentaires de ces modalités trop rares doit être faite.

[30] Pierre Bourdieu, *La distinction*, Paris, Ed. de Minuit, 1979.

III Des types idéaux

On a déjà évoqué le problème plus tôt (page 72) en montrant qu'un type idéal manifesté par une analyse des correspondances ne devait pas être considéré comme un type réel. Revenons sur cette question à partir d'un exemple traité antérieurement[31]: il s'agit d'une enquête sur les ouvriers français faite à la suite des évènements de 1968[32]. En ne prenant en compte que les affiliations politiques et syndicales repérées par la déclaration d'appartenance et le vote pour un parti politique et un syndicat on voit sur la figure 18 une opposition entre :

-un pôle CGT-PC à gauche sur le graphique : appartient et vote CGT, se sent proche du Parti Communiste et a voté J.Duclos, candidat du PCF au premier tour des élections présidentielles de 1969, toutes modalités ayant une contribution supérieure à 150 pour mille alors que la moyenne (1000 divisé par 32 modalités) est de 31 pour mille.

- un pôle gauche non communiste en haut : CFDT, partis de gauche et candidats de gauche.

- en bas à droite, un pôle de droite : UNR, Pompidou, parti et candidat gaullistes de l'époque.

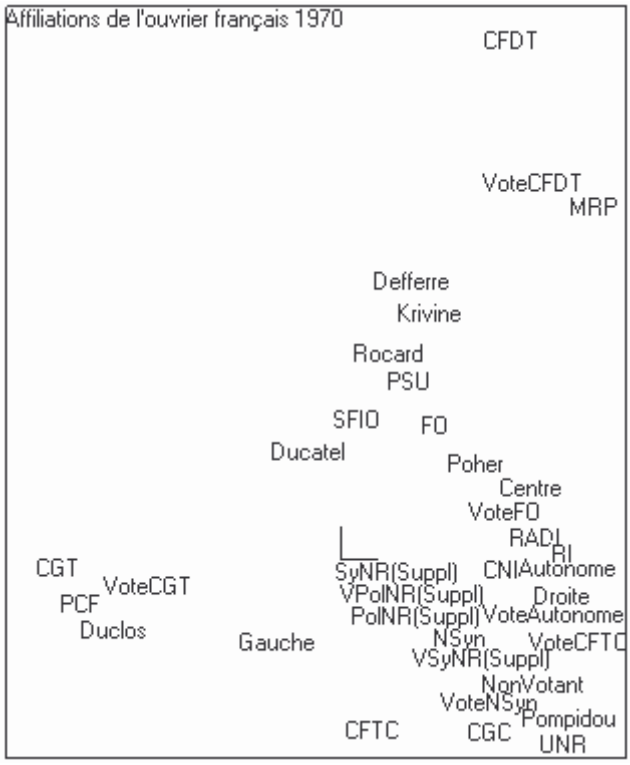


Figure 18 Affiliations de l'ouvrier français 1970

Le paradoxe est que si l'on compte combien d'individu sur un total de 1116 ont les 4 modalités de type PC-CGT (alors que chaque modalité du type représente de 200 à 300 personnes), on n'aboutit qu'à 81 individus, soit 7,3%

---

[31] Philippe Cibois, *L'analyse des données en sociologie*, Paris, PUF, 1984 : cet ouvrage n'est plus édité car il ne correspond plus à l'état des techniques actuelles. Je pense en avoir gardé l'essentiel dans le présent ouvrage.

[32] Gérard Adam, Frédéric Bon, Jean Capdevielle, René Mouriaux, *L'ouvrier français en 1970*, Pris, Presses de la FNSP, 1970.

Avec les mêmes critères (4 modalités du type) on classe 9 individus en gauche non communiste, ce qui explique pourquoi ces faibles effectifs entrainent un fort éloignement au centre, et 41 individus à droite. On classe donc avec ces types 131 individus sur 1116 soit 11,7% du total.

Au vu de ces chiffres on comprend bien en quoi l'analyse des corespondances est un procédé qui produit des types-idéaux et non des types numériquement importants. Pour avoir des effectifs suffisants, il faut prendre le principe déjà évoqué, qui consiste à prendre pour un type approché, l'appartenance à au moins la moitié du nombre d'éléments constitutifs du type, ici 2. Avec ce critère et en prenant des précautions pour ne pas faire de doubles comptes, on peut classer environ 70% de la population.

Bernard Lahire, en reprenant ces résultats[33], rappelle que ceci explique pourquoi Pierre Bourdieu utilisait beaucoup l'analyse des correspondances car elle visualisait des types-idéaux qui correspondaient à sa pensée en termes de champs et il critique ce qu'il appelle l'usage paresseux de la méthode idéaltypique. En effet il montre que, si on fait un comptage soigneux des types de répondants, on s'aperçoit que les dissonnances culturelles (c'est le sous-titre de son livre) sont nombreuses et que tel qui va à l'opéra peut tout aussi bien suivre avec passion le Tour de France, ou pire, aux yeux d'une culture légitime intégriste.

Si l'on a bien repéré que l'analyse des correspondances propose des types-idéaux et non des types statistiquement bien attestés, on pourra se servir de cette méthode sans courir le risque de projeter sur la réalité sociale les propres stéréotypes de cette société, en particulier par le biais de ses pratiques distinctives. Pour ce faire il suffit de compter[34] en construisant ce que j'appelle des variables idéal-typiques, c'est-à-dire des indicateurs qui comptent combien chaque individu à de modalités du type.

[33] Bernard Lahire, *La culture des individus*, Paris, La découverte, 2004, p.132-136.

[34] "Compter ses hommes" était la devise de l'adjudant de compagnie rappelait plaisamment Georges Guilbaud : cela reste la devise du sociologue (à condition de compter tout autant les individus de sexe féminin que les personnes de sexe masculin).

Chapitre 5 : les techniques d'analyse « toutes choses égales par ailleurs »

Ces techniques, comme l'analyse des correspondances sont des *approximations* des données mais les régularisations qu'elles effectuent rendent les données beaucoup plus facile à interpréter, ce qui est intéressant si l'on dispose d'indicateurs qui nous permettent de nous rendre compte de la validité des résultats.

Parler de techniques « toutes choses égales par ailleurs » fait appel à l'idée que si un phénomène social est du à plusieurs causes, il peut être intéressant de voir l'effet propre de chacune des causes, indépendamment des autres.

Pour reprendre un exemple traité par ailleurs[35], si le fait d'avoir un fort niveau d'étude encourage à la lecture, et si l'on sait que les femmes lisent plus que les hommes : comme les deux aspects vont ensemble, on souhaite avoir une méthode qui neutralise l'effet de sexe pour isoler l'effet de niveau d'étude (et inversement qui neutralise l'effet de niveau d'étude pour avoir l'effet de sexe).

I Analyse tabulaire multivariée

On dispose d'une méthode simple et ancienne pour traiter de genre de question dans des tableaux croisés, c'est ce qu'on appelle l'analyse multivariée dont la base est d'abord de disposer d'une répartition de la population en ligne qui croise tous les cas de figure possible : dans l'exemple sur la lecture, 2 pour le sexe et 2 pour le niveau d'étude, ce qui fait 4 cas de figure que l'on va croiser avec la variable à expliquer. Nous allons traiter ici un exemple analogue pour chercher à savoir comment est adopté le nouveau style éducatif, précisément en tenant compte du sexe de l'enfant et du niveau d'étude du père[36]. On a les 4 lignes suivantes :

	Nouveau style éducatif		
	Oui	Non	Total
Masc. NivInf	18 13,33	117 86,67	135 100
Féminin NivInf	28 19,58	115 80,42	143 100
Masc. NivSup	23 21,10	86 78,90	109 100
Féminin NivSup	31 24,80	94 75,20	125 100
Total	100 19,53	412 80,47	512 100

Tableau 11 : analyse multivariée

[35] Philippe Cibois, "Modèle linéaire contre modèle logistique en régression sur données qualitatives", *Bulletin de méthodologie sociologique*, n°64, 1999, p.5-24.

[36] Recodé en *supérieur* pour ceux qui ont un niveau Bac+5 et grandes écoles et *inférieur* pour les autres (essentiellement bac+3 ou 4).

Le plus bas niveau de choix du nouveau style éducatif se trouve dans la première ligne : pour des garçons de la part d'un père de niveau d'étude inférieur. La proportion est de 13,33%.

Nous allons neutraliser successivement l'effet du sexe et du niveau d'étude. Commençons par le niveau d'étude où deux situations sont possibles pour voir l'effet du sexe, dans le cas du niveau inférieur (deux premières lignes) ou du niveau supérieur (deux dernières lignes).

1) effet du sexe : entre les deux premières lignes, toutes deux de niveau inférieur, la seule différence est que, en passant du sexe masculin au sexe féminin, la proportion de nouveau style éducatif passe de 13,33% à 19,58% soit une augmentation de 19,58 – 13,33 = 6,25 points.

Refaisons le même calcul pour le niveau supérieur (les deux dernières lignes) : la différence est cette fois de 24,80 – 21,10 = 3,70. On constate donc qu'il y a dans les deux cas un effet féminin qui fait monter la proportion de nouveau style : l'idée d'approximation sera introduite ici en prenant la moyenne des deux effets : l'effet féminin est de (6,25 + 3,70)/2= 5,0 points de pourcentage.

2) effet du niveau d'étude : nous réutilisons les mêmes lignes du tableau mais de façon différente. Pour le sexe masculin (1ère et 3e ligne), l'effet de niveau supérieur fait que l'on passe de 13,33% à 21,10% soit une augmentation de 7,77.

Pour les sexe féminin (2e et 4e ligne) l'augmentation est de 24,80 – 19,58 = 5,22. Les deux effets vont dans le même sens et l'effet moyen est de 6,5

En utilisant ces effets moyens, il devient possible de présenter les données d'une manière spécifique à l'analyse « toutes choses égales par ailleurs » : par rapport à la situation masculin et niveau inférieur, de 13,3, l'effet féminin ajoute +5,0 et indépendamment, l'effet niveau supérieur ajoute +6,5. On résume l'information en donnant la situation d'où l'on est parti comme référence et l'on donne séparément les deux effets.

Situation de référence : masculin niveau inf. :	13,3
Effet féminin	+5,0
Effet niveau supérieur	+6,5

Les deux effets vont dans le même sens : quand on a un enfant de sexe féminin, la propension à choisir le nouveau style d'éducation augmente de 5% toutes choses égales par ailleurs, c'est-à-dire quelque soit le niveau d'étude du père. De même le niveau supérieur favorise une augmentation de 6,5%. Le choix du point de départ est sans importance. Si on avait pris comme situation de référence le sexe féminin, l'effet masculin aurait simplement été inversé, il aurait fait baisser de 5%, de même pour le niveau d'étude. D'une manière pratique, il faut choisir comme référence ce qui est le plus clair et le plus intelligible. Parler d'effet féminin est clair car on sait qu'il s'agit d'un effet sur les études qui a déjà été repéré[37].

Cette manière simplifiée de présenter les données (situation de référence + effets séparés) permet de reconstruire une approximation des données, dont on vérifiera, dans le cas présent, qu'on ne comment pas trop d'erreur en prenant l'approximation plutôt que la réalité qui, dans ce cas simple, est entièrement connue.

---

[37] Christian Baudelot et Roger Establet, *Allez les filles !* Seuil, 1998

L'erreur, malgré les simplifications apportées par l'utilisation de la moyenne simple, n'atteint pas 2%.

Présence du nouveau style éducatif				
		Modèle	Obs.	Err.
Masc. NivInf	Référence	13,3	13,3	0
Féminin NivInf	Ref.+effet Féminin	13,3+5,0 =18,3	19,6	-1,3
Masc. NivSup	Ref.+effet Nivsup	13,3+ 6,5 =19,8	21,1	-1,3
Féminin NivSup	Ref.+Fémi +NivSup	13,3+5,0 +6,5=24,8	24,8	0

Tableau 12 : comparaison modèle et observation

Dans la suite on utilisera une moyenne pondérée, c'est-à-dire que chaque élément de la moyenne vaudra au prorata de l'effectif du groupe. Par exemple l'effet féminin qui était calculé simplement en prenant la moyenne ordinaire (6,25+3,70)/2= 5,0 sera calculé en pondérant le premier sous-effet de 6,25 par 278 (effectif des deux premières lignes d'où est tirée la différence et qui correspond à l'effectif total du niveau inférieur) et de même 3,70 sera pondéré par l'effectif du niveau supérieur). Le total général est de 512. Le calcul de moyenne pondérée est le suivant :

(6,25 x 278/512) + (3,70 x 234/512) = 5,1 : ici la différence est peu sensible car les deux groupes sont équilibrés. J'appelle cette manière de faire l'analyse tabulaire car tout est issu de calculs à l'intérieur de tableaux croisés[38]. Comme on va le voir, ses résultats sont toujours très proches de la manière de faire la plus utilisée, la régression logistique sur des modalités de réponses.

Il faut parler de la régression logistique pour comprendre sa logique mais, comme ses résultats sont très semblables à ceux de l'analyse tabulaire, je ne chercherai pas dans un ouvrage d'initiation à la présenter comme telle : il vaut mieux interpréter les résultats dans une logique d'analyse tabulaire. Ce qui suit en montre la logique pour qui a déjà une idée de la régression en général.

II La régression multiple

L'idée de la régression multiple (linéaire aussi bien que logistique) est d'avoir une variable à expliquer (y de y=ax+b de l'équation ordinaire d'une droite) et plusieurs variables explicatives (x_1, x_2, x_n pour une régression multiple de la forme

$y = a_1x_1 + a_2x_2 + \ldots + a_nx_n + b$) où les $\underline{x}$ sont la présence (notée 1) ou l'absence (notée 0) d'une modalité explicative (dans l'exemple précédent, le fait d'être de sexe féminin ou de niveau supérieur) et où les $\underline{a}$ sont des coefficients numériques qui vont être calculés. Le coefficient b est appelé l'ordonnée à l'origine (en anglais *intercept*).

[38] Philippe Cibois, "Modèle linéaire contre modèle logistique en régression sur données qualitatives", *Bulletin de méthodologie sociologique*, 1999, n°64, p.5-24.

Le nombre de coefficients a et de modalités x dans le cas de l'analyse précédente est de deux comme on l'a vu. En effet, dans le cas d'une question à deux modalités, définir l'effet féminin, c'est rendre compte de la question en entier car la deuxième modalité a servi de repère, de référence. Quand nous aurons une question à trois modalités (ou davantage), une seule servira de référence et les autres seront toutes un effet spécifique. Si l'on veut utiliser une variable reflétant une orientation politique en droite / centre / gauche, il faudra par exemple choisir le centre comme référence et l'on aura un effet *gauche* et un effet *droite*. On prend souvent une modalité intermédiaire comme référence (par exemple pour les tranches d'âge) mais ce n'est pas une obligation : le but du choix est de rendre l'interprétation plus aisée. Il faut prendre une situation de référence pour chaque question mais on a le choix.

Prendre une modalité comme référence, c'est *ne pas l'utiliser* dans les données car on utilise toutes les autres modalités de la même question qui suffisent donc à l'information. Pour reprendre le codage d'une affiliation politique en trois modalités et qu'on prenne le centre comme référence, si un individu n'est ni de gauche ni de droite, c'est qu'il est du centre, même si cette modalité n'est pas indiquée, l'information qu'elle comporte est portée par les deux autres.

Les données qui sont traitées sont un tableau où en ligne se trouvent tous les individus de l'enquête et où à chaque colonne correspond une modalité (qui n'est pas la référence). Chaque modalité est codée en présence / absence, c'est-à-dire en 0 / 1. Pour l'exemple précédent, les trois cas possibles sont codés de la manière suivante :

Individus	Gauche	Droite
de gauche	1	0
de droite	0	1
du centre	0	0

Tableau 13 : exemple de codage

Si on veut garder la possibilité de non-réponse, il faut créer une modalité supplémentaire et on aurait alors le codage suivant (en conservant le centre comme référence)

Individus	Gauche	Droite	Non-réponse
de gauche	1	0	0
de droite	0	1	0
du centre	0	0	0
non-rép.	0	0	1

Tableau 14 : ajout de la modalité de non-réponse

Comme on le verra, la multiplication du nombre de modalités a des effets plutôt négatifs sur la fiabilité des résultats et il vaut mieux donc recoder les non-réponses.

Pour la question à expliquer, et quelque soit le nombre de modalités, seule est utilisée la modalité qui est précisément à expliquer.

Quand un individu ne prend que les modalités qui sont toutes de références, pour lui, tous les x sont nuls. Le coefficient b correspond alors à la situation de référence et y = b.

Pour mieux comprendre prenons le cas de la régression linéaire appliqué aux trois variables précédentes : la variable à expliquer (Y) est le Nouveau style éducatif, les variables explicatives sont l'effet féminin (X_1) et l'effet niveau supérieur d'éducation (X_2). La manière linéaire d'écrire l'équation de régression multiple $Y = a_1X_1 + a_2X_2 + b$ devient :

NouvStyle = a_1Féminin + a_2NivSup + b

En régression linéaire, les coefficients ont les valeurs suivantes, (entre parenthèse, les valeurs correspondantes de l'analyse tabulaire avec pondération)

a_1 = 5,2 (5,1) effet féminin

a_2 = 6,6 (6,4) effet niveau supérieur

b = 13,8 (13,3) situation de la référence (masc, inf)

Finalement l'équation de régression s'écrit :

NS (:NouvStyle) = 5,2 Féminin+6,6 NivSup + 13,8

Selon qu'on donne la valeur zéro ou un à chaque modalité, selon qu'elle est présente ou absente, le modèle linéaire conduit ici à 4 situations :

Si Fém=1 et NivSup=1 NS=5,2+6,6 +13,8= 25,6%

Si Fém=1 et NivSup=0 NS=5,2 +13,8= 19,0%

Si Fém=0 et NivSup=1 NS=6,6 +13,8= 20,4%

Si Fém=0 et NivSup=0 NS= 13,8 = 13,8%

On voit que là aussi, les paramètres de la régression linéaire sont proches de ceux de l'analyse tabulaire et de l'observation. Dans les résultats de la régression linéaire, les paramètres ne sont pas présentés en pourcentage comme ici, mais en proportion, ce qui n'est pas difficile à transformer.

En régression logistique, ce n'est plus la simple proportion p qui est estimée mais le rapport

p / (1-p) appelé en anglais *odds*, que l'on peut traduire par *chances* (on itilisera toujours le mot *risques* si le contexte le nécessite : on parle des chances d'avoir un examen et du risque d'être malade)

III Chances et rapport des chances

En anglais, *odd*, sans s, désigne "la petite chose qui s'ajoute" : soit au nombre pair (*even*) et c'est alors le nombre impair, soit à un nombre quelconque : *odd* désigne alors ce qui est en plus du nombre rond (*odd change* désigne la monnaie faite à partir d'un billet), d'où par extension, ce qui est dépareillé ou non usuel. Passant de l'adjectif au nom au pluriel, *odds* passe de l'idée d'imparité à la désignation de l'inégalité, des avantages, des chances. L'usage le plus connu du mot est celui utilisé par les turfistes pour parler des *chances* d'un cheval, de sa *cote* : quand on dit que tel cheval est coté à 3 contre 1, on signifie que sa probabilité de

gagner est 3 fois plus grande que sa probabilité de perdre et donc par conséquence que si l'on parie sur lui (et qu'il gagne) on obtiendra 3 fois la somme pariée alors que s'il pert, on perdra la mise. Les *odds*, les *chances*, mettent en rapport une situation dissymétrique : au numérateur on a la probabilité de la réussite, et, plus largement de la "bonne situation" et au dénominateur, la probabilité de l'échec, de la mauvais issue. Evidemment, la relation entre la probabilité de la réussite et celle de l'échec est la complémentarité à l'unité. Si la probabilité pour un cheval d'arriver gagnant est de 0,75, celle de son échec est de

(1 – 0,75) = 0,25 et sa cote est de p / (1 – p) soit 0,75 / 0,25 c'est à dire 3 contre 1.

Une difficulté de vocabulaire vient du fait que l'on parle aussi de chances pour désigner simplement la probabilité : ce qui lève l'ambigüité est le fait que les chances au sens de cote sont toujours suivies de la mention de *contre*.

Ceci s'applique aussi dans le cas des cotes inférieures à 1 (car des chances supérieures à l'unité ne peuvent être confondues avec des probabilités toujours comprises entre 0 et 1). Par exemple si nous reprenons la première ligne du tableau 11 qui croise le choix d'un nouveau style éducatif avec la situation de l'élève, on a :

	Nouveau style éducatif		
	Oui	Non	Total
Masc. NivInf	18 13,33	117 86,67	135 100

Tableau 15

Les chances de recevoir un nouveau style éducatif sont le rapport de la probabilité de l'avoir

(18 / 135) = 0,1333 rapporté à son complément, la probabilité de ne pas l'avoir

(117 / 135) = 0,8667 = (1 – 0,1333). Ces chances sont de 0,13333 / 0,8667 = 0,154 contre 1. Chances qui peuvent être calculées plus simplement en faisant le rapport des effectifs : 18 / 117 = 0,154.

Plutôt que le rapport 0,154 / 1 qui ne parle pas à l'imagination, on le multipliera par 10 (ou par 100) et l'on dira que les chances de recevoir un nouveau style d'éducation pour ces garçons de faible niveau est de 1,54 contre 10 (de ne pas le recevoir) ou de 15,4 contre 100. Evidemment si l'on considérait les risques plutôt que les chances, ils seraient de l'inverse soit 117 / 18 = 6,5 de *ne pas* recevoir un nouveau style éducatif contre 1 (de le recevoir).

Le tableau suivant nous donne le calcul pour les quatre situations :

Chances du nouveau style éducatif			
Situation	Effectif oui	Effectif non	Chances = Oui/Non
Masc. NivInf	18	117	0,1538
Féminin NivInf	28	115	0,2435
Masc. NivSup	23	86	0,2674
Féminin NivSup	31	94	0,3298

Tableau 16

Examinons les chances pour les deux situations extrêmes : pour les garçons de niveau inférieur, les chances de recevoir un nouveau style éducatif sont de 15,38 chances contre 100 de ne pas en recevoir un tandis que pour les filles de niveau supérieur elles sont de 32,98 contre 100, soit le double. Ceci nous introduit au rapport des chances (en anglais *odds ratio* d'où l'abréviation OR souvent utilisée aussi en français).

Si l'on prend comme référence, c'est-à-dire comme dénominateur du rapport, la situation masculine, le rapport des chances est le suivant :

OR : Chances FemiNivSup / Chances MascNivInf

0,3298 / 0,1538 = 2,1 : les chances féminines (de niv. sup) sont le double des chances masculines (de niv. inf.).

Le rapport des chances (OR = *Odds Ratio*) est toujours un nombre positif tantôt inférieur à 1 tantôt supérieur, il servira de multiplicateur pour modifier les chances de la référence.

IV Equation de la régression logistique

L'équation de la régression logistique décrit une situation générale à gauche du signe égal sous la forme des chances de l'obtenir (variable à expliquer, équivalent du y d'une régression linéaire). Il est fréquent d'écrire ces chances sous la forme p / (1 – p) car quand on aura les chances d'une situation particulière, on pourra en déduire la probabilité p.

A droite du signe égal on a les chances de la situation de référence multipliées par un ensemble de multiplicateurs qui dépendent de toutes les situations. Ces multiplicateurs sont les *Odds Ratio*.

p / (1 – p) = chances de la référence x produit de multiplicateurs dépendants des situation. Ici les chances de la référence sont estimées à 0,1659.

Attention, cette valeur ne correspond pas strictement à l'observation 0,1538 (donnée par l'analyse tabulaire) car la régression logistique est un modèle estimé à partir de l'ensemble des données.

Dans le cas présent, il y a deux OR multiplicateurs qui modifient ces chances de la référence : un relatif au sexe féminin (valeur estimée = 1,39) et un relatif au niveau supérieur (valeur estimée = 1,51). L'équation de régression logistique s'écrit donc :

p/(1–p) = 0,1659 x 1,39 (si fémi) x 1,51 (si NivSup)

Cette équation permet d'examiner tous les cas de figures :

1) Féminin et Niveau supérieur : les chances de la référence (0,1659) sont multipliées par le produit des deux multiplicateur 1,39 x 1,51 = 2,09 : on retrouve le fait que les chances observées soient multipliées par deux.

p / (1 – p) = 0,1659 x 1,39 x 1,51 = 0,348 d'où l'on tire p = 0,348 (1 – p) soit p = 0,348 – 0,348p et donc p + 0,348p = 0,348 et p (1 + 0,348) = 0,348 et finalement p = 0,348 / 1,348. Concrètement quand on a des chances *Ch*, pour retrouver la proportion correspondante p, on prend p = *Ch* / (1 + *Ch*) formule appelée dans la suite "p issue des chances". On a ici p = 0,348 / 1,348 = 0,258 soit 25,8%

2) Masculin et niveau inférieur : c'est la situation de référence, il n'y a pas de coefficient multiplicateur, les chances 0,1659 ne sont pas modifiées et la proportion estimée est de p = 0,1659 / 1,1659 = 0,142 soit 14,2%.

3) Féminin seul : c'est là tout l'intérêt de la méthode qui consiste à voir l'effet d'une modalité seule, indépendamment des autres. Dans ce cas, les chances de la référence sont multipliées par le seul coefficient multiplicateur 1,39 qui correspond au sexe féminin.

p / (1 – p) = 0,1659 x 1,39 = 0,231

d'où p = 0,231 / 1,231 = 0,187 soit 18,7% c'est-à-dire 4,5 points de pourcentage de plus que pour la situation de référence (18,7 – 14,2). On dit que l'effet marginal en pourcentage est de 4,5 points (ou, d'une manière discutable mais courante de 4,5%)

4) Niveau supérieur seul :

p / (1 – p) = 0,1659 x 1,51 = 0,251

d'où p = 0,251 / 1,251 = 0,200 soit 20,0%. Le gain de pourcentage est de 20,0 – 14,2 = 5,8%. L'effet niveau supérieur est plus important que l'effet féminin.

Plusieurs remarques sont à faire :

1) les résultats de la régression logistique sont du même ordre que ceux de l'analyse tabulaire donnée plus haut : c'est toujours le cas. Ils sont également de même ordre que ceux de la régression linéaire.

2) si l'on compare la situation de référence, les effets simples et la situation où ses effets vont ensemble on a les 4 cas étudiés plus haut :

- masculin niv.inf. (référence) 14,2%

- féminin seul : 18,7% soit un effet de 4,5%

- niveau sup. seul : 20,0% soit un effet de 5,8%

- féminin et niv.sup. : 25,8 soit un effet de 11,6% qui n'est pas égal à la somme des deux effets isolés car 4,5 + 5,8 = 10,3. Ce phénomène est général : si l'on veut calculer l'effet conjoint il faut multiplier entre eux les OR puis appliquer ce coefficient aux chances de la référence plutôt que d'ajouter algébriquement les effets marginaux

en pourcentage. Cependant, il faut discuter ce principe car les précisions sont illusoires : ce que nous donne la régression logistique est un modèle qui comme tout modèle est simplificateur de l'observation. Vouloir isoler l'effet pur comme étant une réalité existante, alors que c'est une modélisation simplificatrice, risque de transformer une démarche exploratoire en création artificielle qui semble plus exacte que l'observation : c'est une démarche risquée dont je souligne le danger.

3) les formules multiplicatrices que l'on rencontre souvent sont du type suivant

$p/(1-p) = OR_1^{x1} \times OR_2^{x2} \times ChancesRef$

où ici OR_1 et OR_2 sont les OR de "féminin" et de "niveau supérieur" qui servent de coefficient multiplicateur.

Les exposants X_1 et X_2 correspondent au codage des données en présence / absence, c'est-à-dire en 0 /1 étudié plus haut. Pour l'OR = 1,39 correspondant à féminin, $1{,}39^1 = 1{,}39$ correspond au fait qu'on traite la présence codée 1 de la modalité féminin et $1{,}39^0 = 1$ correspond au fait que l'on traite l'absence codée 0 de la modalité féminin. Le multiplicateur 1 est neutre et sans effet sur le reste. D'une manière plus imagée, on peut écrire.

$p/(1-p) = 1{,}39^{Féninin} \times 1{,}51^{NivSup} \times ChancesRef$

4) pour des raisons diverses, théoriques et historiques, une transformation logarithmique est souvent faite de la formule multiplicative. Cette transformation remplace le produit des OR et des chances de la référence par une somme où les exposants deviennent des multiplications. Dans le cas présent on a :

log(p/(1-p))=

Féminin x log(1,39)

+ NivSup x log(1,51)

+ log(ChancesRef)

où "Féminin" ou "NivSup" comme précédemment ne prennent que les valeurs 1 (présence) ou 0 (absence). D'une manière générale on note ces indicateurs de présence/absence par x_1, x_2 etc., les résultats numériques des logarithmes (naturels) des OR par des coefficients a_1, a_2 et le log des chances de la référence par un coefficient b. On retrouve ainsi le symbolisme de la régression linéaire multiple :

$\log (p/(1-p) = a_1x_1 + a_2x_2 + b$ soit ici

$\log (p/(1-p) = 0{,}329\ x_1 + 0{,}409\ x_2 - 1{,}7964$

Ce qui ne simplifie pas la gestion pratique des résultats vient du fait que les programmes usuels donnent soit les coefficients a (log des OR) et b (log des chances de la référence), soit les OR et les chances de la référence, soit l'évaluation en pourcentage de la référence et les effets en pourcentage associés à un OR. Plutôt que de donner des formules directes qui permettent de passer des coefficients a et b aux autres résultats, la manière la plus simple est de se ramener aux OR.

Pour passer des coefficients a aux OR, il suffit de prendre la fonction inverse du logarithme naturel, c'est-à-dire la fonction exponentielle.

Par exemple pour la modalité "féminin" : les programmes donnent un coefficient a = 0,329. Pour retrouver l'OR on prend la fonction exponentielle (inverse du logarithme naturel ln).

Exp (0,329) = 1,39 (inversement ln (1,39) = 0,329)

De même pour passer du coefficient b aux chances de la référence, il suffit de prendre l'exponentielle de b car exp(– 1,7964) = 0,1659

V Utilisation de la régression logistique

Après cette utilisation sur un exemple simplifié, utilisons la méthode en introduisant davantage de modalités, toujours pour expliquer le nouveau style d'éducation.

Il faut cependant être prudent et ne pas introduire un trop grand nombre de modalités : comme le montre l'analyse tabulaire, introduire une nouvelle question (à plusieurs modalités), c'est faire un tri de profondeur supplémentaire, donc émietter les données et rendre les résultats instables (et non significatifs). On peut, pour faire ce choix des modalités à introduire, faire une analyse des correspondances préalable à partir d'une nouvelle variable d'intérêt, ici le niveau éducatif, et introduire toute les variables explicatives pensables : religion, politique, niveau scolaire des grands-parents, etc.

A partir d'une analyse de cette sorte, on va retenir les indicateurs suivants :

- le sexe de l'enfant,

- le fait qu'il soit scolairement en retard ou non,

- l'opinion politique du répondant (qui est la mère), que l'on regroupera en quatre modalités : droite, gauche, écologistes et le regroupement de ceux qui refusent de se positionner ou qui se qualifient au centre (dans la suite "ni gauche ni droite"),

- un indicateur de pratique religieuse : on qualifie un répondant de lié à la religion s'il a un des caractères suivants : pratique religieuse régulière de la mère, enseignement religieux donné à l'enfant, communion solennelle faite par l'enfant.

La variable à expliquer est toujours le fait d'avoir choisi un nouveau style éducatif.

Pour chaque question, il faut examiner l'effet d'une des modalités et donc en prendre une comme référence. Pour le sexe on prendra "masculin" comme référence pour voir s'il y a un effet "féminin" : comme il n'y a que deux modalités, prendre l'option inverse correspondrait à inverser simplement le signe de l'effet, ce qui n'a pas beaucoup d'importance. Pour le retard, on prendra le fait de ne pas y être comme référence (donc à l'heure ou en avance). Pour la politique, le fait de n'avoir pas d'opinion marquée (centre ou refus), pour la religion, le fait de ne pas en avoir.

Les résultats sont donnés directement par les programmes mais le détail est important pour comprendre les résultats :

1) évaluation de la situation de référence : le fait d'être masculin, à l'heure, ni gauche ni droite, sans religion. Le modèle de la régression logistique donne un paramètre qui indique 1) le logarithme des chances de la référence = -1,3987 qui permet d'obtenir 2) les chances de la référence : exp (-1,3987) = 0,2469 et donc 3) par la formule *p issue des chances* vue plus haut p = *Ch* / (1 + *Ch*) la proportion estimée = 0,2469 / 1,2469 = 0,198 soit 19,8%

2) pour chacune des modalités ayant un effet (autre que la référence), on a, par exemple pour le cas de la religion, le logarithme *du rapport* des chances (Odds Ratio ou OR) = -0,5118 : comme il est négatif, son exponentielle sera inférieure à 1 : exp (-0,5118) = 0,60. Quand donc la religion est présente, les chances de la référence sont multipliées par 0,60 soit 0,60 x 0,2469 = 0,1481 et la proportion dans cette situation est égal à 0,1481 / 1,1481 = 0,1290 soit 12,9%. On voit qu'on a baissé de 6,9 points de pourcentage. On note cet "effet marginal" en négatif soit -6,9.

D'une manière analogue, les différents effets marginaux sont les suivants :

Religion	-6,9
Droite	-9,7
Gauche	0,0
Ecologiste	8,3
Féminin	4,6
En retard	18,1

Dans les programmes usuels, ces résultats sont jugés significatifs ou non en utilisant un indicateur qui suit une distribution du Khi-deux. Ici tous les effets sont significatifs sauf celui de la gauche.

Les programmes usuels donnent plus ou moins de détails, mais une version complète peut donner les coefficients logarithmiques, les OR, les effets marginaux, leur seuil de signification : dans le tableau 6 ci-dessous, trois étoiles correspondent à un seuil de 1%, deux à 5%, une à 10% et *ns* veut dire non-significatif.

Si seuls les coefficients sont indiqués, il faut se souvenir que l'exponentielle d'un coefficient négatif correspond à un OR inférieur à un qui correspond à un effet marginal négatif. L'influence est négative et symétriquement elle est positive quand le coefficient est positif. De même un OR plus grand que un correspond à un effet marginal positif (et réciproquement s'il est inférieur à un).

Modalité à expliquer : nouveau style éducatif					
		Coeff.	Chances	Pourc.	
Sit.réf.		-1,40	0,2469	19,8	
			Odds-ratio	Effet marg.	Test
Relig.					
	Oui	-0,51	0,60	-6,9	***
	Non	ref.			
Pol.					
	Droite	-0,79	0,45	-9,7	**
	Non marqué	ref.			
	Gauche	0,003	1,00	0,0	ns
	Ecolog.	0,46	1,59	8,3	*
Sexe					
	Masc	ref.			
	Fémi	0,27	1,31	4,6	*
Retard					
	Oui	0,90	2,47	18,1	**
	Non	ref.			

Tableau 17 : modèle complet de régression logistique

L'interprétation que l'on peut faire est qu'il y a des caractéristiques qui sont plus ou moins importantes dans le choix d'un nouveau style éducatif et que ces caractéristiques peuvent agir indépendamment les unes des autres. En ce qui concerne l'enfant, le fait qu'il soit en retard agit puissamment (+18%), beaucoup plus que le fait qu'il soit de sexe féminin (+5%) : on retrouve des résultats déjà étudiés. Par contre, les opinions touchant les parents sont également à prendre en compte : si l'option de gauche parait neutre, l'option écologiste, c'est-à-dire une certaine non-conformité au modèle des classes moyennes supérieures agit dans le sens d'un choix éducatif non-conformiste (+8%). Inversement, le choix de droite (-10%) ou le choix religieux (-7%) manifestent bien ce refus d'un choix éducatif non traditionnel. Le choix d'un style éducatif nouveau fait de confiance en l'enfant, de respect de son développement s'impose plus quand il est en difficulté mais peut être aussi choisi au nom d'options politiques et sociales. La régression logistique manifeste que ces choix sont faits "toutes choses égales par ailleurs" c'est-à-dire d'une manière indépendante.

VI Variations

On peut présenter les mêmes résultats sous forme d'une formule, soit additive en utilisant les coefficients logarithmiques, soit multiplicative en utilisant les odds-ratios. A gauche du signe égal on a la variable à expliquer, les chances de la situation générale dans le cas multiplicatif, son logarithme dans le cas additif.

1) manière de faire multiplicative : les chances du cas général sont égales aux chances de la référence multipliée par les rapport des chances (OR) des différentes modalités autres que de référence. Dans l'exemple du tableau 6 on a :

$p/(1-p) = 0{,}60^{\text{ReligOui}} \times 0{,}45^{\text{Droite}} \times 1{,}0^{\text{Gauche}} \times 1{,}59^{\text{Ecolog}} \times 1{,}31^{\text{Fémi}} \times 2{,}47^{\text{RetardOui}} \times 0{,}2469$

Quand une modalité est prise (codage logique = 1), son rapport de chance est pris, si elle n'est pas prise (codage logique = 0), le multiplicateur vaut 1 donc est neutre. Cette formule permet de cumuler plusieurs situations : étudions par exemple le cas d'une mère écologiste (OR = 1,59) dont l'enfant est en retard (OR=2,47). Le produit des rapports appliqué aux chances de la référence est égal à (calculé avec 4 décimales):

1,5858 x 2,4682 x 0,2469 = 0,9664 d'où l'on tire par *p issue des chances* :

p = 0,9664 / 1,9664 = 0,491 soit 49,1%. L'écart à la situation de référence est de 49,1 – 19,8 = +29,3. On remarquera que cet écart est différent de la somme des deux effets marginaux correspondants 8,3 + 18,1 = 26,4 ce résultat est général et entraine la consigne "qu'on n'a pas le droit d'additionner algébriquement les effets marginaux", ce qui est exact mais qui ne doit pas laisser penser que les résultats calculés avec le passage aux OR sont contradictoires avec la dite somme. De toute façon, comme déjà dit, les résultats obtenus sont les résultats de l'estimation d'un modèle, non d'une observation (que peut donner le détail d'une analyse tabulaire).

2) manière de faire additive : le logarithme (naturel) des chances du cas général sont égales au logarithme des chances de la référence additionné des logarithmes des OR des modalités autres que la référence. Dans l'exemple du tableau 6 on a :

ln(p/(1-p)) (quantité appelée aussi logit)

ln(p/(1-p)) = -0,51ReligOui -0,79Droite +0,003Gauche +0,46Ecolog +0,27Fémi +0,90RetardOui –1,40

Quand une modalité est prise (codage logique = 1), son coefficient logarithmique est pris, si elle n'est pas prise (codage logique = 0), le coefficient logarithmique vaut 0 donc est neutre. Si nous reprenons l'exemple précédent (Ecologiste et retard) et en prenant les valeurs exactes à la 4^{e} décimale, ln(p/(1-p) = 0,4611 + 0,9035 – 1,3987 = -0,0341

l'exponentiel du membre de gauche nous donne les chances p/(1-p), exp(-0,0341) =0,9665. On retrouve à l'arrondi près le coefficient multiplicateur précédent qui conduit donc au même résultat.

Cette gymnastique de calculs permet de s'assurer de la bonne compréhension des résultats mais la lecture rapide qui est faite des résultats porte soit sur le signe des coefficients logarithmiques, soit sur la position par rapport à l'unité des OR, soit sur le signe des effets marginaux, tout en vérifiant qu'ils soient significatifs.

Annexe au chapitre 2 : Algorithme

Le lecteur qui souhaiterait ouvrir la "boite noire" qui était évoquée dans la préface de cet ouvrage, trouvera ici des éléments de réponse à la question de savoir comment fait l'analyse factorielle pour trouver un couple des vecteurs, propres à des données, et qui permettent d'en construire une approximation.

Ce que nous cherchons c'est, à partir d'un tableau quelconque, de trouver un jeu de coefficients pour des lignes et des colonnes qui permettent, par multiplication terme à terme, de trouver un tableau connu par ses marges. Pour montrer comment peut se faire cette recherche, nous allons utiliser un tableau à trois lignes (marquées A, B et C) et deux colonnes (I et II) : il s'agit d'un exemple choisi pour sa simplicité, mais qui ne correspond à aucune donnée précise.

	I	II
A	0	1
B	1	2
C	3	3

Recherche de coefficients lignes et colonnes

Examinons les colonnes du tableau : dans les deux cas, le premier élément est inférieur au deuxième, lui-même inférieur au troisième. La suite de coefficients colonnes que nous recherchons, et désormais nous appellerons ces suites de nombres des *vecteurs*, ce vecteur colonne donc, qui doit être un résumé des deux colonnes, doit avoir leur structure et doit donc ressembler à quelque chose comme (1, 2, 4) ou (1, 5, 10) mais certainement pas (10, 5, 1).

L'algorithme que nous allons utiliser (et qui est plus simple que celui de l'analyse des correspondances que nous verrons ensuite) suppose une valeur de départ, même imprécise, qui sera améliorée dans la suite.

Nous prendrons donc comme point de départ à améliorer le vecteur colonne (1, 2, 4). Ici la suite des opérations consiste à multiplier *scalairement* le vecteur colonne à chacune des deux colonnes. Cette multiplication scalaire nous est familière mais dans le registre de l'opération "facture", qui consiste, pour chacun des éléments achetés, à multiplier chacun par son prix individuel et à additionner le tout. Le résultat de la multiplication des deux vecteurs n'est pas un vecteur mais un résultat sur l'échelle numérique (*scala* est l'échelle en italien).

Faisons l'opération en appelant le vecteur initial du nom de F0 et le résultat final en ligne du nom de F1:

	I	II	F0
A	0	1	1
B	1	2	2
C	3	3	4
F1	14	17	

Le premier élément de F1 s'obtient en multipliant scalairement la colonne I et F0, le détail du calcul est le suivant

I		F0		
0	x	1	=	0
1	x	2	=	2
3	x	4	=	12
			Total=	14

En faisant de même pour la colonne II, on obtient le nouveau vecteur F1, constitué à partir des deux résultats. On constate que ce vecteur respecte la structure des trois lignes où le premier élément est inférieur ou à la limite égal au deuxième. Sans prétendre justifier l'algorithme, on voit qu'il intègre progressivement la structure des données du tableau. Pour continuer, il faut répéter la multiplication scalaire du vecteur F1 mais cette fois avec chacune des lignes du tableau.

	I	II	F0	F2
A	0	1	1	17
B	1	2	2	48
C	3	3	4	93
F1	14	17		

Pour la ligne C le détail du calcul est le suivant :

C		F1		
3	x	14	=	42
3	x	17	=	51
			Total=	93

La structure de F2 est comparable à celle de F0, notre point de départ arbitraire (mais choisi avec vraisemblance), en arrondissant on peut dire que F2 a pour structure (20, 50, 90), soit, en divisant chaque élément par 20, ce qui ne modifie pas la structure (1 2,5 4,5) assez proche du point de départ à une multiplication par 20 près. Pour pouvoir voir le phénomène avec plus de précision, examinons la structure en proportion de chacun des vecteurs. Par exemple pour F2, le premier élément 17 représente 17 / 158 = 0,108 soit 10,8%.

	I	II		F0	PropF0	F2	PropF2
A	0	1		1	0,143	17	0,108
B	1	2		2	0,286	48	0,304
C	3	3		4	0,571	93	0,589
			Total	7	1,000	158	1,000
F1	14	17	31				
PropF1	0,452	0,548	1,000				

On voit que de F0 à F2, la proportion du premier élément à baissé, ceux des autres a augmenté. Continuons les *itérations* de l'algorithme, c'est dire reprenons les étapes précédentes en prenant la valeur de F2 à la place de celle de F0. Nous multiplions

scalairement chacune des colonnes du tableau par F2 et nous obtenons F3 puis à partir de F3 multiplié par chacune des lignes nous obtenons F4.

	I	II		F0	PropF0	F2	PropF2	F4	PropF4
A	0	1		1	0,143	17	0,108	392	0,107
B	1	2		2	0,286	48	0,304	1111	0,304
C	3	3		4	0,571	93	0,589	2157	0,589
			Total	7	1,000	158	1,000	3660	1,000

F1	14	17	31	
PropF1	0,452	0,548	1,000	
F3	327	392	719	
PropF3	0,455	0,545	1,000	
F5	7582	9085	16667	
PropF5	0,455	0,545	1,000	Stop

En comparant les proportions de F2 et F4, on constate que, pour une précision de trois chiffres significatifs, les proportions sont égales sauf pour le premier élément qui passe de 10,8% à 10,7%. On voit ce qu'on appelle la *convergence* de l'algorithme qui se stabilise pour une précision donnée. Il suffit de faire une itération supplémentaire et passer de F4 à F5 pour retrouver strictement les proportions de F3. L'algorithme est terminé. Nous nous sommes affranchis de la valeur arbitraire du point de départ, les vecteurs sont maintenant *propres* aux données. Pour s'en rendre compte il suffit de changer F0 et de prendre par exemple la valeur la plus neutre possible (1, 1, 1).

	I	II		F0	PropF0	F2	PropF2	F4	PropF4	F6	PropF6
A	0	1		1	0,333	6	0,115	128	0,107	2958	0,107
B	1	2		1	0,333	16	0,308	362	0,304	8384	0,304
C	3	3		1	0,333	30	0,577	702	0,589	16278	0,589
			Total	3	1,000	52	1,000	1192	1,000	27620	1,000

F1	4	6	10	Stop
PropF1	0,400	0,600	1,000	
F3	106	128	234	
PropF3	0,453	0,547	1,000	
F5	2468	2958	5426	
PropF5	0,455	0,545	1,000	

Prendre un vecteur initial quelconque a modifié tous les effectifs mais non les proportions, on voit seulement qu'il a fallu une itération supplémentaire (PropF6 = PropF4) pour arriver à la convergence de l'algorithme. De même, si on prend un point de départ (qui peut tout aussi bien être pris en ligne), complètement erroné comme (10, 5, 1), on constate que la convergence n'est pas assurée à l'itération 6.

Prendre un mauvais point de départ a pour effet simplement d'augmenter le nombre d'itération. Dans une programmation en machine, on prend toujours le point de départ le plus neutre possible, soit (1, 1, 1)

Nous avons donc maintenant un couple de coefficients lignes et colonnes, des *vecteurs propres* aux données, qui expriment le mieux possible la structure du tableau, à condition qu'ils soient pris ensemble, par multiplication.

Reconstitution du tableau d'approximation

La reconstitution se fait donc par multiplication terme à terme des coefficients marginaux lignes et colonnes. Il faut prendre les vecteurs propres (donc après convergence de l'algorithme), c'est-à-dire à l'étape 5 pour le vecteur en ligne et à l'étape 6 pour le vecteur en colonne. Se pose simplement le problème de savoir quel vecteur propre choisir, celui en effectifs ou celui en proportions ? Comme ils sont proportionnels, ils expriment tous la même structure et il en existe donc une infinité de semblables. Pour rendre plus clair les opérations (mais on sort du cadre d'une analyse standard), il s'agit de faire l'approximation d'un tableau d'origine dont la somme des éléments est égale à 10 (cf. le tableau ci-dessous où les marges du tableau et sont total sont calculés).

Tableau d'origine

	I	II	Total
A	0	1	1
B	1	2	3
C	3	3	6
Total	4	6	10

	I	II	F6
A	0,049	0,058	0,107
B	0,138	0,166	0,304
C	0,268	0,321	0,589
F5	0,455	0,545	1

Proportion

Approximation

	I	II	F6
A	0,49	0,58	0,107
B	1,38	1,66	0,304
C	2,68	3,21	0,589
F5	0,455	0,545	10

Multiplié par 10

On calcule d'abord l'approximation en proportion par multiplication terme à terme (par exemple pour la première case A-I 0,107 x 0,455 = 0,049), puis, pour rendre la comparaison possible, on multiplie le résultat obtenu par 10. On voit alors que cette première case est approximée par 0,49. On voit que l'approximation est plutôt "bonne". Pour la dernière ligne, pour la première colonne, il manque 3 – 2,68 = 0,32 et pour la deuxième, il y a 0, 21 en trop. Examinons toutes les erreurs en étendant le calcul par soustraction à l'ensemble : on obtient le tableau du *reste*, ce qu'il faut ajouter à l'approximation pour retrouver le tableau d'origine.

Tableau d'origine =

	I	II
A	0	1
B	1	2
C	3	3

Approximation +

	I	II
A	0,49	0,58
B	1,38	1,66
C	2,68	3,21

Reste

	I	II
A	-0,49	0,42
B	-0,38	0,34
C	0,32	-0,21

On voit sur cet exemple que l'approximation a beaucoup plus d'importance que le reste : la plus petite valeur qu'on y rencontre, 0,49 est la plus grande (en valeur absolue) du reste. L'algorithme utilisé nous a permis de décomposer un tableau en deux tableaux dont le premier est une bonne approximation du tableau d'origine.

Mais il y a plusieurs types d'algorithmes : celui qui est le plus utilisé aujourd'hui est l'algorithme de l'analyse des correspondances qui, pour ne pas que les colonnes ou les lignes les plus importantes en effectif imposent le choix de l'élément prépondérant du facteur, introduit une *pondération par les marges*. A chaque pas de l'algorithme, quand un vecteur est obtenu, il est pondéré par les marges, c'est à dire divisé par elles. Reprenons l'exemple précédent en utilisant le point de départ le plus neutre possible, c'est à dire (1, 1, 1).

	I	II	Total	F0	F2NonPond	F2Pond
A	0	1	1	1	1	1
B	1	2	3	1	3	1
C	3	3	6	1	6	1
Total	4	6				
F1NonPond	4	6				
F1Pond	1	1				

Comme on l'a vu plus haut, le résultat obtenu pour F1 est (4, 6). Il est encore ici non pondéré, le pondérer, c'est le diviser par les marges et trouver comme vecteur F1 pondéré la valeur (1, 1). Le processus se répète dans l'autre sens et en multipliant le vecteur F1 pondéré par le tableau on obtient un vecteur F2 non pondéré égal à la marge en colonne. En pondérant on retrouve le vecteur F0 de départ et l'algorithme se termine ici puisque la convergence est immédiate.

Pour la reconstitution, on se sert des vecteurs non pondérés (identiques aux marges) et le produit des marges est (à la division par le total près) identique à l'effectif théorique correspondant à l'indépendance.

Tableau d'origine

	I	II	Total
A	0	1	1
B	1	2	3
C	3	3	6
Total	4	6	10

	I	II	F2NP
A	4	6	1
B	12	18	3
C	24	36	6
F1NP	4	6	

Approximation

	I	II
A	0,40	0,60
B	1,20	1,80
C	2,40	3,60

Divisé par 10

Dans ce cas particulier, la première approximation correspond à l'indépendance est le reste constitue les écarts à l'indépendance.

	Tableau d'origine =			Indépendance +			Ecart à l'indépendance	
	I	II		I	II		I	II
A	0	1	A	0,40	0,60	A	-0,40	0,40
B	1	2	B	1,20	1,80	B	-0,20	0,20
C	3	3	C	2,40	3,60	C	0,60	-0,60

Cette particularité est un des atouts de l'analyse des correspondances : la première approximation du tableau est l'indépendance ce qui veut dire que l'information pertinente se trouve dans le tableau des écarts à l'indépendance.

En résumé, nous avons vu qu'un tableau quelconque pouvait par le biais d'un algorithme être décomposé en une série de plusieurs tableaux : le premier, reconstitué par multiplication terme à terme des coefficients obtenus après convergence de l'algorithme, est une bonne approximation du tableau d'origine. Nous allons étudier maintenant la méthode la plus couramment utilisée en analyse factorielle, l'analyse des correspondances.

L'algorithme de l'analyse des correspondances

Pour montrer son fonctionnement, nous l'appliquons à des données réelles déjà vues, le tableau 1, intérêt vis-à-vis de la religion en fonction de la proximité politique. Comme en analyse des correspondances la première approximation est le tableau correspondant à l'indépendance (tableau 2 plus haut), ce qui reste du tableau initial après soustraction de cette première approximation est le tableau des écarts à l'indépendance (tableau 3). Pour la suite des calculs il est mis en proportion : les marges, qui vont servir de pondération dans la suite, sont également en proportion et le total général est de 1.

Ecarts	Intérêt pour la religion			
	Fort	Moyen	Nul	Total
Droite	0,0161	0,0133	-0,0294	0,1406
Centre	0,0034	0,0067	-0,0101	0,1094
Gauche	-0,0269	-0,0032	0,0301	0,3730
NiGniD	0,0074	-0,0168	0,0094	0,3770
Total	0,2188	0,4746	0,3066	1

Tableau des écarts à l'indépendance en proportion

Dans l'extrait de tableur ci-dessous, on présente les deux premières itérations de l'algorithme où 4 opérations sont utilisées :

1) le produit scalaire : par exemple entre le point de départ V0 (point de départ arbitraire fait d'unités avec cependant des signes plus et moins aléatoires, pour accélérer la convergence) et chacune des colonnes du tableau. Par exemple -0,0391 est égal à -1 x 0,0161 + -1 x 0,034 + 1 x -0,0269 + 1 x 0,0074[39].

[39] Les résultats affichés le sont à une certaine précision mais les calculs sont faits avec la précision maximum

	Fort	Moyen	Nul	Pond	V0	V2CNP	V2CPnd	CarPnd	V2RNP	V2RPnd
Droite	*0,0161*	*0,0133*	*-0,0294*	*0,1406*	-1	-0,07	-0,470	0,03	-0,29	-2,10
Centre	*0,0034*	*0,0067*	*-0,0101*	*0,1094*	-1	-0,02	-0,197	0,00	-0,10	-0,88
Gauche	*-0,0269*	*-0,0032*	*0,0301*	*0,3730*	1	0,07	0,196	0,01	0,33	0,88
NiGniD	*0,0074*	*-0,0168*	*0,0094*	*0,3770*	1	0,01	0,038	0,00	0,06	0,17
Pond	*0,2188*	*0,4746*	*0,3066*				Somme	0,05		
V1CNP	-0,0391	-0,0400	0,0791				Norme	0,2243		
V1CPnd	-0,1786	-0,0844	0,2580	Somme	Norme					
CarPnd	0,0070	0,0034	0,0204	0,0308	0,1754					
V1RNP	-0,22	-0,23	0,45					Itération 1		
V1RPnd	-1,02	-0,48	1,47							

	Fort	Moyen	Nul	Pond	V2RPnd	V4CNP	**V4CPnd**	CarPnd	V4RNP	V4RPnd
Droite	*0,0161*	*0,0133*	*-0,0294*	*0,1406*	-2,10	-0,07	**-0,470**	0,03	-0,29	-2,08
Centre	*0,0034*	*0,0067*	*-0,0101*	*0,1094*	-0,88	-0,02	**-0,191**	0,00	-0,09	-0,85
Gauche	*-0,0269*	*-0,0032*	*0,0301*	*0,3730*	0,88	0,08	**0,205**	0,02	0,34	0,91
NiGniD	*0,0074*	*-0,0168*	*0,0094*	*0,3770*	0,17	0,01	**0,029**	0,00	0,05	0,13
Pond	*0,2188*	*0,4746*	*0,3066*				Somme	0,05		
V3CNP	-0,06	-0,04	0,10				Norme	0,2260		
V3CPnd	**-0,270**	**-0,083**	**0,322**	**Somme**	Norme					
CarPnd	0,02	0,00	0,03	**0,0510**	0,2258					
V3RNP	-0,26	-0,18	0,44					Itération 2		
V3RPnd	-1,20	-0,37	1,42							

Analyse des correspondances, 1er facteur, Itérations 1 et 2

Le résultat de chaque produit scalaire est noté par la première ligne du vecteur V1 qui est "calibré" (c'est à dire qui a subit une augmentation de son importance) et qui n'est pas encore pondéré par les marges (notation : C pour calibré; Np pour non pondéré et dans la suite Pnd pour pondéré). Tous les résultats successifs des produits scalaires sont calibrés et non pondérés CNP.

2) la deuxième opération est la pondération qui permet de relativiser l'importance des lignes ou colonnes trop importantes. Chaque élément d'un vecteur calibré est divisé par l'élément marginal correspondant (noté Pond dans le tableau). Le vecteur résultant est noté Pnd pour pondéré : en ligne il est sous le précédent non pondéré, en colonne à droite. Par exemple le premier élément de V1CPnd -0,1786 est égal à -0,0391 / 0,2188.

3) la troisième opération est nécessaire pour réduire le vecteur, c'est-à-dire le rendre de "longueur" unitaire. A cette fin on calcule le carré de chaque élément non pondéré et on le divise par la pondération : cela revient à faire le produit d'un élément non pondéré par un élément pondéré. Un ligne (ou une colonne) présente ces carrés pondérés : elle est notée CarPnd. Sa somme est donnée ainsi que sa racine carrée appelée norme. Quand l'algorithme à convergé, cette somme est la valeur propre du facteur. Par exemple le premier élément 0,0070 est égal à -0,0391² / 0,2188. La somme des trois éléments est égale à 0,308 et la norme 0,1754 en est la racine carrée.

4) muni de cette norme on va réduire les vecteurs calibrés, c'est-à-dire diviser chaque élément par la norme. On a ainsi une version réduite notée R des vecteurs

non pondérés et pondérés. Par exemple le premier élément -0,22 de V1Rnp = -0,0391 / 0,1754 et le premier élément -1,02 de V1RPnd = -0,1786 / 0,1754.

Quand on a pondéré V1, calculé sa norme et qu'on l'a normé, on recommence le produit scalaire mais cette fois entre les lignes du tableau et le vecteur réduit pondéré V1RPnd : on obtient le vecteur 2, en colonne cette fois, V2CNP dont par exemple le premier élément -0,07 est égal à 0,0161 x -1,02 + 0,0133 x -0,48 + -0,0294 x 1,47. On le pondère, on calcule sa norme, on le réduit et le vecteur V2RPnd sert maintenant de vecteur initial pour la 2e itération. On peut effectuer simplement les itérations suivantes avec un tableur en dupliquant la première itération et en remplaçant les valeurs de V0 par celle de V2RPnd.

On répète les opérations jusqu'à fixité des valeurs (ici pas tout à fait obtenue entre V2 et V4, par contre V6, non indiqué ici, redonne les mêmes valeurs que V4). Les valeurs données par les programmes sont les vecteurs calibrés pondérés (ici en gras).

Ce sont les vecteurs non pondérés qui multipliés termes à termes et divisés par leur norme commune redonnent l'approximation[40]. Par exemple la première case du tableau (Droite, fort intérêt) 0,018 est égale à -0,470 x 0,1406 x -0,274 x 0,2188 / 0,2260

	Fort	Moyen	Nul	Pond	V6CPnd
Droite	0,018	0,011	-0,029	0,1406	**-0,470**
Centre	0,006	0,004	-0,009	0,1094	**-0,191**
Gauche	-0,020	-0,013	0,033	0,3730	**0,205**
NiGniD	-0,003	-0,002	0,005	0,3770	**0,028**
Pond	0,2188	0,4746	0,3066		Norme
V5CPnd	**-0,274**	**-0,081**	**0,321**		0,2260

Reconstitution des écarts : premier facteur.

Si l'on veut retrouver ce tableau en effectif tel qu'il est présenté plus haut au tableau 4, il suffit de multiplier chaque case par l'effectif total 512 pour passer des proportions aux effectifs (aux arrondis près). [41]

En ôtant cette reconstitution des écarts à l'indépendance initiaux on trouve un nouveau reste sur lequel on réitère le processus pour avoir le 2e facteur (ici le dernier).

[40] Cependant comme les programmes usuels donnent la version pondérée de ces vecteurs, on utilise ici cette version et on dépondère en multipliant par la pondération

[41] Les coefficients marginaux du tableau 4 sont une simplification des calculs : chaque élément est égal au vecteur calibré non pondéré divisé par la racine carrée de la norme et multiplié par la racine carrée de l'effectif. Par exemple l'élément "droite" du coefficient -3,147 est égal à -0,066 (affiché -0,07) élément correspondant de V4CNP, divisé par la racine carrée de la norme 0,2258 soit 0,4752 multiplié par la racine carrée de l'effectif total 512 soit 22,63. Du fait de la multiplication terme à terme, on retrouve la formule générale donnée ici.

Annexe au chapitre 5 : algorithme du maximum de vraisemblance

Ici encore nous proposons pour le lecteur qui veut ouvrir la "boite noire" de l'algorithme qui permet de calculer les éléments d'une régression logistique une présentation qui lui permettra d'en comprendre la logique. L'algorithme utilisé est dit du *maximum de vraisemblance*. Pour en comprendre la logique nous commencerons par un apologue.

Chicago

Nous sommes à Chicago dans un salon de jeu où l'on parie sur la sortie d'un six aux dés. Au bout de 83 coups, le six est sorti 19 fois. Est-ce suspect aux yeux de la police des jeux ? Pour en avoir le cœur net, les inspecteurs Neyman et Pearson appliquent le test du khi-deux avec une proportion théorique de $1/6^{e}$ pour le six et de $5/6^{e}$ pour les autres cas. Cela donne un khi-deux à un degré de liberté de 2,3 inférieur au seuil critique à 10% qui est de 2,7. Donc pas de dé pipé, pas d'infraction. Pourtant le commissaire Fisher se pose une question : est-ce que les données que nous avons sous les yeux ne nous conduisent pas à supposer une proportion de sortie du six différente de $1/6^{e}$ et plus vraisemblable? Pour cela il propose d'utiliser la méthode mise au point par son homonyme statisticien[42].

Dans le cas présent, le dé est peut-être truqué et la proportion de $1/5^{e}$ ne serait-elle pas plus vraisemblable que celle de $1/6^{e}$? Pour répondre à cette question on compare les probabilités de la situation observée selon les deux hypothèses.

La probabilité de 19 sorties d'un six au jeu de dé qui aurait une probabilité inconnue θ sur 83 tirages est de $\mathbf{k}\ \theta^{19}\ (1 - \theta)^{83-19}$ où **k** est le nombre de manières différentes d'avoir 19 évènements six dans 83 tirages soit le nombre 83! / 19! (83-19)! (c'est à dire le nombre de pascal $({}_{19}{}^{83})$.)

1) hypothèse $1/6^{e}$ notée H6 .La probabilité de l'éventualité observée dans cette hypothèse est de $\mathbf{k}\ (1/6)^{19}\ (1 - 1/6)^{83-19} = \mathbf{k}\ 1.4045\ 10^{-20}$

2) hypothèse $1/5^{e}$ notée H5. La probabilité de l'éventualité observée dans cette hypothèse est de $\mathbf{k}\ (1/5)^{19}\ (1 - 1/5)^{83-19} = \mathbf{k}\ 3.2910\ 10^{-20}$

Le rapport des probabilités de H5 par rapport à H6 est :
$\mathbf{k}\ 3.2910\ 10^{-20}$ / $\mathbf{k}\ 1.4045\ 10^{-20}$ soit 2.3 et l'on dit donc que H5 est plus de deux fois vraisemblable que H6 puisque les rapports des probabilités sont dans ce rapport. On préférera donc H5 à H6 mais une autre hypothèse sera peut-être encore plus vraisemblable. Notons bien qu'à l'inverse de ce qui se fait classiquement où l'on teste des données par rapport à une hypothèse, ici on fait varier les hypothèses pour un jeu de données fixe.

Le rapport des probabilités prises deux à deux est appelé *rapport de vraisemblance*. Par convention on prend comme dénominateur le cas le plus vraisemblable qui a pour paramètre θ_{exact} et pour probabilité $\mathbf{k}\ \theta_{exact}{}^{19}\ (1 - \theta_{exact})^{83-19}$. La probabilité individuelle du cas observé est fonction de θ qui est le paramètre recherché et égale à $\mathbf{k}\ \theta^{19}\ (1 - \theta)^{83-19}$. Pour trouver la valeur de θ qui maximise le

[42] Fisher, R. A. *Contributions to Mathematical Statistics*. New York Wiley, 1950

rapport à la valeur 1 qui sera obtenue quand $\theta = \theta_{exact}$ on a à maximiser le rapport suivant :

$$\mathbf{k}\ \theta^{19}\ (1 - \theta)^{83-19}\ /\ \mathbf{k}\ \theta_{exact}^{19}\ (1 - \theta_{exact})^{83-19}$$

ce rapport se simplifie par **k** et comme θ_{exact} est fixé (bien qu'inconnu), le rapport peut s'écrire :

$$(1 / (\theta_{exact}^{19}\ (1 - \theta_{exact})^{83-19}))x\ (\theta^{19}\ (1 - \theta)^{83-19})$$

Le premier terme est constant et pour maximiser le rapport, il suffit de maximiser le deuxième terme : $\theta^{19}\ (1 - \theta)^{83-19}$. Cette quantité est appelée *vraisemblance* et en général nommée L (comme *likelihood*, vraisemblance)

$L = \theta^{19}\ (1 - \theta)^{83-19}$

On notera que ce produit est égal au produit de tous les coups joués : les 19 de probabilité θ et les (83-19) de probabilité $(1-\theta)$. Ces évènements étant indépendants, la probabilité de l'ensemble est égal au produit des probabilité individuelles $\theta\ \theta\ \theta\ \theta\ \theta$ [19 fois] $(1-\theta)\ (1-\theta)\ (1-\theta)$ [83-19] fois. Cette situation est générale, la vraisemblance est toujours le produit des probabilités de chacun des cas observés.

Pour deux raisons allant dans le même sens on ne cherche pas à maximiser cette expression mais son logarithme naturel qui rend maximum θ au même moment :

1) parce que le calcul est plus simple : $\log(L) = 19 \log(\theta) + (83-19) \log (1- \theta)$ qui est une expression simple à maximiser. Cette première raison est traditionnelle mais perd de sa valeur avec l'informatique.

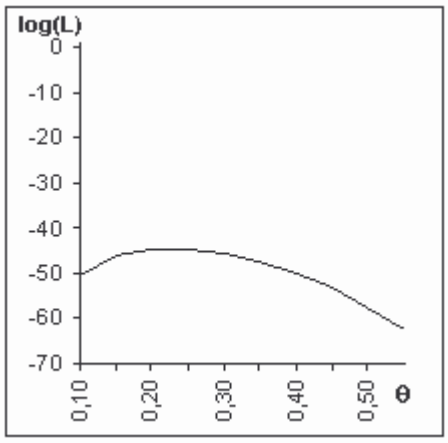
2) log (L), expression appelée la log-vraisemblance a d'autres propriétés intéressantes en rapport avec ce qu'on appelle information ou entropie. A un facteur -2 près, la log-vraisemblance est proche du khi-deux.

On peut maximiser directement log(L). Si un essai fait augmenter log(L), nous sommes automatiquement dans la bonne direction, si nous montons puis que la tendance s'inverse c'est que nous sommes passés par le maximum. Cette propriété est vraie même quand plusieurs paramètres sont à explorer (sauf minimum local, il faut donc prendre au départ des valeurs proches du résultat final).

Soit plusieurs valeurs de θ (croissantes par pas de 0,05) et ce qu'elles donnent pour log (L) :

θ	log(L)
0,10	-50,5
0,15	-46,4
0,20	**-44,9**
0,25	**-44,8**
0,30	**-45,7**
0,35	-47,5
0,40	-50,1

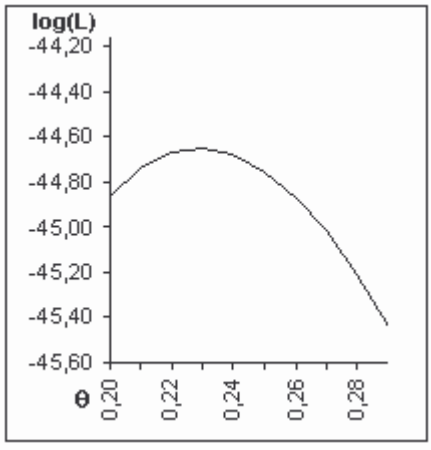
On voit que le maximum se situe pour des valeurs de θ comprises entre 0,20 et 0,30 (cf figure 19)



Il suffit de refaire l'expérience précédente en prenant un incrément plus petit (pas de 0,01 pour θ)et une précision plus grande pour le résultat:

θ	log(L)
0,20	-44,86
0,21	-44,74
0,22	**-44,67**
0,23	**-44,65**
0,24	**-44,68**
0,25	-44,75 (inutile de continuer)

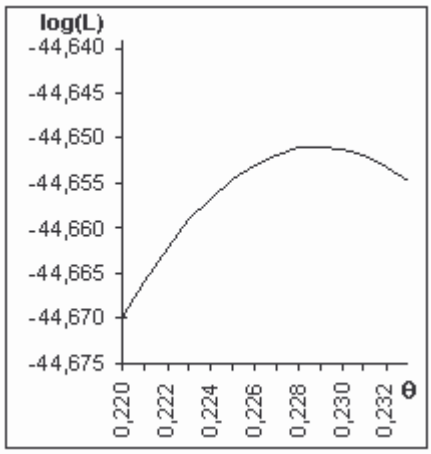
Le maximum correspond à des valeurs situées entre 0,22 et 0,24 (figure 20).



Recommençons le processus (pas de 0,005):

θ	log(L)
0,220	-44,670
0,225	-44,655
0,230	**-44,651**
0,235	-44,660
0,240	-44,679

Le maximum correspond à des valeurs situées entre 0,225 et 0,235 (figure 21)



Prenons 0,23 comme valeur a deux décimales : l'algorithme du maximum de vraisemblance nous a permis de trouver la valeur la plus vraisemblable. Evidemment, pour ce cas il existe une méthode plus simple, la moyenne, puisque cette valeur est simplement égale à 19/83 mais l'algorithme est utilisable quand ce calcul direct n'est pas possible ce qui arrive quand on a plusieurs paramètres à estimer et que leur liaison n'est pas linéaire. C'est ce que nous allons faire maintenant avec la recherche des paramètres d'une régression logistique.

Application

Reprenons les données du début de ce chapitre où l'on veut expliquer le choix du niveau style éducatif en fonction du sexe et du niveau d'étude (données du tableau 11).

Le calcul est fait de la manière suivante : on a ici trois paramètres à trouver, celui de l'effet féminin (le sexe masculin est de référence), l'effet du niveau supérieur (le niveau inférieur est de référence) et la situation de référence (masculin, niveau inférieur). Pour initier l'algorithme, on part des valeurs données par l'analyse tabulaire dont on a constaté qu'elles étaient toujours proches de celles données par la régression logistique. Ces valeurs sont les suivantes :

Situation de référence masc., inf. (observée) : 13,3%

Effet féminin : 5,1%

Effet Niveau Supérieur d'études : 6,4%

Puisque nous passons en régression logistique, il faut passer aux chances (pour la références) et aux rapports de chances (Odds Ratios) pour les deux effets.

Chances de la référence : la proportion est p = 0,133 et les chances sont p / (1-p) = 0,1534

Pour les deux effets, il s'agit du rapport des chances de la situation avec effet féminin, aux chances de la référence précédemment calculée.

La proportion dans le cas féminin est égal à la référence 0,133 + l'effet féminin 0,051 = 0,184. Les chances correspondantes sont 0,184 / (1 – 0,184) = 0,2255

L'OR correspondant = 0,2255 / 0,1534 = 1,4700

Le même calcul pour l'OR de niveau supérieur donne 1,5991.

L'algorithme consiste à calculer la vraisemblance correspondante aux données puis à faire varier un par un les paramètres : si la vraisemblance augmente, il faut

continuer à faire varier les paramètres, si elle ne bouge plus car arrivée à son maximum, les paramètres sont maintenant les bons.

Comme le tableau des données (tableau 11) a huit cases, nous n'avons que 8 situations élémentaires, chacune devant être répétée autant de fois que l'effectif de la case correspondante. Nous allons utiliser à cette fin un tableur :

ChRef 0,15	OR 1,47	OR 1,60					
Nouv Style	Fémi	Niveau Sup	Effectif	Produits ChxOR	Prop.	Ln (Prop)	Log x Eff
1	1	1	31	0,3528	0,261	-1,3440	-41,665
1	1	0	28	0,2205	0,181	-1,7111	-47,911
1	0	1	23	0,2400	0,194	-1,6422	-37,771
1	0	0	18	0,1500	0,130	-2,0369	-36,664
0	1	1	94		0,739	-0,3022	-28,405
0	1	0	115		0,819	-0,1993	-22,915
0	0	1	86		0,806	-0,2151	-18,500
0	0	0	117		0,870	-0,1398	-16,352
		Total	512			LogVrais=	-250,183

Tableau 18 algorithme : étape initiale

En tête on trouve les 3 paramètres avec des valeurs arrondies. Ce sont les chances de la référence et les OR des deux effets féminin et niveau sup.

Chaque ligne correspond à une ligne du tableau des données : la première ligne correspond à la case d'effectif 31 : féminin, niveau supérieur ayant choisi le nouveau style éducatif (les 1 et les 0 correspondent au choix ou au non-choix).

La colonne "produit" correspond au produit des chances de la référence avec les effets présents. Pour la première ligne 0,15 x 1,47 x 1,60 = 0,3528.

A partir de la formule "p issue des chances" on tire la proportion estimée = 0,3528 / 1,3528 = 0,2608.

On fait de même pour les 4 premières lignes qui correspondent au choix du nouveau style. Pour les 4 lignes suivantes où ce choix n'est pas fait, la proportion est le complément à 1 de la ligne correspondante.

Féminin – Niveau sup. Nouveau style p = 0,2608

Féminin – Niveau sup. Ancien style p = 1- 0,2606 = 0,7392

Comme la vraisemblance est le produit de toutes les proportions (31 fois la première x 28 fois la seconde etc.), la log-vraisemblance est calculée plus facilement en prenant le logarithme naturel de chaque proportion et en le multipliant par l'effectif. La somme de toutes les lignes donne la log-vraisemblance globale. Elle est négative ce qui veut dire qu'une croissance fera diminuer la valeur absolue.

Avec les paramètres initiaux, la log-vraisemblance est égale à -250,183.

Nous allons faire varier les paramètres initiaux et examiner si la log-vraisemblance monte ou diminue. Si elle monte on continue si elle baisse on revient en arrière.

La situation est comparable à celle d'un randonneur qui monterait sur un dôme volcanique régulier dans le brouillard. Tant qu'il monte, il est dans la bonne direction, s'il descend, c'est qu'il est sur un flan et il doit changer de direction (équivalent de changer de paramètre). Quand dans toutes les directions, il ne monte plus, c'est qu'il est au sommet.

Nous allons faire de même (mais le déroulement complet de l'algorithme doit être programmé). Commençons par faire évoluer le paramètre du niveau supérieur :

OR NivSup		Log-Vraisemblance	
Début	1,60	-250,183	
-0,01	1,59	-250,185	Décroissance Stop
+0,01	1,61	-250,182	Croissance
+0,01	1,62	-250,182	Palier : Stop. Changement de paramètre
Chances référence			
Début	0,150	-250,182	
-0,001	0,149	-250,200	Décroissance Stop
+0,001	0,151	-250,168	Croissance
+0,001	0,152	-250,158	Croissance
+0,001	0,153	-250,151	Croissance
+0,001	0,154	-250,147	Croissance
+0,001	0,155	-250,147	Palier : Stop
Changement de paramètre			
OR Féminin		Log-Vraisemblance	
Début	1,47	-250,147	
-0,01	1,46	-250,140	Croissance
-0,01	1,45	-250,135	Croissance
-0,01	1,44	-250,131	Croissance
-0,01	1,43	-250,131	Palier : Stop

Bien que les trois paramètres aient été pris en compte, l'algorithme n'est pas terminé car il faut reprendre le premier paramètre avec les valeurs des autres jusqu'à fixité des 3, puis augmenter la précision en prenant des incréments plus faibles.

A la fin on a la situation suivante qui correspond à ce que donnent les programmes.

ChRef	OR	OR					
0,166	1,389	1,506					
Nouv Style	Fémi	Niveau Sup	Effectif	Produits ChxOR	Prop.	Ln (Prop)	Log x Eff
1	1	1	31	0,347244	0,258	-1,3558	-42,029
1	1	0	28	0,230574	0,187	-1,6747	-46,891
1	0	1	23	0,249996	0,200	-1,6095	-37,017
1	0	0	18	0,166000	0,142	-1,9493	-35,088
0	1	1	94		0,742	-0,2981	-28,018
0	1	0	115		0,813	-0,2075	-23,860
0	0	1	86		0,800	-0,2231	-19,190
0	0	0	117		0,858	-0,1536	-17,969
		Total	512			LogVrais=	-250,062

Tableau 19 algorithme : état final

A partir des valeurs de ces paramètres, on peut retrouver par les formules vues les proportions et les effets.

L'originalité de cet algorithme du maximum de vraisemblance réside dans le fait que les estimations ne sont pas faites directement à partir des données mais en trouvant les valeurs les plus compatibles avec les données : l'algorithme fait intervenir les données à chaque étape.

Chapitre 6 : comment faire ?

Le but d'un ouvrage comme celui-ci est d'inciter à l'action, à dépouiller des enquêtes déjà faites[43]. Je voudrais pour conclure résumer la suite des opérations pour l'analyse des correspondances et donner quelques indications complémentaires pour l'analyse "toutes choses égales par ailleurs"

I Etapes préparatoires

La première étape est le rassemblement de la documentation nécessaire sur l'enquête traitée : questionnaire de base, documents donnant les raisons de la rélisation de l'enquête, hypothèses posées à ce moment, publications déjà faites sur les données s'il s'agit d'une analyse secondaire. Par contre, il est trop tôt pour lire la littérature sur la question traitée elle-même : cette étape ne sera fructueuse qu'après de premiers contacts avec l'enquête.

Les données de l'enquête ayant été dupliquées, on utilisera un logiciel qui ne modifie jamais les données de base mais qui crée des fichiers auxiliaires après chaque modification. On constituera alors un tri à plat de toutes les variables de l'enquête (comptage de chacune des modalités de toutes les questions de l'enquête) et l'on reportera sur un questionnaire l'effectif de chaque réponse afin de voir comment chaque question a été reçue. On ouvrira un journal de l'enquête pour y porter les résultats qui vont maintenant être obtenus.

Une première familiarisation avec l'enquête se fera en croisant un nombre limités de questions correspondant à des hypothèses préalables ou à des questions simples que l'on peut se poser en utilisant des modalités explicatives classiques (sexe, âge, niveau d'étude, catégorie sociale, politique, religion). Cette première étape permettra d'envisager le recodage de certaines modalités trop dispersées : la gestion des recodages faits doit en priorité figurer dans le journal de bord de l'enquête. En cas d'hésitation, il faut prévoir deux recodages d'une même question : un encore assez détaillé et un autre plus énergique. Pour une modalité jugée stratégique, une analyse des correspondances de plusieurs questions permet de faire le choix des regroupements. En général le recodage se fait par proximités jugées à partir de l'intitulé lui-même et, en cas d'hésitation, par un tri croisé qui permet de voir comment se sont opérées les associations de modalités.

On préparera avec soin les questions qui sont au cœur de l'enquête, les variables d'intérêt, qui vont servir de base pour l'étape suivante.

II Analyse globale

Par le biais d'une variable d'intérêt, on déterminera les questions qui serviront pour l'analyse globale de l'enquête. Cette liste est donnée en prenant les questions

---

[43] On trouvera des exemples d'application des présentes méthodes dans les deux articles suivants : Philippe Cibois, "La bonne volonté scolaire. Expliquer la carrière scolaire d'élèves issus de l'immigration" *in* Philippe Blanchard et Thomas Ribémont (Dir.), *Méthodes et outils des sciences sociales. Innovation et renouvellement*, Paris, L'Harmattan, 2002, coll. "Cahiers politiques", p.111-126 ; et Philippe Cibois, "Technique d'analyse des données d'enquête. Exemple avec l'insémination artificielle et anonymat du donneur, ou comment éclairer un débat de société", *RSI Recherche en Soins Infirmiers*, n°85, juin 2006, p.22-35.

les plus liées globalement (PEM global) à la variable d'intérêt. On prendra suffisamment de questions pour que le total des modalités de ces questions atteignent environ 200 modalités. Dans le logiciel, il sera alors temps de donner un nom aux questions, mais aussi aux modalités retenues, mais l'opération pourra se faire en plusieurs étapes en prenant d'abord les questions les plus liées à la variable d'intérêt.

Si le premier plan factoriel obtenu à une allure de "comète", on mettra en supplémentaires les quelques points qui ont créé le ou les premiers facteurs. En quelques opérations on arrivera à procéder à la "fission" du cœur de la comète et à arriver ainsi à un premier plan factoriel où les points sont suffisamment répartis.

Comme 200 points ne sont pas simultanément lisibles (en particulier quand les points proches n'ont pas encore été désintriqués) et afin de s'assurer de la contribution suffisante des points affichés, on diminuera progressivement le niveau de contribution des points affichés par doublements approximatifs successifs (1 pour mille, puis 2, puis 5, 10, 20, 50, etc.). On se souviendra que la contribution moyenne est obtenue en divisant 1000, la base du calcul, par le nombre de modalités *actives*. Avec 200 modalités, la moyenne 5 est vite atteinte. Pour les supplémentaires qui par définition sont moins contributives que les actives (puisque issue d'autres dimensions), le seuil peut être différent et est souvent plus faible que pour les actives.

En modifiant la variable d'intérêt on pourra, si l'on dispose du temps nécessaire, explorer de façon analogue plusieurs dimensions de l'enquête : c'est à ce moment que l'examen de la littérature sur le sujet deviendra fructueuse car elle permettra de confronter les résultats de l'enquête avec ce qui était su auparavant sur le domaine de recherche.

C'est également à ce moment que la rédaction du commentaire sur les plans factoriels obtenus sera utile : ils peuvent être mis dans le journal en vue d'une réexploitation ultérieure. Ne pas rédiger au moment de l'analyse conduit bien souvent à devoir recommencer le travail fait, ou à perdre beaucoup de résultats. Inversement, l'examen ultérieur des commentaires faits à chaud permet souvent de se rendre compte qu'une analyse est trop entrée dans le détail et qu'il faudra finalement n'en retenir que les aspects les plus importants. En tout état de cause, le principe (mais il est difficile à suivre) est de toujours rédiger au moment où l'on fait des analyses.

III Retour aux données

Comme l'analyse des correspondances propose des types idéaux, ce qui est sa qualité, il faut, pour éviter que cette qualité ne se transforme en piège, retourner aux données par le biais de comptages de nouvelles variables créées à partir des types-idéaux (variables idéaltypiques) ou en explorant sur le plan factoriel, les graphes de liens de PEM local associés à une modalité précise.

La création de variables idéaltypiques permet de se faire une idée statistique du type : on prendra souvent comme faisant partie du type les individus qui ont au moins la moitié des modalités constitutives du type idéal.

Pour l'examen des graphes de PEM, on acceptera les graphes de faible intensité (PEM <10%) et l'on se souviendra que les PEM les plus élevés (PEM

>50%) sont souvent révélateurs d'une redondance, de l'appartenance des deux modalités à un même univers.

IV Les variables explicatives

Ayant déterminé des types suffisamment importants du point de vue de leurs effectifs, on pourra les mettre en relation de plusieurs façons avec des modalités explicatives.

Une première manière consiste à mettre en supplémentaires de l'analyse globale les modalités explicatives. Une manière complémentaire est ensuite de regarder les liens entre modalités explicatives, en les mettant en actives, et à projeter en supplémentaires, les variables idéaltypiques obtenues dans l'étape antérieure. Ceci permettra de préparer les analyses ultérieures "toutes choses égales par ailleurs".

Ces analyses sont actuellement privilégiées dans les publications car leur présentation peut être brève, elles semblent faciles à comprendre et ayant un fort pouvoir explicatif.

Il est exact que la présentation d'une analyse des correspondances suppose, pour être convenablement comprise, un espace rédactionnel suffisant. On peut faire l'hypothèse que l'analyse des correspondances joue, pour une analyse donnée, le rôle d'un échafaudage qui a permis de construire une démarche d'exploration et qu'il peut être démonté après usage. De ce fait les seuls résultats présentés sont les types de répondants bien attestés, vérifiés par des comptages et qui sont ensuite "expliqués" par une régression logistique.

V Les régressions multiples

Le problème se complique du fait qu'il existe plusieurs types de régressions multiples : j'en ai présenté deux : l'analyse tabulaire, simple dans son principe et qui donne des résultats proches de la régression logistique, la méthode la plus utilisée mais non la plus simple. J'ai fait allusion à la régression linéaire sur les mêmes données qui donne aussi des résultats très proches et il existe enfin des variantes de la régression logistique qui permettent par exemple d'éliminer la nécessité d'une modalité de référence[44].

Dans une première étape de régressions multiples, l'analyse tabulaire présente beaucoup d'intérêt : comme ce n'est pas un modèle mais une observation des données dans toute leur complexité, elle permet de se rendre compte de plusieurs phénomènes :

1) elle permet de vérifier si les données sont suffisantes pour prendre en compte en même temps beaucoup de questions. Par exemple dans l'analyse présentée page 106 et suivantes, les données de base sont données à la page suivante : il s'agit d'un ordre lexicographique où chaque ligne est un précroisement d'une modalité de sexe, d'engagement religieux, d'âge scolaire, de proximité

[44] Henri Leridon et Laurent Toulemon, *Démographie. Approche statistique et dynamique des populaitions*, Paris, Economica, 1997, p. 252. Exemple d'utilisation dans le numéro 415 (septembre 2005) de *Population & Sociétés*.

politique. Chaque ligne a un effectif et correspond à un pourcentage de choix de la variable à expliquer, ici le style nouveau d'éducation. Il s'agit en sorte d'un tableau croisé à deux colonnes où par exemple pour la première ligne, les 26 individus de sexe masculin, de mère pratiquante, d'enfant à l'heure scolairement, de mère de droite se répartissent en 7,7% de nouveau style éducatif (et 100 – 7,7 = 92,3 qui ne choisissent pas ce style). On voit que le fait que l'on fasse intervenir 4 questions à 2 modalités (sexe, religion et âge scolaire) et à 4 pour la politique, émiette déjà beaucoup les données. En fait toutes les données ne sont pas présentes, il manque 16 individus car sur les 32 lignes possibles (2 x 2 x 2 x 4), seuls 22 sont complets, c'est-à-dire ayant un effectif non nul pour le style nouveau et son complément, ce qui permet de calculer un pourcentage différent de zéro ou de 100 (qui seraient très incertains). On constate déjà que les effectifs de chaque ligne sont faibles et que la situation de référence (toutes les modalités marquées du R correspondant à la référence) ne regroupe que 32 individus.

	Sexe	Rel	AgeS	Pol	Eff.	%
01	Masc R	Rel		Al'h R	Droi	26 7.7
02	Masc R	Rel		Al'h R	NiNi R	34 11.8
03	Masc R	Rel		Al'h R	Gauc	25 8.0
04	Masc R	Rel		Al'h R	Ecol	8 12.5
05	Masc R	NRel R	EnRe		Gauc	4 50.0
06	Masc R	NRel R	EnRe		Ecol	2 50.0
07	Masc R	NRel R	Al'h R	Droi		10 10.0
08	Masc R	NRel R	Al'h R	NiNi R		32 25.0
09	Masc R	NRel R	Al'h R	Gauc		63 15.9
10	Masc R	NRel R	Al'h R	Ecol		28 35.7
11	Fémi	Rel	EnRe	Droi		3 33.3
12	Fémi	Rel	EnRe	Gauc		9 33.3
13	Fémi	Rel	EnRe	Ecol		5 20.0
14	Fémi	Rel	Al'h R	Droi		19 5.3
15	Fémi	Rel	Al'h R	NiNi R		41 19.5
16	Fémi	Rel	Al'h R	Gauc		21 19.0
17	Fémi	Rel	Al'h R	Ecol		17 29.4
18	Fémi	NRel R	EnRe	NiNi R		2 50.0
19	Fémi	NRel R	Al'h R	Droi		10 10.0
20	Fémi	NRel R	Al'h R	NiNi R		48 18.8
21	Fémi	NRel R	Al'h R	Gauc		67 28.4
22	Fémi	NRel R	Al'h R	Ecol		22 27.3

Population concernée= 496 soit 96.9% de l'effectif total

On peut évidemment construire un modèle de régression logistique avec davantage de questions et de modalités et il pourra donner des résultats significatifs mais les données de base nous manifestent que ce serait bien risqué car les observations des situations correspondraient à peu de lignes complètes.

Ce que permet d'observer aussi l'analyse tabulaire, c'est la présence ou non d'interactions qui vont à l'encontre du postulat du modèle "toutes choses égales par ailleurs".

Prenons par exemple le cas de l'effet "gauche" dont la régression logistique nous dit qu'il est nul et non significatif. L'analyse tabulaire nous en propose la lecture suivante :

```
Effet Gauch 4  sous-effets (s-e)
Sous-population
                    s-e Eff Tot
A Masc Rel  Al'h  -3.8  59 331
B Masc NRel Al'h  -9.1  95 331
C Fémi Rel  Al'h  -0.5  62 331
D Fémi NRel Al'h   9.6 115 331
moyenne pondérée des sous-effets =  0.0
Ecart type pondéré en pourcentage =  7.7
*** attention
    l'écart-type est plus grand que la valeur absolue de la
moyenne
    l'effet moyen est peu fiable du fait des interactions
```

En analyse tabulaire, la moyenne des sous-effets est effectivement nulle mais ce résultat vient de résultats divergents avec des sous effets tantôt positifs (ligne D), tantôt négatifs (lignes A, B et C). Ces sous-effets se déduisent des données de base de la manière suivante : l'effet A oppose l'orientation politique à gauche par rapport à la référence (ni gauche ni droite) pour la sous-population "masculin, participation religieuse, à l'heure scolairement". Ceci correspond aux lignes 1 à 4 des données de base pour chacune des options politiques à l'intérieur desquelles l'opposition entre "gauche" et "ni gauche ni droite" correspond à la différence entre les lignes 3 et 2. Pour la gauche de ligne 3, la proportion de nouveau style est de 8,0 % ; pour l'orientation ni gauche ni droite de la ligne 2, cette même proportion est de 11,8%, le sous-effet pour les lignes 2 et 3 est de 8,0 – 11,8 = - 3,8.

On voit que dans la population "féminin, sans religion à l'heure" (lignes 20 et 21), l'effet D est lui positif : l'effet de gauche est positif ou négatif selon le contexte mais non nul. Nous sommes en présence d'interactions.

On constatera aussi que seulement 4 effets sont calculés sur les 8 possibles car le fait d'être en retard scolairement est trop peu représenté dans les données de base (cf. les lignes 5 et 12 de faible effectif et qui n'ont pas de situation de référence observable).

En conclusion c'est donc au vu de l'analyse tabulaire que l'on choisira un modèle de régression logistique qui puisse conduire, par des effectifs suffisants des données de base, à des résultats fiables.

VI Annexe

On trouvera ci-dessous le détail complet des calculs de l'analyse tabulaire pour l'effet gauche avec le détail des tableaux croisés correspondants (Nouv. désigne le nouveau style éducatif, *reste* l'ancien).

```
Effet Gauch 4  sous-effets (s-e)
Sous-population
                   s-e  Eff  Tot
 Masc Rel  Al'h  -3.8   59  331
       Nouv Reste Tot   Nouv Reste
NiNi      4   30   34   11.8 88.2 100
Gauc      2   23   25    8.0 92.0 100
  Tot     6   53   59   10.2 89.8 100
Sous-population
                   s-e  Eff  Tot
 Masc NRel Al'h  -9.1   95  331
       Nouv Reste Tot   Nouv Reste
NiNi      8   24   32   25.0 75.0 100
Gauc     10   53   63   15.9 84.1 100
  Tot    18   77   95   18.9 81.1 100
Sous-population
                   s-e  Eff  Tot
 Fémi Rel  Al'h  -0.5   62  331
       Nouv Reste Tot   Nouv Reste
NiNi      8   33   41   19.5 80.5 100
Gauc      4   17   21   19.0 81.0 100
  Tot    12   50   62   19.4 80.6 100
Sous-population
                   s-e  Eff  Tot
 Fémi NRel Al'h   9.6  115  331
       Nouv Reste Tot   Nouv Reste
NiNi      9   39   48   18.8 81.2 100
Gauc     19   48   67   28.4 71.6 100
  Tot    28   87  115   24.3 75.7 100
moyenne pondérée des sous-effets = 0.0
```

Bibliographie

Benzécri, Jean-Paul, et *al.*; *L'analyse des données*, Paris, Dunod, 1973, vol.1 : *La Taxinomie*, vol. 2 : *Correspondances.* Constitue la référence pour l'analyse des correspondances mais se situe, pour la partie théorique à un niveau élevé de compétences mathématiques.

Cibois, Philippe, "Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence", *Bulletin de méthodologie sociologique*, 1993, n°40, p.43-63.

Cibois, Philippe, "Les pièges de l'analyse des correspondances", *Histoire & Mesure*, 12 (3/4), 1997, pp. 299-320.

Cibois, Philippe, "Modèle linéaire contre modèle logistique en régression sur données qualitatives", *Bulletin de méthodologie sociologique*, n°64, 1999, p.5-24. Présentation de l'analyse tabulaire.

Escofier, Brigitte, Pagès, Jérôme, *Analyses factorielle simples et multiples*, Paris, Dunod, 1988. Présentation géométrique mais relativement accessible.

Lebaron, Frédéric, *L'enquête quantitative en sciences sociales*, Paris, Dunod, 2006. Recueil et analyse des données sont expliqués dans le cadre de l'analyse géométrique des données. Des études de cas permettent de voir comment la démarche est mise en œuvre.

Menard, Scott, *Applied Logistic Regression Analysis*, Thousand Oaks CA, Sage University Paper series on Quantitiative Applications in the Social Sciences 106, 1995. Beaucoup des présentations de méthodes de cette collection, partent d'exemples et sont souvent plus compréhensibles que celles de certains auteurs de langue française (qui parlent surtout le langage mathématique). Cette présentation, associée à celle de Pampel dans la même collection permet de comprendre la régression logistique.

Pampel, Fred C., *Logistic Regression. A primer*, Thousand Oaks CA, Sage University Paper series on Quantitiative Applications in the Social Sciences 132, 2000

Rouanet, Henry, Le Roux, Brigitte, *Analyse des données multidimensionnelles*, Paris, Dunod, 1993. Présentation de l'analyse factorielle et des méthodes dérivées dans une approche "géométrico-formelle".

Logiciel

On trouvera l'accès libre et gratuit au logiciel Trideux à partir du site de l'auteur (utiliser un moteur de recherche à partir de "prénom + nom")

Table des matières

Introduction

Chapitre 1 : repérer les questions pertinentes

I Première étape : les préalables - II Sélectionner les questions pertinentes

Chapitre 2. L'analyse factorielle des correspondances

I Décomposition des écarts à l'indépendance - II Contributions des modalités, des tableaux. - III Procédure de codage en tableau de Burt - IV Modalités supplémentaires - V Résumé - Annexe

Chapitre 3 : rechercher des types de répondants avec l'analyse des correspondances

I Première analyse : la queue de comète - II Analyse finale - III Type : bon niveau scolaire - IV Type : difficultés scolaires - V Type : style éducatif ancien - VI Type : nouveau style éducatif - VII Retour aux hypothèses de départ - VIII Education nouvelle et société - IX Retour à l'analyse locale - X Retour à la méthode - XI Construire une nouvelle variable d'un type

Chapitre 4 : les figures de l'analyse des correspondances

I Parabole de l'effet Guttman - II Effets des faibles effectifs - III Des types idéaux -

Chapitre 5 : les techniques d'analyse « toutes choses égales par ailleurs »

I Analyse tabulaire multivariée - II La régression multiple – III Chances et rapport des chances - IV Equation de la régression logistique - V Utilisation de la régression logistique - VI Variations - Annexe

Chapitre 6 : comment faire ?

I Etapes préparatoires - II Analyse globale - III Retour aux données - IV Les variables explicatives - V Les régressions multiples - VI Annexe.

Bibliographie