Traitement statistique d'un gros fichier de données. Enchaînement de quelques méthodes. Deuxième partie

Joseph SAINT PIERRE,
Centre Interuniversitaire
de Calcul de Toulouse
118, Route de Narbonne
31062 Toulouse Cedex 04
Tél: 05 61 36 60 21
Fax: 05 61 52 14 58
e-mail: Joseph.Saint-Pierre@cict.fr

Vendredi 26 Février 1999

Cet exposé comme le précédent est consacré à l'exploration statistique d'un même fichier de données. Le but de ces exposés, plus que de montrer des méthodes statistiques est de montrer une démarche possible en utilisant plusieurs types de méthodes.

Ce document correspond à la deuxième partie de l'exposé et est donc incomplet, le plan et la rédaction de cette partie sont aussi très informels. Éventuellement un document complet sera rédigé en intégrant les deux parties et en faisant une structuration. Il faut donc voir ce document comme un outil permettant de suivre l'exposé oral correspondant. De plus ces données font l'objet d'une étude plus poussée par une équipe de recherche à laquelle appartiennent les données.

Il est inutile de rappeler par écrit le contexte de l'étude, il suffit de reprendre le document du premier exposé pour se remettre dans le bain.

En premier lieu, il est sans doute intéressant de revenir sur certaines questions qui ont été posées lors du premier exposé. La question la plus gênante fut de savoir si le but de ces exposés était d'apprendre aux auditeurs à faire des statistiques. La réponse fut bien évidemment oui. Mais il s'agit de préciser certains points, il est hors de question de comprendre en quelques heures toutes les méthodes statistiques présentées. Le but original de ces conférences n'est surtout pas de se substituer à des cours de statistiques traditionnels, mais plutôt de sensibiliser l'auditoire à l'intérêt des statistiques. Par ailleurs il s'agit de conférences s'adressant à un public assez hétérogène et il est difficile de connaître le point de départ de chacun des auditeurs, l'organisation de cours avec séance de TP sur ordinateurs se fera, à partir de liste d'inscription permettant de mieux cibler le contenu.

Il paraît bon, aussi de signaler que l'année dernière, dans le cadre des conférences pluridisciplinaires de l'IED, plusieurs exposés ont été consacrés à des méthodes statistiques de manière plus précise. Le 15 Mai 1998, notamment, il y a eu un exposé sur les méthodes d'analyses factorielles (ACP, AF, AFC etc.). Dans le cadre des exposés du 15 Janvier 1999 et d'aujourd'hui, il est largement fait appel à ces techniques mais il est hors de question de refaire un exposé sur celles-ci. Les personnes intéressées par une présentation des analyses factorielles peuvent récupérer le texte de la conférence du 15 Mai 1998 à l'IED.

Une question a été posée dès le début pour demander comment avait été notée la modalité «non réponse» dans le traitement quantitatif. L'option qui a été choisie fut de ne pas prendre en compte celles-ci et de ne faire les calculs de moyennes et de variances uniquement avec les données 1, 2, 3, 4 et 5. Ce qui était suggéré dans la question était de remplacer la «non réponse» par un 0. J'ai depuis le dernier exposé refait les calculs en prenant cette option, cela donne des changements qui pour certains mots sont assez nets, à titre d'exemple le mot mansuétude (16) passerait de la place 165 à la place 185 si on comptait 0 pour les 73 réponses manquantes, le mot décentration (7) passerait de la place 154 à la place 172 et enfin le mot ascèse passerait de la place 204 à la place 210 soit l'avant dernière. De par un jugement simple, on peut émettre l'idée que mansuétude et ascèse sont des mots frappés d'obsolescence, alors que décentration serait plutôt un néologisme, de ce point de vue là les raisons de la forte absence de réponse apparaissent comme diamétralement opposées. Dans une première lecture du questionnaire, le seul mot que je ne connaissais pas était le mot décentration et c'est le seul auquel je n'aurais pas mis de réponse mais, après enquête, le concept m'a paru important et ne pas mériter un 0. C'est peut être à partir de cette impression purement personnelle qu'il m'a paru préférable de ne pas mettre 0 pour les données manquantes. Une alternative aurait été d'éliminer de l'étude les mots ayant trop de réponses manquantes, mais cela ne faisait pas partie de mon travail.

Il y a eu, au moment de la présentation de l'AFC simple, une question sur l'interprétation de la proximité entre les modalités 2 et les modalités «non réponse» et le relatif éloignement entre les modalités 1 et les modalités «non réponse». Cette question pourrait presque être reliée à la question précédente, en effet de façon naïve on peut penser que les mots inconnus représentent des choses pas du tout importantes pour les répondants, si tel était le cas on devrait observer une ligne droite et les modalités «non réponse» devraient se trouver dans l'alignement des modalités 5, 4, 3, 2 et 1. L'AFC simple nous indique que cette linéarité n'est vraisemblablement pas pertinente pour décrire ces données. Ceci étant, une remarque faite lors de l'exposé précédent me paraît expliquer en partie ce décalage, il y a une différence évidente entre le mot et le concept (valeur) associé. Quant à savoir ce qu'ont pensé, des divers mots, les répondants il n'est pas possible de le savoir.

Par ailleurs j'ai reçu, par courrier électronique, une contribution très intéressante à l'analyse du dernier tableau présenté dans le document de l'exposé du 15 Janvier 1999, où était donnés les classement des 211 mots pour les quatre groupes définis par le sexe et la filière. Le contributeur s'est proposé l'objectif suivant: «Le but est de trouver les mots avec le plus grand contraste entre les groupes filles Mirail d'un coté et filles IUT puis entre les groupes garçons Mirail et garçons IUT.» Par un petit calcul élémentaire que je ne détaille pas on peut s'apercevoir que les mots qui sont les plus caractéristiques des filles de l'IUT par rapport aux filles du Mirail sont: dynamisme (39), joie (93), combativité (59), séduction (209), plaisir (36), loisir (11), ambition (166), travail (160), organisation (72), persévérance (4) et progrès (20). Les mots les plus caractéristiques des filles du Mirail par rapport à celles de l'IUT sont: tolérance (152), bonté (137), affection (19), authenticité (150), espérance (77), sensibilité (32) et patience (123). Les mots les plus caractéristiques des garçons de l'IUT par rapport à ceux du Mirail sont: plaisir (36), ambition (59), combativité (59), passion (174), sincérité (183), sécurité (162), indépendance (122), loisir (11), complicité (69), innovation (40), intelligence (66), dynamisme (39), intimité (71), joie (93), action (58) et séduction (209). Enfin les mots qui caractérisent le plus les garçons du Mirail par rapport à ceux de l'IUT sont: tolérance (152), autocontrôle (80), respect (170), loyauté (1), sagesse (177), intuition (168), curiosité (100), féminité (97), altruisme (176), douceur (82), espérance (77), égalité (65), bonté (137) et connaissance (81). L'interprétation qui m'était fournie avec cette comparaison était que les étudiants de l'IUT paraissaient plus ludiques que ceux du Mirail, cela m'a rappelé que le mot «jeu» était absent de la liste des 211.

Enfin une dernière question très importante lors de l'exposé était de savoir si il y avait des problèmes à faire de l'ACP en l'absence de la normalité des observations. Cette question a un rapport avec la démarche de cet exposé. Les variables correspondant à des échelles en 5 points ne sont pas qualitatives mais elles ne sont pas vraiment quantitatives, certainement pas continues et souvent très éloignées de la normalité. Nous avons à faire à du qualitatif ordonné et il y a même un doute sur l'ordre... Il n'y a pas de méthode d'analyse factorielle parfaitement adaptée à ce type de variables. L'ACP fait implicitement l'hypothèse que les variables sont normalement distribuées, en effet l'ACP utilise les corrélations ou les covariances et ces quantités là ont du sens pour des variables normalement distribuées. Mais nombreuses sont les situations où l'on utilise l'ACP sur des variables de ce style, c'est ce que j'ai fait dans le cadre de l'exposé précédent.

Lors de ce précédent exposé j'ai insisté sur certaines dualités d'approche et notamment celle concernant la vision quantitative et la vision qualitative. En présentant le premier tableau avec simultanément les fréquences de chacune des modalités et les moyennes et variance de chacune des 211 variables, je montrais la double approche simple sur ces variables. L'ACP a correspondu à un traitement pour tenter de résumer les 211 variables, en considérant qu'elles étaient quantitatives, si l'on considère que ces variables sont qualitatives une approche possible voisine de l'ACP est l'analyse factorielle des correspondances multiples ou AFCM. Fondamentalement, une AFCM est une ACP particulière, il s'agit d'une adaptation de l'ACP. L'AFCM n'existe pas, à ma connaissance, avec des rotations type «varimax», au moins dans les logiciels standard que j'utilise. En anglais, le mot «factorial» est réservé pour ce que l'on appelle parfois en français analyse en facteurs. L'analyse factorielle des correspondances s'appelle correspondence analysis et non, comme je l'ai parfois vu, factorial correspondence analysis. Voici ce qu'a fourni l'AFCM sur ce fichier de données. La représentation graphique est avec les deux premiers axes mais il faut considérer que le nombre de modalités est énorme, il y a 820 points cachés.

```
                             AXE 1  *  AXE 2
 AXE 2
       ---------------------- 58_ 94_ 70_146_-------------------------
      !                       10_     !                               !
      !                               !                               !
      !                               !                               !
      !                               !                               !
      !                              140_                             !
      31_                 93_102_161_ !      1874 53_ 51_0235         !
      !                          179_ !          119_                 !
      !              174_             !   52_062403040175    0185     !
  .6  +                          18730234203320531674     25_         !
      !  163_         99_ 59_195_ 48_13342034172_1094                 !
      !              134_184_131415730224134415941174                 !
      !   33_    101_    2052069414631734128418141414                 !
      !       84_168_162_1504136416941654190415931914 41_10542054      7_
      !           96_ 81_071408341724114413240313048408541444         !
      !  1382    1264 26_12540272157200742113040401840395             !
      !           71_182_116410520513002 4204_ 17_12450515026507850675 !
      !          0563036204531953052316 0_056507550115020505551485     !
  .0  +----------05021692007217530293 1831fam80655142507951125015502451905
      !          08841843115318231173 54_fam506951115131519250875 1175 50_
      !          1583172303541473sco2res1 74_04251365123504351545 01651285
     1082        00831632011308721 87_2031116_0174 98_148_1305115513550225
      57_    215_095316821202132202310731171_    10451695046504450605 1945
     0052    05030252129313730522211111912051    00920315168509951215 0485
     1162200_1523 79_136314530692  9_192_017110311795007521 1_202518551275
     169_196_20120842044216721002056201011811    18150525096518751885 2055
     216_ 28_ 82_129206420742092312310301 19_0521    0985 13_156504952105
 -.6  +  203_13832003147216020282099200711421 0421    211513250125 60_1785
     178_    068213520863035205310911060117611371    1445        02811535
     188_131212720412046219421 13_0471029117310751    0021        165_1095
      !      0892207_0722040118010851158111511 28_                    21_
      86_    08020812093318611232015102512041 1061    0081            !
      !  0582014317221902196117911131054107911111                     !
     141_071212221492118218211721062101110631193 10921    1691        !
      !      08910702160116710371194117810811189120711 07_           0351
      !  0332    0801    114114510571197112011011 1291               1221
-1.2  +      13621262    032200110901076107711631 44_1161        0431 !
      !          1702 11_154106911901150 11071        157_           1021
     1381    056102610641088204512021166113 91sco5209_15210581        !
     13610832213_00910862046114312101082120910131197_17010941---------+
    -1.2             -.6              .0               .6            1.2
                                                                   AXE 1
```

Il n'est pas possible de voir grand chose, on peut seulement repérer quelques amas de modalités semblables, dans la représentation le numéro de la variable est sur les 3 premiers chiffres le quatrième chiffre sert à donner la modalité, la donnée manquante se repère par le symbole souligné «_». On peut apercevoir que le premier axe oppose un petit peu les modalités 2 et 3 d'un coté et 1 et surtout 5 de l'autre alors que le deuxième axe oppose la modalité 4 aux autres modalités et surtout 1. Il est nécessaire pour lire ces résultats d'intégrer d'autres tableaux que je n'ai pas mis ici pour économiser de la place. Le but n'est pas de lire les résultats d'une AFCM.

Afin de présenter un résumé du tableau j'ai calculé les moyennes pondérées par les effectifs correspondants des modalités. Ainsi on peut mieux voir sur le graphique où se trouvent, plutôt, l'ensemble des modalités. Les modalités 5 se trouvent à droite, les modalités 4 en haut, les modalités 3 et 2 se trouvent à gauche, les 2 étant un peu plus bas et plus à gauche, les modalités 1 se trouvent en bas et les 6 au milieu.

```
                            AXE 1  *  AXE 2
AXE 2
      ----------------------IUTgCélC---------------------------------
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                       4IMP!                   !
     !                                           !                   !
     !                                           !                   !
 .2  +                                           !  IUT              !
     !                                           !                   !
     !            garçons                        !          res3     !
     !                                           !          IUTf     !
     !                                           !      4ème         !
     !                                parents    !                   !
     !                                           !                   !
     !                          fam7             !                    7_
     !                                           !                 Couple
     !  3ème                                RFam !                   !
 .0  +----------------3MOY-----------------------+--1ère-------------!
     !                                  Céli 6NR!fillesfam8         5IMP
     !                                           !                   !
     !                                    Mirail !              seul !
     !                                          Âge MirF             !
    MirG                                         !                   !
     !                                           !                   !
     !                                           !  Indp             !
     !                              2ème         !                   !
     !                                           !                   !
-.2  +                                           !                   !
     !  2PEU                                     !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
-.4  +                                           !                   !
     !                                          Doct                 !
      -------+-----------+-----------+---1PAS----+--mari-----+-------
            -.36        -.24        -.12         .00         .12
                                                                  AXE 1
```

On peut voir certaines associations entre les variables illustratives et les modalités des 211 variables. Mais surtout ce que confirme cette analyse c'est la non linéarité des mesures, les modalités sont distribuées de manière presque circulaire. L'AFCM permet comme l'ACP de regrouper les variables mais aussi les modalités, comme pour l'approche développée lors du premier exposé nous allons nous intéresser à la possibilité de discriminer des groupes définis à priori. Il se trouve que le graphique de l'AFCM suggère que la différence entre garçons et filles est assez importante dans le choix des modalités de réponse.

Il existe une méthode assez simple pour détecter les liaisons entre modalités de variables qualitatives qui consiste à croiser systématiquement les variables entre elles. Il est fastidieux et long à analyser d'énormes quantités de tableaux croisés, avec 211 variables cela serait énorme. Dans le logiciel SPAD il existe une procédure qui s'appelle DEMOD (pour description de modalités) qui permet de rationaliser ce travail en ne gardant que les liaisons qui sont significatives au sens d'un certain test. Dans les anciennes versions de SPAD, celle que j'ai commencé à utiliser en 1984 par exemple, la procédure s'appelait TAMIS, un certain nombre de statisticiens appellent encore ce principe de croisement par le nom ancien de cette procédure. Ce principe de croisement se ramène toujours à des croisements de variables à deux modalités. Si par exemple on voulait caractériser les modalités de la variable scolarité par les modalités des 211 variables associées aux mots, le logiciel commencerait par créer virtuellement autant de variables que de modalités en jeu dans la procédure de description, chacune de ces variables est binaire présence contre absence. Dans l'exemple de la scolarité on a 5 variables, Scolarité1 qui vaut 1 si l'étudiant est en première année et 2 sinon, puis une variable Scolarité2 qui vaut 1 si l'étudiant est en deuxième année et 2 sinon et ainsi de suite. On fait de même pour les variables mots, chaque variables créant 5 ou 6 variables suivant qu'il y ait ou non des «non réponses». En croisant systématiquement chacune des 5 variables associées à la scolarité avec chacune des 1201 variables associées aux 211 mots, le logiciel calcule une statistique de liaison qui n'est pas le χ^2 en ce qui concerne SPAD mais qui en est voisin. SPAD utilise la loi hyper-géométrique qui est plus justifiée dans le cas des petits effectifs. Ce que donne sur un cas précis cette méthode c'est qu'il y a une certaine liaison entre la troisième année d'étude et la modalité 2 (peu important) de sécurité (176). Cela apparaît dans un tableau de la façon suivante:

```
Sécurité   Peu imp.  25.00   20.00  20  2.59  .005
```

Par ailleurs le logiciel SPAD peut fournir le tableau et le χ^2 correspondant au croisement entre scolarité et sécurité:

EFFECTIF % EN LIGNE % EN COLONNE	sco1	sco2	sco3	sco4	sco5
Pas imp.	1 16.67 .47	2 33.33 2.90	2 33.33 8.00	0 .00 .00	1 16.67 100.00
Peu imp.	4 20.00 1.88	6 30.00 8.70	5 25.00 20.00	5 25.00 5.05	0 .00 .00
Moy. imp.	42 47.19 19.72	17 19.10 24.64	3 3.37 12.00	27 30.34 27.27	0 .00 .00
Imp.	96 54.86 45.07	27 15.43 39.13	11 6.29 44.00	41 23.43 41.41	0 .00 .00
Très imp.	70 60.34 32.86	17 14.66 24.64	4 3.45 16.00	25 21.55 25.25	0 .00 .00

```
CHI2 = 105.78
20 DEGRES DE LIBERTE
17 EFFECTIFS THEORIQUES INFERIEURS A 5
PROBA ( CHI2 > 105.78 ) =  .000
V.TEST =   7.33
```

Les deux résultats ont un rapport direct entre eux, il apparaît sur la ligne le nom de la variable en ligne dans le tableau (sécurité), puis le nom de la modalité retenue comme significativement liée à la troisième année de scolarité, il s'agit de la modalité 2 (peu important), le premier nombre marqué est 25.00 ce qui correspond au pourcentage en ligne, cela veut dire 25% des gens qui ont dit que la sécurité étaient peu importante sont en troisième année. Le deuxième nombre correspond au pourcentage colonne, cela veut dire que 20% des gens de troisième année ont dit que la sécurité c'était peu important. Le troisième nombre indique quel est sur l'ensemble de la population le nombre de répondants qui ont choisi la modalité 2 (peu important) pour la sécurité, il y en a 20. Il y a ensuite la valeur du test fondé sur la loi hyper-géométrique celle ci vaut 2.59, puis enfin une probabilité associée. Cette probabilité se lit comme dans tous les tests... L'exemple que j'ai pris n'est qu'une ligne parmi bien d'autres. Si on calcule le résidu standardisé de la cellule correspondant à la troisième année et la modalité 2 (peu important) dans le tableau croisé entre scolarité et sécurité on trouve 3.4 et c'est le plus fort résidu du tableau cela correspond à une valeur significative pour un test fondé sur le χ^2.

C'est en utilisant cette technique sur les variables illustratives de l'AFCM, que j'ai essayé de détecter les variables associées aux mots qui pouvaient discriminer le mieux les groupes définis par les modalités des variables illustratives. Le but de cet exposé n'est surtout pas de rentrer dans les détails de cette analyse par contre il m'a paru intéressant de s'attarder sur une de ces analyses en raison de l'étrangeté des résultats. Voilà ce qu'a donné la description de la variable *sexe* par les 211 variables mots, je n'ai retenu que les modalités pour lesquelles la probabilité est inférieure à $0,010$. J'ai ajouté dans le tableau le numéro de la variable et le numéro de la modalité, les modalités sont présentées dans l'ordre décroissant de significativité du test. D'abord il y a les modalités les plus caractéristiques des garçons puis les modalités les plus caractéristiques des filles.

Garçons

028 solidarité	3 Moy. imp.	37.50	36.84	56	4.70	.000
195 Promesse	3 Moy. imp.	26.04	43.86	96	3.54	.000
019 affection	4 Imp.	21.88	61.40	160	3.49	.000
102 générosité	3 Moy. imp.	30.00	31.58	60	3.39	.000
003 Fidélité	4 Imp.	23.39	50.88	124	3.35	.000
124 profession	3 Moy. imp.	29.82	29.82	57	3.24	.001
093 joie	4 Imp.	23.28	47.37	116	3.14	.001
064 coopération	3 Moy. imp.	24.24	42.11	99	3.07	.001
141 Propriété	5 Très imp.	34.29	21.05	35	3.04	.001
205 Sainteté	3 Moy. imp.	29.09	28.07	55	3.02	.001
027 désintéressement	2 Peu imp.	24.72	38.60	89	2.98	.001
097 féminité	1 Pas imp.	42.11	14.04	19	2.88	.002
173 Masculinité	4 Imp.	22.73	43.86	110	2.83	.002
038 pluralisme	4 Imp.	24.42	36.84	86	2.82	.002
021 fraternité	3 Moy. imp.	24.42	36.84	86	2.82	.002
050 amour	3 Moy. imp.	50.00	10.53	12	2.78	.003
119 ascèse	4 Imp.	50.00	10.53	12	2.78	.003
023 tout-puissance	3 Moy. imp.	26.15	29.82	65	2.72	.003
168 Intuition	5 Très imp.	25.35	31.58	71	2.70	.004
107 simplicité	3 Moy. imp.	24.68	33.33	77	2.68	.004
119 ascèse	3 Moy. imp.	22.92	38.60	96	2.61	.005
062 rénovation	4 Imp.	25.37	29.82	67	2.60	.005
082 douceur	3 Moy. imp.	26.79	26.32	56	2.59	.005
171 Ruse	4 Imp.	24.05	33.33	79	2.57	.005
194 Compétence	4 Imp.	18.69	64.91	198	2.52	.006
056 santé	3 Moy. imp.	50.00	8.77	10	2.48	.007
143 Amabilité	3 Moy. imp.	22.58	36.84	93	2.45	.007
108 vérité	4 Imp.	21.19	43.86	118	2.45	.007
029 acceptation	3 Moy. imp.	21.70	40.35	106	2.42	.008
088 honnêté	4 Imp.	21.50	40.35	107	2.37	.009

Filles

003 Fidélité	5 Très imp.	92.83	66.57	251	4.83	.000
019 affection	5 Très imp.	93.88	52.57	196	4.39	.000
143 Amabilité	5 Très imp.	96.30	29.71	108	3.78	.000
092 partage	5 Très imp.	95.76	32.29	118	3.77	.000
093 joie	5 Très imp.	91.09	67.14	258	3.68	.000
183 Sincérité	5 Très imp.	91.56	62.00	237	3.65	.000
108 vérité	5 Très imp.	90.74	70.00	270	3.63	.000
075 fécondité	5 Très imp.	93.96	40.00	149	3.54	.000
088 honnêté	5 Très imp.	90.14	73.14	284	3.39	.000
063 spontanéïté	5 Très imp.	95.37	29.43	108	3.37	.000
190 Assurance	5 Très imp.	98.41	17.71	63	3.35	.000
028 solidarité	5 Très imp.	92.43	48.86	185	3.34	.000
023 tout-puissance	1 Pas imp.	90.80	64.86	250	3.33	.000
103 affectivité	5 Très imp.	91.92	52.00	198	3.25	.001
107 simplicité	5 Très imp.	93.94	35.43	132	3.21	.001
086 sympathie	5 Très imp.	92.44	45.43	172	3.14	.001
184 Protection	5 Très imp.	95.05	27.43	101	3.08	.001
136 Autonomie	5 Très imp.	91.71	47.43	181	2.88	.002
126 tendresse	5 Très imp.	90.83	56.57	218	2.88	.002
124 profession	5 Très imp.	91.91	45.43	173	2.87	.002
118 famille	5 Très imp.	89.67	69.43	271	2.80	.003
069 complicité	5 Très imp.	90.74	56.00	216	2.80	.003
102 générosité	5 Très imp.	92.47	38.57	146	2.76	.003
171 Ruse	1 Pas imp.	95.18	22.57	83	2.73	.003
035 dialogue	5 Très imp.	89.84	65.71	256	2.73	.003
192 Pardon	5 Très imp.	93.69	29.71	111	2.73	.003
115 serviabilité	5 Très imp.	95.12	22.29	82	2.69	.004
065 égalité	5 Très imp.	90.04	62.00	241	2.67	.004
012 révolution	2 Peu imp.	93.22	31.43	118	2.65	.004
039 dynamisme	5 Très imp.	92.59	35.71	135	2.65	.004
064 coopération	5 Très imp.	95.06	22.00	81	2.65	.004
142 Mariage	5 Très imp.	93.58	29.14	109	2.64	.004
044 méthode	2 Peu imp.	96.61	16.29	59	2.59	.005
153 Rectitude	0 Manquant	97.87	13.14	47	2.57	.005
160 Travail	5 Très imp.	91.95	39.14	149	2.55	.005
131 harmonie	5 Très imp.	91.95	39.14	149	2.55	.005
168 Intuition	3 Moy. imp.	92.42	34.86	132	2.53	.006
082 douceur	5 Très imp.	91.41	42.57	163	2.48	.007
074 surêté	5 Très imp.	95.45	18.00	66	2.42	.008

La constatation évidente c'est que sur 39 modalités retenues pour les filles il y a 33 fois la modalité «très important» (5) et pas une seule fois la modalité «important» (4), la modalité «moyennement important» (3) n'apparaîssant qu'une fois. Par contre chez les garçons sur 30 modalités retenues, il y a 15 fois la modalité «moyennement important» (3) et 11 fois la modalité «important» (4). La modalité «pas important» (1) n'est pratiquement pas présente chez les garçons, le seul mot pour lequel elle apparaît est féminité. Le fait que les garçons soient plutôt du côté des modalités 3 et 4 va dans le même sens que ce que suggérait l'AFCM. Cela pose un problème intéressant, l'AFCM comme la description de modalités présentée plus haut donne plus une information sur le type de notation que sur les mots caractéristiques des garçons ou des filles. C'est un problème assez similaire de celui de l'ACP présentée dans le premier exposé, la façon de contourner ce problème dans l'ACP avait été de recentrer les notes. C'est là que j'ai pensé qu'il pouvait y avoir une situation ressemblant au paradoxe de Simpson, j'ai donc pensé à croiser deux variables, le *sexe* et la façon de noter. Mais pour cela il a fallu réorganiser le fichier et considérer que l'on avait un fichier de $407 * 211 = 85877$ lignes chaque ligne correspondant à un croisement répondant par *mot* et une seule variable réponse la note. En ayant fait cela j'ai calculé un simple tableau croisé qui est présenté ici:

Eff obs % Lig Eff thé % Col Résidu stand.	Garçons 1		Filles 2		Total Ligne
0	81	7,6	983	92,4	1064
	149	0,7	915	1,3	1,2%
	-5,6		2,2		
1	908	12,8	6203	87,2	7111
	996	7,5	6115	8,4	8,3%
	-2,8		1,1		
2	1306	14,7	7581	85,3	8887
	1245	10,9	7642	10,3	10,3%
	1,7		-0,7		
3	3321	15,6	17994	84,4	21315
	2985	27,6	18330	24,4	24,8%
	6,1		-2,5		
4	4205	15,6	22775	84,4	26980
	3779	35,0	23201	30,8	31,4%
	6,9		-2,8		
5	2206	10,8	18314	89,2	20520
	2873	18,3	17646	24,8	23,9%
	-12,5		5,0		
Total Colonne	12027	14%	73850	86%	85877

```
Khi2 de Pearson=329,00826
Degrés de liberté=5
Niveau de signification=0,00000
```

Ce tableau contient 5 informations par cellule, l'effectif observé, l'effectif théorique, le pourcentage ligne, le pourcentage colonne et enfin le résidu standardisé, c'est à partir de cette valeur qu'on détermine les cellules contribuant le plus au χ^2. Ce tableau extrêmement simple se caractérise dès le premier coup d'oeil par un résidu standardisé négatif très important $-12,5$ correspondant à la modalité «très important» pour les garçons. Cette modalité représente près de 24% pour l'ensemble de la population et seulement

18% pour les garçons. En raison de la beaucoup plus grande proportion de filles que de garçons les résidus dans la colonne des garçons sont beaucoup plus grands que ceux de la colonne des filles, mais les résidus sont vraiment dans le même ordre inversé, c'est la modalité «très important» (5) qui est la plus sur représentée chez les filles suivie de la modalité «pas important» (1). Ce tableau met en évidence une liaison très forte entre la variable *sexe* et la façon de noter les mots de ce questionnaire et ce pour l'ensemble des 211 mots, ce qui rend la lecture de la description des modalités douteuse, on est là dans une situation similaire à celle du paradoxe de Simpson. Nous avons ici 3 variables qui sont le *sexe*, la *note* et le *mot* . La variable *sexe* a deux modalités, la variable *note* a 6 modalités et la variable *mot* a 211 modalités. Dès le premier tableau récapitulatif des données nous avons présenté celles-ci exactement comme si nous croisions la variable *mot* avec la variable note, je redonne ici un extrait de ce tableau:

Pas	Peu	Moy	Imp	Très	NR	Num	Intitulé
3	8	57	141	198	0	V001	Loyauté
10	40	178	133	42	4	V002	Charité
0	5	26	124	251	1	V003	Fidélité
1	11	82	180	131	2	V004	persévérance
1	1	26	87	289	3	V005	Liberté
19	42	142	132	68	4	V006	charme
28	55	150	101	34	39	V007	Décentration
3	12	76	200	112	4	V008	Détermination
1	3	21	106	275	1	V009	Franchise
100	86	141	42	14	24	V010	piété
7	15	86	162	137	0	V011	Loisir
68	118	131	64	22	4	V012	révolution
1	4	31	176	195	0	V013	volonté
0	0	27	120	259	1	V014	sentiment
13	33	142	130	78	11	V015	développement
28	56	149	76	25	73	V016	mansuétude
164	101	113	20	3	6	V017	célibat
142	114	100	43	7	1	V018	patriotisme

Quand on a fait une AFC simple sur ce tableau on a étudié la liaison entre les notes et les mots, le tableau de χ^2 précédent nous a permis d'étudier la liaison entre la variable *sexe* et la variable *note* . La dernière liaison que l'on pourrait étudier est celle entre la variable *mot* et la variable *sexe*, mais cela n'a aucun sens. En effet, compte tenu du fait qu'il y a une modalité «non réponse», tous les mots ont 407 réponses et pour chaque mot il y a 350 filles et 57 garçons, il n'y a donc, par construction, aucune liaison entre ces deux variables. Le but éventuel de l'AFCM et aussi de la description de modalités telle qu'elle a été faite est de mesurer une liaison entre ces trois variables. La similitude avec le paradoxe de Simpson vient du fait que la liaison entre le *sexe* et la note agit de manière parasite pour lire la liaison entre *sexe*, *mot* et *note*. Plusieurs méthodes ont été envisagées pour tenter de résoudre ce «paradoxe». Une des idées possibles est d'utiliser l'AFC simple mais en regroupant deux tableaux constitués un en considérant uniquement les filles et un uniquement en considérant les garçons. Deux façons de regrouper, soit en collant les tableaux côte à côte en faisant un tableau de 211 lignes et 12 colonnes soit en les collant à la file, en faisant un tableau de 422 lignes et 6 colonnes. C'est la représentation de cette AFC sur les deux premiers axes qui est donnée, de manière incomplète, toujours pour des raisons de place. L'idée de cette méthode consiste à repérer les différences de représentations des mots pour les filles et pour les garçons.

```
                           AXE 1  *  AXE 2
    F108F116G116------------------------------------------------F23 F203
    F9  G5                   !                                  F205F31
    F3  G50                  !                                  G22 F22
    F14 F93                  !                                  F159G73
    G139G36                  !                                  F73 G31
    F65 G56                  !                                  PAS1G159
    F71                      !                                  G181 !
    F152F125F36              !                              F157    G205
    F126F69 G9               !                                  F27  !
    TRE5F76 G14              !                                  G23 F109
 .4 G88 F103G118             !                          G187        F17
    F19 F122G125        F75  !      F142                    F18 F55 F96
    F28 F138F13 F124         !                          F187G78 F181 !
    F136G76 F111             !                      G157        G17  !
    G170G108G1  G71          !                                       !
    F86 G35 G65 F77          !          G211        G18              !
     !  G93 G83 G152         !              G142F211F51 F30 G27 G109 !
     !  G13 F160F107G174    G97                             F105    F119
     !  F149F32 F21 F166F195 !          G98     F78         F148     !
     !  G183F39 G80 F150     !          F173F98 G12 G96              !
     !  F129F33 F11 F97     F59 G53 G75     G134G146G55 G30 F10      !
 .0  +--G89 G138F163G168G166G100--------F161--------F171----G105-----!
     !  F92 G3  G95 F112G177F42 F53 F208        F146PEU2G148         !
     !  F81 F63 G66 G61 F115G208F114G104F47 F134    G10              !
     !     G81 G149F68 G21 F61 F164G141G161G51     F12 G144         !
     !     G112F46 G160G131G6  G114G171G24 G47         F144         !
     !     G8  G200F154G40 F25 G197F186F60 F91 F48                  !
     !     G19 G32 G70 G185F74 G195F24 F38 F113    G85 F85 G119     !
     !     G33 G162F106F40 F135F168G37 G186G87 F79 F62              !
     !  F56     G39 G102G137G180F44 G91 F141G204G153F132             !
     !     F50 F194F189G107G135G44 MOY3F117F204G113F128G48          !
     !              G58 G72 G29 F155G167G156G62 G110F153             !
-.4  +              G158F127F210G155G120F7  F178G79 G128             !
     !     F88     G198G92 G175G34 G191G101G49 F201                 !
     !                  F2  G41 F34 F49 G117                         !
     !              G194G189G127G74 G206G121F16                      !
     !                  G64  !  G2          G7          NR6          !
     !                  G143 !              G201                     !
     !                      G190            G132                     !
      -------+---------------+---------------+------G16 -----+-------
            -.5              .0              .5             1.0
```

Compte tenu du nombre de points un tel graphique est difficile à lire dans sa globalité. Les variables mots sont représentées avec leur numéro précédé d'un F pour les filles et d'un G pour les garçons. On peut toutefois repérer certains petits détails, en bas à gauche il y a, par exemple F56, F50 et F88 qui correspondent aux mots santé, amour et honnêteté pour les filles alors que G56, G50 et G88 correspondant aux mêmes mots pour les garçons sont en haut en gauche. Attention cela ne veut pas dire, malgré les apparences que ces mots soient plus loin de la modalité 5 pour les filles. F75 et F142 correspondant à la fécondité et au mariage pour les filles sont bien au milieu du «fer à cheval» alors que G75 et G142 sont un peu plus noyés dans la masse. Cela correspond au fait que ces mots ont une plus grande variance pour les filles.

Pour essayer d'analyser ce paradoxe j'ai essayé d'utiliser le modèle log-linéaire, mais pour ce fichier de données le nombre de cellules est absolument gigantesque $211 * 6 * 2 = 2532$. J'ai donc mis la même table que celle utilisée pour l'AFC précédente dans un modèle. Le principe du modèle consiste à dire que le nombre de personnes d'un sexe ayant attribué une note à un mot dépend de plusieurs types d'effets. De manière évidente cela dépend du nombre de personnes de chacun des deux sexes, cela dépend du nombre de réponses pour le mot, mais ici ce nombre est le même pour tous les mots (407) et cela dépend du nombre de fois où la note a été donnée sur l'ensemble des mots. Ces trois choses représentent ce que l'on appelle les effets simples. Le principe est une généralisation de la table de contingence (tableau croisé simple) ici à un tableau à 3 entrées comme si on empilait deux tableaux l'un sur l'autre. On peut ensuite envisager les effets d'interaction de variables deux à deux, puis d'interaction entre les trois variables. Le principe est similaire à une analyse de variance à 3 facteurs sauf que la variable à expliquer (variable dépendante pour certains logiciels) est une variable de comptage qui ne suit pas une loi normale mais plutôt une loi de Poisson, ici c'est le nombre de personnes dans une cellule du tableaux à trois entrées défini par le sexe, le mot et la note. Dans la première AFC simple nous avons étudié l'interaction entre la variable mot et la variable note, dans le tableau de χ^2 précédent nous avons étudié l'interaction entre la variable sexe et la variable note, dans le modèle log-linéaire nous allons mettre ces deux interactions ensemble. L'interaction entre la variable sexe et la variable mot est sans aucun intérêt d'un point de vue qualité d'ajustement du modèle, mais par contre cela est utile pour l'estimation des paramètres et cela est intéressant. Si l'on met dans le modèle l'interaction entre les trois variables on obtient un modèle dit saturé. Pour traiter les données il a été plus simple de créer un fichier de 2532 lignes et 4 variables, la première variable étant le mot, la deuxième le sexe, la troisième la note et enfin la quatrième étant l'effectif correspondant à la cellule, cette variable a d'ailleurs était appelée *eff* . Voici à titre exemple un extrait de ce fichier correspondant au mot numéro 36.

```
 36  1  1       0
 36  1  2       2
 36  1  3       6
 36  1  4      13
 36  1  5      36
 36  1  6       0
 36  2  1       3
 36  2  2       3
 36  2  3      29
 36  2  4     126
 36  2  5     189
 36  2  6       0
```

La note 6 correspond ici à la «non réponse». Cela indique par exemple qu'il y a 13 garçons qui ont donné la note 4 (important) au mot 36 ou qu'il y a 29 filles qui ont donné la note 3 (moyennement important) au même mot. La syntaxe pour faire du modèle log-linéaire avec SPSS peut être la suivante, c'est celle que j'ai retenue. Cette syntaxe ne peut hélas pas être obtenue par des menus et la commande obtenue par les menus ne m'a pas convaincu, pour le moment et je reste donc fidèle à celle là.

```
weight by eff
loglinear note (1,6) mot (1,211) sexe (1,2)
 /design= note mot sexe note by mot sexe by note sexe by mot
 /print=freq estim
```

Attention il s'agit d'un calcul excessivement long à effectuer, je ne l'ai d'ailleurs pas fait avec SPSS pour le fichier en entier, mais avec GLIM, voici la syntaxe de ce logiciel.

```
$error poisson$
$yvariate eff$
$fit note+mot+sexe+sexe.note+mot.note+sexe.mot$
```

Je l'ai fait aussi avec SAS voici la syntaxe de ce logiciel.

```
proc genmod;
class  mot sexe note;
model eff = mot sexe  note note*sexe mot*note mot*sexe
    /dist =poi link=log ;
run;
```

À partir de comparaisons poussées SAS et GLIM donnent les mêmes résultats. Sur bien des exemples SPSS et SAS donnent les mêmes résultats. Sur des sous ensembles de ce fichiers SPSS et GLIM donnent bien les mêmes résultats sur les modèles log-linéaires.

Si j'ai choisi GLIM c'est parce que c'est le logiciel qui produit les résultats les plus compacts ce qui compte tenu de la taille du fichier était très important, il a fallu plusieurs heures à un gros calculateur pour ajuster ce modèle. Il est évident qu'il faut être très prudent pour lancer ce genre de calcul sur un micro ordinateur surtout avec SPSS.

Dans le modèle avec toutes les interactions d'ordre 2 mais sans celle d'ordre 3 j'ai regardé où il y avait des résidus importants, c'est à dire inférieurs à -2 ou supérieurs à 2 je n'ai gardé que les cellules concernant les garçons et voici un tableau récapitulatif des résidus importants.

Mots	PAS 1	PEU 2	MOY 3	IMP 4	TRES 5	NR 6
Propriété					4.67	
Intuition			-2.26		4.16	
Ruse				2.14	3.38	
patriotisme					2.77	
raison					2.64	
révolution		-2.16			2.63	
chance					2.45	
intelligence					2.32	
Méditation					2.29	
mysticisme					2.18	
honneur					2.17	
objectivité					2.15	
courage					2.14	
Loisir					2.04	
plaisir				-2.23	2.03	
partage		2.16			-2.04	
Assurance					-2.07	3.44
affection	2.30				-2.09	
Amabilité					-2.11	
Fidélité					-2.11	3.07
égalité		2.01				
tout-puissance			2.01			
sympathie		2.10				2.94
solidarité			3.76			
santé			2.22			
rénovation				2.27		
professionnalisme	2.05					
profession			2.42			
pluralisme				2.12		
méthode		-2.21				
mansuétude						3.44
honnêté	2.17					

Mots	PAS 1	PEU 2	MOY 3	IMP 4	TRES 5	NR 6
générosité			2.64			
féminité	3.51					
désintéressement		2.35				
discrétion	2.28					
coopération			2.26			
compétitivité						3.40
cohérence						2.22
beauté						3.46
autorité						3.49
ascèse				2.86		
amour			2.42			
Travail	2.01					
Tolérance	2.26					
Sincérité		2.03				
Sainteté			2.39			
Promesse			2.47			
Modestie						4.79
Masculinité				2.04		
Indulgence		2.43				
Franchise	2.18					
Diplomatie		2.10				
Compromis	2.63					

Il est facile de voir où se trouvent les mots avec les modalités correspondantes. Là où se trouve un résidu positif c'est que le nombre de garçons ayant choisi cette modalité est important par rapport aux nombre de filles mais surtout en tenant compte de la liaison existant entre le sexe et la façon de noter. On peut à partir de là, comparer avec ce tableau avec celui de la description des modalités obtenu par DEMOD. J'ai repris les 30 modalités les plus représentatives des garçons obtenues par DEMOD en accolant au tableau la valeur du résidu obtenu dans le modèle log-linéaire, et en mettant *abs* si le résidu n'est pas assez important.

028	solidarité	3 Moy. imp.	37.50	36.84	56	4.70	3.76
195	Promesse	3 Moy. imp.	26.04	43.86	96	3.54	2.47
019	affection	4 Imp.	21.88	61.40	160	3.49	abs
102	générosité	3 Moy. imp.	30.00	31.58	60	3.39	2.64
003	Fidélité	4 Imp.	23.39	50.88	124	3.35	abs
124	profession	3 Moy. imp.	29.82	29.82	57	3.24	2.42
093	joie	4 Imp.	23.28	47.37	116	3.14	abs
064	coopération	3 Moy. imp.	24.24	42.11	99	3.07	2.26
141	Propriété	5 Très imp.	34.29	21.05	35	3.04	4.67
205	Sainteté	3 Moy. imp.	29.09	28.07	55	3.02	2.39
027	désintéressement	2 Peu imp.	24.72	38.60	89	2.98	2.35
097	féminité	1 Pas imp.	42.11	14.04	19	2.88	3.51
173	Masculinité	4 Imp.	22.73	43.86	110	2.83	2.04
038	pluralisme	4 Imp.	24.42	36.84	86	2.82	2.12
021	fraternité	3 Moy. imp.	24.42	36.84	86	2.82	abs
050	amour	3 Moy. imp.	50.00	10.53	12	2.78	2.42
119	ascèse	4 Imp.	50.00	10.53	12	2.78	2.86
023	tout-puissance	3 Moy. imp.	26.15	29.82	65	2.72	2.01
168	Intuition	5 Très imp.	25.35	31.58	71	2.70	4.16
107	simplicité	3 Moy. imp.	24.68	33.33	77	2.68	abs
119	ascèse	3 Moy. imp.	22.92	38.60	96	2.61	abs
062	rénovation	4 Imp.	25.37	29.82	67	2.60	2.27
082	douceur	3 Moy. imp.	26.79	26.32	56	2.59	abs
171	Ruse	4 Imp.	24.05	33.33	79	2.57	2.14
194	Compétence	4 Imp.	18.69	64.91	198	2.52	abs
056	santé	3 Moy. imp.	50.00	8.77	10	2.48	2.22
143	Amabilité	3 Moy. imp.	22.58	36.84	93	2.45	abs
108	vérité	4 Imp.	21.19	43.86	118	2.45	abs
029	acceptation	3 Moy. imp.	21.70	40.35	106	2.42	abs
088	honnêté	4 Imp.	21.50	40.35	107	2.37	abs

L'analyse ce tableau permet de voir qu'il y a des différences importantes, 12 modalités qui avaient été considérées par DEMOD sont absentes dans le modèle log-linéaire réciproquement des résidus importants pour le modèle log-linéaire correspondent à des modalités qui sont absentes, la hiérarchie des résidus n'est pas vraiment la même que celle de DEMOD.

Pour mieux présenter les sorties de SPSS sur le modèle log-linéaire, je me suis limité à l'étude d'un seul mot. en partant du fichier de 2532 lignes j'ai recodé la variable mot en deux modalités la valeur 36 est devenue la valeur 1 et toutes les autres valeurs sont devenues 2 cela fournit un fichier de seulement 24 lignes. Cela donne le fichier suivant:

```
1 1 1       0
1 1 2       2
1 1 3       6
1 1 4      13
1 1 5      36
1 1 6       0
1 2 1       3
1 2 2       3
1 2 3      29
1 2 4     126
1 2 5     189
1 2 6       0
2 1 1     908
2 1 2    1304
2 1 3    3315
2 1 4    4192
2 1 5    2170
2 1 6      81
2 2 1    6200
2 2 2    7578
2 2 3   17965
2 2 4   22469
2 2 5   18125
2 2 6     983
```

La syntaxe de SPSS devient, en changeant le 211 par 2:

```
weight by eff
loglinear note (1,6) mot (1,2) sexe (1,2)
 /design= note mot sexe , note by mot, sexe by note
 /print=freq estim
```

Et obtient grosso modo les résultats suivant que j'ai un petit peu allégés...

```
* * * * * * * * * * L O G   L I N E A R   A N A L Y S I S * * * *

DATA   Information

         24 unweighted cases accepted.
          0 cases rejected because of out-of-range factor values.
          0 cases rejected because of missing data.
      85697 weighted cases will be used in the analysis.


FACTOR Information

   Factor  Level  Label
   NOTE       6
   MOT        2
   SEXE       2
 Correspondence Between Effects and Columns of Design/Model 1

  Starting  Ending
   Column   Column   Effect Name

      1        5     NOTE
      6        6     MOT
      7        7     SEXE
      8       12     NOTE BY MOT
     13       17     SEXE BY NOTE
     18       18     MOT BY SEXE
```

Factor	Code	OBS. count	EXP. count
NOTE	Pas important		
MOT	plaisir		
SEXE	Garçon	0	.42
SEXE	Fille	3	2.58
MOT	autres		
SEXE	Garçon	908	907.58
SEXE	Fille	6200	6200.42
NOTE	Peu important		
MOT	plaisir		
SEXE	Garçon	2	.80
SEXE	Fille	3	4.20
MOT	autres		
SEXE	Garçon	1304	1305.20
SEXE	Fille	7578	7576.80
NOTE	Moyennement important		
MOT	plaisir		
SEXE	Garçon	6	5.90
SEXE	Fille	29	29.10
MOT	autres		
SEXE	Garçon	3315	3315.10
SEXE	Fille	17965	17964.90
NOTE	Important		
MOT	plaisir		
SEXE	Garçon	13	23.60
SEXE	Fille	126	115.40
MOT	autres		
SEXE	Garçon	4192	4181.40
SEXE	Fille	22469	22479.60
NOTE	Très important		
MOT	plaisir		
SEXE	Garçon	36	26.29
SEXE	Fille	189	198.71
MOT	autres		
SEXE	Garçon	2170	2179.71
SEXE	Fille	18125	18115.29
NOTE	Non réponse		
MOT	plaisir		
SEXE	Garçon	0	.00
SEXE	Fille	0	.00
MOT	autres		
SEXE	Garçon	81	81.00
SEXE	Fille	983	983.00

NOTE	Pas important			
MOT	plaisir			
SEXE	Garçon	-.4158	-.6449	-.6976
SEXE	Fille	.4158	.2587	.6976
MOT	autres			
SEXE	Garçon	.4158	.0138	.6976
SEXE	Fille	-.4158	-.0053	-.6976
NOTE	Peu important			
MOT	plaisir			
SEXE	Garçon	1.2039	1.3492	1.4822
SEXE	Fille	-1.2039	-.5872	-1.4822
MOT	autres			
SEXE	Garçon	-1.2039	-.0333	-1.4822
SEXE	Fille	1.2039	.0138	1.4822
NOTE	Moyennement important			
MOT	plaisir			
SEXE	Garçon	.0971	.0400	.0463
SEXE	Fille	-.0971	-.0180	-.0463
MOT	autres			
SEXE	Garçon	-.0971	-.0017	-.0463
SEXE	Fille	.0971	.0007	.0000
NOTE	Important			
MOT	plaisir			
SEXE	Garçon	-10.5988	-2.1818	-3.1076
SEXE	Fille	10.5988	.9866	3.1076
MOT	autres			
SEXE	Garçon	10.5988	.1639	3.1076
SEXE	Fille	-10.5988	-.0707	-3.1076
NOTE	Très important			
MOT	plaisir			
SEXE	Garçon	9.7137	1.8946	2.7976
SEXE	Fille	-9.7137	-.6891	-2.7976
MOT	autres			
SEXE	Garçon	-9.7137	-.2081	-2.7976
SEXE	Fille	9.7137	.0722	2.7976
NOTE	Non réponse			
MOT	plaisir			
SEXE	Garçon	.0000	-.0002	-.0002
SEXE	Fille	.0000	-.0007	-.0007
MOT	autres			
SEXE	Garçon	.0000	.0000	.0000
SEXE	Fille	.0000	.0000	.0000

```
Goodness-of-Fit test statistics

   Likelihood Ratio Chi Square =    12.97674    DF = 6  P =  .043
             Pearson Chi Square =    12.52975    DF = 6  P =  .051
```

SPSS commence par présenter le modèle qui a été utilisé puis donne les valeurs observées et les valeurs ajustées, parfois appelées valeurs théoriques ou valeurs attendues, de toutes les cases du tableau ensuite il fournit les résidus du modèle et enfin donne les statistiques d'ajustement et notamment le χ^2. Les résidus importants sont pour les modalités 4 négatif et 5 positif pour les garçons, cela est cohérent avec le modèle global quand on considère les 211 mots. À titre de comparaison le tableau croisé entre la variable *sexe* et la variable *V36* ne donne pas de résidu important et le χ^2 n'est pas significatif. Le tableau du modèle log-linéaire donne trois types de résidus les résidus bruts, les résidus standardisés et les résidus ajustés, cela renforce la ressemblance entre le modèle log-linéaire et l'analyse d'un tableau croisé, toutefois dans ce tableau je n'ai mis que l'effectif et le résidu standardisé.

```
                SEXE
        Eff     |
        Résidu  |Garçons  Filles
                |                  Row
                |    1   |    2   | Total
V36     --------+--------+--------+
             1  |    0   |    3   |     3
  Pas imp.      |   -.6  |    .3  |   .7%
                +--------+--------+
             2  |    2   |    3   |     5
  Peu imp.      |  1.6   |   -.6  |  1.2%
                +--------+--------+
             3  |    6   |   29   |    35
  Moy. imp.     |    .5  |   -.2  |  8.6%
                +--------+--------+
             4  |   13   |  126   |   139
  Imp.          |  -1.5  |    .6  | 34.2%
                +--------+--------+
             5  |   36   |  189   |   225
  Très imp.     |    .8  |   -.3  | 55.3%
                +--------+--------+
         Column     57       350      407
          Total   14.0%    86.0%   100.0%
```

Chi-Square	Value	DF	Significance
Pearson	6.82190	4	.14560

On a donc ici une situation correspondant à un paradoxe qui montre qu'il faut se méfier des croisements de variables deux à deux. Mais bien sûr à partir de cet exemple on peut se douter que par exemple la variable *filière* pourrait avoir un rôle important dans le modèle alors qu'elle n'y est pas. Mais le fait de voir la petite contradiction entre le croisement de deux variables et le croisement de trois variables permet de se poser des questions ce qui est souvent encore plus important que de donner des réponses. L'intérêt de cette dernière approche est de montrer que le traitement des données en qualitatif permet d'apprécier des différences entre modalités très voisines, le mot plaisir (36) est particulièrement intéressant de ce point de vue là.

Il peut être pertinent d'effectuer un retour en arrière pour voir le comportement d'une telle variable dans diverses analyses précédentes. Premièrement c'est un mot avec une très forte moyenne, plus de 50% pour la modalité 5, les modalités 4 et 5 représentant 90% des réponses. Dans la première AFC simple il apparaît du côté de la modalité 5 mais légèrement décalé du groupe de tête vie (116), santé (56), amour (50) ce qui correspond au fait que le mot est classé 19 pour la moyenne. Dans l'ACP le mot est plutôt gauche sur le graphique légèrement en bas. Une lecture naïve des variables illustratives de l'ACP mettrait ce mot du côté des filles, du côté du Mirail etc. Le dernier traitement simple effectué pour le premier exposé nous montrait quelque chose de contraire, le mot est mieux classé chez les garçons que chez les filles, et il est mieux classé à l'IUT qu'au Mirail. Dans ce petit traitement ce qui apparaissait nettement c'est que le mot était classé deuxième juste après la vie par les garçons de l'IUT alors qu'il n'arrivait qu'à la place 25 chez les filles du Mirail. Dans l'AFCM le plus simple est de refaire un graphique pour visualiser où se trouvent les 5 modalités du mot numéro 36, elles ont été notées $0361, 0362, 0363, 0364, 0365$.

```
                            AXE 1  *  AXE 2
AXE 2
      ----------------------IUTgCélC---------------------------------
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                       4IMP!                   !
     !                                           !                   !
     !                                           !                   !
 .2  +                                           !  IUT              !
     !                                           !                   !
     !            garçons                        !          res3     !
     !                                           !          IUTf     !
     !                                           !      4ème         !
     !                                parents    ! 0365              !
     !                                           !                   !
     !                          fam7             !                    7_
     !    0364                                   !                 Couple
     !  3ème                                RFam !                   !
 .0  -0362------------3MOY-----------------------+--1ère-------------!
     !                                  Céli  6NR!fillesfam8         5IMP
     !                                           !                   !
     !                                    Mirail !              seul !
     !                                          Âge MirF             !
    MirG                                         !                   !
     !                                           !                   !
     !                                           !  Indp             !
     !                              2ème         !                   !
     !                                           !                   !
-.2  +                                           !                   !
     !  2PEU                                     !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
     !                                           !                   !
-.4  +                                           !                   !
   0363                                         Doct              0361
      -------+-----------+-----------+---1PAS----+--mari-----+-------
            -.36        -.24        -.12         .00         .12
                                                                  AXE 1
```

La représentation du centre des modalités 5 est trompeuse car elle se trouve loin à l'extérieur du graphique, ainsi la modalité 5 du mot 36 est beaucoup plus proche du centre mais c'est sur le deuxième axe qu'apparaît un léger décalage. Le centre des modalités 4 se trouve en haut de l'axe alors que la modalité 4 du mot 36 se trouve plus bas que la modalité 5 du même mot, ce qui est en général le contraire; ce deuxième axe oppose d'ailleurs le Mirail et l'IUT et aussi les garçons et les filles. Dans l'AFC simple avec représentation différenciée pour les garçons et les filles G36 apparaît dans le groupe de tête alors que F36 est légèrement plus bas.

Traitements quantitatifs ou traitements qualitatifs, traitements visant à regrouper ou à discriminer, traitements simples ou complexes, traitements descriptifs ou traitement inférentiels, j'ai essayé au long du traitement de ce fichier de données de choisir un certains nombres d'approches différentes. J'ai effectué un grand nombre d'autres traitements statistiques que je n'ai pas jugés intéressants et que je je n'ai pas intégrés dans ces exposés, classifications sur les sujets ou sur les variables notamment. Il y a d'autres traitements que j'ai utilisé comme aides ou compléments et que je n'ai pas intégrés au document faute d'espace, de temps et peut être aussi à cause de la difficulté. Par ailleurs il me paraît utile de rappeler que les traitements auraient pu être différents si des éliminations de variables avaient été effectuées suite aux premiers traitements, une telle décision ne peut venir uniquement du traitement statistique mais doit être appuyée par des considérations propres au domaine d'étude. Ces exposés n'avaient pas d'autre prétention que de présenter certaines méthodes et surtout, comme l'indiquait le titre, comment il est possible de les enchaîner ou de les articuler. Comme pour les exposés précédents ce document représente un support écrit qui sert essentiellement pour les auditeurs de l'exposé, ce document doit se voir comme un complément ou une aide pour suivre et non comme quelque chose d'autonome. De plus il pourrait y avoir des interprétations plus poussées des divers résultats statistiques présentés mais il s'agit là du travail des praticiens.