Faut-il se méfier de l'utilisation des statistiques en sciences humaines ? Présentation de quelques méthodes sur un exemple simple

Joseph SAINT PIERRE, CICT

Vendredi 16 Janvier 1998

Le CICT qu'est ce que c'est ?

Pour commencer cet exposé, il me semble intéressant de dire que j'interviens ici en tant qu'ingénieur au CICT (Centre Interuniversitaire de Calcul de Toulouse), ingénieur responsable de l'utilisation des logiciels de statistiques du CICT. Le CICT a été créé en 1972, suite à la restructuration des facultés de l'Université de Toulouse, la constitution des universités actuelles, UT1, UTM et UPS, l'autonomisation des grandes écoles de l'Université de Toulouse regroupées dans l'INP. Il existait dans la faculté des sciences de l'Université un Institut de Calcul Numérique, cet Institut fut le premier endroit à disposer d'ordinateur dans une université française en 1957. Cet Institut a disposé lors de la création du campus de Rangueil de locaux importants dans un bâtiment conçu spécialement pour abriter les gros ordinateurs et assez vite de personnels pour s'occuper de ces ordinateurs. Le CICT a été constitué en utilisant les locaux de l'Institut de Calcul Numérique (le bâtiment s'appelle toujours l'ICN), en continuant l'utilisation des ordinateurs de cet Institut et en gardant la majeure partie de ses personnels. Ainsi le CICT a pu prolonger l'activité de calcul qui était de l'ICN, et ainsi a pu fournir des services de Calcul aux universitaires Toulousains, ces services se sont même étendus à d'autres établissements. C'est en fait de là que vient le titre d'interuniversitaire du CICT, dès sa création il était donc prévu que les moyens de calculs importants tant en matériel qu'en personnel devaient rester disponibles pour l'ensemble des universitaires Toulousains y compris ceux des universités de Sciences Sociales juridiques et économiques ou sciences humaines et littératures, il était donc évident qu'il fallait garder un aspect pluridisciplinaire au calcul, il est donc normal, dans cette optique, qu'un spécialiste du calcul du CICT intervienne dans le cadre d'une telle conférence.

L'évolution des moyens de calcul dans le passé récent.

Depuis 1972 la situation dans le domaine du calcul a grandement évolué et ce essentiellement grâce aux progrès techniques dans le domaine. Les ordinateurs se sont très largement répandus, ont baissé de prix, sont devenus beaucoup plus performants et surtout beaucoup plus faciles à utiliser. Une date à retenir de ce point de vue là est 1984 année de sortie du Macintosh. Les micro ordinateurs ont rendu possible l'accès à des moyens de calculs à un nombre beaucoup plus important d'utilisateurs, notamment à des utilisateurs issus des sciences humaines. Depuis 1985 le réseau à commencé à s'implanter fortement dans les Universités Toulousaines rendant plus facile le travail à distance et les communications entre utilisateurs et spécialistes, en 1990 l'arrivée d'Internet à accéléré ce processus. L'extension du réseau continue d'être phénoménale et favorise l'accès aux moyens de calculs et surtout favorise la circulation de l'information sur ces moyens. Par ailleurs et c'est là le point le plus important les logiciels ont fait d'énormes progrès tant du point de vue des performances que du point de vue de la facilité d'utilisation mais aussi de la disponibilité. Les logiciels de calcul statistique qui avaient été conçus pour des gros orinateurs dans les années 1960-1970 tels que BMDP, SAS et SPSS ont évolué pour s'adapter à la micro informatique des années 1980-1990. Un certain nombre d'autres logiciels spécifiques aux micro ordinateurs sont apparus (SYSTAT, STATVIEW). Dans le domaine des sciences sociales et humaines SPSS a assis sa position de référence, au départ SPSS avait été développé par des étudiants en science politiques et lors du début de sa commercialistaion en 1967 voulait dire Statistical Package for Social Sciences. A partir de la version 6 sur PC, SPSS est passé dans la catégorie des logiciels faciles à utiliser, les menus ont même été traduits en français et en 7 langues. En 1993 SPSS a eu aussi un changement de politique commerciale fort avec le CICT lors de l'ouverture de l'Agence SPSS France. Le CICT qui avait SPSS sur gros ordinateurs depuis 1988 a obtenu une licence de site ńinfinież. Ces élèments ont accru le nombre d'utilisateurs de méthodes statistiques mais a induit une forte hétérogénéïté des niveaux d'utilisation. Hélas l'augmentation de la disponibilité des logiciels de calcul en statistique ne s'est pas accompagnée d'une augmentation du niveau des formations, il est sûr que ces exposés dans le cadre de l'IED sont là pour tenter de pallier ce décalage entre les performances des logiciels et le niveau de leurs utilisateurs.

Liens entre mathématiques et sciences humaines. L'affaire ńSokalż

Attention, cependant à la façon dont souvent est présentée l'utilisation des statistiques dans les sciences humaines. Le titre que j'ai donné à cet exposé est volontairement provocateur : ńFaut il se méfier de l'utilisation des statistiques en sciences humaines ?ż. les statistiques font partie des mathématiques, sont donc rattachées aux sciences dures. ńL'évidence mathématique est le symbole de l'idée claire et distincte ; ces idées s'opposent aux perceptions obscures de l'imagination et des sens. ż René Descartes (1596-1650). La séparation universitaire telle qu'elle est pratiquée à Toulouse, Mirail et Rangueil indique qu'il y a une différence fondamentale si ce n'est une opposition entre les mathématiques et les sciences humaines. Dans la présentation habituelle des mathématiques, celles ci sont adaptées à la modélisation du monde physique et non au discours sur l'humain. Il y a eu en 1996 une affaire qui a relancé fortement le débat sur les relations entre d'un côté les mathématiques et la physique et de l'autre la sociologie, la philosophie et la psychanalyse, il s'agit de l'affaire ńAlan Sokalż. Alan Sokal professeur de physique à l'université de New-York a envoyé un article canular à une revue de sciences sociales ńSocial Textż, article parlant de physique et de mathématiques, mais truffé d'erreurs volontaires, l'article ayant été accepté dans le numéro de Juin 1996, Sokal a ensuite fait un autre article dénonçant la supercherie et ridiculisant ainsi le comité de lecture. Dans la bibliographie de l'article canular figuraient un certain nombre de philosophes, sociologues, psychanalystes français parmi lesquels : Deleuze, Guattari, Lacan, Latour, Virilio, Derrida, Kristeva, Irigaray etc... ce fait là a porté le débat en France qui était ainsi montrée du doigt comme un lieu où la liaison entre mathématiques physique et sciences humaines pêchait par manque de rigueur. La polémique s'est d'ailleurs poursuivie en France et dans la francophonie par la publication, en septembre 1997, d'un livre intitulé ńimpostures intellectuellesż écrit par Alan Sokal et Jean de Bricmont (professeur de physique à Louvain en Belgique). Je laisse de côté la critique sur l'affaire Sokal et je vous renvoie aux sociologues Toulousains qui l'ont étudiée, Franck Cochoy et Michel Grossetti par exemple, il existe aussi une abondante bibliographie sur le sujet sur Internet. Mais il est évident que cette affaire a mis en évidence la difficulté qu'il y a à faire coexister certain types de discours scientifiques, ce qui constitue parfois l'essence de mon métier.

Une présentation différente de l'histoire des mathématiques

Ce qu'ont oublié les deux camps qui s'opposent dans la polémique induite par l'affaire Sokal, c'est que de bonnes parties des mathématiques ne sont pas nées de la modèlisation physique du monde mais de pratiques humaines. Abu Abd-Allah ibn Musa al'Khwarizmi a écrit en 830 Hisab al-jabr w'al-muqabala ce qui a donné le mot algèbre. Le livre en question décrit l'utilisation des techniques de calcul que l'on appelle algèbre pour des situations de partages d'argent et de biens pour des héritages ou des conflits commerciaux. L'arithmétique est en grande partie liée à l'utilisation des chiffres indiens (que l'on appelle arabes) et au système de numération décimal. L'utilisation de ces chiffres et de ce système de numération est essentiellement due à des problèmes de comptabilité. Quant à l'astronomie qui a longtemps été intimement liée aux mathématiques elle a souvent été utilisée pour mettre au point les calendriers pour l'agriculture mais aussi pour des raisons politiques et religieuses. Le mot ńcomputż qui a donné computer, computing, la commande compute de SPSS désignait le calcul de la date de Pâques dans le calendrier chrétien. Les probabilités ont deux origines les jeux de hasard et les problèmes de répartition avec des situations aléatoires ou les problèmes liés à l'assurance. Les statistiques ont leur éthymologie dans le mot état et son usage important a été le fait d'états organisés, à fin de recensement, impôts, problèmes de santé publique (épidémies de pestes). De par leur origine les statistiques ont, à mon avis, autant leur place dans l'économie, la géographie ou la sociologie que dans les mathématiques. À l'intérieur des statistiques une technique telle que l'analyse factorielle a été mise au point pour des raisons issues du champ de la psychologie. Pour une histoire de l'analyse factorielle je vous renvoie à l'excellent livre de Stephen Jay GOULD ńLa mal-mesure de l'hommeż. Le but de cette introduction est de vous faire comprendre que c'est souvent un processus de formalisation de l'approche du social et de l'humain qui a influencé le développement des mathématiques même si on peut légitimement admettre que les mathématiques existent indépendamment de l'activité humaine. Ce qui est absurde c'est de croire que toutes les mathématiques existent depuis que le monde physique qui nous entoure a été à peu près décrit correctement par des formules mathématiques par Galilée (1564-1642) ou Newton (1643-1727) et que les mathématiques n'ont jamais servi qu'à décrire la physique et non des entités sociales ou humaines.

Renversement de la polémique :

J'aurais pu intituler cet exposé ńFaut il se méfier de l'absence de statistiques dans les sciences humaines ?ż ou plutôt ń Faut il se méfier de l'absence de culture statistique chez les spécialistes de sciences humaines ?ż Le sous titre de l'exposé laisse supposer que c'est évidemment mon point de vue et qu'il faut présenter des méthodes pour augmenter le savoir statistiques des auditeurs... Mais je dirais aussi que lorsqu'il m'arrive (eh oui çà m'arrive) de lire des ouvrages de sciences humaines, je préfère qu'il y ait une justification quantitative. Cela est particulièrement vrai pour l'histoire les formulations vagues du genre ńune grande quantitéż ou ńparfoisż sont moins rassurantes qu'un chiffre ou une date, et l'absence de chiffres sérieux en histoire me semble être la porte ouverte à tous les mensonges et à tous les révisionnismes. Il est hors de question de tomber dans l'utilisation systèmatique des chiffres mais lorsqu'on les utilise, il me semble qu'il faut prendre des précautions, un des élèments dont je pense qu'il faut tenir compte c'est l'incertitude toujours liée aux informations recueillies. Les probabilités sont, à l'intérieur des mathématiques, la discipline qui étudie le hasard et l'incertitude, il me semble donc important d'intégrer un minimum de vision probabilisée dans les modélisations statistiques, de manière à tenter intégrer l'erreur dans les calculs plutôt que de nier la possibilité qu'il y ait une erreur. Mais surtout il faut considérer que cet exposé est préalable à d'autres exposés et que le but cette série est d'arriver à un exposé qui sera consacré aux modèles de type ńLISRELż ces modèles entrent pleinement dans le cadre d'une modèlisation probabiliste des données et il est donc nécessaire de s'imprégner, au préalable des concepts de bases de la statistique mathématique.

J'aimerais insister sur un nombre de points concernant la lecture des résultats mais aussi, avant cela, sur certains principes de codage des données. Le premier point à noter concernant le codage des données je relève que très souvent les données sont codées de manière discrète alors qu'elles sont susceptibles d'être codées de manière continue, les exemples les plus remarquables sont l'âge ou le revenu. Il est fréquent même que ces variables soient constituées de deux catégories les jeunes et les vieux ou les riches et les pauvres ce qui permet ainsi de constituer un tableau de khi2 particulièrement simple. Mais attention cette façon de coder fait perdre beaucoup d'information, surtout elle oblige à utiliser une méthode sur les variables qualitatives ce qui est moins simple d'un point de vue mathématique. Mais surtout, il faut savoir, qu'à l'intérieur d'un logiciel comme SPSS le recodage d'une variable quantitative en variable qualitative est hyper simple alors que l'inverse n'a pas de sens. Donc avec une variable quantitative il est possible éventuellement de faire des traitements sur variables qualitatives, après recodage préalable, mais aussi bien sur, des traitements sur variables quantitatives

La méthode qui en arrive en tête du hit parade des demandes statistiques auxquelles j'ai eu à faire face cette année 1997 c'est l'utilisation de projections de variables illustratives en ACP et AFC. Énormément d'utilisateurs m'ont effectivement demandé de faire une ACP ou une AFC sur leur variables avec représentation des groupes définis par d'autres variables qualitatives sur les graphiques de l'ACP ou de l'AFC afin de mener une interprétation globale de leurs données. En principe, pour faire des graphiques faciles à lire j'utilise le logiciel SPAD, mais je sais que l'utilisation de telles méthodes peut amener à de graves erreurs d'interprétation des données.

La première source d'erreur vient de l'absence de stabilité ństatistiqueż des méthodes, en effet la quasi totalité des résultats d'une ACP ou d'une AFC ne tient pas compte de la taille de l'échantillon (nombres de lignes dans le fichier de données, nombre de sujets, de cas etc...) alors que sur un fichier de 30 cas un seul individu peut définir le premier axe ! ! ! Les concepteurs du logiciels SPAD (Lebart et Morineau) ont mis au point le concept de valeur test qui permet de réintroduire un peu de statistique dans l'AFC mais ce concept est souvent négligé par les utilisateurs.

La source d'erreur la plus importante est toutefois plus profonde, plus inhérente à la construction mathématique de ces méthodes qui ne s'intéressent qu'aux liaisons deux à deux des variables. Si l'on a trois variables A, B et C on peut voir la liaison entre A et B entre A et C et entre B et C mais jamais l'action conjointe de A et B sur C. C'est en s'articulant sur ce genre de principe que l'on trouve de ńfaussesż liaisons et des interprétations douteuses.

Il existe un fichier de données réelles célébres pour montrer les erreurs que l'on peut commettre avec trois variables. Il s'agit du fichier concernant 4764 homicides jugés en Floride de 1973 à 1979. Ces données ont été publiés dans le New-York Times du 11 Mars 1979 et maintes fois utilisées par des statisticiens. Il y a 3 variables, la race du meurtrier, la race de la victime, et la condamnation à la fin du jugement. Il n'y a que deux races, tant pour les meurtriers que pour les victimes, les noirs et les blancs, il n'y que deux modalités pour la sentence, la peine de mort et les autres peines. On se trouve dans un cas très simple avec trois variables et deux modalités pour chaque variable. Ce fichier peut se présenter de la façon suivante :

Meurtrier	Victime	Peine de Mort	Autre Peine
Blanc	Blanche	72	2074
Blanc	Noire	0	111
Noir	Blanche	48	239
Noir	Noire	11	2209

On peut se poser 3 questions simples, celle qui était posée par le journal était la suivante ńqui est le plus souvent condamné à mort, parmi les meurtriers, les noirs ou les blancs ? ż

Note sociologique et éthique : la nature du problème posé, les variables relevées, l'appellation des variables ńrace du meurtrierż ou ńrace de la victimeż révélent la conception différentialiste de la société américaine qui s'oppose à ma vision personnelle des choses beaucoup plus universaliste. Pour une discussion anthropologique étendue sur le sujet je conseille vivement la lecture du livre d'Emmanuel Todd, ńLe destin des immigrészż. D'un point de vue statistique, j'aurais plutôt tendance à considérer la couleur de la peau comme une variable continue que comme une variable discrète à 2 modalités. Ceci étant les données sont traitées en ignorant mon point de vue personnel sur la question.

Le calcul est fort simple il y a 72 blancs condamnés à mort parmi 2257 meurtriers blancs soit 3,19% des meurtriers blancs qui sont condamnés à mort, il y a 59 noirs condamnés à mort parmi 2507 meurtriers noirs soit 2,35% des meurtriers noirs qui sont condamnés à mort. Comme 2,35% $<$ 3,19% la réponse est limpide les meurtriers blancs sont plus sévérement punis que les meurtriers noirs. Il est possible de calculer le Khi2 en croisant les deux variables :

```
MEURTRIER  by  SENTENCE

                    SENTENCE
            Count  |
           Exp Val |
           Row Pct |mort     autre
           Col Pct |                    Row
           Std Res |     1  |     2  | Total
MEURTRIER  --------+--------+--------+
               1   |    72  |  2185  |  2257
  blanc            |  62.1  |2194.9  | 47.4%
                   |  3.2%  | 96.8%  |
                   | 55.0%  | 47.2%  |
                   |   1.3  |   -.2  |
                   +--------+--------+
               2   |    59  |  2448  |  2507
  noir             |  68.9  |2438.1  | 52.6%
                   |  2.4%  | 97.6%  |
                   | 45.0%  | 52.8%  |
                   |  -1.2  |    .2  |
                   +--------+--------+
            Column     131     4633     4764
             Total    2.7%    97.3%   100.0%

      Chi-Square                   Value          DF              Significance
--------------------         -----------        ----              ------------

Pearson                          3.10905          1                  .07786
```

Il n'est pas significatif au niveau 5% mais il l'est presque et la conclusion est bien que les blancs sont plus souvent condamnés à mort que les noirs en cas d'homicide.

Mais attention si on fait deux tableaux séparés en fonction de la race de la victime on obtient le premier tableau si la victime est blanche :

Meurtrier	Victime	Peine de Mort	Autre Peine
Blanc	Blanche	72	2074
Noir	Blanche	48	239

Il y a 72 meutriers blancs condamnés à mort sur 2146 meurtiers de victimes blanches soit 3,36% des meurtriers. Pour les meurtriers noirs de victimes blanches il y 48 condamnés à mort sur 287 jugés soit 16,72%. Incontestablement 16,72 % > 3,36% donc si la victime est blanche les meurtriers noirs sont beaucoup plus sévérement punis que les meurtriers blancs. On peut dans ce cas là faire le tableau de Khi2 et on obtient la chose suivante :

```
MEURTRIER  by  SENTENCE

                    SENTENCE
           Count  |
           Exp Val |
           Row Pct |mort     autre
           Col Pct |                    Row
           Std Res |     1  |     2  | Total
MEURTRIER  --------+--------+--------+
               1  |    72  |  2074  |  2146
  blanc           | 105.8  |2040.2  | 88.2%
                  |  3.4%  | 96.6%  |
                  | 60.0%  | 89.7%  |
                  |  -3.3  |    .7  |
                  +--------+--------+
               2  |    48  |   239  |   287
  noir            |  14.2  | 272.8  | 11.8%
                  | 16.7%  | 83.3%  |
                  | 40.0%  | 10.3%  |
                  |   9.0  |  -2.0  |
                  +--------+--------+
            Column     120     2313     2433
             Total    4.9%    95.1%   100.0%

      Chi-Square                  Value          DF              Significance
--------------------          -----------       ----              ------------

Pearson                         96.50227         1                   .00000
```

Le Khi2 est très hautement significatif, et de manière évidente la peine de mort est surtout appliquée aux meurtriers noirs.

En se restreignant aux cas où la victime est noire on obtient le tableau suivant :

Meurtrier	Victime	Peine de Mort	Autre Peine
Blanc	Noire	0	111
Noir	Noire	11	2209

Il y a 0 meurtriers blancs qui ont été condamnés à mort pour 111 jugés soit 0% de condamnation à mort. Il y a 11 condamnés à mort parmi les 2310 meurtriers noirs jugés soit 0,48% de condamnation à mort. Il apparait donc que 0,48% > 0% et que donc si la victime de l'homicide est de race noire la justice de Floride condamne plus souvent les meurtriers noirs à la peine de mort que les meurtriers blancs. En faisant le calcul du Khi2 on obtient la chose suivante :

```
MEURTRIER  by  SENTENCE

                    SENTENCE
            Count  |
           Exp Val |
           Row Pct |mort     autre
           Col Pct |                     Row
           Std Res |     1  |     2  | Total
MEURTRIER  --------+--------+--------+
                1  |     0  |   111  |   111
  blanc            |    .5  | 110.5  |  4.8%
                   |   .0%  |100.0%  |
                   |   .0%  |  4.8%  |
                   |   -.7  |    .0  |
                   +--------+--------+
                2  |    11  |  2209  |  2220
  noir             |  10.5  |2209.5  | 95.2%
                   |   .5%  | 99.5%  |
                   |100.0%  | 95.2%  |
                   |    .2  |    .0  |
                   +--------+--------+
            Column      11     2320     2331
             Total     .5%    99.5%   100.0%

      Chi-Square                 Value         DF              Significance
--------------------          -----------     ----             ------------

Pearson                           .55261        1                  .45725
```

Le Khi2 n'est pas du tout significatif. Mais la situation est paradoxale, sur l'ensembe des jugements les meurtriers blancs sont plus souvent condamnés à mort, mais si on considère la variable ńrace de la victimeż, on constate que pour les deux races de victimes ce sont les meurtriers noirs qui sont plus le souvent condamnés à mort. La considération de la variable ńrace de la victimeż n'était pas évidente à priori et celle ci pourrait être considérée comme parasite par rapport à la question initiale : ńqui des meurtriers noirs ou blancs sont le plus souvent condamnés à mort ?ż

Si le résultat est paradoxal c'est bien parce que la variable, ńrace de la victimeż est très importante pour s'en convaincre il suffit de regarder les liaisons qui existe entre cette variable et les deux variables ńrace du meurtrierż et sentence. Pour cela il suffit de faire les deux tableaux de Khi2 correspondant.

D'abord observons la liaison entre ńrace de la victimeż et ńrace du meurtrierż.

```
MEURTRIER  by  VICTIME

                    VICTIME
            Count  |
           Exp Val |
           Row Pct |blanc    noir
           Col Pct |                      Row
           Std Res |     1  |     2  | Total
MEURTRIER  --------+--------+--------+
                 1 |  2146  |   111  |  2257
  blanc            |1152.7  |1104.3  | 47.4%
                   | 95.1%  |  4.9%  |
                   | 88.2%  |  4.8%  |
                   |  29.3  | -29.9  |
                   +--------+--------+
                 2 |   287  |  2220  |  2507
  noir             |1280.3  |1226.7  | 52.6%
                   | 11.4%  | 88.6%  |
                   | 11.8%  | 95.2%  |
                   | -27.8  |  28.4  |
                   +--------+--------+
            Column    2433     2331     4764
             Total   51.1%    48.9%   100.0%

      Chi-Square                Value          DF              Significance
--------------------         -----------      ----             ------------

Pearson                      3324.59864         1                 .00000
```

Le Khi2 est exceptionnellement fort et ultra significatif et la conclusion est évidente il y a surtout des homicides à l'intérieur de chaque race, les meurtriers blancs tuent essentiellement des victimes blanches et les meutriers noirs des victimes noires.

Regardons maintenant la liaison entre ńrace de la victimeż et sentence.

```
VICTIME  by  SENTENCE

                    SENTENCE
            Count  |
           Exp Val |
           Row Pct |mort     autre
           Col Pct |                     Row
           Std Res |     1  |     2  | Total
VICTIME    --------+--------+--------+
                1  |   120  |  2313  |  2433
  blanc            |  66.9  |2366.1  | 51.1%
                   |  4.9%  | 95.1%  |
                   | 91.6%  | 49.9%  |
                   |   6.5  |  -1.1  |
                   +--------+--------+
                2  |    11  |  2320  |  2331
  noir             |  64.1  |2266.9  | 48.9%
                   |   .5%  | 99.5%  |
                   |  8.4%  | 50.1%  |
                   |  -6.6  |   1.1  |
                   +--------+--------+
            Column     131     4633     4764
             Total    2.7%    97.3%   100.0%

      Chi-Square                  Value          DF             Significance
--------------------          -----------       ----            ------------

Pearson                         88.56195          1                .00000
```

Le Khi2 est fort, très significatif, beaucoup plus que pour la liaison entre ńrace du meutrierż et sentence, et il est remarquable de constater que le meurtrier d'un blanc est condamné à mort dans près de 5% des cas alors que le meurtrier d'un noir est condamné à mort dans 0,5% des cas soit 10 fois moins souvent.

La liaison entre ńrace du meurtrierż et sentence n'est qu'une ńconséquenceż des deux liaisons précédentes, les meurtriers blancs tuant surtout des victimes blanches et ce type d'homicide étant beaucoup plus souvent puni de la peine de mort les meurtriers blancs sont plus souvent punis de la peine de mort.

Si on fait une analyse factorielle des correspondances multiples (AFCM) en mettant les 3 variables on obtient sur deux axes (95% de l'inertie)

Le premier axe est défini par les deux variables ńrace du meurtrierż et ńrace de la victimeż le deuxième est presqu'uniquement défini par la modalité ńpeine de mortż de la variable sentence.

```
                          AXE 1  *  AXE 2
 AXE 2
  6 Peine de Mort -----------------------------------------------
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
  3  +                              !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
    !                               !                           !
  0  +                              !                           !
    !--Victime Blanche--------------+--------Meurtrier Noir VictimeNoire
    +--Meurtrier Blanc--------Autre condamnation----+-----------
   -1.2            -.6              .0              .6
                                                             AXE 1
```

On retrouve dans l'AFCM la relation la plus forte parmi les 3 variables sur le premier axe, c'est à dire la liaison entre la race du meurtier et la race de la victime. Cela n'est pas étonnant sachant que l'AFCM est calculé à partir du tableau de Burt que voici :

```
     ! MeBl MeNo ! ViBl ViNo ! Mort Autr !
-----+-----------+-----------+-----------+
MeBl ! 2257    0 !
MeNo !    0 2507 !
-----+-----------+-----------+
ViBl ! 2146  287 ! 2433    0 !
ViNo !  111 2220 !    0 2331 !
-----+-----------+-----------+-----------+
Mort !   72   59 !  120   11 !  131    0 !
Autr ! 2185 2448 ! 2313 2320 !    0 4633 !
-----+-----------+-----------+-----------+
     ! MeBl MeNo ! ViBl ViNo ! Mort Autr !
```

On a dans ce tableau les trois tableaux précédemment utilisés pour les tableaux de khi2 des variables prises deux à deux. L'AFCM n'est que l'analyse factorielle des correspondances simple (AFC) de ce tableau particulier. Les résultats de l'AFCM ne permettent pas de voir que les meurtriers noirs sont plus souvent condamnés à mort que les meurtiers blancs si on considére la variable ńrace de la victimeż alors que cette variable figure dans l'analyse. En jouant sur le statut de variable active ou de variable illustrative dans l'AFCM , ce qui fait 3 AFCM possibles on obtient toujours des résultats similaires

Voila ce que l'on obtient avec la sentence en illustratif :

```
                         AXE 1  *  AXE 2
AXE 2
    ViNo--------------------------------------------------------MeBl
     !                            !                                !
     !                            !                                !
     !                            !                                !
 .0  +--------------------------Autr-------------------------------!
     !                            !                                !
     !                            !                                !
     !                            !                                !
     !                            !                                !
    MeNo                          !                           ViBl !
     !                            !                                !
     !                            !                                !
     !                            !                                !
     !                            !                                !
-.6  +                            !                                !
     !                            !                                !
     !                            !                                !
     !                            !                                !
     !                            !                                !
     !                            !                                !
     !                            !                                !
     !                            !                                !
      -----------+----------------+-------------Mort-------------+
                -.5              .0               .5              1.0
                                                                AXE 1
```

Voila ce que l’on obtient avec ńla race de la victimeż en illustratif :

```
                           AXE 1  *  AXE 2
 AXE 2
       ------------------MeBl-------------------------------------
      !           !      ViBl                                     !
      !           !                                               !
      !      Autr !                                               !
  .0  +-----------+-----------------------------------------------!
      !           !                                               !
     ViNoMeNo     !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
-2.5  +           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
      !           !                                               !
       -----------+---------------+---------------+--------------Mort
                  .0             1.5             3.0
                                                               AXE 1
```

et voila ce que l'on obtient avec ńla race du meurtrierż en illustratif :

```
                          AXE 1  *  AXE 2
AXE 2
 4  +----------------------------------------------------------Mort
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
 2  +       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
    !       !                                                   !
   ViNoMeno !                                                   !
    !       !                                                   !
    !       !                                                   !
    !--Autr-+---------------------------------------------------!
 0  +       !                                                   !
    !       !                                                   !
     -------+--MeBlViBl-----+---------------+---------------+---
            .0             1.5             3.0             4.5
                                                             AXE 1
```

Dans tous les cas la ńpeine de mortż apparait du côté des meurtriers blancs et des victimes blanches, même si elle est toujours loin du centre, cela est dû au fait que la sentence peine de mort est quand même assez rare, par contre la modalité ńautre sentenceż est toujours près du centre car très fréquente. Les variables avec répartition très déséquilibrées présente souvent ce type de situation en AFCM. Il apparait que les variables ńrace de la victimeż et ńrace du meurtrierż sont d'un type et que la variable sentence est d'un autre type, de manière évidente les races sont connues avant le procès et la sentence après (en principe...).

Une première façon de contourner la difficulté posée par cette situation consiste à créer une variable à 4 modalités que l'on appellera ńtype d'homicideż, le type 1 correspondant à la situation Meurtrier Blanc et Victime Blanche, le type 2 : Meurtrier Blanc et Victime Noire, le type 3 : Meurtrier Noir et Victime Blanche et enfin le type 4 Meurtrier Noir et Victime Noire. On peut alors étudier la liaison entre cette variable et la sentence du jugement. Cela donne le tableau suivant.

```
TYPE  Type d'homicide  by  SENTENCE

                    SENTENCE
            Count  |
          Exp Val  |
          Row Pct  |mort     autre
          Col Pct  |                    Row
          Std Res  |     1  |     2  | Total
TYPE      --------+--------+--------+
                1  |    72  |  2074  |  2146
  MeBla/ViBla      |  59.0  |2087.0  | 45.0%
                   |  3.4%  | 96.6%  |
                   | 55.0%  | 44.8%  |
                   |   1.7  |   -.3  |
                   +--------+--------+
                2  |     0  |   111  |   111
  MeBla/ViNoi      |   3.1  | 107.9  |  2.3%
                   |   .0%  |100.0%  |
                   |   .0%  |  2.4%  |
                   |  -1.7  |    .3  |
                   +--------+--------+
                3  |    48  |   239  |   287
  MeNoi/ViBla      |   7.9  | 279.1  |  6.0%
                   | 16.7%  | 83.3%  |
                   | 36.6%  |  5.2%  |
                   |  14.3  |  -2.4  |
                   +--------+--------+
                4  |    11  |  2209  |  2220
  MeNoi/ViNoi      |  61.0  |2159.0  | 46.6%
                   |   .5%  | 99.5%  |
                   |  8.4%  | 47.7%  |
                   |  -6.4  |   1.1  |
                   +--------+--------+
           Column     131     4633     4764
            Total    2.7%    97.3%   100.0%
```

Chi-Square	Value	DF	Significance
Pearson	257.86670	3	.00000

Il apparait nettement qu'il y a une liaison entre les 2 variables, le Khi2 est très significatif, l'interprétation est simple si la victime est noire le risque de peine de mort est peu élevé, si la victime est blanche le risque de peine de mort est élevé et si le meurtrier est noir le risque devient encore plus important.

Une autre approche assez différente du Khi2 consiste à considérer la variable sentence comme une variable à expliquer et les deux variables ńrace de la victimeż et ńrace du meurtrierż comme deux variables explicatives, la variable sentence ayant deux modalités on est typiquement dans le cas d'application de la régression logistique. On dit que la probabilité d'être condamné à mort est fonction de la race du meurtrier et de la race de la victime.

Pour traiter les données dans SAS il suffit de rentrer un fichier avec 4 lignes, colonnes

```
meurtrier  victime  peinmort jugement
Blanc Blanche 72  2146
Blanc Noire    0   111
Noir  Blanche 48   287
Noir  Noire   11  2220

proc genmod;
class meurtrie victime ;
model peinmort/jugement=   meurtrie victime
                 /dist =bin link=logit waldci type3 obstat;
run;
```

On obtient les résultats suivants :

```
                  Criteria For Assessing Goodness Of Fit

           Criterion              DF           Value       Value/DF

           Deviance                1          0.1898         0.1898
           Scaled Deviance         1          0.1898         0.1898
           Pearson Chi-Square      1          0.0955         0.0955
           Scaled Pearson X2       1          0.0955         0.0955
           Log Likelihood          .       -514.2241              .
```

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-5.3111	0.3023	308.7567	0.0001
MEURTRIE	Blanc	1	-1.7590	0.1980	78.8855	0.0001
MEURTRIE	Noir	0	0.0000	0.0000	.	.
VICTIME	Blanche	1	3.7081	0.3404	118.6915	0.0001
VICTIME	Noire	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

Normal Confidence Intervals For Parameters

Two-Sided Confidence Coefficient: 0.9500

Parameter	Confidence Limits	
PRM1	Lower	-5.9035
PRM1	Upper	-4.7186
PRM2	Lower	-2.1472
PRM2	Upper	-1.3709
PRM4	Lower	3.0410
PRM4	Upper	4.3753

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
MEURTRIE	1	67.5702	0.0001
VICTIME	1	168.3467	0.0001

Observation Statistics

PEINMORT	EFF	Pred	Xbeta	Std	HessWgt	Lower
72	2146	0.0335	-3.3619	0.1199	69.4964	0.0267
0	111	0.000849	-7.0701	0.3604	0.0942	0.000419
48	287	0.1676	-1.6029	0.1579	40.0348	0.1287
11	2220	0.004912	-5.3111	0.3023	10.8521	0.002723

Observation Statistics

Upper	Resraw	Reschi	Resdev
0.0420	0.0943	0.0113	0.0113
0.001720	-0.0943	-0.3072	-0.4343
0.2153	-0.0943	-0.0149	-0.0149
0.008848	0.0943	0.0286	0.0286

L'ajustement du modèle est excellent, les deux variables ńrace de la victimeż et ńrace du meurtrierż ont un effet très fort, mais c'est la ńrace de la victimeż qui est la plus importante. Les risques de condamnation à la peine sont prédits avec une grande précision les résidus sont ridiculement faibles.

Le choix du modèle de régression logistique a été fait avec glim4. Les variables ńrace du meurtrierż et ńrace de la victimeż ont été codées 1 pour blanche et 2 pour noir, dans l'estimation des paramètres c'est la valeur 1 qui est prise comme référence et l'effet de la variable se mesure par l'effet différentiel entre la valeur 1 et la valeur 2.

Voici la démarche on essaie les 4 modèles suivants : 1) Aucun effet des variables 2) Effet de la variable ńrace du meurtrierż 3) Effet de la variable ńrace de la victimeż 4) Effets simples de ces deux variables ensemble.

Le modèle avec interaction est nécessairement ńsaturéż et n'offre donc aucun intérêt.

```
  $fit$
 scaled deviance =  171.64 at cycle 4
    residual df =    3

  $display e$
          estimate        s.e.     parameter
     1      -3.566     0.08826      1

  $display r$
   unit   observed    out of    fitted   residual
      1         72      2146     59.01      1.714
      2          0       111      3.05     -1.772
      3         48       287      7.89     14.477
      4         11      2220     61.05     -6.495
```

Le modèle sans effet des variables est surtout intéressant pour les résidus qui permettent de voir l'analogie entre le modèle logit et le Khi2, en effet les résidus du modèle logit sont pratiquement égaux aux résidus standardisés du tableau croisé entre la variable à quatre modalités ńtype d'homicideż et sentence en considérant uniquement la modalité ńpeine de mortż. En regardant ces résidus on arrive exactement à la même conclusion que précédemment.

```
  $fit +meurtr$
 scaled deviance =  168.54 (change =   -3.105) at cycle 5
    residual df =    2    (change =   -1    )

  $display e$
          estimate        s.e.     parameter
     1      -3.413      0.1198      1
```

```
  2     -0.3128       0.1778      MEURTR(2)


$display r$
 unit   observed    out of    fitted   residual
    1         72      2146     68.46      0.435
    2          0       111      3.54     -1.913
    3         48       287      6.75     16.060
    4         11      2220     52.25     -5.775
```

En ajoutant la variable ńrace du meurtrierż on n'améliore que faiblement l'ajustement du modèle et cette variable n'apparait pas vraiment significative, il apparait raisonnable d'essayer l'autre variable ńrace de la victimeż.

```
  $fit -meurtr+victime$
scaled deviance =  67.760 (change =   -100.8) at cycle 4
    residual df =   2     (change =      0  )

  $display e$
        estimate        s.e.     parameter
    1      -2.959     0.09361      1
    2      -2.393      0.3164      VICTIME(2)


  $display r$
   unit   observed    out of    fitted   residual
      1         72      2146    105.84     -3.374
      2          0       111      0.52     -0.725
      3         48       287     14.16      9.226
      4         11      2220     10.48      0.162
```

Enlever la variable ńrace du meurtrierż et ajouter la variable ńrace de la victimeż fait chuter vertigineusement la déviance et améliore condidérablement la qualité d'ajustement. La variable ńrace de la victimeż est très significative mais on constate que les résidus des observations 1 et 3 sont élevés suggérant que la race du meurtrier pourrait avoir un effet dans le cas où la victime est un blanc. L'interprétation étant pour la variable ńrace de la victimeż que l'homicide d'une victime blanche entraine plus souvent la peine de mort.

```
  $fit +meurtr$
scaled deviance =  0.18983 (change =   -67.57) at cycle 3
    residual df =  1       (change =    -1   )

  $display e$
        estimate        s.e.     parameter
    1      -3.362      0.1199      1
    2      -3.708      0.3403      VICTIME(2)
    3       1.759      0.1980      MEURTR(2)


  $display r$
   unit   observed    out of    fitted   residual
      1         72      2146     71.91      0.011
      2          0       111      0.09     -0.307
```

3	48	287	48.09	-0.015
4	11	2220	10.91	0.029

On voit qu'avec les deux variables ńrace de la victimeż et ńrace du meurtrierż on a un excellent modèle les résidus sont extrêmement faibles. La déviance vaut pratiquement 0 et il est inutile de mettre l'interaction. D'un point de vue modèlisation, on voit que c'est à l'intérieur de ce modèle que l'on doit faire l'interprétation. Les meurtres de victimes blanches sont plus souvent punis de mort et les meurtriers noirs sont plus souvent punis de mort.

Dans le cadre d'une comparaison entre l'AFCM et une méthode de modélisation il serait judicieux d'utiliser une approche par le modèle log-linéaire. En effet si on considère que les trois variables sont sur le même plan et qu'il n'y a pas de variables à expliquer et de variables explicatives, le modèle log-linéaire est plus approprié que le modèle logit. J'ai mené cette modèlisation de façon non conforme en partant du modèle blanc en allant vers le modèle saturé. J'ai codé la variable sentence 1 pour la peine de mort et 2 pour les autres peines les autres variables ont été codées comme précedemment, la recherche du modèle optimal a été faite avec glim4 voici ce que cela a donné.

```
? $f$
scaled deviance =  9524.4 at cycle 5
    residual df =     7

? $display e$
          estimate        s.e.     parameter
     1       6.389     0.01448      1

? $display r$
   unit   observed    fitted   residual
      1         72    595.50    -21.452
      2       2074    595.50     60.587
      3          0    595.50    -24.403
      4        111    595.50    -19.854
      5         48    595.50    -22.436
      6        239    595.50    -14.609
      7         11    595.50    -23.952
      8       2209    595.50     66.119

? $f +meurtr+victime+sentence$
scaled deviance =  4104.7 (change =   -5420.) at cycle 5
    residual df =     4   (change =      -3 )

? $display e$
          estimate        s.e.     parameter
     1       3.456     0.08980      1
     2      0.1051     0.02901      MEURTR(2)
     3    -0.04283     0.02898      VICTIME(2)
     4       3.566     0.08857      SENTENCE(2)

? $display r$
   unit   observed    fitted   residual
      1         72     31.70      7.159
      2       2074   1120.97     28.465
```

3	0	30.37	-5.511
4	111	1073.97	-29.384
5	48	35.21	2.156
6	239	1245.13	-28.513
7	11	33.73	-3.914
8	2209	1192.93	29.418

```
? $f +meurtr.victime$
scaled deviance =  171.64 (change =   -3933.) at cycle 4
    residual df =    3    (change =      -1 )

? $display e$
          estimate        s.e.     parameter
     1       4.078     0.08837      1
     2      -2.012     0.06284      MEURTR(2)
     3      -2.962     0.09733      VICTIME(2)
     4       3.566     0.08815      SENTENCE(2)
     5       5.008      0.1158      MEURTR(2).VICTIME(2)

? $display r$
   unit   observed    fitted   residual
      1         72     59.01      1.690
      2       2074   2086.99     -0.284
      3          0      3.05     -1.747
      4        111    107.95      0.294
      5         48      7.89     14.277
      6        239    279.11     -2.401
      7         11     61.05     -6.406
      8       2209   2158.95      1.077

? $f +meurtr.sentence$
scaled deviance =  168.54 (change =   -3.105) at cycle 5
    residual df =    2    (change =   -1    )

? $display e$
          estimate        s.e.     parameter
     1       4.226      0.1179      1
     2      -2.316      0.1838      MEURTR(2)
     3      -2.962     0.09734      VICTIME(2)
     4       3.413      0.1198      SENTENCE(2)
     5       5.008      0.1158      MEURTR(2).VICTIME(2)
     6      0.3127      0.1776      MEURTR(2).SENTENCE(2)

? $display r$
   unit   observed    fitted   residual
      1         72     68.46      0.428
      2       2074   2077.54     -0.078
      3          0      3.54     -1.882
      4        111    107.46      0.342
      5         48      6.75     15.870
      6        239    280.25     -2.464
```

7	11	52.25	-5.707
8	2209	2167.75	0.886

```
? $f -meurtr.sentence+victime.sentence$
scaled deviance =  67.760 (change =   -100.8) at cycle 4
    residual df =   2     (change =      0  )

? $display e$
          estimate        s.e.     parameter
     1       4.662     0.09157      1
     2      -2.012     0.06285      MEURTR(2)
     3      -5.309      0.3284      VICTIME(2)
     4       2.959     0.09361      SENTENCE(2)
     5       5.008      0.1158      MEURTR(2).VICTIME(2)
     6       2.393      0.3164      VICTIME(2).SENTENCE(2)

? $display r$
   unit   observed    fitted   residual
      1         72    105.84     -3.290
      2       2074   2040.16      0.749
      3          0      0.52     -0.724
      4        111    110.48      0.050
      5         48     14.16      8.996
      6        239    272.84     -2.049
      7         11     10.48      0.162
      8       2209   2209.52     -0.011

? $f +meurtr.sentence$
scaled deviance =  0.18983 (change =   -67.57) at cycle 3
    residual df =  1       (change =    -1   )

? $display e$
          estimate        s.e.     parameter
     1       4.275      0.1179      1
     2     -0.4022      0.1860      MEURTR(2)
     3      -6.637      0.3537      VICTIME(2)
     4       3.362      0.1199      SENTENCE(2)
     5       5.153      0.1188      MEURTR(2).VICTIME(2)
     6      -1.759      0.1980      MEURTR(2).SENTENCE(2)
     7       3.708      0.3403      VICTIME(2).SENTENCE(2)

? $display r$
   unit   observed    fitted   residual
      1         72     71.91      0.011
      2       2074   2074.09     -0.002
      3          0      0.09     -0.307
      4        111    110.91      0.009
```

```
5          48     48.09      -0.014
6         239    238.91       0.006
7          11     10.91       0.029
8        2209   2209.09      -0.002
```

Le modèle auquel on aboutit est le modèle avec toutes les interactions d'ordre 2 mais sans l'interaction d'ordre 3. On constate que le critère d'adéquation du modèle (la déviance) vaut exactement la même chose que dans le cas du modèle logit et qu'il en est de même pour les valeurs ajustées et l'interprétation des résultats est similaire. La similitude des résultats entre le modèle logit et le modèle log-linéaire est explicable par le lien très fort entre la loi de Poisson et la loi multinomiale en général et donc la loi binomiale en particulier et par le fait que dans cet exemple toutes les variables ont exactement 2 modalités. Le modèle log-linéaire permet facilement de classer les interactions entre les 3 variables.

Il va de soi que l'on n'a pas dans ces données des variables essentielles telles que type de meurtre, préméditation ou non, récidive ou non, âge des victimes ou des meurtriers etc. Ces données n'ont que valeur d'exemple pour montrer diverses approches statistiques. J'ai vu, pour la première fois, ces données dans le cadre d'une présentation des modèles graphiques mais ne disposant pas de logiciel me permettant de faire des modèles graphiques je n'en ai pas fait. Il est sans doute possible de traiter ces données différemment mais tant pis. Je n'ai pas intégré dans ce texte toutes les sorties des logiciels utilisés, mais seulement celles qui m'ont parues pertinentes. Les tableaux de Khi2 ont été faits avec SPSS, les AFC avec SPAD, la recherche de modèle logit avec Glim et la présentation détaillé des résultats du modèle logit avec SAS. Le but de ce texte est de mettre en garde, les utilisateurs qui me demandent des ACP et des AFC, contre certaines dérives liées à des méthodes fondées sur les liaisons deux à deux entre variables. Il va de soi qu'il existe des tas d'articles en statistiques dédiés à la comparaison de méthodes statistiques sur les mêmes fichiers de données, il en existe notamment sur la comparaison entre l'AFC et le modèle log-linéaire. La plupart de ces articles s'intéressent aux comparaisons mathématiques entre les méthodes, ce texte est dénué de réflexions mathématiques et s'adresse à un public d'utilisateurs non mathématiciens.

Pour donner une idée simple qui permette de voir une analogie avec le ńparadoxeż des données traitées j'ai pensé à l'exemple suivant : 2 élèves A et B passent leur baccalauréat, l'élève A a des meilleures notes que l'élève B à toutes les épreuves mais A est collé et B reçu car B a plus de 10 de moyenne et A collé car il a moins de 10 de moyenne. L'explication résidant dans le fait qu'ils ne sont pas dans la même filière. Simplifions à l'extrême la réalité et imaginons un baccalauréat scientifique avec un coefficient de 5 pour les épreuves de sciences (notées globalement) et un coefficient de 2 pour les épreuves de littérature (elles aussi notées globalement) et un baccalauréat littéraire avec les cofficients inversés (i.e. 2 pour les sciences et 5 pour les lettres). Supposons que l'élève B a eu 15/20 en sciences et 1/20 en littérature cela lui fait $(5*15+2*1)/7=11$ de moyenne et il est reçu, supposons que A a eu 19/20 en sciences et 5/20 en littérature cela lui fait $(2*19+5*5)/7=9$ et il est collé alors que $19>15$ et $5>1$.

Il existe aussi fréquemment en régression linéaire simple, des paradoxes de même nature que celui étudié ici mais plus faciles à visualiser.

La Floride a une proportion de noirs qui s'élève à 14%. À l'époque considérée la population devait s'élever à environ 6 millions d'habitants. Pour mémoire la France avait à cette époque plus de 50 millions d'habitants, la peine de mort a été abolie en 1981, la France était le dernier pays de la CEE à avoir la peine de mort, le nombre de condamnations à la peine capitale en France sur la période 1973-1979 ne doit pas dépasser la dizaine. Il est évident qu'un simple fichier de données peut amener à se poser des questions sociologiques, historiques ou anthropologiques sur la comparaison entre l'Europe et les États Unis d'Amérique.