Confusion de variables

Joseph Saint Pierre, Université Toulouse

Rencontre ingénieurs statisticiens- Lundi 2 Février 2015

Un problème important et souvent rencontré dans ma pratique des statistiques se situe dans l'utilisation des croisements de variables deux à deux, même lorsque les variables sont plus nombreuses. Le cas le plus simple consiste à traiter 3 variables A, B et C et à regarder 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 ou de A et C sur B etc...

Ainsi on trouve parfois de fausses liaisons et cela peut conduire à des interprétations douteuses.

Un exemple populaire consiste à trouver une liaison entre le fait de dormir avec ses chaussures et d'avoir mal à la tête au réveil en oubliant la consommation alcoolique qui peut expliquer plus raisonnablement les deux autres.

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 immigrésż. 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.

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.

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. 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.

Il s'agit là de données très fortement utilisées, mais j'ai souvent rencontré des phénomènes comparables sur des données réelles.

Un exemple sur l'université Toulouse 3

Un document intitulé « La différenciation sexuée des filières de formation et des carrières des personnels à l'Université Toulouse III - Paul Sabatier »

J'ai regardé les deux tableaux de la page 44 de ce document Je me suis particulièrement intéressé à la différence entre deux domaines, la physique et la médecine. Ces deux domaines se caractérisent par les deux plus faibles proportions de femmes parmi les professeurs 2 femmes sur 30 en physique soit 6,7% et 20 femmes en médecine sur 170 soit 11,8%. Première remarque il est possible de calculer un intervalle de confiance pour des proportions et cela serait souhaitable pour les sondages électoraux. Les deux intervalles pour ces proprtions se chevauchent très fortement et la différence entre 6,7% et 11,8% n'est donc pas significative cela peut se calculer facilement en ligne (hommes/femmes) avec le statut (professeurs/maîtres de conférences) pour les deux disciplines.

Pour la physique on obtient le tableau :

	H	F
P	28	2
MC	29	6

Pour la médecine on obtient le tableau :

	H	F
P	150	20
MC	37	53

La différence entre les deux tableaux est très forte, compte tenu des faibles effectifs j'ai utilisé le test exact de Fisher pour mesurer la liaison. Il n'y a pas de liaison significative en physique entre le sexe et le statut alors qu'elle est extrêment forte en médecine. Il y a peu de femmes professeurs en physique mais il y a peu de femmes maîtres de conférences en physique, en revanche il y a beaucoup d'hommes professeurs en médecine et il y assez peu d'hommes maîtres de conférences en médecine.

Pour analyser les trois dimensions il faut regarder le tableau croisant le statut avec le domaine obtient le tableau

	Méd	Phy
P	170	30
MC	90	35

Ce tableau donne une liaison significative il y a proportionnellement beaucoup plus de professeurs en médecine. Cette liaison ne concerne pas directement l'égalité entre les hommes et les femmes mais elle me semble pertinente indirectement. Les effectifs d'enseignants en médecine est très important et c'est là qu'existe la plus forte discrimination entre les hommes et les femmes pour l'accès aux postes les plus prestigieux.

Ces tableaux peuvent être regardés avec les tableaux concernant les étudiants, d'après le tableau de la page 9 du document le pourcentage de femmes étudiantes en médecine est passé entre 1998 et 2012 de 54% à 66,2% pour la physique la proportion de femmes est passé dans la même période de 34% à 23,3%.

J'ai vu un phénomène de confusion de variables sur un fichier de données en 1995 avec de nombreuses variables, la confusion portait sur trois variables, sexe, âge et niveau d'étude. En effet dans la population les personnes âgées sont surtout des femmes en raison de leur plus grande longévité, et le niveau d'étude est plus élevé chez les jeunes générations, ce phénomène étant plus net en 1995 qu'en 2015. Dans l'échantillon représentatif de la population régionale, les femmes avaient plutôt des niveaux d'étude faibles, comparativement en raison des personnes âgées, alors que déjà en 1995 chez les jeunes le niveau d'étude des femmes était plutôt supérieur à celui des hommes.

On attribue à David Spiegelhalter la citation suivante :

« There are a lot of small data problems that occur in big data, they don't disappear because you've got lots of the stuff. They get worse. »

Notamment dans l'article « Big Data : Are we making a big mistake ? »de Tim Harford

Je pense que la confusion des variables est facile à détecter sur des nombres réduits de variables, mais l'augmentation du nombre de variables dans les mégadonnées n'a aucune chance de faire disparaitre ce phénomène.

La page wikipedia en anglais qui aborde le sujet de la confusion en statistique, présente quelques idées qui m'avaient guidé lors de mon exposé sur les statistiques et les données massives du 27 janvier 2014

http ://en.wikipedia.org/wiki/Confounding

J'avais pensé à l'utilisation des modèles graphiques. C'est lors d'un exposé sur les modèles graphiques que pour la première fois j'ai entendu parler du paradoxe de Simpson avec les données de l'exemple sur les meurtres en Floride,