Confusion de variables et données massives

Joseph Saint Pierre

Université de Toulouse

rencontre ingénieurs statisticiens 2 février 2015

des petites données

4764 homicides jugés en Floride de 1973 à 1979
3 variables
Couleur du meurtrier (Blanc/Noir)
Couleur de la victime (Blanc/Noir)
Peine (Mort/autres)

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

Question :
qui est le plus souvent condamné à mort, parmi les meurtriers, les noirs ou les blancs ?
Blancs 72 sur 2257 3,19%
Noirs 59 sur 2507 2,35%

Les meurtriers blancs seraient plus souvent condamnés à mort.
Mais la différence n'est pas significative.

La variable couleur de la victime n'est prise en compte
3 variables :
Tableau pour les victimes blanches

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

Blancs 72 sur 2146 3,36%
Noirs 48 sur 287 16,72%

Si la victime est blanche, les meurtriers noirs sont plus souvent condamnés à mort.

Le test est très significatif.

Tableau pour les victimes noires.

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

Blancs 0 sur 111 0%
Noirs 11 sur 2209 0,48%

Si la victime est noire, les meurtriers noirs sont plus souvent condamnés à mort. Le test n'est pas significatif.

Les deux tableaux sont différents, les pourcentages sont plus faibles. Le tableau initial permet aussi de percevoir la liaison entre Tableau entre couleur victime et sentence

Victime	Peine de Mort	Autre Peine
Blanche	120	2313
Noire	11	2320

Blancs 120 sur 2433 4,93%
Noirs 11 sur 2331 0,47%

Le test est hautement significatif, il y a beaucoup plus de peines de morts pour les meurtiers de blancs.

Tableau entre couleur meurtrier et couleur victime

Meurtrier	Victime	
Blanc	Blanche	2146
Blanc	Noire	111
Noir	Blanche	287
Noir	Noire	2220

Lien vraiment très fort. Les blancs tuent essentiellement des blancs, les noirs tuent surtout des noirs.

Le lien très fort entre couleur du meurtrier et couleur de la victime avec le lien très fort entre sentence et couleur de la victime, explique le lien entre sentence et couleur du meurtrier.

On peut analyser ce tableau par diverses méthodes statistiques pour voir les liens entre les diverses variables. Les petits tableaux ont montré certaines liaisons.

On peut utiliser de la régression logistique pour mesurer l'effet de la couleur du meurtrier et de la couleur de la victime sur la sentence.

Modèles graphiques.

Analyse des correspondances ne donne rien.

Une approche simple. Une nouvelle variable 4 valeurs : type de meurtres
Meutrier Blanc et Victime Blanche
Meutrier Blanc et Victime Noire
Meurtrier Noir et Victime Blanche
Meurtier Noir et Victime Noire

Cela consiste à regarder le tableau initial de manière différente.

Type Meurtre	Peine de Mort	Autre Peine	% mort
Blanc-Blanche	72	2074	3,4%
Blanc-Noire	0	111	0 %
Noir-Blanche	48	239	16,7%
Noir-Noire	11	2209	0,5%

L'analyse du tableau de contingence est assez simple. Il y a une forte liaison entre le type de meutre et la sentence. Le risque de peine de mort est fort si la victime est blanche et très fort si de plus le meutrier est noir.

Un autre petit tableau université Toulouse 3 Égalité Femme-Homme
Nombres de femmes et d'hommes parmi les professeurs et maîtres de conférence suivant les disciplines.
Encore trois variables, sexe, niveau, discipline.
Seule deux disciplines, physique et médecine sont retenues.

Niveau	Disicipline	Homme	Femme
Professeur	Physique	28	2
Professeur	Medecine	150	20
MCF	Physique	29	6
MCF	Medecine	37	53

La physique est la discipline avec la plus faible proportion de femmes pour les professeurs et les maîtres de conférence dans l'université Toulouse 3. Dans le document sur l'égalité il y a tableau pour les professeurs et un tableau pour les maîtres de conférence.

La juxtaposition des deux tableaux permet de voir un autre aspect des données, l'écart important entre la proportion de femmes professeurs de médecine et la proprortion de femmes maîtres de conférence, dans cette discipline.
Dans cette étude il y a d'autres variables.

Ce phénomène de confusion de variables se rencontre dans les analyses comme l'analyse en composantes principales ou l'analyse des correspondances fondées sur les liaisons de variables deux à deux. Exemple avec beaucoup de variables dont trois pour lesquelles la confusion est explicable, sexe, âge et niveau d'étude.

Pas d'exemple pour des données massives n'ayant pas d'outils pour les traiter.

Mais les choix proposés par les moteurs de recherche et autres sites internet semblent fondés sur des croisements de variables deux à deux et sont très souvent peu judicieux.

Les corrélations peuvent être totalement fortuites.
Il y a peut être des confusions de variables...
Comment les chercher ? Graphes...