Leçon 9 Analyse des Correspondances Multiples

L'analyse des correspondances nous permet d'analyser des tableaux résultats du croisement de deux variables qualitatives. Cette technique a été étendue au traitement de fichiers dans lesquels les individus sont caractérisés par plusieurs variables qualitatives. Comme d'habitude, nous présenterons cette méthode à partir d'exemples.

I Un exemple de fichier de données d'enquête

1) les données

En 2003, l'INSEE a fait une enquête appelée "Participation culturelle et sportive" dont le but est de mieux cerner les différentes formes de participation à la vie culturelle et sportive. Cette enquête avait plusieurs intérêts en particulier :

- observer l'interaction entre ces deux formes prépondérantes d'usage du temps libre,
- mieux connaître les personnes pratiquant telle ou telle activité
- mieux comprendre les raisons de la non pratique

De cette enquête, nous avons extrait un échantillon de 50 individus et leurs réponses à 5 questions qui ont permis de définir 5 variables qualitatives. Nous n'allons pas utiliser cet échantillon ridicule pour essayer de rendre compte des réponses à ces préoccupations. Nous allons simplement essayer de montrer comment on fait en pratique pour analyser un tel fichier.

les variables de l'analyse sont :

VELO	Pratiquez vous la natation, le vélo ou la randonnée ?	VELO1 Jamais VELO2 Moins d'une fois par semaine VELO3 Une fois par semaine ou plus
PETA	Pratiquez-vous la Pétanque, la Chasse ou la Pêche ?	PETA1 Jamais PETA2 Moins d'une fois par mois PETA3 Une fois par mois ou plus
LIVR	Combien de livres avez-vous lu au cours des 12 derniers mois ?	LIVR1 aucun LIVR2 1 à 9 livres LIVR3 10 à 24 livres LIVR4 25 ou plus
SEXE	Etes vous un Homme ou une Femme ?	SEXE1 Homme SEXE2 Femme
DIPL	Quel est le niveau de votre plus haut diplôme ?	DIPL1 aucun ou CEP DIPL2 inférieur au Bac DIPL3 Bac DIPL4 supérieur au Bac

Les réponses des 50 individus sont données dans le tableau ci dessous :

IND	VELO	PETA	TELE	LECT	SEXE	DIPL
ind1	1	1	1	2	2	1
ind2	1	1	1	1	2	2
ind3	1	1	1	2	2	2
ind4	2	2	2	2	1	2
ind5	1	1	1	1	2	4
ind6	1	1	3	1	1	2
ind7	3	1	1	4	1	2
ind8	1	1	3	2	2	1
ind9	3	3	2	2	1	2
ind10	3	3	1	4	2	4
ind11	2	1	1	1	1	3
ind12	1	1	2	4	2	2
ind13	2	1	1	3	1	3
ind14	2	2	1	4	1	2
ind15	3	2	2	3	1	4
ind16	1	1	1	4	2	2
ind17	2	1	1	2	1	2
ind18	3	2	2	2	1	3
ind19	2	3	1	2	2	2
ind20	1	1	1	3	1	1
ind21	1	3	1	2	1	1
ind22	2	3	1	4	1	3
ind23	1	1	3	1	2	4
ind24	1	1	1	1	1	1
ind25	2	2	1	3	2	4
ind26	2	1	2	3	1	2
ind27	1	1	3	2	2	2
ind28	2	1	2	4	2	3
ind29	1	1	2	2	2	2
ind30	3	1	3	2	2	1
ind31	1	1	3	2	1	1
ind32	1	1	1	1	1	1
ind33	1	1	2	2	1	2
ind34	3	1	1	3	2	2
ind35	3	3	1	1	1	1
ind36	1	1	1	3	2	1
ind37	3	1	1	3	2	4
ind38	2	2	1	4	2	2
ind39	3	3	1	2	1	4
ind40	3	1	2	1	2	4
ind41	1	1	1	2	1	1
ind42	1	1	1	3	1	1
ind43	2	3	1	4	1	4
ind44	2	2	2	1	2	1
ind45	3	1	1	3	1	2
ind46	1	1	3	3	2	2
ind47	3	1	1	3	2	3
ind48	1	1	1	4	2	2
ind49	1	1	2	2	2	1
ind50	2	2	1	2	1	2

2) le tableau disjonctif complet et le tableau de Burt

Avant toute calcul, il faut être conscient du fait qu'il n'est pas question d'appliquer brutalement une méthode d'analyse des données à ce tableau : les nombres dans le tableau représentent simplement des codages des réponses individuelles, on n'a pas le droit de les additionner ou de les multiplier.

On utilise deux astuces pour transformer ce tableau de données en tableau "analysable"

La première astuce pour analyser un tel tableau consiste à définir pour chaque variable autant de réponses possibles qu'il y a de modalités. Prenons la variable VELO par exemple. Elle admet 3 modalités de réponses. On peut la représenter par un vecteur 3 questions-modalités différentes (VELO1, VELO2, VELO3). Un individu qui aurait

la réponse 1 à la variable VELO aura les réponses

VELO1	VELO2	VELO3
1	0	0

la réponse 2 à la variable VELO aura les réponses

VELO1	VELO2	VELO3
0	1	0

la réponse 3 à la variable VELO aura les réponses

VELO1	VELO2	VELO3
0	0	1

En reproduisant ce procédé pour les 5 variables, on obtient ce que l'on appelle un **tableau disjonctif complet (**l'ensemble des valeurs des variables-modalités d'une même variable comporte une fois (*complet*) la valeur 1 et une seule (*disjonctif*).

	VELO			PETA			TELE			LECT				SEXE		DIPL			
ind	VELO1	VELO2	VELO3	PETA1	PETA2	PETA3	TELE1	TELE2	TELE3	LECT1	LECT2	LECT3	LECT4	SEXE1	SEXE2	DIPL1	DIPL2	DIPL3	DIPL4
ind1	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	1	0	0	0
ind2	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	0
ind3	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0
ind4	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
ind5	1	0	0	1	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1
ind6	1	0	0	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0
ind7	0	0	1	1	0	0	1	0	0	0	0	0	1	1	0	0	1	0	0
ind8	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	1	0	0	0
ind9	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0	1	0	0
ind10	0	0	1	0	0	1	1	0	0	0	0	0	1	0	1	0	0	0	1
ind11	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0
ind12	1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0
ind13	0	1	0	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0
ind14	0	1	0	0	1	0	1	0	0	0	0	0	1	1	0	0	1	0	0
ind15	0	0	1	0	1	0	0	1	0	0	0	1	0	1	0	0	0	0	1
ind16	1	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0
ind17	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0
ind18	0	0	1	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0
ind19	0	1	0	0	0	1	1	0	0	0	1	0	0	0	1	0	1	0	0
ind20	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0
ind21	1	0	0	0	0	1	1	0	0	0	1	0	0	1	0	1	0	0	0
ind22	0	1	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	1	0
ind23	1	0	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	1
ind24	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0
ind25	0	1	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0	0	1
ind26	0	1	0	1	0	0	0	1	0	0	0	1	0	1	0	0	1	0	0
ind27	1	0	0	1	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0
ind28	0	1	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0	1	0
ind29	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0
ind30	0	0	1	1	0	0	0	0	1	0	1	0	0	0	1	1	0	0	0
ind31	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0
ind32	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0	0
ind33	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	1	0	0
ind34	0	0	1	1	0	0	1	0	0	0	0	1	0	0	1	0	1	0	0
ind35	0	0	1	0	0	1	1	0	0	1	0	0	0	1	0	1	0	0	0
ind36	1	0	0	1	0	0	1	0	0	0	0	1	0	0	1	1	0	0	0
ind37	0	0	1	1	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1
ind38	0	1	0	0	1	0	1	0	0	0	0	0	1	0	1	0	1	0	0
ind39	0	0	1	0	0	1	1	0	0	0	1	0	0	1	0	0	0	0	1
ind40	0	0	1	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1
ind41	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	1	0	0	0
ind42	1	0	0	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	0
ind43	0	1	0	0	0	1	1	0	0	0	0	0	1	1	0	0	0	0	1
ind44	0	1	0	0	1	0	0	1	0	1	0	0	0	0	1	1	0	0	0
ind45	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0	0	1	0	0
ind46	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0	1	0	0
ind47	0	0	1	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
ind48	1	0	0	1	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0
ind49	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	1	0	0	0
ind50	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1	0	0
TOTAL	*23*	*14*	*13*	*34*	*8*	*8*	*31*	*12*	*7*	*10*	*18*	*12*	*10*	*25*	*25*	*14*	*21*	*6*	*9*

La somme des éléments d'une colonne représente le nombre d'individus ayant choisi la modalité représentée par cette colonne : ainsi dans l'échantillon, 23 personnes ne pratiquent jamais le vélo ou la randonnée, 10 personnes lisent plus de 25 livres dans l'année.

La somme des éléments d'une ligne est constante et est égale au nombre de variables étudiées.

La technique de **l'analyse des correspondances multiples** consiste à appliquer les techniques de l'analyse des correspondances des leçons précédentes à un tel fichier.

La seconde astuce consiste à calculer pour chaque couple de variables, le tableau de leur croisement. On construit un tableau qui regroupe tous les croisements possibles :

	VELO1	VELO2	VELO3	PETA1	PETA2	PETA3	TELE1	TELE2	TELE3	LECT1	LECT2	LECT3	LECT4	SEXE1	SEXE2	DIPL1	DIPL2	DIPL3	DIPL4	*total*
VELO1	23	0	0	22	0	1	13	4	6	6	10	4	3	9	14	11	10	0	2	*138*
VELO2	0	14	0	5	6	3	10	4	0	2	4	3	5	9	5	1	7	4	2	*84*
VELO3	0	0	13	7	2	4	8	4	1	2	4	5	2	7	6	2	4	2	5	*78*
PETA1	22	5	7	34	0	0	20	7	7	8	11	10	5	14	20	11	15	4	4	*204*
PETA2	0	6	2	0	8	0	4	4	0	1	3	2	2	5	3	1	4	1	2	*48*
PETA3	1	3	4	0	0	8	7	1	0	1	4	0	3	6	2	2	2	1	3	*48*
TELE1	13	10	8	20	4	7	31	0	0	6	8	9	8	17	14	9	12	4	6	*186*
TELE2	4	4	4	7	4	1	0	12	0	2	6	2	2	6	6	2	6	2	2	*72*
TELE3	6	0	1	7	0	0	0	0	7	2	4	1	0	2	5	3	3	0	1	*42*
LECT1	6	2	2	8	1	1	6	2	2	10	0	0	0	5	5	4	2	1	3	*60*
LECT2	10	4	4	11	3	4	8	6	4	0	18	0	0	10	8	7	9	1	1	*108*
LECT3	4	3	5	10	2	0	9	2	1	0	0	12	0	6	6	3	4	2	3	*72*
LECT4	3	5	2	5	2	3	8	2	0	0	0	0	10	4	6	0	6	2	2	*60*
SEXE1	9	9	7	14	5	6	17	6	2	5	10	6	4	25	0	8	10	4	3	*150*
SEXE2	14	5	6	20	3	2	14	6	5	5	8	6	6	0	25	6	11	2	6	*150*
DIPL1	11	1	2	11	1	2	9	2	3	4	7	3	0	8	6	14	0	0	0	*84*
DIPL2	10	7	4	15	4	2	12	6	3	2	9	4	6	10	11	0	21	0	0	*126*
DIPL3	0	4	2	4	1	1	4	2	0	1	1	2	2	4	2	0	0	6	0	*36*
DIPL4	2	2	5	4	2	3	6	2	1	3	1	3	2	3	6	0	0	0	9	*54*
total	*138*	*84*	*78*	*204*	*48*	*48*	*186*	*72*	*42*	*60*	*108*	*72*	*60*	*150*	*150*	*84*	*126*	*36*	*54*	*1800*

Le **tableau de Burt** que l'on obtient ici n'est pas un tableau de contingence mais plutôt une juxtaposition de tableaux croisant les variables entre elles. S'il y a 6 questions dans le fichier à analyser, le tableau de Burt est la juxtaposition de 6*6 = 36 tableaux de contingence. Chaque individu y est compté 36 fois et la somme des nombres inscrits dans ce tableau vaut donc 36*50 = 1800.

Ce tableau est clairement symétrique par rapport à sa diagonale, les tableaux de la diagonale en grisé donnent les effectifs de chacune des modalités.

Il se trouve qu'en appliquant les techniques de l'analyse des correspondances sur ce tableau, on obtient pratiquement les mêmes résultats que ceux de l'analyse du tableau disjonctif complet. Aussi l'analyse des correspondances multiples désigne l'analyse du **tableau disjonctif complet** aussi bien que celle du **tableau de Burt.** .

Pour illustrer l'équivalence de ces deux méthodes, nous allons les appliquer à l'exemple du sport et de la culture, puis à un second exemple qui traitera des pratiques de tri.