L'analyse statistique des données catégorielles en Sciences humaines et sociales

Mardi 6 juin 2006

L'analyse booléenne des questionnaires

Alain Degenne et
Marie-Odile Lebeaux

L'essentiel de cette présentation prend appui sur l'ouvrage de Claude Flament :

***L'analyse booléenne de questionnaire*, Paris : EHESS & Mouton, 1976.**

Un programme permettant de calculer les projections ultimes et de mettre en œuvre la méthode de la case vide a été écrit par Marie-Odile Lebeaux. Il est disponible sur le site :

http://lasmas.iresco.fr/logiciels.php

La compétition des femmes dans le mariage

	La dot existe	Pas de dot	Ensemble
Monogamie	27 (7)	45 (65)	72
Polygamie	5 (25)	263 (243)	268
Ensemble	32	308	340

D'après Gaulin et Boster (1990) et Lang (1993)

L'interprétation fait intervenir un modèle

L'observé

	Dot	Pas de dot	Total
Monogamie	27	45	72
Polygamie	5	263	268
Ensemble	32	308	340

Le modèle

Dot	Pas de dot	Total
32	40	72
0	268	268
32	308	340

La dot implique la monogamie

Le principe de l'analyse booléenne

Travailler sur l'opposition entre

ce qui apparaît et

ce qui n'apparaît pas,

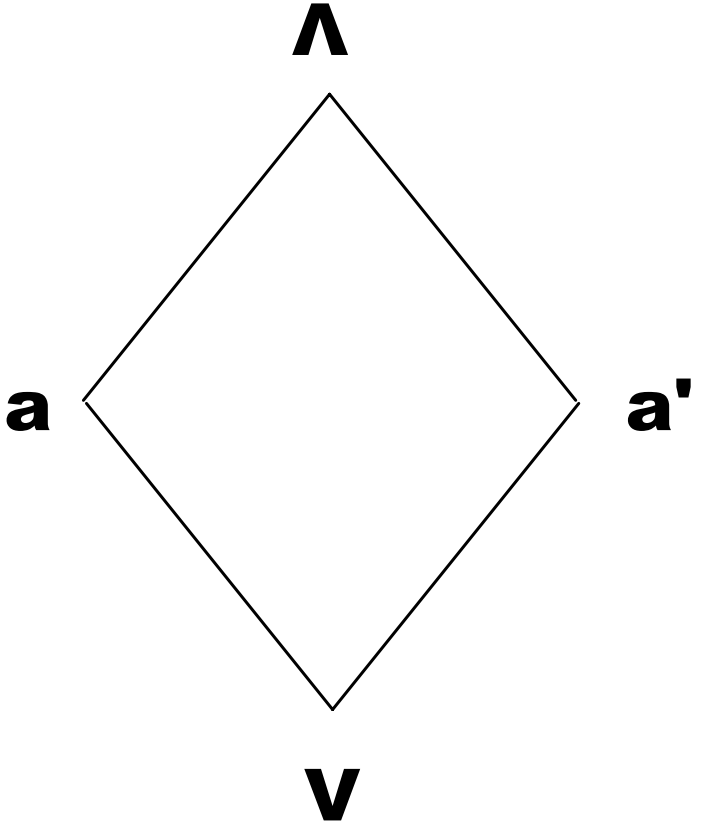
et chercher à en déduire des implications

Les algèbres de Boole

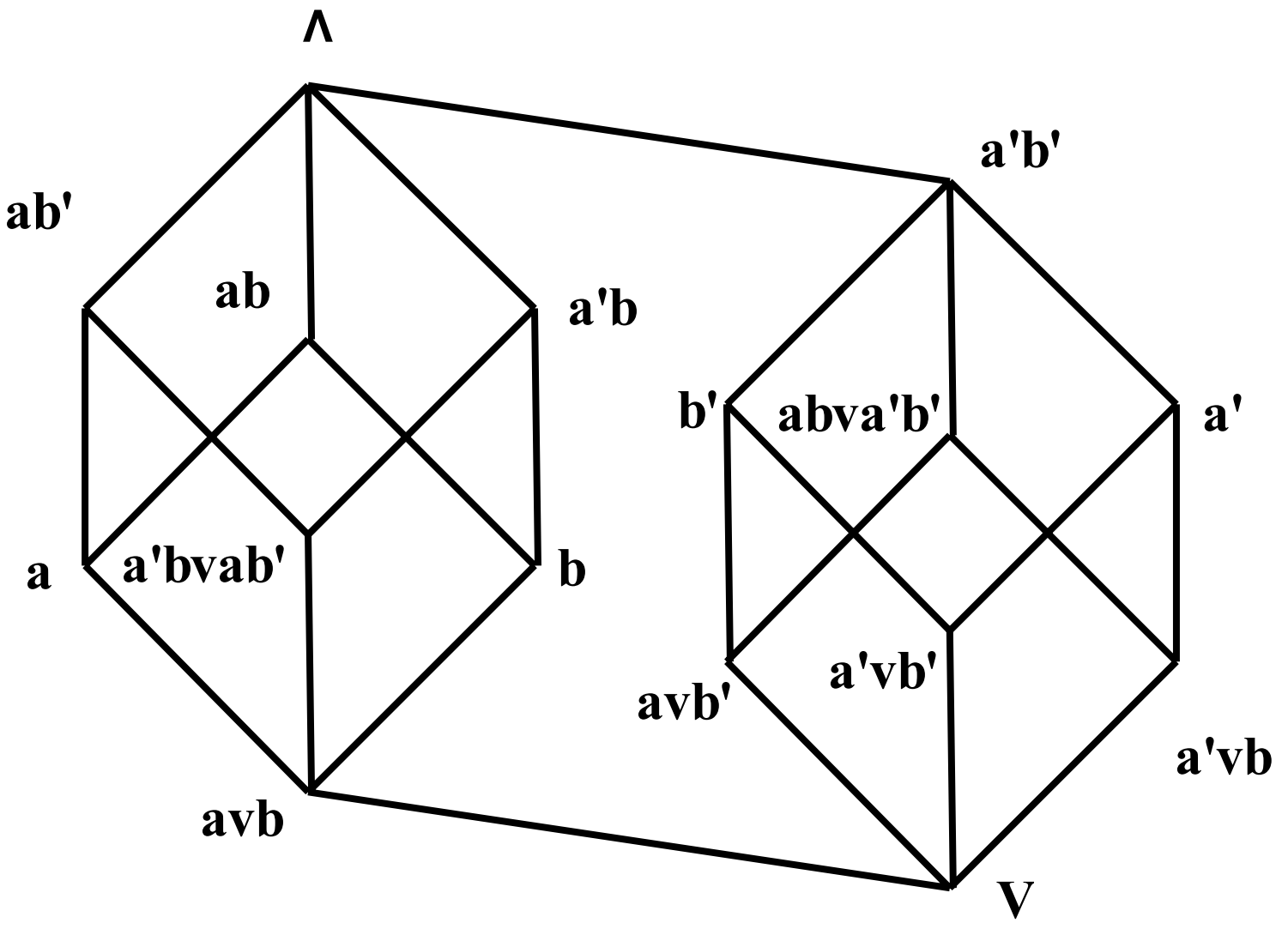
- **Une algèbre de Boole est construite sur des générateurs a, b, c etc.**
- **Chaque générateur peut prendre deux états : a et son conjugué a', b et b', c et c' etc.**
- **Une première opération permet de passer d'un élément à son conjugué : a donne a'. a'' = a.**

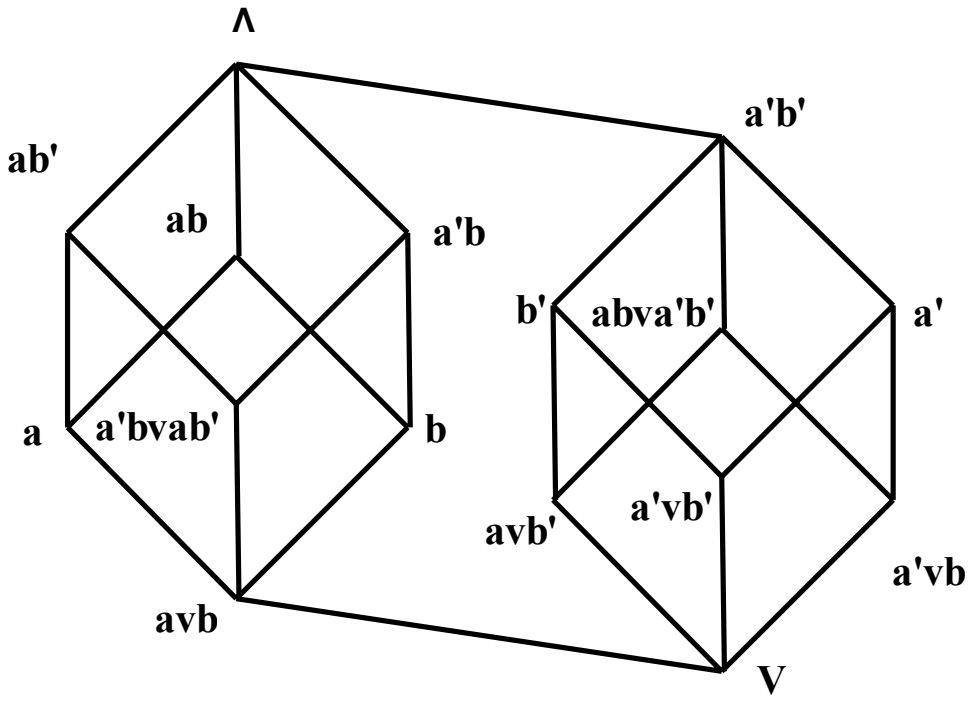
- **Il existe deux opérations binaires**
  - ***et,*** **qu'on notera par ∧, ou, plus simplement, par la concaténation : ab' signifie a et b', comme a ∧ b'.**
  - ***ou,*** **qu'on notera par v ; avb' signifie a ou b'**
- **Il y a deux éléments distingués**
  - **Le supremum V ; V = ava'**
  - **L'infimum Λ ; Λ = a∧a'**
- **Règles de calcul :**
  - **avV=V ; a∧V = a ; avΛ =a ; a ∧ Λ = Λ**
  - **(avb)' = a' ∧ b' = a'b' ; (a∧b)' = a' v b'**
  - **av(b ∧ c) = (avb) ∧ (avc)**
  - **a ∧ (bvc) = (a ∧ b) v (a ∧ c) = ab v ac**

L’algèbre de Boole à un seul générateur



Algèbre de Boole libre à deux générateurs





- **Les éléments qui sont au plus près de l'infimum (ici ab', a'b, ab, a'b'), sont appelés mintermes. les autres sont des inftermes.**
- **Les mintermes sont les configurations de réponses que l'on observe directement dans une enquête, les "patrons de réponse". Leur écriture fait intervenir tous les générateurs, sous forme simple ou accentuée, reliés uniquement par des Λ.**

Une algèbre de Boole libre à n générateurs comporte $2^{(2^n)}$ éléments.

n=1	4 éléments
n=2	16 éléments
n=3	256 éléments
n=4	65536 éléments

Algèbre quotient

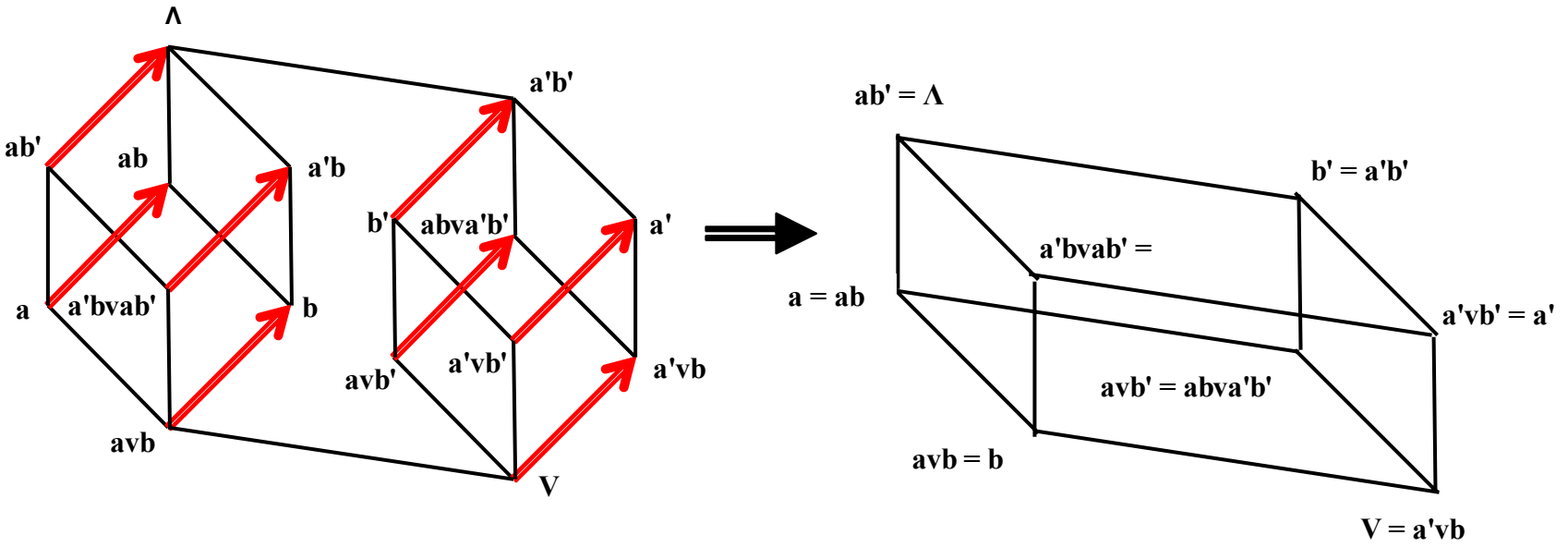
- **En constatant ou en décidant que certaines réponses possibles n'apparaissent pas, on crée une équivalence entre un ensemble d'éléments de l'algèbre et l'infimum (Λ).**

- **Ceci engendre un endomorphisme de l'algèbre et crée une algèbre de Boole quotient.**

- **Les éléments rendus équivalents à Λ engendrent un *idéal* de l'algèbre. Le supremum de cet idéal résume l'information.**

- **Nous utilisons les ressources du calcul booléen pour simplifier l'expression de ce**

L’algèbre de Boole à deux générateurs (monogamie et dot), liée par la condition ab’ = Λ



Principe d'une analyse de questionnaire

- **Soit un questionnaire de n questions,**
- **Coder en 0 et 1 les réponses à chaque question,**
- **En déduire les configurations de réponse ou *mintermes* et leur fréquence; (2^n *mintermes* possibles).**
- **Décider d'une partition de l'ensemble des réponses en *mintermes apparus* et *mintermes non apparus*. Ces derniers engendrent l'idéal de l'algèbre,**
- **Rechercher *l'expression la plus simple* du supremum de l'idéal.**

Un exemple sur 3 questions

Mintermes	Statistique
a b c	166
a’ b’ c’	340
a b c’	200
a’ b c’	260
a b’ c	4
a’ b c	10
a b’ c’	8
a’ b’ c	12
Total	1000

- Les configurations de réponse sont appelées *mintermes*
- Au vu des effectifs, on décide quels sont les mintermes « apparus » et les mintermes « non apparus », c’est-à-dire d’effectif négligeable.

On travaille sur les mintermes non apparus. On les classe en fonction du nombre croissant de lettres accentuées

Mintermes non apparus

a b’ c
a’ b c
a b’ c’
a’ b’ c

Expression qui va engendrer l’endomorphisme

ab’c v a’bc v ab’c’ v a’b’c = Λ

Réduction (algorithme de Lagrange) :

- ab’c v ab’c’ = ab’ (cvc’) = ab’ V = ab’
- ab’c v a’b’c = (a v a’) b’c = V b’c = b’c
- a’bc v a’b’c = (b v b’) a’c = V a’c = a’c

Forme réduite : ab’ v b’c v a’c = Λ

Les cases en jaune représentent les mintermes non apparus. L'opération consiste à exprimer cet ensemble comme union de surfaces qui se définissent au moyen d'un minimum de lettres, deux si possibles. Ici ab', a'c suffisent mais on constate que b'c doit aussi être retenue. C'est la règle dite de fermeture.

	a		a'	
	b'	b	b	b'
c				
c'				

Les implications simples

ab = Λ

	b	b′
a	0	y
a′	z	t

$a \Rightarrow b'$

$b \Rightarrow a'$

ab′= Λ

	b	b′
a	x	0
a′	z	t

$a \Rightarrow b$

$b' \Rightarrow a'$

a′b = Λ

	b	b′
a	x	y
a′	0	t

$b \Rightarrow a$

$b' \Rightarrow a'$

a′b′= Λ

	b	b′
a	x	y
a′	z	0

$a' \Rightarrow b$

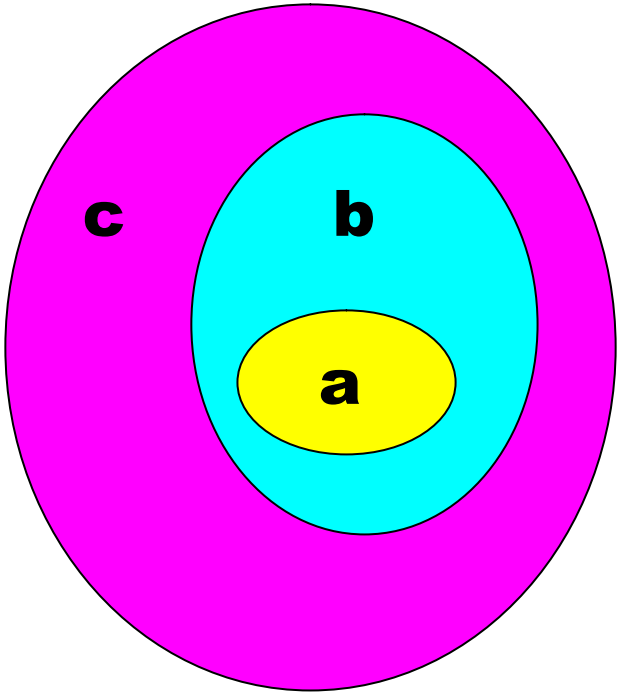
$b' \Rightarrow a$

Traduction

ab’ v b’c v a’c = Λ

- **a implique b**
- **b implique c**
- **a implique c**

C’est un une échelle cumulative du type Guttman.



Cas plus complexes

$abc' = \Lambda$

$ab \Rightarrow c$ $\qquad$ $c' \Rightarrow a'vb'$

$ac' \Rightarrow b'$ $\qquad$ $b \Rightarrow a'vc$

$bc' \Rightarrow a'$ $\qquad$ $a \Rightarrow b'vc$

Algorithme de Kuntzman

- **Dans un questionnaire, le nombre de mintermes « non apparus » peut se révéler beaucoup plus grand que celui des mintermes apparus. Dans ce cas, on utilise l'algorithme de Kuntzman qui fournit la forme réduite à partir des mintermes apparus. C'est celui qui est mis en œuvre dans le programme de Marie-Odile Lebeaux.**

Méthode de la case vide

- **Il n'est pas toujours évident de décider quels sont les mintermes apparus et les mintermes non apparus.**
- **On choisit souvent de prendre les décisions directement sur les projections (inftermes) de 2 ou 3 éléments, en fonction de leurs effectifs. C'est la méthode dite de la case vide.**
- **Il faut alors s'assurer de la fermeture du système.**

Quatre questions extraites de l'enquête PCV97 de l'INSEE

- (a) Est-ce que depuis un an, vous-même ou une autre personne du foyer avez eu l'occasion d'entrer chez des ménages voisins **oui** **non**
- (b) Depuis un an, des voisins vous ont-ils rendu un service comme par exemple : garder vos enfants ou garder vos clefs, vos plantes, vos animaux, vous prêter des outils ou des produits de cuisine, etc. ? **oui** **non**
- (c) Y a-t-il des voisins avec lesquels votre ménage a des relations plus poussées encore (sorties en commun, confidences, liens d'amitié, entraide pour des travaux, etc.) ? **oui** **non**
- (d) Avez-vous parfois, vous-même, ou d'autres personnes de votre foyer, des conversations avec des voisins (échange de nouvelles, de conseils, de recettes, etc.) ? **oui** **non**

Liste des min-termes du protocole

Minterme	Effectif réel	Effectif probable	Réel/ probable	Minterme	Effectif réel	Effectif probable	Réel/ probable
abcd	944	200	470	abCd	12	79	15
aBcd	113	217	52	aBCd	9	86	10
Abcd	208	337	62	AbCd	29	134	22
ABcd	195	366	53	ABCd	56	145	39
abcD	696	526	132	abCD	46	208	22
aBcD	253	570	44	aBCD	42	226	19
AbcD	571	886	64	AbCD	217	351	62
ABcD	1086	960	113	ABCD	1198	380	315

		A		a'	
		b'	B	B	b'
C	d'	29	56	9	12
	D			42	46
c'	D				
	d'				

Projections ultimes après élimination des mintermes dont l'effectif est inférieur à 30 : aCd v bCd = Λ

- ***C ⇒ A ou D (coût : 21)***

Des relations plus poussées impliquent d'entrer chez les voisins ou d'avoir des conversations

- ***C ⇒ B ou D (coût : 41)***

Des relations plus poussées impliquent des conversations ou que les voisins aient rendu service

Projections ultimes après élimination des mintermes dont l'effectif est inférieur à 50 : aC v bCd= Λ

- ***C ⇒ A (coût : 109)***

Des relations plus poussées impliquent d'entrer chez les voisins

- ***C ⇒ B ou D (coût : 41)***

Des relations plus poussées impliquent des conversations ou des services rendus par les voisins

Projections ultimes après élimination des mintermes dont l'effectif est inférieur à 60 : aC v Cd= Λ

- ***C ⇒ A (coût : 109)***

Des relations plus poussées impliquent d'entrer chez les voisins

- ***C ⇒ D (coût : 106)***

Des relations plus poussées impliquent des conversations

Méthode de la case vide
inftermes de longueur 2

ab**	1698	1014	167	Ab**	1025	1708	60
aB**	417	1100	38	AB**	2535	1851	137
a*c*	2006	1515	132	A*c*	2060	2550	81
a*C*	109	599	18	A*C*	1500	1009	149
a**d	1078	583	185	A**d	488	982	50
a**D	1037	1531	68	A**D	3072	2577	119
bc	2419	1951	124	*Bc*	1647	2115	78
bC	304	772	39	*BC*	1305	837	156
*b*d	1193	751	159	*B*d	373	814	46
*b*D	1530	1971	78	*B*D	2579	2137	121
**cd	1460	1122	130	**Cd	106	444	24
**cD	2606	2944	89	**CD	1503	1165	129

Methode de la case vide inftermes de longueur 3

abc*	1640	727	226
aBc*	366	788	46
abC*	58	287	20
aBC*	51	311	16
ab*d	956	280	341
aB*d	122	303	40
ab*D	742	734	101
aB*D	295	796	37
a*cd	1057	418	253
a*Cd	21	165	13
a*cD	949	1097	86
a*CD	88	434	20
*bcd	1152	538	214
*bCd	41	213	19
*bcD	1267	1412	90
*bCD	263	559	47

Abc*	779	1223	64
ABc*	1281	1326	97
AbC*	246	484	51
ABC*	1254	525	239
Ab*d	237	471	50
AB*d	251	511	49
Ab*D	788	1236	64
AB*D	2284	1340	170
A*cd	403	703	57
A*Cd	85	278	31
A*cD	1657	1846	90
A*CD	1415	730	194
*Bcd	308	583	53
*BCd	65	231	28
*BcD	1339	1531	87
*BCD	1240	606	205

La question de l’approximation

- **Il n’existe pas de technique standard d’approximation d’un modèle d’implication. Le choix d'un modèle est une prise de position théorique.**
- **Ici nous avons tenu compte du coût du modèle qui correspond au nombre de mintermes qui sont apparus et qui sont en contradiction avec le modèle.**
- **Nous avons également comparé la fréquence d'apparition des inftermes avec leur effectif probable.**

Modèle cumulatif de Guttman

- ***Il est caractérisé par un ensemble fermé d'implications ab'= Λ, ac'= Λ, ad'= Λ, bc'= Λ, bd'= Λ, cd'= Λ***
- ***Il représente une famille de questions de difficulté croissante : a $\Rightarrow$ b $\Rightarrow$ c $\Rightarrow$ d***

- **(a) : J'accepterais volontiers qu'un de mes enfants épouse un(e) immigré(e)**
- **(b) : J'accepterais volontiers qu'un de mes enfants fréquente un(e) immigré(e)**
- **(c) : J'accepterais volontiers un immigré comme voisin**
- **(d) : J'accepterais volontiers que mon pays accueille plus d'immigrés**

Modèle de connexité (*unfolding* de Coombs)

- **Il correspond à l'idée d'un continuum tel que si l'on répond oui à deux questions, on répond oui également à celles qui sont situées entre celles-ci sur le continuum :**

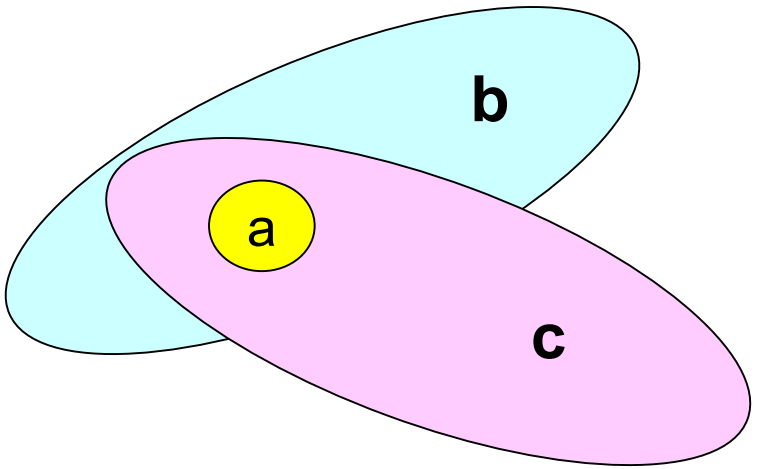
a — b — c — d — e

ab'c v ab'd v ab'e v ac'd v ac'e v ad'e = Λ

- **(a) : L'État ne doit pas intervenir dans la politique des entreprises**
- **(b) : Par une planification indicative, L'État peut orienter le développement de certains secteurs**
- **(c) : En jouant sur certains moyens d'intervention, l'État doit contrôler la marche générale de l'économie**
- **(d) : L'État doit pouvoir contrôler l'action des grands moyens de production, en particulier pour obtenir la réalisation des objectifs du plan**
- **(e) : Il est souhaitable que les grands moyens de production soient des entreprises nationalisées**

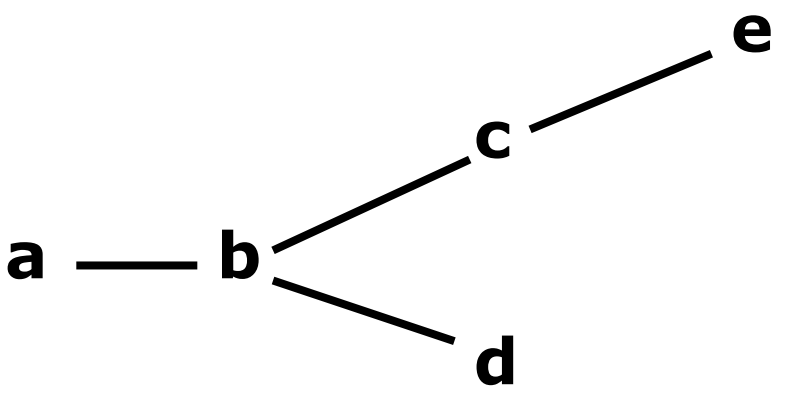
- **Dans la logique cumulative, on peut viser des modèles d'ordre partiel, moins exigeants que l'ordre total**

$$a \Rightarrow b \quad a \Rightarrow c$$



- **Il est possible de construire des modèles de connexité sur un arbre :**

  ab'c v ab'e v bc'e v ab'd = Λ



L'analyse booléenne permet d'ajuster à un questionnaire tout une classe de modèles caractérisés par des implications.

- **Degenne A., Lebeaux M.-O., (1996), "Boolean analysis of questionnaire data", *Social Networks*, n° 18, p. 231-245.**
- **Flament C., (1976), *L'analyse booléenne de questionnaire*, Paris, EHESS & Mouton.**
- **Gaulin S.J.C. & Boster J.S., (1990), Dowry and Female Competiton, *American Anthropologist*, 93, 994-1005**
- **Lang H., (1993), Dowry and Female Competition: a Boolean Reanalysis, *Current Anthropology*, 14(5), 775-778.**
- **Theuns P., (1994), " A dichotomisation method for Boolean analysis for quantifiable cooccurence data ", in G.H. Fisher &D. Laming (Eds), *Contributions to Mathematical Psychology, Psychometrics and Methodology*, New York: Springer.**
- **Theuns P., (1998), " Building a knowledge space via Boolean analysis of co-occurrence data ", in C. Dowling, F. Roberts & P. Theuns (Eds), *Recent Progress in Mathematical Psychology, Psychophysics, Knowledge, Representation, Cognition and Measurement*, New York: Lawrence Erlbaum.**
- **White D. R., (2000), Manual for Statistical Entailment Analysis 2.0, *World Culture,* 11(1), 77-90.**
- **White D. R., McCann H.G., (1988), "Cites and Fights: Material Entailment Analysis of the Nineteenth-Century Chemical Revolution ", in B. Wellman & S. D. Berkowitz, (Eds), *Social Structures, a Network Approach*, Cambridge, Cambridge University Press.**