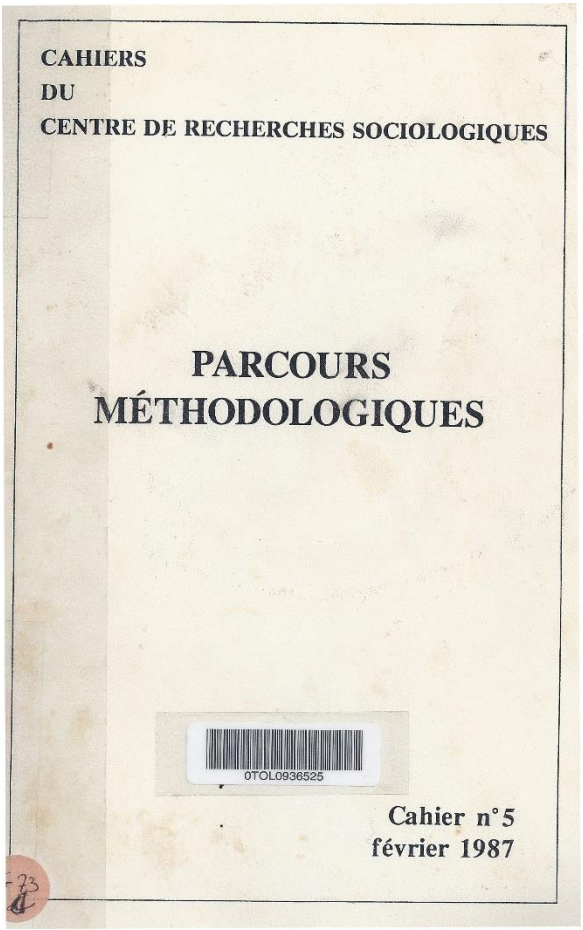
Le Chi2 : un usage exploratoire d'un test classique

Michel Grossetti

Michel Grossetti

Le Chi2 : un usage exploratoire d'un test classique

Cahiers du Centre de Recherches Sociologiques, n°5, 1987, pp. 133-141.



Les principaux outils statistiques de la sociologie en France ont longtemps été les comptages simples, les tables de contingence, quelques calculs de moyennes et, bien sûr, l'inévitable ennemi perpétuel des étudiants du 1er Cycle de sociologie : le test du Khi 2. Même si, çà et là, des chercheurs poussaient plus loin le recours à la statistique (C. Thélot ou A. Degenne par exemple), le lot de la plupart de ceux qui utilisaient les enquêtes par questionnaire restait l'examen fastidieux de longues listes de tableaux avec le test du Khi 2 et l'aspirine pour seule aide.

Et puis sont arrivées les méthodes multidimensionnelles dont la plus connue des sociologues, l'analyse factorielle des correspondances. Ces méthodes n'ont pas toujours fait, loin de là, l'unanimité, tant auprès des statisticiens que des utilisateurs, soit qu'on leur reproche leur manque de rigueur et la variabilité des interprétations qu'elles peuvent susciter, soit que l'on s'effraie de la difficulté même de ces interprétations et des pièges qu'elles recèlent. On sait par exemple que l'AFC visualise la structure des écarts à l'indépendance d'une table de contingence mais non la force de ces écarts (ainsi un tableau pour lequel un test du Khi 2 conclurait à l'indépendance fera l'objet d'un graphique factoriel tout à fait habituel et on pourra, si l'on n'y prend garde, bâtir des interprétations sophistiquées sur des écarts si négligeables qu'ils sont peut-être dus à trois erreurs de codage ou de saisie). D'autre part, interpréter une analyse factorielle n'est pas une mince affaire : il faut évaluer la qualité du résumé obtenu (valeurs-propres), la qualité des projections des diverses modalités (contributions, valeurs-tests), donner un sens aux axes (contributions), cerner l'effet sur la détermination des axes d'éventuelles corrélations structurelles, etc. Autrement dit, interpréter une AFC, c'est du travail, qui peut se révéler aussi long et fastidieux que les analyses de tables de contingences que nous évoquions plus haut.

Tout cela n'a nullement empêché les sociologues de faire un usage de plus en plus intensif de ces méthodes et de délaisser les techniques plus classiques, soit que la séduction de l'aspect graphique des analyses factorielles se soit exercée sur des chercheurs plutôt littéraires et peu attirés par les chiffres, soit que, tout simplement, ces techniques s'adaptassent bien à ce qu'ils attendaient des statistiques, une vision d'ensemble plutôt que l'accumulation d'éléments précis mais partiels (1).

Le problème est que, emportés par leur enthousiasme, ceux d'entre les statisticiens qui défendaient ces méthodes, n'ont pas toujours insisté assez sur les pièges qu'elles peuvent recéler sur leurs limites, ou sur la technique de l'interprétation, et que l'on a pu voir des interprétations, uniquement fondées sur les graphiques (le plus souvent sur UN graphique), renvoyer presque ceux-ci au statut de tache dans le test de Rorschach: polarisation d'une projection imaginaire plus que réel outil scientifique (1).

Tout statisticien a au moins une fois entendu cette histoire du chercheur en Sciences Humaines qui fait réaliser une analyse factorielle par un statisticien et repart avec son listing. Le statisticien s'aperçoit après coup que son programme comporte une erreur et que les résultats remis sont faux. Vite, il corrige son programme, recommence l'analyse et téléphone tout penaud au chercheur, lequel est très réticent à recommencer son interprétation tant le résultat vérifie bien ses hypothèses. 11 le fait toutefois et explique au statisticien que le nouveau résultat valide encore mieux ses hypothèses !

J'ai entendu plusieurs variantes de cette histoire, dont je pense qu'elle est significative d'au moins une chose : la confiance qu'ont les statisticiens dans l'interprétation des analyses factorielles par des non spécialistes à qui ils présentent encore trop souvent cette méthode comme sans danger.

On pourrait ajouter beaucoup d'histoires sur l'analyse factorielle comme celle de cet étudiant de troisième cycle qui me demandait de lui réaliser une analyse factorielle ; sur quelles variables ? Heu ... je ne sais pas ... ; mais qu'est-ce que vous cherchez exactement à avoir ? ; une analyse factorielle ... ; oui, certes, mais dans quel but ? ; pour avoir ces graphiques, là, vous savez bien ...

Cela signifie-t-il qu'il ne faut pas utiliser l'analyse factorielle ? Certes non ! Cela suggère qu'il serait bon de cesser d'en faire une panacée et de l'utiliser à tort et à travers. Si l'on cherche à

obtenir une typologie, ne vaut-il pas mieux utiliser directement une technique de classification au lieu de chercher à déduire des classes d'un graphique factoriel ? Si on veut expliquer une variable par d'autres, n'y a-t-il pas intérêt à faire usage d'une technique de régression ou de segmentation ? Enfin, rien n'interdit, avant de se jeter dans la complexité de méthodes multidimensionnelles, de débroussailler un peu le problème à l'aide de techniques simples mais sûres, et cela sans revenir nécessairement à l'aspirine et aux longues listes de chiffres. Il y a plus d'une façon d'utiliser les tests classiques et notamment celui du Khi2. M.C .Viguier, P. Cibois (2) et surtout A. Morineau (3) entre autres s'en sont aperçus depuis longtemps et utilisent le test du Khi 2 sur des tableaux d'indicatrices.

Une indicatrice, c'est une variable à deux modalités (oui/non, présence/absence, etc.). Par exemple, supposons que j'ai la variable qualitative classique "profession du père" en 8 catégories: artisan ou commerçant, cadres, profession intermédiaires, techniciens ou contremaîtres, employés, ouvriers, inactifs, et l'inévitable catégorie 11non réponse".

Chacune de ces modalités peut constituer une variable dite indicatrice. On obtient donc 8 variables : artisan ou commerçant (oui ou non), cadre (oui ou non), etc.

Si toutes les variables sont transformées en indicatrices, on peut calculer le Khi 2 entre une modalité et une autre. Supposons que j'aie des lycéens de terminale et que je veuille savoir s'il y a corrélation entre être en terminale C et avoir un père cadre. Je teste la corrélation entre les deux indicatrices associées à ces deux situations sur le tableau suivant :

	Terminale C	Autre Section
Père cadre		
Autres Professions		

Le Khi 2 de ce tableau me dira s'il y a corrélation au seuil que je choisis (le plus simple est de faire calculer par le programme le risque alpha). Je pourrai aussi savoir si cette corrélation marque une attirance (les fils de cadres sont plus souvent en C que la moyenne) ou une répulsion (l'inverse). Jusque-là rien de neuf. Ce qui est intéressant, c'est que je peux faire la même chose entre "cadre" et une autre modalité, par exemple "mention bien au baccalauréat" ; comme les deux khi 2 sont calculés avec·1a même population et le même degré de liberté (1), on peut les comparer.

De plus, comme les modalités sont comparées séparément, fini le problème des modalités mal définies (non réponses) ou d'effectif trop faible qui venaient troubler le test du khi 2 appliqué à un tableau de contingence traditionnel.

Enfin et surtout, l'opération peut être réitérée pour toutes les modalités d'une série de variables, ce qui permettra de balayer rapidement une grande quantité d'information et de ne retenir que ce qui est corrélé à la modalité que l'on cherche à "expliquer", et d'obtenir ainsi un portrait contrasté de la sous-population qu'elle recouvre. On aura ainsi, immédiatement et sans effort ce qu'il fallait de longues heures pour obtenir par le passé puisque l'"interprétation" est toute faite.

Attention, cette méthode n'est rien de plus qu'un débroussaillage, elle ne tient aucun compte des interactions dans l'explication d'une modalité. De plus, elle est axée sur les corrélations et fait ressortir les différences entre une sous-population et la population totale et ne retient pas des caractéristiques qui seraient dominantes dans la sous-population sans que cela entraîne de différence importante avec ce qui se passe pour la population totale. Enfin, les programmes dont nous nous servons (logiciels SPAD et SICLA) ne retiennent que les modalités attirées par celle que l'on cherche à expliquer et non celles qui sont repoussées par elle.

Toutefois, cette technique s'avère bien utile dans le premier temps d'un dépouillement d'enquête. On peut ainsi très vite vérifier les corrélations avec les variables fondamentales (sexe, âge, etc.) ou toute autre pour laquelle on cherche à avoir une liste exhaustive des attirances.

L'exemple qui suit est tiré d'une enquête effectuée au Centre de Recherches Sociologiques sous la direction de J. M. Berthelot dans le cadre d'une recherche sur les transitions dans le système éducatif (4).

524 bacheliers de l'académie de Toulouse. On dispose d'indicateurs de leur situation scolaire de terminale (SECTION, RETARD SCOLAIRE, NIVEAU au Ier Trimestre), de leur mention au bac, d'indicateurs divers (SEXE, PROFESSION DU PERE). On sait ce qu'ils sont devenus après le bac et on voudrait, par un balayage rapide, caractériser chaque orientation suivie. On fait donc ce que je viens de décrire plus haut. Pour chaque situation (la première ligne sert de titre) on a la liste des modalités les plus corrélées au sens du Khi 2 dans le sens d'une attirance, par ordre décroissant de corrélation, avec la valeur du Khi 2 ("Khi 2"), la probabilité d'indépendance ("proba"), l'effectif commun des deux modalités ("effe"), la proportion que cela représente pour la première (me/cl), pour la seconde (mc/mt) ainsi que la proportion de la seconde dans la population totale (mt/n).

Par exemple, dans les tableaux qui suivent, la modalité la plus corrélée avec l'orientation "grande école prépa" est "TC". Le khi 2 est de 127, la probabilité d'indépendance nulle. Il y a 55 élèves de C parmi les 79 étudiants de classes préparatoires, soit 55/79 = 70%. Dans la population totale, il y a 22% de TC seulement. Enfin, 49 % des TC sont en classe préparatoire.

En fait, tous ces chiffres ne sont que des raffinements permettant de détailler la corrélation. Une simple lecture des libellés des modalités corrélées avec chaque orientation suffit à avoir une première vision déjà nette du problème.

Ainsi dans notre exemple, il n'est pas surprenant de voir se polariser sur !'orientation en classes préparatoires tous les signes de l'excellence scolaire (section C, bon niveau, mentions, avance ou absence de retard) mais aussi l'appartenance aux catégories sociales favorisées ou au sexe masculin (qui reste très corrélé à la section C). L'orientation en I.U.T. ou STS reste liée aux secteurs techniques et à leurs caractéristiques. L'orientation vers les écoles d'accès direct après le bac représente finalement une sorte de juste milieu sans corrélations particulières avec quoi que ce soit.

Plus intéressant est de voir l'orientation en Université corrélée à des performances scolaires médiocres ou de voir que le fait de faire un 2ème bac ou de recommencer une Formation de niveau terminale est souvent le fait d'élèves de D (cherchant le plus souvent par ce biais à améliorer leurs chances par l'obtention d'un bac C).

Cet exemple très simple et "pédagogique" n'a pas d'autre but ici que d'illustrer clairement la méthode dont le principal intérêt est de permettre des hiérarchisations de corrélations et surtout de trier rapidement un grand nombre de variables (5).

NOTES

(1) cf. P. CIBOIS : "Tri-deux : une méthode post-factorielle de dépouillement d'enquête", *L'Année sociologique*, 1982, vol. 32.

(2) "Tri-deux une méthode post factorielle de dépouillement d'enquête, "L'Analyse des données en sociologie", PUF, 1984.

(3) "Note complémentaire sur les valeurs-tests", bulletin technique du CESIA, volume 2 1-2, 1984.

(4) cf. "Détermination Sociétales et hiérarchisation des choix" rapports 1 à 5. Centre de Recherches Sociologiques. Toulouse.

(5) Le nombre des variables est à priori illimité (ce qui n'est pas vraiment le cas dans l'A. F. C.), puisque les corrélations sont calculées indépendamment les unes des autres et hiérarchisées a postériori.

classe : 1 effectif : 79, 15%

vari able	libelle	moda lite	libelle	khi2	proba	effe	mc cl	mt n	mc mt
iS	inscription en 85-86	iS01	grande ecole prepa	524	0.000	79	100	15	100
SEC	SECTION	SEC3	TC	127	0.000	55	70	22	49
BAC	resultat au bac	BAC1	Bien ou Tres Bien	54	0.000	20	25	6	59
BAC	resultat au bac	BAC3	Assez Bien	49	0.000	35	44	17	39
IRS	niveau scol 1er trim terminale	IRS1	IRS = tres bon ou bon	40	0.000	69	87	54	24
RED	REDOUBLEMENT	RED4	aucun redoublement	36	0.000	71	90	59	23
PP	PROFESSION DU PERE	PP 3	chef d'ent cadre prof lib etc	28	0.000	36	46	23	31
AGE	ANNEE DE NAISSANCE	AGE5	68 = 1 an avance	24	0.000	17	22	8	41
AGE	ANNEE DE NAISSANCE	AGE4	67 = a l'heure	13	0.000	57	72	53	20
SEX	SEXE	SEX1	homme	8	0.005	44	56	41	20

classe : 2 effectif : 125, 24%

vari able	libelle	moda lite	libelle	khi2	proba	effe	mc cl	mt n	mc mt
iS	inscription en 85-86	iS02	iut ou bts	524	0.000	125	100	24	100
SEC	SECTION	SEC5	TE	27	0.000	19	15	6	63
SEC	SECTION	SEC6	TF	22	0.000	28	22	11	49
BAC	resultat au bac	BAC4	sans mention	11	0.001	74	59	46	31
RED	REDOUBLEMENT	RED1	1 redoublement au moins	10	0.001	61	49	36	32
SEX	SEXE	SEX1	homme	10	0.001	67	54	41	31

classe : 3 effectif : 25, 5%

vari able	libelle	moda lite	libelle	khi2	proba	effe	mc cl	mt n	mc mt
iS	inscription en 85-86	iS04	ecole	524	0.000	25	100	5	100

classe : 4 effectif : 244, 47%

vari able	libelle	moda lite	libelle	khi2	proba	effe	mc cl	mt n	mc mt
iS	inscription en 85-86	iS05	universite	524	0.000	244	100	47	100
IRS	niveau scol 1er trim terminale	IRS3	moyen insuff nul ou non rep	17	0.000	135	55	46	56
SEC	SECTION	SEC2	TB	17	0.000	51	21	14	69
SEX	SEXE	SEX2	femme	9	0.002	161	66	59	52
SEC	SECTION	SEC1	TA	9	0.002	55	23	17	61
SEC	SECTION	SEC4	TD	7	0.006	73	30	24	57
BAC	resultat au bac	BAC5	admis rattrappage	7	0.008	88	36	30	55

classe : 5 effectif : 39, 7%

vari able	libelle	moda lite	libelle	khi2	proba	effe	mc cl	mt n	mc mt
iS	inscription en 85-86	iS06	SN ou ANPE ou TUC ou vie active	524	0.000	39	100	7	100
SEC	SECTION	SEC7	TG	28	0.000	10	26	6	31
SEC	SECTION	SEC6	TF	27	0.000	14	36	11	25
RED	REDOUBLEMENT	RED1	1 redoublement au moins	11	0.001	24	62	36	13
AGE	ANNEE DE NAISSANCE	AGE3	66 = 1 an retard	10	0.001	19	49	26	14
AGE	ANNEE DE NAISSANCE	AGE1	64 ou 65 = 2ans retard au mins	8	0.004	10	26	11	17
RED	REDOUBLEMENT	RED5	RED = non rep	7	0.008	5	13	4	22
BAC	resultat au bac	BAC5	admis rattrappage	4	0.026	18	46	30	11
SEX	SEXE	SEX2	femme	4	0.040	29	74	59	9

classe : 6 effectif : 12, 2%

vari able	libelle	moda lite	libelle	khi2	proba	effe	mc cl	mt n	mc mt
iS	inscription en 85-86	iS11	2e bac ou autres situations	524	0.000	12	100	2	100
AGE	ANNEE DE NAISSANCE	AGE8	non rep AGE	7	0.008	1	8	1	20
BAC	resultat au bac	BAC5	admis rattrappage	4	0.033	7	58	30	4
SEC	SECTION	SEC4	TD	4	0.037	6	50	24	5