

STATISTIQUE ET ANALYSE DES DONNÉES

BRIGITTE ESCOFIER

Quelques indices pour comparer des tableaux de contingence

Statistique et analyse des données, tome 6, n° 1 (1981), p. 39-51

http://www.numdam.org/item?id=SAD_1981__6_1_39_0

© Association pour la statistique et ses utilisations, 1981, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

Statistique et Analyse des données
1981 - 1 - pp. 39-51

QUELQUES INDICES POUR COMPARER DES
TABLEAUX DE CONTINGENCE

Brigitte ESCOFIER

Résumé : *Nous proposons ici quelques indices de comparaison permettant d'avoir rapidement une idée générale des ressemblances et des différences entre deux ou plusieurs tableaux de contingence définis sur le même couple de caractère. Un exemple commenté en montre l'utilisation.*

Abstract : *We propose in this paper, some measures of comparison which enable to have rapidly a general idea about the resemblances and the differences between two or more contingency tables, defined on the same pair of characters. A commented example shows it's application.*

Mots clés : *Tableau de contingence, Analyse des correspondances, Comparaison de tableaux de données.*

I - LES TABLEAUX DE CONTINGENCE

Rappelons qu'un tableau de contingence se construit en considérant deux caractères qualitatifs, qu'on appellera I et J. Ces deux caractères sont définis sur la même population. Ils ont chacun un ensemble fini de modalités qu'on notera aussi I et J. Les modalités étant repérées par les indices i et j , le tableau de contingence est formé des nombres k_{ij} , qui sont les nombres d'individus de la population ayant à la fois la modalité i de I et la modalité j de J.

On dispose souvent de plusieurs tableaux de contingence définis par les mêmes caractères I et J. Par exemple une suite de tableaux indexés par le temps, ou des tableaux définis par des populations de type différent. Citons, pour fixer les idées, l'exemple traité et commenté dans la suite : deux tableaux répartissent l'un, la population féminine, l'autre, la population masculine suivant les deux caractères : niveaux de diplôme obtenu et type d'emploi occupé.

Il est naturel alors de vouloir comparer ces tableaux. Nous proposons ici quelques indices globaux de comparaison permettant d'avoir rapidement une idée générale des ressemblances et des différences entre deux ou plusieurs tableaux. On peut se contenter de ces indices ou bien les utiliser comme aides à l'interprétation de traitements plus ponctuels comme celui qui consiste à mettre en éléments supplémentaires lignes et colonnes des tableaux à comparer dans l'analyse d'un tableau moyen.

II - COMPARAISON DES TABLEAUX

Pour définir ces indices, nous considérons et nous comparons les populations à partir desquelles sont construits les tableaux de contingence. Chacune d'entre elles peut être représentée, dans l'espace vectoriel $R_{I \cup J}$, par un nuage de points : ce nuage qui décrit le comportement de la population vis à vis des deux caractères I et J est celui qui est étudié dans l'analyse des correspondances du tableau disjonctif complet croisant $I \cup J$ avec la population. On sait (cf. 1 et 7) que cette analyse est équivalente à celle du tableau de contingence.

II.a - Les centres de gravité

Le comportement moyen d'une population est représentée par le centre de gravité de son nuage. Ce point a pour coordonnées dans $R_{I \cup J}$ les marges sur I et sur J du tableau de contingence.

Pour voir si ces comportements moyens diffèrent, et en quoi ils diffèrent, nous allons comparer les positions de ces centres de gravité en les projetant sur les axes d'inertie du nuage représentant l'ensemble de toutes les populations.

A cet ensemble, décrit par le tableau disjonctif complet juxtaposant les tableaux de chaque population, correspond un tableau de contingence qui est la somme de tous les tableaux considérés. Un calcul simple (voir les formules en III) basé sur l'équivalence des analyses du tableau disjonctif complet et du tableau de contingence associé montre qu'il suffit de faire l'analyse de ce dernier, en mettant en éléments supplémentaires les marges de tous les tableaux

de contingence pour obtenir ces projections : le centre de gravité d'une population est situé au milieu du segment joignant les projections de ses deux marges.

Pour compléter ce résultat, qui ne fait qu'indiquer, pour chaque facteur de la population entière la position des centres de gravité de chaque sous population, nous calculons leur qualité de représentation. Ce qui permet de repérer les facteurs différenciant bien les moyennes des populations. Nous calculons aussi le pourcentage d'inertie d'un facteur apportée par les centres de gravité, ou "inertie inter". Plus ce pourcentage, qui est compris entre 0 et 1 est grand, plus ce facteur est un facteur d'opposition entre les populations.

II.b - Les axes d'inertie

Nous voulons comparer en plus du comportement moyen, les structures des tableaux. Pour cela, nous proposons de comparer les facteurs des différents tableaux et du tableau somme. Cette comparaison se fera en termes d'angles entre facteurs. Elle pose un problème de métrique car les métriques induites par chaque tableau sont liées à leurs marges et sont généralement différentes.

La solution que nous proposons est de se placer, comme pour les centres de gravité dans l'espace $\mathbb{R}_{I \cup J}$. On sait que, (cf. 1 et 5), dans cet espace, les axes d'inertie d'un nuage, calculés pour sa métrique propre, ont pour image par cette métrique les facteurs sur $I \cup J$ du tableau associé. Nous calculerons les cosinus des angles entre les premiers axes d'inertie de chaque sous nuage et ceux du nuage entier, en munissant $\mathbb{R}_{I \cup J}$ de la métrique moyenne associée à la somme de tous les tableaux. Nous représentons graphiquement les projections des vecteurs unitaires des axes d'inertie des sous nuages sur les plans factoriels du nuage entier. Ceci permet de repérer facilement les facteurs des sous populations proches des facteurs de la population entière, et lorsqu'ils sont bien représentés, les facteurs des sous populations proches entre eux.

III - LES FORMULES

Nous donnons ici les formules permettant de calculer les indices introduits ci-dessus.

Notons R le nombre de tableaux à comparer.

Notons k_{ij}^r le terme général du r -ième tableau de contingence, k_i^r l'effectif de la population qui le définit, k_i^r et k_j^r ses marges. Notons λ_s^r sa valeur propre d'ordre s et \mathcal{F}_s^r et \mathcal{G}_s^r les facteurs associés à λ_s^r .

Notons de même, mais sans l'indice supérieur r , k_{ij} le terme général de la somme des R tableaux, k_i et k_j ses marges, k l'effectif total, λ_s , \mathcal{F}_s et \mathcal{G}_s ses valeurs propres et ses facteurs.

Notons \mathcal{H}_s le facteur d'ordre s sur $I \cup J$ du tableau disjonctif complet équivalent. La valeur propre associée à \mathcal{H}_s vaut $\mu_s = (1 + \sqrt{\lambda_s})/2$ (cf. 1 et 5). Le facteur \mathcal{H}_s s'obtient en multipliant la fonction obtenue en juxtaposant \mathcal{F}_s et \mathcal{G}_s par $\sqrt{\mu_s/\lambda_s}$ afin que sa norme soit égale à μ_s .

. Le centre de gravité G_r de la population associée a pour coordonnées sur \mathcal{H}_s :

$$\mathcal{H}_s(G_r) = \frac{1}{\sqrt{\lambda_s}} \left\{ \sum_{i \in I} \frac{k_i^r \mathcal{F}_s(i)}{2k^r} + \sum_{j \in J} \frac{k_j^r \mathcal{G}_s(j)}{2k^r} \right\}$$

. L'inertie totale de G_r qui permet de calculer sa qualité de représentation sur \mathcal{H}_s vaut :

$$\text{Inertie de } G_r = \frac{k^r}{k} \left\{ \sum_{i \in I} \left(\frac{k_i^r}{2k^r} \right)^2 \frac{2k}{k_i} + \sum_{j \in J} \left(\frac{k_j^r}{2k^r} \right)^2 \frac{2k}{k_j} - 1 \right\}$$

. Le pourcentage d'inertie inter de \mathcal{H}_s vaut :

$$\sum_{r=1}^R \frac{k^r}{k} \{ \mathcal{H}_s(G_r) \}^2 \frac{1}{1 + \sqrt{\lambda_s}}$$

. Le cosinus de l'angle entre l'axe d'inertie d'ordre s du nuage entier et celui d'ordre t de la population définissant le r -ième tableau vaut :

$$\cos(h_s, h_t^r) = \frac{1}{\sqrt{\lambda_s} a_t^r} \left\{ \sum_{i \in I} \mathcal{F}_s(i) \frac{k_i^r}{2k^r} \mathcal{F}_t^r(i) + \sum_{j \in J} \mathcal{G}_s(j) \frac{k_j^r}{2k^r} \mathcal{G}_t^r(j) \right\}$$

où a_t^r qui est la norme de $(\mathcal{F}_t^r, \mathcal{G}_t^r)$ vaut :

$$a_t^r = \sum_{i \in I} \left(\frac{k_i^r}{k^r} \mathcal{F}_t^r(i) \frac{k}{k_i} \right) + \sum_{j \in J} \left(\frac{k_j^r}{k^r} \mathcal{G}_t^r(j) \frac{k}{k_j} \right)$$

IV - EXEMPLE

IV.a - Les données

Ces données sont extraites de [7] où elles illustrent l'analyse factorielle sphérique appliquée à la comparaison de deux tableaux de contingence.

Il s'agit de 2 petits tableaux répartissant la population des élèves sortis du système éducatif et ayant trouvé un emploi suivant les deux caractères : I emploi occupé et J niveau du diplôme. L'un concerne la population féminine, l'autre la population masculine.

	Agriculteur	Ingénieur	Technicien	Ouvrier qualifié	Ouvrier non qualifié	Cadre Supérieur	Cadre Moyen	Employé qualifié	Employé non qualifié
Sans diplôme	15068	0	302	10143	59394	596	2142	5445	4879
B.E.P.C.	2701	337	1697	3702	8087	298	289	7348	4987
B.E.P. ou C.A.P.	5709	309	2242	30926	17862	892	672	4719	1514
Bac général	297	917	1969	314	2887	1227	6495	4353	3478
Bac Technique	1242	0	1399	1861	1696	298	924	1280	886
DEUG	0	308	357	0	0	2362	2807	614	1326
DUT ou BTS	322	0	1943	0	0	318	2301	982	0
Diplome Supérieur	0	4383	381	337	323	6781	4030	0	661

Population féminine

Sans diplôme	5089	0	281	7470	29997	0	1577	21616	19849
B.E.P.C.	1212	0	0	1859	4334	0	1806	19915	7325
B.E.P. ou C.A.P.	1166	0	320	4017	4358	0	4549	32452	6484
Bac général	0	316	320	1752	1882	2236	17063	16137	5111
Bac Technique	0	0	283	657	0	595	875	5865	898
DEUG	0	0	0	0	0	911	4152	1256	294
DUT ou BTS	0	304	683	285	0	569	15731	3332	635
Diplome Supérieur	0	1033	0	0	0	6780	3491	1286	0

Population masculine

IV.b - Les indices obtenusProjections des centres de gravité sur les facteurs et qualité de représentation

	F ₁	QLT	F ₂	QLT	F ₃	QLT	F ₄	QLT
Homme	127	068	260	286	-036	006	-334	474
Femme	-119	068	-246	286	035	006	318	474

Cosinus des angles entre les axes d'inertie

		\mathcal{F}_1	\mathcal{F}_2	\mathcal{F}_3	\mathcal{F}_4
Hommes + Femmes					
Homme	H ₁	0.95	-0.14	-0.03	-0.21
	H ₂	-0.07	0.84	0.20	-0.36
	H ₃	-0.08	-0.30	0.85	-0.33
	H ₄	-0.09	-0.10	-0.28	-0.60
Femme	F ₁	0.95	0.13	0.05	0.21
	F ₂	-0.01	0.95	0.08	0.15
	F ₃	0.03	-0.04	0.80	0.46
	F ₄	-0.07	-0.01	-0.19	0.54

IV.c - L'interprétation des résultats

Le plan des deux premiers facteurs de la somme des deux tableaux (voir figure 1) traduit un seul phénomène : le classement des diplômes du plus élevé au plus faible, associé au classement des catégories d'emplois suivant le même critère. En effet, les points sont répartis sur une parabole, suivant un effet classique appelé "effet Guttman".

Les centres de gravité

Nous avons projeté les centres de gravité des populations masculines et féminines. L'importance de l'écart de ces centres de gravité par rapport au centre de gravité général est mesuré par les pourcentages d'inertie "inter" de chaque facteur. Pour le premier facteur, ce pourcentage est faible, il est de l'ordre de 1 %, pour le second, il est déjà plus important et vaut 5,5 %.

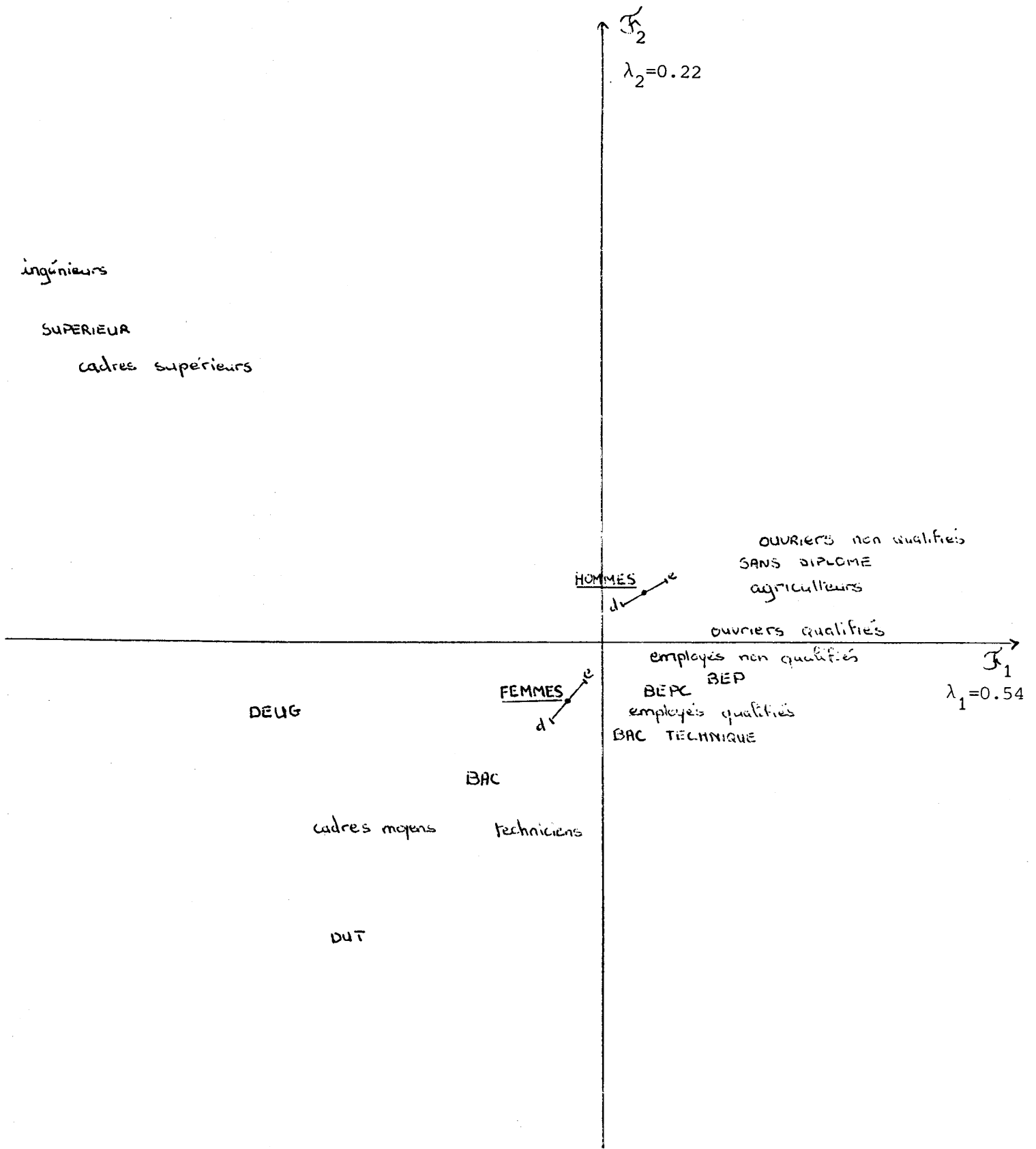


Figure 1 - Plan des deux premiers facteurs :

Projection des centres de gravité HOMMES et FEMMES au milieu des segments joignant les projections des marges sur les diplômes (notés d) et des marges sur les emplois (notés e).

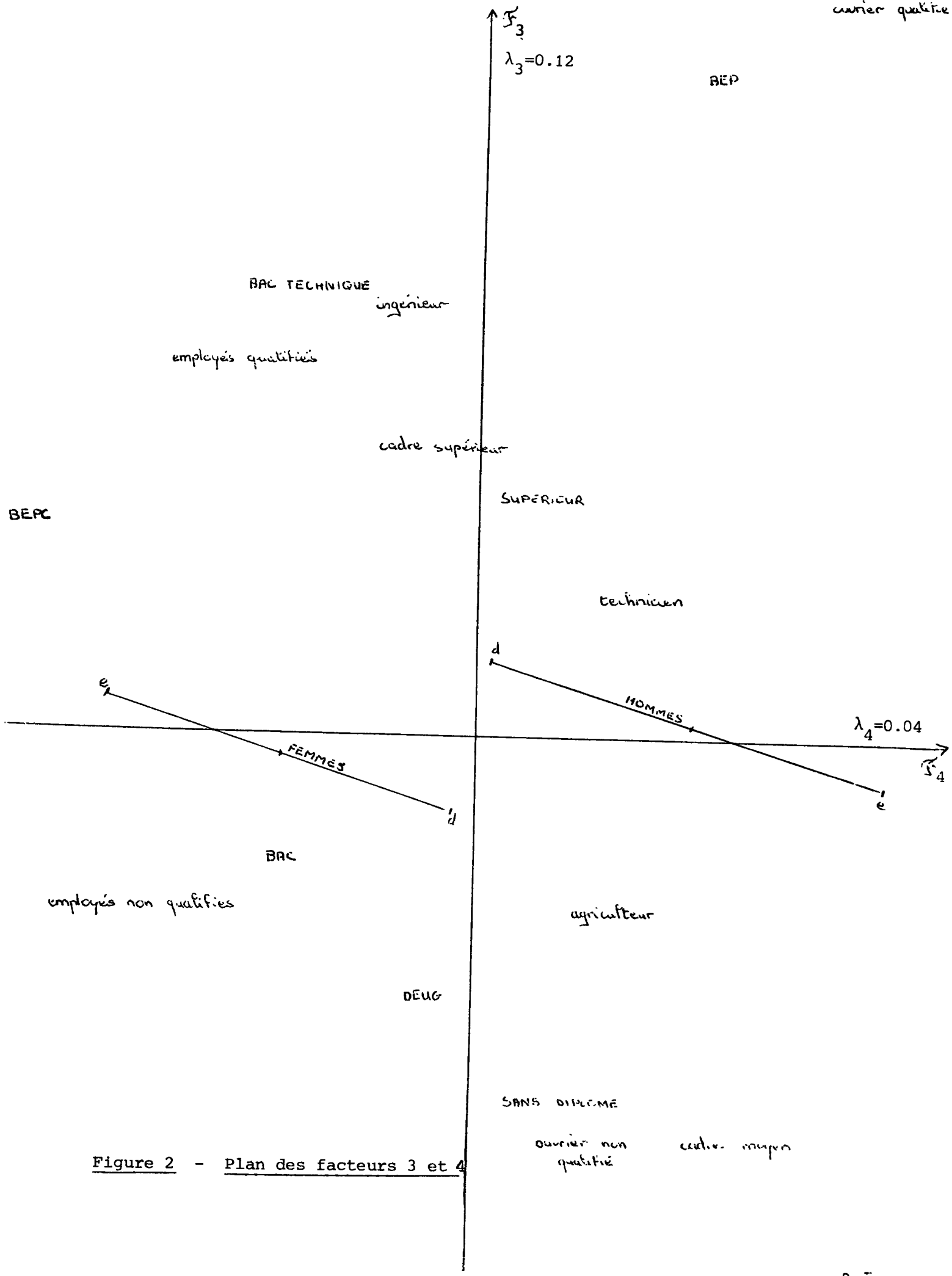


Figure 2 - Plan des facteurs 3 et 4

Le centre de gravité des hommes est situé à droite sur le premier facteur, ce qui signifie, qu'en moyenne, les hommes ont des catégories d'emplois et des diplômes légèrement moins élevés. Sur le second facteur, où les deux populations se différencient plus, les femmes sont situées en-dessous, ce qui indique que les femmes ont plus d'emplois et de diplômes moyens, les hommes étant proportionnellement plus nombreux aux deux extrémités des échelles des emplois et des diplômes.

Ces centres de gravité sont situés au milieu d'un segment que l'on a figuré sur le graphique. Les extrémités de ce segment s'interprètent dans le cadre de l'analyse des tableaux de fréquence, alors que le milieu s'interprétait dans le cadre de l'analyse du tableau disjonctif complet. Ce sont les projections des profils moyens des populations pour les catégories d'emplois d'une part, pour les types de diplômes d'autre part. La position de ces points permet de préciser les conclusions précédentes : les extrémités des segments concernant les catégories d'emplois étant les plus éloignées du centre de gravité, les tendances indiquées ci-dessus sont plus marquées pour les catégories d'emplois que pour les diplômes.

Sur le plan des 3ème et 4ème facteurs, cet écart entre les catégories d'emplois et les diplômes est encore plus net, surtout le 4ème facteur qui oppose ouvriers et employés. Les hommes occupent en moyenne beaucoup plus d'emplois d'ouvriers et les femmes beaucoup plus d'emplois d'employés ; ceci ne peut être attribué aux diplômes possédés, puisque sur ce facteur les moyennes des hommes et des femmes ne se différencient pas.

La qualité de représentation des centres de gravité est faible sur le 1er et le 3ème facteur, et assez importante (voir tableau) sur le 2ème et surtout le 4ème. Ceci montre que l'opposition moyenne entre les hommes et les femmes est due, d'abord, au fait que les femmes occupent plus d'emplois d'employés tandis que les hommes occupent plutôt les emplois d'ouvriers, (4ème facteur) ; ensuite, au fait que les hommes sont plus nombreux aux deux extrémités de l'échelle des emplois et des diplômes, cette tendance étant plus marquée au niveau des emplois que des diplômes (2ème facteur).

Remarquons qu'une simple analyse des centres de gravité, proposée par exemple en [2] pour des tableaux de mesure et en [5] pour des tableaux de contingence n'aurait pu mettre en évidence ces deux facteurs indépendants. En effet pour deux tableaux, nous n'aurions qu'un facteur.

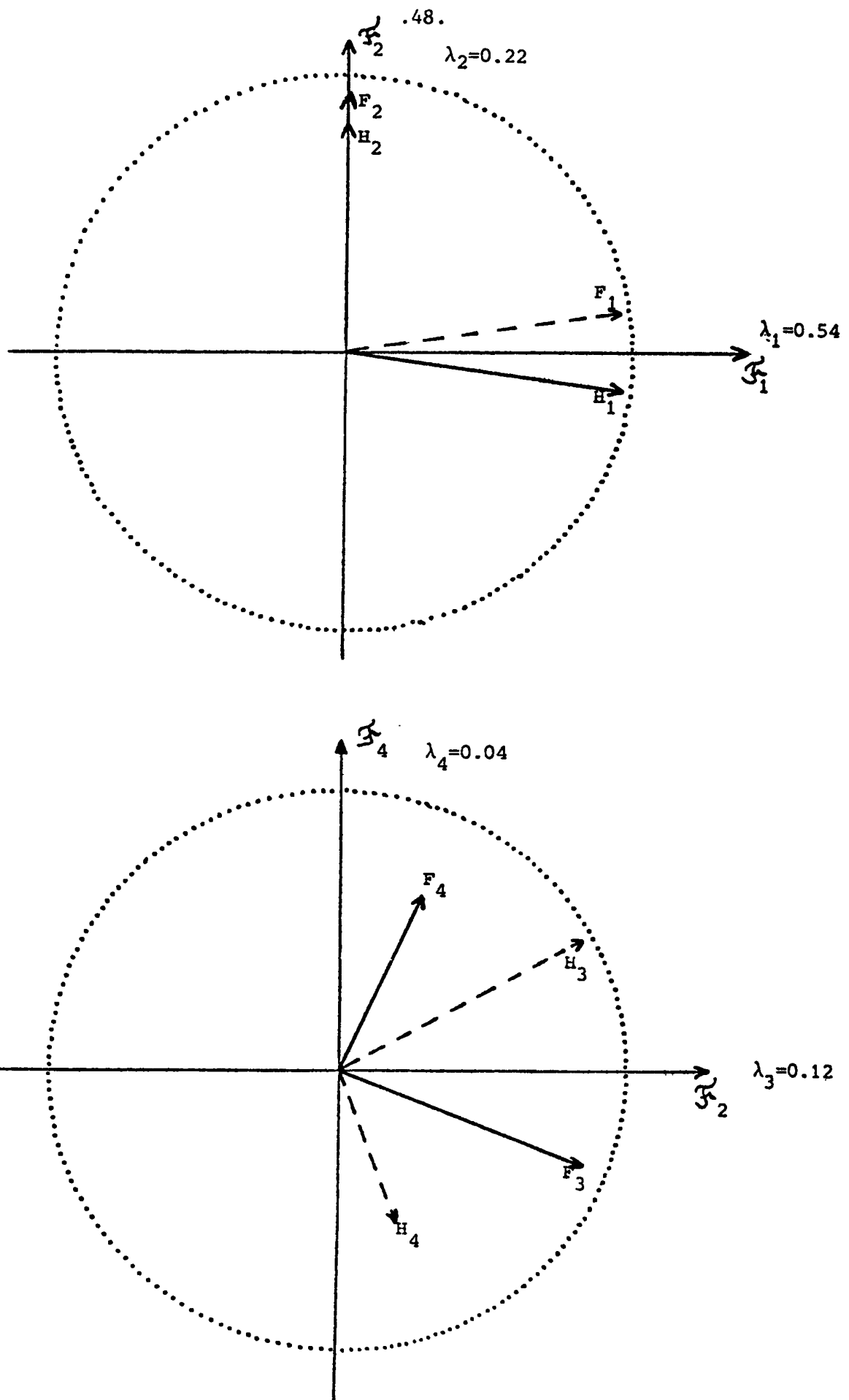


Figure 3 - Projection des vecteurs unitaires des axes d'inertie de la population masculine (H) et féminine (F) sur les plans principaux d'inertie de la population entière.

Les axes d'inertie

Nous avons projeté les vecteurs unitaires des axes d'inertie des nuages de chacune des populations sur les axes d'inertie du nuage de la population entière (figure 3). Sur ce graphique, nous avons tracé le cercle unité qui permet de juger de la qualité de représentation des vecteurs. Seuls les vecteurs assez bien représentés ont été tracés pour rendre la lecture du graphique plus simple.

a) Le plan des deux premiers axes d'inertie.

Les deux premiers axes des trois nuages sont très proches, les deux premiers facteurs sont donc très proches, l'effet Guttman traduit dans le premier plan est un effet commun dans les tableaux.

b) Le plan des axes 3 et 4.

L'axe 3 du nuage entier est une moyenne entre les axes 3 du nuage des deux populations ; l'angle entre ces deux derniers est assez grand, les facteurs d'ordre 3 des deux tableaux sont assez différents. L'axe 4 n'est proche d'aucun des axes des deux tableaux. C'est un facteur qui n'apparaît qu'avec la population entière.

V - CONCLUSION. COMPARAISON AVEC D'AUTRES METHODES

Il semble que les indices proposés permettent de dégager rapidement les grandes lignes d'une comparaison entre plusieurs tableaux de fréquence.

Pour l'étude des tendances moyennes, il paraît déjà clair sur un exemple aussi simple que celui-ci, que l'étude de leur projection sur un tableau moyen est plus riche et plus instructive que l'analyse simple de ces centres de gravité proposée en [2] pour des tableaux de contingence. De plus, les projections simultanées des deux marges, permettent d'affiner considérablement l'interprétation des positions de ces centres de gravité. Les pourcentages d'inertie "inter" donnent immédiatement une bonne idée de l'importance de leur écart à la moyenne.

Les projections des vecteurs unitaires des axes d'inertie ont permis de repérer les facteurs communs aux deux tableaux.

Le choix du tableau somme comme base de représentation est assez logique, car il concerne l'union des populations. De plus, c'est un tableau peu sensible aux aléas. Les formules sont encore valables si l'on pondère différemment chaque tableau, en particulier si l'on donne un poids nul à un certain nombre

d'entre eux qui sont ainsi traités en "tableaux supplémentaires".

Il existe d'autres techniques de comparaison de tableaux de contingence qui peuvent éventuellement compléter -ou être complétées- par celle que nous proposons ici car les points de vue sont différents.

Citons deux d'entre elles qui mettent l'accent sur les différences entre les tableaux. L'une d'elles est une analyse des différences des profils de deux tableaux. C'est l'analyse factorielle sphérique [cf. 7]. Elle s'applique lorsqu'il n'y a que deux tableaux. L'autre, au contraire, n'a d'intérêt que si les tableaux sont en assez grand nombre. Elle permet, en définissant une distance entre les tableaux, et en les représentant ensuite par des points, de décrire leurs proximités. C'est l'analyse des opérateurs associés aux tableaux [cf. 6].

Remarque :

Nous avons proposé comme sous produit d'une technique d'analyse rapide de grands tableaux par division en sous tableaux, des indices de comparaison de sous tableaux [cf. 4]. Ceci en est une application directe. En effet, les tableaux disjonctifs complets associés à chaque population apparaissent comme des sous tableaux de celui qui représente la population entière. Les indices de [4] peuvent donc être calculés ; l'équivalence entre les analyses des tableaux disjonctifs complets et des tableaux de contingence permet de les obtenir en travaillant seulement sur ces derniers.

BIBLIOGRAPHIE

- [1] BENZECRI J.P. et collaborateurs
"L'analyse des données" - Dunod 1973.
- [2] BOUROCHE S.M.
"Analyse des données ternaires ; la double analyse en composantes principales" - Thèse de 3ème cycle, Paris 6, 1975.
- [3] ESCOPIER B.
"Stabilité et approximation en analyse factorielle" - Thèse d'Etat, Paris 6, 1979.
- [4] FOUCART T.
"Sur les suites de tableaux de contingence échelonnés dans le temps" - Statistique et Analyse des Données n° 2, 1978.
- [5] LEBART L., MORINEAU A., TABARD N.
"Technique de la description statistique" - Dunod 1977.
- [6] PAGES J.P., ESCOUPIER Y., CAZES P.
"Opérateurs et analyse des tableaux à plus de deux dimensions" - Cahiers du BURO, 1976.
- [7] VOLLE M., DOMENGENES D.
"Analyse factorielle sphérique : une exploration" - Annales de l'I.N.S.E.E. n° 35, 1979.