Année psychol.
1976, *76*, 133-144

NOTE

A PROPOS DE « L'ANALYSE DES DONNÉES » SELON BENZÉCRI[1]

par Henry ROUANET
U.E.R. de Mathématiques, Université René-Descartes[2]

et Dominique LÉPINE
Laboratoire de Psychologie expérimentale et comparée[3]
Université René-Descartes et E.P.H.E., 3e section
(associé au C.N.R.S.)

suivi d'une Lettre de commentaires
de J.-P. BENZÉCRI

Les méthodes de Benzécri connaissent une telle diffusion, surtout l'analyse des correspondances qui fait désormais partie des grandes « méthodes de routine » d'analyse des tableaux d'effectifs, que la parution d'un ouvrage d'ensemble consacré à ces méthodes nous a paru devoir susciter davantage qu'un simple compte rendu. Sans viser à une critique épistémologique, il nous a semblé opportun de présenter quelques réflexions, issues d'une « lecture » ne prétendant nullement à une « objectivité » assez illusoire ; ces réflexions, nous les proposerons à un lecteur familier avec la tradition psychométrique en psychologie et désireux d'approfondir sa connaissance des méthodes de Benzécri, voire éventuellement de les acquérir (à cet égard, nous pensons que le présent ouvrage peut constituer la meilleure introduction existant à ce jour).

Quiconque a fréquenté tant soit peu Benzécri sait que la psychologie n'est que l'un de ses multiples domaines d'intérêt ; dans la « somme » qu'est le présent ouvrage, on trouvera des applications aux secteurs les plus variés des sciences de la nature et des sciences de l'homme

1. J.-P. BENZÉCRI et coll., *L'analyse des données*, Paris, Dunod, 1973 ; t. I : *La taxinomie*, VIII + 615 p. ; t. II : *L'analyse des correspondances*, VI + 619 p.
2. 12, rue Cujas, 75005 Paris.
3. 28, rue Serpente, 75006 Paris.

(lesquelles, à vrai dire, pour Benzécri comme pour Aristote, font partie des sciences de la nature), mais assez peu de travaux se situant dans le domaine de la psychologie proprement dite (les beaux exposés sur la psychophysique, la vision des couleurs, etc., que Benzécri avait partiellement diffusés sous forme ronéotée, n'y figurent malheureusement pas ; sans doute feront-ils l'objet de publications ultérieures). En ce qui concerne la psychologie, l'intérêt du présent ouvrage réside surtout dans les méthodes.

Si l'on distingue, avec Cronbach et bien d'autres, dans la psychologie dite scientifique, la tradition « expérimentaliste », qui met l'accent sur le *design* et les hypothèses, et ce qu'on peut appeler la tradition « psychométrique », qui met l'accent sur la recherche des structures, on peut dire sans crainte de se tromper que Benzécri se situe à l'intérieur de cette seconde tradition, et ce n'est qu'en s'y plaçant également, croyons-nous, qu'on peut formuler des commentaires un peu féconds. Manifestement, l'auteur a une connaissance de première main des travaux essentiels de cette école anglo-saxonne qui, sur les traces de Thurstone, a élaboré à peu près toutes les méthodes dont le but est de dégager les structures d'un ensemble de données multivariées — méthodes qui n'ont pas attendu, bien sûr, pour être mises en œuvre avec fruit, l'avènement du FØRTRAN, et qui, depuis cet avènement, n'en sont que davantage florissantes. Malheureusement, le lecteur désireux de situer avec précision, soit les techniques spécifiques préconisées par Benzécri, soit ses options méthodologiques fondamentales, parmi celles des auteurs psychométriciens familiers, devra souvent se contenter de remarques incidentes et d'allusions (à l'exception notable de Guttman, qui fait l'objet de références nombreuses)[1]. Sur certains points techniques importants, parce que soulevant des problèmes méthodologiques fondamentaux, tels que le problème des communautés, celui des rotations, etc., il restera sans doute sur sa faim. En revanche, d'autres points essentiels sont largement développés et illustrés, notamment l'intérêt méthodologique de conjuguer les deux approches corrélatives que sont la typologie (on regroupe les individus en classes selon des partitions ou des hiérarchies) et l'analyse factorielle (on dérive de nouvelles variables, éventuellement mais non nécessairement en nombre plus restreint, de manière à faire apparaître des structures), la deuxième approche étant, dans l'esprit de Benzécri, sensiblement privilégiée (cf. par ex. t. 1, p. 98).

Cela dit, la lecture de l'ouvrage, surtout s'il s'agit d'un premier contact avec les œuvres de l'auteur, risque d'être assez malaisée pour plusieurs raisons. Tout d'abord, l'organisation d'ensemble n'est guère « cartésienne » ; les différents chapitres reprennent des exposés adressés à

1. Dans un texte récent, actuellement sous forme ronéotée, *Histoire et préhistoire de l'analyse des données*, BENZÉCRI vient de combler, en grande partie, cette lacune.

des publics extrêmement divers tant par leurs intérêts que par leur niveau mathématique ; la division en deux tomes, consacrés à la typologie (resp. à l'analyse des correspondances) et dédiés à Linné (resp. à Huyghens) est surtout pertinente pour l'informaticien ; le lecteur avant tout désireux de se faire une idée des lignes méthodologiques essentielles risque de trouver le texte assez touffu. Ensuite, une difficulté que nous nous garderons de sous-estimer (mais qui une fois surmontée ouvrira des perspectives inattendues ; nous y reviendrons) réside dans le langage mathématique adopté par l'auteur ; contrairement à la plupart des auteurs psychométriciens, qui utilisent la « langue vulgaire » des mathématiques appliquées (à base essentiellement de calcul matriciel), Benzécri s'exprime dans la « langue liturgique » des mathématiciens, que ceux-ci réservent en général aux exposés de « mathématiques pures ». C'est pourquoi avant d'entreprendre le « voyage », il pourra être recommandé de se munir d'un manuel de Terminale C (qui contiendra tout le vocabulaire de base) pour le consulter en cas de besoin. Dans le même esprit, il pourra être utile d'adopter un « itinéraire » autre que la lecture linéaire ; personnellement, nous avons adopté l'itinéraire suivant, que nous reproduisons ici à titre indicatif :

1re étape : Lire, en guise d'introduction d'ensemble, le chap. t. I, A, n° 2 (p. 15 à 61) ;

2e étape : Examiner un chapitre portant sur une application concrète ; le lecteur aura l'embarras du choix entre les « Peurs enfantines », t. I, p. 513 ; les « Professions de foi des députés élus en 1881 », t. II, p. 326 ; les « Niveaux et conditions de vie au Liban », t. II, p. 344, etc.

3e étape : Sur le tableau de contingence le plus proche procéder, sans plus tarder, à une « analyse des correspondances », la fameuse « méthode-carrefour » de Benzécri. Il est essentiel que le lecteur acquière une bonne familiarité avec cette méthode, tellement plus informative, dans tant de cas, que le khi-deux traditionnel (dont elle constitue d'ailleurs, du point de vue descriptif, une généralisation), même si les problèmes d'interprétation que pose la « représentation simultanée » sont parfois délicats. Bien entendu, pour que l'apprentissage se fasse avec fruit, on évitera de faire fonctionner, sur l'exemple choisi, un « programme en boîte », mais on procédera progressivement, en contrôlant les étapes successives du calcul, et en ne confiant à l'ordinateur (au moins dans cette phase d'apprentissage) que les parties vraiment malaisées des calculs, c'est-à-dire essentiellement la diagonalisation (et même, grâce aux récentes calculatrices de poche, la diagonalisation devrait pouvoir être effectuée à la main si la plus petite dimension du tableau analysé ne dépasse pas quatre).

4e étape : Etudier de près le chap. t. II, A, n° 2 : « Pratique de l'analyse des correspondances ».

5e étape : Parcourir les deux tomes au gré de ses intérêts et de sa fantaisie, en se gardant, au cours du voyage, d'une attitude de « touriste

pressé », qui risquerait de faire manquer les panoramas grandioses et les échappées pittoresques pas toujours bien signalés ; savoir admirer, le moment venu, ce style hardi et singulier, dosage inimitable des trois registres : littéraire, haute mathématique et informatique.

Sur le terrain, pour s'orienter, on cherchera quel est le sens, au point où on se trouve, à attribuer au terme d' « inertie » que Benzécri emploie pour désigner tout bonnement tantôt une somme de carrés (appelée également moment centré d'ordre 2), tantôt une moyenne de carrés, c'est-à-dire une variance (la distinction étant bien entendu sans importance lorsqu'on considère des « fractions d'inertie » ; pour une discussion d'ensemble, voir par ex. t. I, p. 185).

En revanche, il conviendra de ne pas se laisser égarer par toute une catégorie de panneaux trompeurs, mais heureusement faciles à identifier, puisque tous marqués du signe « probabilité ». En effet, le terme de « probabilité » est employé par Benzécri pour désigner indistinctement toute mesure positive de masse totale égale à l'unité, quel que soit son statut méthodologique, alors même que cette mesure peut n'avoir rien à voir avec la formalisation d'un processus faisant intervenir le « hasard », comme c'est le cas par exemple pour une distribution de fréquences conditionnelles. Entraîné par cette terminologie « panprobabiliste », l'auteur en arrive à parler parfois de la corrélation comme d'une « notion probabiliste » ; mais à ce compte, pourquoi pas également la moyenne ? Curieuse « confusion » entre le langage probabiliste et le langage proprement statistique, d'autant plus curieuse, du moins à première vue, chez un auteur qui, par ailleurs, consacre tant de soin et d'énergie à développer la distinction épistémologique (qu'il voudrait radicale) entre « probabilité » (entendue alors bien sûr comme une formalisation du hasard) et « statistique » (au sens précis de procédures à base de dénombrements), au point d'en faire son premier principe : « Statistique n'est pas probabilité » (t. II, p. 3). L'origine de ce petit mystère n'est pas très difficile à découvrir : si l'on se place avec Benzécri dans la tradition « psychométrique » la plus stricte, la description statistique d'une population revêt, en quelque sorte automatiquement, une interprétation probabiliste dès que cette population est regardée, ne serait-ce qu'implicitement, comme un échantillon plus ou moins représentatif d'une population de référence (techniquement parlant, une statistique descriptive reçoit alors automatiquement le statut d'une statistique estimatrice). Du moment que, dans cette perspective, l'accent n'est pas mis sur les procédures effectives d'échantillonnage, il n'y a pas de contre-indication absolue à en rester à un certain niveau de syncrétisme entre les langages probabiliste et statistique, et à parler d'une distribution statistique comme d'une sorte de distribution de probabilité en acte. Du point de vue rhétorique, cette façon de parler procure même le bénéfice secondaire de donner au discours plus de « champ » ou de profondeur apparents. Mais en contrepartie, quelle source de confusion

et de contresens, surtout quand on sait à quels abus de la terminologie probabiliste se livrent tant de mathématiciens, qui eux sont naïfs et dont le niveau de conscience méthodologique n'est pas la vertu dominante... Mais nous n'irons pas plus loin dans cette voie, de peur de passer, vis-à-vis des abus de la probabilité, pour plus vigilants que Benzécri lui-même...

Le lecteur qui aura fait un certain effort d'adaptation appréciera rapidement, nous en sommes persuadés, ce qui constitue selon nous l'apport le plus authentique de Benzécri : l'éclairage qu'apporte, pour la bonne compréhension des méthodes factorielles ou typologiques, l'explicitation des structures mathématiques sous-jacentes et, plus précisément, la conjugaison judicieuse des deux types de discours que rend possible l'usage de la « langue noble » des mathématiques, à savoir le discours « géométrique » (plus synthétique et intuitif) et le discours « algébrique » (plus analytique et discursif) ; alors que l'habituel discours matriciel dégénère trop souvent en un simple sabir, capable seulement de communiquer un savoir-faire (la « recette » !) sans donner la moindre lumière sur le savoir qui le sous-tend. Nous pensons que le gain en compréhension ne devrait pas avoir que des avantages d'ordre esthétique : la prise de conscience du caractère surdéterminé des structures mathématiques devrait, à notre avis, contribuer puissamment à donner aux utilisateurs un sens plus juste des contraintes réelles de l'outil mathématique, d'où par implication, des véritables points de choix méthodologiques.

C'est de ce point de vue qu'il faut apprécier tant de commentaires de Benzécri, qu'il s'agisse des réflexions sur le caractère privilégié des structures euclidiennes (par ex. t. II, p. 31 et p. 76), ou des déclarations telles que celle-ci, qui est bien plus qu'une boutade : « Une analyse de données n'est, en bonne mathématique, qu'une recherche de vecteurs propres ; toute la science, ou tout l'art... étant de savoir quelle matrice traiter » (t. II, p. 23), etc. Ce sont bien ces conceptions qui conduisent l'auteur à tant de vues renouvelées sur mainte méthode ou « idée » courante sur laquelle on aurait pu penser que « tout avait déjà été dit », comme les échelles de Guttman (t. II, p. 192), les modèles de processus proposés par le même Guttman (t. II, p. 203), le rôle de la distribution normale (t. I, p. 406 à 417), etc. Laissant au lecteur le soin d'allonger cette liste, nous proposerons maintenant quelques réflexions d'ordre plus général, suscitées par l'expression d' « analyse des données » chère à l'auteur.

Que faut-il entendre par « analyse des données » ? L'expression est, au moins depuis une quinzaine d'années, très courante chez les Anglo-Saxons *(data analysis)*, aussi bien chez les théoriciens que chez les usagers de la statistique ; toujours utilisée, outre-mer, de façon très souple, elle désigne, non pas vraiment un ensemble de techniques, et encore moins une « doctrine établie », mais plutôt « une certaine idée

de la statistique », selon laquelle il est légitime *en principe* (même si dans la pratique cela ne va pas toujours sans problèmes) d'examiner les données pour les interpréter, quelles que soient les intentions et les modalités qui ont pu présider à leur recueil, et sans avoir à s'enfermer dans un modèle ou des hypothèses restrictives. Cette conception, qui pour un psychologue (surtout de tempérament un peu clinicien) pourrait aller presque de soi, a dû en fait se constituer et s'affirmer en réaction contre les excès de l'école statistique « décisionniste » naguère dominante, laquelle, selon une déviation certes peu conforme à l'esprit des pères fondateurs de la statistique moderne, en arrivait à ne plus voir dans les données qu'une sorte d'intermédiaire destiné à permettre de prendre mécaniquement une « décision » (celle-ci d'ailleurs en général toute formelle) dont tous les termes (modèle probabiliste, mais aussi, le cas échéant, fonction de coût, probabilités *a priori*, etc.) devaient (ou auraient dû), toujours en principe, être posés au départ. Bien sûr, dans leurs pratiques, les usagers avertis ne manquaient pas de prendre leurs distances vis-à-vis de telles thèses, mais ce qu'a apporté le courant de l' « analyse des données », bien plus encore que de nouvelles techniques, c'est l'assurance sans cesse grandissante que de telles pratiques, loin d'être « honteuses », peuvent souvent recevoir un fondement théorique, et que c'était peut-être, au contraire, l'école décisionniste qui poursuivait une chimère. Outre-mer, on ne compte plus, aujourd'hui, les statisticiens dont la problématique est directement inspirée des perspectives de l' « analyse des données ». Adhésion d'autant plus large que là-bas, adopter ces perspectives apparaît en parfaite harmonie avec d'une part la tradition de la statistique dans l'expérimentation et d'autre part l'utilisation de méthodes inférentielles ; rappelons que Tukey qui, en 1955, en pleine mode « décisionniste », proclamait l'une des thèses essentielles de l' « analyse des données », selon laquelle l'objectif principal de la statistique doit être de conduire à des « conclusions » (plutôt qu'à des décisions), développait à la même époque de nouvelles méthodes inférentielles d'analyse des données expérimentales, portant sur les comparaisons *a posteriori*, c'est-à-dire suggérées par l'examen des données mais non prévues par le plan d'expérience.

Si maintenant nous revenons à Benzécri et à son ouvrage : on y retrouve bien, textuellement, le point de vue de l' « analyse des données » tel que nous venons de le caractériser : il faut, dit-il, « se garder de mêler trop intimement, à ce que nous observons et mesurons, ce que nous pensons en être la structure sous-jacente » (t. II, p. 16). Mais chez Benzécri, ce point de vue prend des allures plus extrêmes (même en laissant de côté les prises de position anti-expérimentalistes) :

— d'une part, le point de vue de l'analyse des données se transforme en un principe radical qui tend à éliminer, (plutôt qu'à contre-balancer) tout autre point de vue, notamment le point de vue décisionnel ;

— d'autre part, l'expression même d' « analyse des données » tend à désigner également les méthodes favorites de l'auteur.

On pourrait discuter assez longuement sur l'origine de ce décalage, dont certainement pourrait rendre compte au moins en partie le contexte assez particulier dans lequel s'était développée et surtout enseignée dans les départements scientifiques français la statistique dite mathématique. Disons simplement qu'en prouvant, par la théorie et par l'exemple, qu'il est possible de développer des méthodes d'analyse des données fondées en droit et applicables en fait, en rendant, par son enseignement, les noms de Thurstone ou de Guttman familiers (mieux encore, respectables, à des étudiants scientifiques, il nous apparaît suffisamment clair que l'action de Benzécri dans le contexte français des rapports entre mathématiques et sciences humaines, et en dépit des aspects parfois un peu provocants qu'elle a pu prendre, a été réellement démystificatrice et partant fondamentalement bénéfique.

Le lecteur qui examinera le texte de Benzécri d'assez près partagera sans doute notre jugement, lorsqu'il constatera la grande prudence manifestée par l'auteur, dès que l'on quitte les généralités et qu'on en vient aux modalités concrètes. Ainsi, lorsqu'il s'agit d'élargir la méthode d'analyse des correspondances, initialement conçue pour l'analyse des tableaux de contingence, pour en faire une méthode « universelle » de traitement des tableaux de nombres positifs. Cette extension n'est possible, nous précise soigneusement l'auteur (t. II, p. 21-23) que si deux exigences fondamentales sont remplies. La première, qu'il appelle « homogénéité », exprime qu'on peut trouver « une unité de mesure qui conserve à peu près le même sens sur toute l'étendue du tableau » ; dans un langage sans doute plus familier au psychologue, c'est donc l'exigence de *comparabilité* ou, si l'on préfère, la nécessité de se donner un espace d'observation. La deuxième, qualifiée d' « exhaustivité » (encore une terminologie singulière, mais toute ambiguïté est ici écartée par le contexte), est tout simplement l'exigence non moins familière de *représentativité* : les individus, voire les variables, doivent pouvoir être regardés comme des échantillons représentatifs... Bien plus, comme pour bien mettre les « points sur les i », l'auteur nous précise encore que les exigences en question ne sauraient généralement être déclarées remplies à partir de critères exclusivement opérationnels, et que c'est donc à l'usager qu'incombe la responsabilité de considérer que dans une situation donnée elles seront, ou non, admissibles ; car nous lisons bien (t. II, p. 23) que, s'agissant de tableaux numériques quelconques, « la part laissée à l'arbitraire (entendons : au choix raisonné) est forcément grande », et que les résultats ne pourront être sûrs que si le spécialiste qui conçoit l'étude « a le sens des exigences de régularité propres à la statistique » *(sic)*. Voilà des déclarations qui devraient rassurer tous ceux que pourrait inquiéter le succès « quasi universel » (t. II, p. 150)

de l'analyse des correspondances. C'est que l'analyse des correspondances est une procédure ni plus ni moins universelle qu'un décilage de test ou que le calcul d'une variance ; mais la structure qu'elle révèle ne sera la bonne que si le protocole auquel on l'applique est défini à un niveau de codage... judicieux ! De telles déclarations, par ailleurs, nous dispenseront d'avoir à nous étendre sur les maladresses de maint « analyseur de correspondances », vis-à-vis desquels, manifestement, Benzécri prend ses distances...

Chez Benzécri, on constate enfin, et peut-être surtout, la même prudence subtile au niveau des principes qui doivent guider l'interprétation de l'analyse. Lisons, par exemple, les réflexions que l'auteur consacre à ce problème au détour de l'enquête « Zaïre » (t. II, p. 479). Ne nous laissons pas abuser par la déclaration provocante : « En analyse multidimensionnelle, la complexité des informations traitées et des réponses obtenues est telle qu'un résultat issu de fluctuations aléatoires a très peu de chances d'être interprétable ; par conséquent, *on peut légitimement admettre que tout ce qui est interprétable est valide* »[1]. En effet, immédiatement après cette déclaration, le lecteur est invité à ne pas trop se bercer d'illusions, puisqu'on lui précise que la maxime précédente ne saurait s'appliquer qu'à la structure factorielle « globale », et non pas « à la validité des particularités de détail des résultats d'analyse, autrement dit de la précision avec laquelle les facteurs calculés expriment une réalité intrinsèque ». Voilà encore une déclaration qui ne cautionne guère les excès interprétatifs d' « analyseurs de données » trop naïfs ou enthousiastes...

Maintenant, quelles méthodes concrètes Benzécri propose-t-il pour mettre à l'épreuve cette « réalité intrinsèque » des facteurs ? Quels éléments de réponse apporte-t-il à des questions (dont il ne conteste pas la légitimité) du type : Quel est le nombre des facteurs à extraire ? De quelle part de variance « vraie » peut-on admettre qu'ils rendent compte ? Etc. Là, il faut l'avouer, le lecteur risque de rester sur sa faim. Dans le chapitre consacré aux épreuves de validité (t. II, p. 480), Benzécri suggère d'engendrer des tableaux fictifs à partir de l'hypothèse d'indépendance et d'examiner si le tableau réel fournit des valeurs propres plus fortes que celles issues des tableaux fictifs. Malheureusement, d'une part c'est là aborder avec des moyens rustiques un problème technique (celui de la distribution des valeurs propres), dont les difficultés sont bien connues des spécialistes ; d'autre part, et surtout, c'est s'aligner sur la méthodologie inférentielle la plus pauvre et la plus inutile (pour les problèmes qui nous occupent ici) et si l'on n'y prend garde, la plus dangereuse : celle des tests de signification. Pourtant, cette pauvreté et ce danger sont bien connus des psychométriciens avertis, qui ont depuis belle lurette dénoncé les abus des tests de signification. Bien

1. C'est nous qui soulignons.

sûr, Benzécri se garde lui-même de ces abus, mais il n'écarte pas assez catégoriquement les interprétations inadmissibles que favorise, en l'absence d'une analyse sérieuse des objectifs de l'inférence, l'ambiguïté sémantique du mot piège : « significatif ». Ce faisant, il laisse la porte ouverte aux pires distorsions des tests de signification. Malheureusement, il ne s'agit pas là de risques imaginaires : alors que, en bonne logique, on aurait pu espérer que la diffusion de bonnes méthodes descriptives allait tempérer les excès des tests de signification, chez les usagers peu avertis on rencontre désormais des travaux où l'analyse des correspondances est purement et simplement surajoutée à la perpétuation des interprétations les plus contestables du test du khi-deux, sans, bien sûr, qu'aucune liaison sérieuse soit faite entre les deux types d'analyse. De telles pratiques sont à proprement parler assez monstrueuses. A s'être trop aligné (peut-être pour se démarquer le plus possible des positions excessives de l'école décisionniste) sur la problématique inférentielle la plus pauvre, Benzécri a pris ici quelques risques, au moins pour les autres...

On n'en dira pas autant, fort heureusement, en ce qui concerne le problème de l' « interprétation » des facteurs (ou des classes lorsqu'il s'agit de typologie) pour lequel nous trouvons une combinaison bien venue de hardiesse et de prudence. Hardiesse, car l'auteur, ici, ne sous-estime pas l'importance du problème. A propos de l'interprétation du questionnaire des « peurs enfantines », nous lisons : « Il ne s'agit ici rien de moins que de la validité du langage commun et de la place des mathématiques dans les sciences de l'homme » (t. I, p. 520). Prudence, car devant l'exemple concret, on constate à quel point les interprétations de Benzécri visent à « coller au plus près » du contenu manifeste du questionnaire, en évitant de renvoyer à une « psychologie des profondeurs » à laquelle on peut croire ou non, mais à laquelle (la prudence est ici de l'honnêteté) l'analyse des données ne saurait apporter une caution. Quelle peut être alors la nature de l'interprétation issue de l'analyse ? « L'analyse mathématique des données doit imposer au raisonnement usuel une contrainte très forte, une épreuve de laquelle il sorte vrai et efficace » (t. I, p. 520). Donc ni « quantomanie », ni psychologie des profondeurs ; n'est-ce pas là la sagesse suprême pour un statisticien ?

Clarté et fermeté au niveau des principes, liberté et prudence au niveau de la pratique ; sans adhérer à toutes les thèses de l'auteur, et notamment à toutes ses prises de position « anti-expérimentalistes », il est difficile de rester insensible à tant de qualités, quand on songe à quel point le domaine des méthodes mathématiques en sciences humaines peut être envahi par tant de théories inconsistantes et de pratiques serviles.

En réponse au texte qui précède, le professeur Benzécri nous a fait parvenir la lettre suivante :

Puisque vous avez eu la bienveillance de me communiquer l'article que vous avez écrit à propos de *L'analyse des données*, je me permets d'apporter à ce texte quelques brèves notes qui, je pense, pourront l'accompagner dans la revue. Somme toute je ne trouve rien à redire sinon que le ton de votre prose me flatte jusqu'à la confusion d'autant plus que vous avez omis de citer seulement les très nombreux collaborateurs dont j'ai dirigé les travaux et sans qui le livre n'existerait pas. A part cela, il ne s'agira ici que de détails relevés au fil de la lecture.

Quand vous écrivez que j'ai « une connaissance de première main des travaux essentiels de cette école anglo-saxonne... », vous vous trompez quelque peu. J'ai fait la plupart de mes travaux en m'inspirant de ce que je savais par ouï-dire de l'analyse factorielle des psychologues et l'on peut m'appliquer à moi-même cette remarque extraite de mon texte *Histoire et préhistoire...*[1] : « ... il n'est pas impossible qu'un statisticien inspiré par la doctrine d'un psychologue instruit lui-même par un astronome ait cru inventer seul la méthode mathématique utilisée par celui-ci ». C'est seulement pour écrire le texte *Histoire...* que j'ai tenté d'acquérir une connaissance de première main des classiques de notre discipline. *A posteriori*, je pense d'ailleurs qu'une trop grande érudition m'aurait plus embarrassé que servi dans mon entreprise méthodologique.

Si le seul auteur souvent cité dans *L'analyse des données* est Guttman, c'est que d'une part il est quasi le seul auteur dont j'ai eu très tôt une « connaissance de première main » et que, d'autre part, ses premiers travaux, à la différence de la plupart des tendances statistiques dont j'avais eu vent, s'inspirent d'une philosophie que j'approuve.

Qu'après avoir écrit que Statistique n'est pas Probabilité, je me permette ce que vous appelez une curieuse « confusion entre le langage probabiliste et le langage proprement statistique », n'est pas un « mystère » dont l'origine soit « très difficile à découvrir ». Une fois admis que le formalisme probabiliste va au-delà de la nature, où les probabilités stables sont rares, on reste libre d'utiliser le langage probabiliste dont la richesse analogique est très grande. Telle est, sinon la justification, au moins l'explication des us et abus que je fais de ce langage. Vous notez justement que « l'organisation d'ensemble n'est guère cartésienne ». C'est que, d'abord, je ne suis point disciple de ce philosophe qui entend élever sur une table rase l'édifice de la connaissance. Je crois que la pensée progresse toujours, non à partir d'un bout mais à partir

1. Ce long texte, que vous citez obligeamment, paraît en feuilleton dans la revue *Les cahiers de l'analyse des données*.

du milieu. Dans un traité qui, comme celui de *L'analyse des données* s'adresse à des lecteurs extrêmement divers, proposer un parcours linéaire exposait l'auteur à se retrouver seul avant la centième page. Vous avez donc bien deviné que ce livre est soumis au lecteur comme ces partitions dont des compositeurs contemporains demandent à l'orchestre une interprétation aléatoire. Et je vous sais gré d'avoir semé dans ma forêt à l'intention du psychologue les cailloux du Petit Poucet. Mais l'indice systématique très détaillé et dont maint article est rédigé comme une petite leçon de catéchisme est là pour aider chacun à trouver sa voie. Il est encore une autre raison pour que ce livre n'ait pas un plan cartésien : c'est qu'il est l'œuvre de 70 auteurs. Je le répète, tout en assumant la responsabilité de l'ensemble, j'aurais été incapable de produire seul une œuvre aussi vivante dans sa polyphonie.

Les calculs de simulation vous paraissent « un moyen bien rustique » d'aborder le problème technique de la distribution des valeurs propres. Et vous craignez que par là on « s'aligne sur la méthodologie pauvre, inutile et dangereuse des tests de signification ». Si j'ai parlé de tests — mot que je n'emploie jamais qu'entre guillemets — c'est d'abord parce qu'ils sont l'instrument obligé de maint statisticien et qu'il me semblait convenable de jeter entre eux et nous un pont, fût-il précaire. Mais je ne pense pas que les méthodes de simulation méritent votre mépris ; car d'une part ce sont les seules praticables lorsqu'on ne dispose que d'hypothèses très faibles, celles qui justement sont les plus sûres ; et d'autre part là même où l'on dispose d'une expression analytique exacte mais très complexe, le calcul par simulation peut être plus rapide : c'est ce que vient de constater L. Lebart.

L. Lebart a montré que la loi des valeurs propres issues de l'analyse d'un tableau de contingence, dans l'hypothèse d'indépendance peut être convenablement *approchée* par la loi des valeurs propres issues de matrices distribuées suivant la loi de Wishart. Cette dernière loi de valeurs propres est connue depuis Fisher ; mais l'exploitation exacte des formules est un problème d'analyse numérique si complexe, que les calculs de simulation sont le moyen le plus simple d'obtenir quant à l'analyse des correspondances les informations statistiques qui vous intéressent.

Vous employez plusieurs fois le terme *d'anti-expérimentalisme* qui ne m'était pas connu. Je pense que vous faites allusion à cette thèse que l'index systématique à l'article « observation » formule ainsi : l'observation reprend le pas sur l'expérimentation. Je n'entends pas par là bannir toute expérimentation. Voici un exemple : mon ami le Dr Rösh, épidémiologiste distingué et peintre de talent, m'a déclaré lundi dernier avoir passé son dimanche à retrouver le nombre d'or et l'épure du dodécaèdre dans un *Couronnement de la Vierge* et une *Pietà*, tous deux œuvres de l'Ecole d'Avignon. « Mais, me dit-il, comment m'assurer qu'à force de solliciter les traits, je n'aie pas une probabilité élevée de

retrouver le dodécaèdre n'importe où ? Voilà le problème de probabilité géométrique que je vous soumets. » A quoi j'ai répondu : « Ce n'est pas un problème de probabilité géométrique, mais d'abord une question de psychologie expérimentale. Il faut soumettre des tableaux à des sujets assez cultivés et leur demander : où voyez-vous une ligne, un sommet ; ainsi on obtiendra le squelette perçu d'une peinture, le réseau de départ de vos cogitations. »

Ceci posé, je vous confirme que les études de psychologie que je juge le plus urgentes sont celles où l'expérimentation se réduirait à préparer à l'observation un cadre régulier et formalisé. Une situation expérimentale éloignée de l'activité naturelle ne me semble que rarement offrir des conclusions dont la généralité et la pertinence satisfassent ceux dont l'objet ultime est de connaître l'âme humaine.

Quant aux us et abus de l'analyse des correspondances, vous avez bien deviné quelle peut être ma réserve, et je conclurai cette lettre sur les mêmes paroles que le texte *Histoire* déjà cité : « A partir de 1965, des paquets de cartes..., ont servi sans que nous sachions ni à qui ni à quoi... L'Analyse des correspondances est une méthode ; elle est aussi un outil. A la philosophie de la méthode l'outil doit son efficacité ; mais marteau sans maître, celui-ci frappe désormais librement. En nous appliquant à instruire des statisticiens philosophes, nous espérons au moins servir ceux qui saisissent l'outil pour dégager de la gangue des données le pur diamant de la véridique nature. »

J.-P. Benzécri,
Professeur de statistique,
Université P.-et-M.-Curie.