



Chapitre 3. L'analyse statistique

Bernard Denni, Jean-François Tchernia

DANS **POLITIQUE EN +** 2007, PAGES 51 À 76

ÉDITIONS **PRESSES UNIVERSITAIRES DE GRENOBLE**

ISBN 9782706114045

DOI 10.3917/pug.marc.2007.01.0051

Article disponible en ligne à l'adresse

<https://www.cairn.info/etudier-lopinion--9782706114045-page-51.htm>



CAIRN.INFO
MATIÈRES À RÉFLEXION

Découvrir le sommaire de ce numéro, suivre la revue par email, s'abonner...

Flashez ce QR Code pour accéder à la page de ce numéro sur Cairn.info.



Distribution électronique Cairn.info pour Presses universitaires de Grenoble.

La reproduction ou représentation de cet article, notamment par photocopie, n'est autorisée que dans les limites des conditions générales d'utilisation du site ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Toute autre reproduction ou représentation, en tout ou partie, sous quelque forme et de quelque manière que ce soit, est interdite sauf accord préalable et écrit de l'éditeur, en dehors des cas prévus par la législation en vigueur en France. Il est précisé que son stockage dans une base de données est également interdit.

Chapitre 3

L'analyse statistique

Bernard Denni, Jean-François Tchernia

L'entretien d'enquête consiste à recueillir de l'information auprès d'individus en mesure de la délivrer. Dans ce contexte, la qualité des données dépend de la valeur du dispositif d'enquête, et en particulier de la relation entretenue entre l'enquêteur et le répondant : la confiance et la compréhension mutuelle sont déterminantes. Lorsque le chercheur entreprend l'analyse statistique des résultats d'une enquête par sondage, ces relations interindividuelles perdent pratiquement toute pertinence : le système de référence est maintenant la population des individus. L'individu est alors ramené à une unité statistique, unité qui n'est pratiquement jamais analysée en tant que telle : la lecture d'un ou de quelques questionnaires est parfois intéressante, voire pittoresque, mais n'a pas d'intérêt pour la statistique¹. Ce changement de système de référence s'accompagne d'un changement des cadres disciplinaires : pour préparer un entretien ou élaborer un questionnaire, le chercheur a recours aux méthodes et aux connaissances de la psychologie sociale, voire de la linguistique, tandis que pour analyser des données il utilise ses compétences statistiques. Au-delà du chercheur lui-même, ce changement de posture est attendu du lecteur ou de l'auditoire auquel il s'adresse : autant la discussion sur la validité du questionnaire pourra prendre en considération la sémantique des questions posées ou leur qualité rédactionnelle, autant la prise de connaissance de résultats statistiques aura tendance à se référer à des comparaisons de moyennes ou de fréquences ou à des tests de significativité. Ce passage de

-
1. Les statisticiens considèrent qu'un groupe dont l'effectif est inférieur à 30 est de taille insuffisante pour que les statistiques produites en observant ce groupe aient une réelle validité.

l'interrogation des individus à l'analyse d'un fichier statistique s'apparente à un véritable changement de la logique de recherche : il s'agit de mesurer des faits sociaux, mettre en évidence et interpréter leurs régularités et leurs relations, dans un cadre intellectuel probabiliste, multifactoriel et systémique (Denni, Lecomte, 1999, tome I, chapitre 2).

Cependant, l'analyste statistique ne doit rien ignorer de la phase qui précède : ni la problématique qui a conduit au choix des indicateurs et à la rédaction du questionnaire, ni les conditions techniques de production des réponses (enquête en face à face, au téléphone, etc.), ni, bien sûr, la formulation précise des questions. Une fréquentation assidue et attentive des données d'enquête montre que, le plus souvent, loin de répondre massivement au hasard ou pour plaire à l'enquêteur, la grande majorité des personnes interrogées répondent à la question posée en étant très attentives, ou du moins sensibles, aux mots employés, aux tournures utilisés, aux modalités de réponses proposées². La distribution statistique des données d'enquête dépend de la « réalité » du phénomène étudié, mais aussi de tous ces éléments constitutifs du processus d'observation. Si celui-ci est mal contrôlé les résultats obtenus refléteront moins les phénomènes que l'on cherche à appréhender que les défauts du dispositif d'enquête³.

Selon quelles logiques se déroule le traitement statistique d'une enquête ? Quels sont les niveaux d'analyse possibles et les objectifs recherchés ? Ces questions sont abordées avant de présenter un exemple de déroulement d'une analyse statistique.

La logique du traitement d'enquête

L'analyse de données d'enquête est un travail complexe qui articule intimement deux logiques intellectuelles : l'une relève des sciences sociales, l'autre de la statistique.

2. Cette sensibilité du public à l'énoncé et à la place des questions fait de la rédaction du questionnaire un art difficile. Pour une introduction à ces aspects, voir François de Singly, 1992, chapitre 3.
3. Tous les professionnels de l'enquête par questionnaire sont très attentifs à ces difficultés. Voir par exemple, pour un panorama général, Association pour la statistique et ses utilisateurs, 1992 ; ou encore en science politique, Grunberg, Mayer, Sniderman, 2002.

La démarche sociologique

De la conception du dispositif d'observation jusqu'à la présentation des résultats, le travail avec des données d'enquête est, plus ou moins consciemment, guidé par les principes de division concrète de la sociologie définis par Marcel Mauss et Émile Durkheim : « En fait il n'y a dans une société que deux choses : le groupe qui la forme, d'ordinaire sur un sol déterminé, d'une part ; les représentations et les mouvements de ce groupe, d'autre part » (Mauss, 1969, chapitre 2). L'analyse sociologique porte donc sur trois grands types de phénomènes collectifs : les structures matérielles de la société, ou les « manières d'être » de ses membres (leur sexe, âge, profession, etc.) ; les pratiques ou « manières de faire », c'est-à-dire leurs actions produites par leur vie en commun ; enfin les représentations, ou les « manières de penser », c'est-à-dire « les idées et sentiments collectifs qui président ou correspondent à ces actes, ou sont tout au moins l'objet de croyances collectives ». Bien que l'enquête par questionnaire relève d'une autre tradition en sciences sociales, centrée sur l'individu plus que sur le groupe, cette division de la sociologie reste pertinente pour tous ceux qui produisent et traitent des données d'enquête.

Quel que soit le domaine de l'enquête, il est aisé de classer les questions dans les trois catégories définies par les fondateurs de l'école française de sociologie. Tout questionnaire bien construit interroge sur :

- la situation matérielle des personnes enquêtées (leur sexe, âge, professions, diplôme, revenu qui constituent des renseignements signalétiques traditionnels) et leur contexte de vie (type de commune, habitat, pays),
- leurs façons de penser (leurs valeurs, opinions, images, etc.),
- leurs façons de faire (par exemple, leurs actions, vote ou achat d'un produit)⁴.

4. Dans ce dernier cas, il est très important de souligner que les enquêtes par questionnaire n'enregistrent pas des actions réelles, mais des déclarations de comportements, qui font appel à la mémoire de l'enquêté, à son intention d'agir, et bien sûr à la sincérité de ses réponses. C'est dire que, contrairement à ce qui est parfois affirmé, l'enquête par questionnaire mesure souvent de façon approximative les actions.

Cette classification des questions fournit un guide intellectuel très utile pour définir une stratégie de dépouillement d'une enquête. Il est prudent, dans un premier temps, d'étudier séparément chacun de ces trois pôles en commençant par les informations relatives aux « manières d'être » : d'une part cette analyse permet de savoir qui a répondu en dessinant le fond de carte sociologique des réponses aux autres questions ; d'autre part elle permet de savoir si les caractéristiques sociales des répondants sont à l'image de celles de la population étudiée. Cette validation est indispensable pour évaluer la portée des résultats de l'enquête, leur possible généralisation. Ensuite, on examine ce que la population enquêtée pense et ce qu'elle dit avoir fait ou vouloir faire. Cette étude de chacune des trois dimensions fondamentales de la vie sociale est d'abord très analytique, conduite question après question ; c'est la seule façon de savoir quelle est la variation des réponses d'une question à l'autre, de repérer des résultats surprenants (par exemple, distribution des réponses qui s'écarte anormalement des mesures précédentes, taux élevé de sans réponses), voire des anomalies (erreur de codage, données manquantes, etc.). Cette première étape franchie, il est souvent utile d'avoir une vision plus synthétique des données de l'enquête. Pour cela, le chercheur a à sa disposition des outils statistiques lui permettant de construire, pôle par pôle, des indices de position sociale, des échelles d'attitudes ou des typologies comportementales.

Mais l'objectif principal d'une enquête dépasse généralement ce stade de la description pour aller vers celui de l'explication. Les dimensions matérielles, idéelles et comportementales d'une société – ou de ses membres – sont en perpétuelles interactions dont l'analyse doit rendre compte. L'intérêt se porte alors sur les relations statistiques entre ces trois domaines de la vie sociale. Une nouvelle catégorisation des données est alors nécessaire. Selon le thème de la recherche, les théories mobilisées et la problématique retenue, l'analyste doit définir les faits qu'il cherche à expliquer et ceux qui sont susceptibles de rendre compte des variations observées. Comme l'analyse statistique étudie des variations – des phénomènes parfaitement uniformes ne donneraient aucune prise à cette méthode –, le vocabulaire méthodologique parle de « variables à expliquer » ou « dépendantes » et de « variables explicatives » ou « indépendantes ». De nombreux outils statistiques permettent d'analyser ces relations entre « variables », dans des perspectives plus ou moins explicatives, selon leur capacité à isoler et à évaluer l'effet de chaque variable indépendante sur le phénomène

étudié. De l'analyse des fréquences des réponses, question par question, aux modélisations les plus sophistiquées de l'analyse causale, c'est toujours une démarche de ce type qui est suivie. Il est temps de présenter les outils statistiques utilisés pour la mettre en œuvre.

La démarche statistique

L'analyse statistique consiste à prendre une connaissance aussi complète que possible des données contenues dans un fichier d'enquête. Ces données constituent un ensemble d'indicateurs⁵ portant sur des faits sociaux très complexes qui souvent ne sont pas directement observables (les attitudes par exemple) et sont toujours multidimensionnels. À ce stade de la démarche, un questionnaire est considéré comme une liste d'indicateurs plus ou moins pertinents⁶.

Dans un fichier de données d'enquête, chaque ligne correspond à un individu et chaque colonne correspond à un caractère mesuré chez les individus. Les caractères sont les informations recueillies auprès des individus, désignés habituellement par le terme de variable. Une variable peut correspondre à une quantité (par exemple l'âge en années), ou à une information identifiée par différentes modalités (par exemple la réponse oui ou non donnée à une question, ou le sexe de la personne interrogée). La totalité des informations recueillies auprès des individus sont ainsi groupées dans un ensemble qui est le fichier d'enquête, ensemble que l'analyse statistique aide à connaître. L'analyse statistique peut se situer à différents niveaux de complexité, en fonction du nombre de variables considérées.

Le niveau le plus élémentaire consiste à décrire la distribution des variables une par une, soit par des graphiques, soit par des tableaux de fréquences, soit par des statistiques récapitulatives, comme la moyenne ou

5. Un indicateur est l'une des manifestations observables possibles des phénomènes. Il y a de bons et de moins bons indicateurs, leur pertinence s'érode parfois avec le temps. L'indicateur permet le passage de la démarche sociologique à la démarche statistique en donnant un contenu empirique aux objets et aux concepts des sciences sociales.
6. Sur le rôle des indicateurs dans la recherche sociologique notamment quantitative, voir Boudon et Fillieule, 2002. Pour un panorama plus complet sur les méthodes quantitatives, voir Lebaron, 2006.

l'écart-type. Dans ces analyses dites univariées, chaque variable est observée indépendamment de toutes les autres : cette réduction du champ de l'observation à une seule variable simplifie la prise de connaissance, mais en contrepartie ne donne qu'une vision lacunaire des données. En effet, lors du recueil d'information, chacun des individus interrogés se définit par la totalité des informations qu'il a communiquées à l'enquêteur. L'observation des variables une par une néglige ce caractère multidimensionnel qui définit le profil des réponses de chaque répondant. Or ces profils sont beaucoup plus riches de significations que les réponses prises une à une. Cette étape est donc très insuffisante pour exploiter des données d'enquête. On a vu cependant qu'il serait imprudent d'en faire l'économie.

Un deuxième niveau de complexité consiste à rapprocher deux variables entre elles et à étudier la fréquence des combinaisons d'informations. Par exemple, les réponses à une question d'opinion seront étudiées conjointement avec le sexe des personnes interrogées, ce qui permettra de calculer une fréquence de réponse pour les hommes, une fréquence de réponse pour les femmes et de les comparer entre elles. Cette méthode d'analyse bivariée, conduite ici à l'aide de tris croisés, constitue un enrichissement par rapport à l'analyse univariée pour deux raisons : d'abord parce que l'analyse prend en compte deux informations à la fois au lieu d'une seule, et ensuite parce que la *comparaison* permet de progresser dans la connaissance des données. La relation entre deux variables n'est d'ailleurs pas seulement observée, mais elle est aussi évaluée : différents indicateurs permettent, en effet, de mesurer l'intensité de la relation entre les deux informations observées. L'existence d'une relation statistique significative entre deux variables indique au chercheur qu'elles ont quelque chose en commun. L'intensité de la relation révélée par des indicateurs statistiques tels que le χ^2 ou le coefficient de corrélation va aider à déterminer l'importance de l'enjeu attaché à ces propriétés sociales communes. Les corrélations guident le chercheur dans son travail d'interprétation sociologique : elles sont une étape dans le travail de validation des hypothèses ; elles font naître de nouvelles interrogations, orientent la recherche dans de nouvelles directions. Par exemple, l'existence d'une relation entre l'intérêt pour la politique et le niveau d'instruction – la politisation augmentant avec le niveau de diplôme – indique que l'investissement dans la politique est lié à une certaine capacité à comprendre la politique qui est plus forte chez les personnes qui ont suivi des études longues. Cela ne signifie pas,

bien sûr, que les études soient le laboratoire de la politisation, mais que celle-ci mobilise des capacités cognitives particulières que les études aident à développer.

Dès le premier niveau de complexité, et encore plus avec le deuxième niveau, l'analyste est confronté au besoin de disposer de critères pour juger les résultats statistiques qui se présentent à lui. Les statisticiens se sont préoccupés depuis longtemps de cette question de l'évaluation statistique, et l'ont résolue notamment en élaborant différents tests destinés à signaler les particularités des distributions observées.

Pour les statistiques univariées, et notamment les distributions de fréquences, les tests ont en général pour objet d'estimer dans quelle mesure la distribution observée se rapproche ou non d'une loi de distribution déjà connue. L'analyste teste, par exemple, la normalité de la distribution de la variable étudiée, c'est-à-dire son ajustement à une distribution normale. L'enjeu ici est double : d'une part, le rapprochement avec une loi de distribution connue permet de prédire la distribution de n'importe quelle valeur de la variable étudiée, même si cette valeur n'a pas été observée lors du sondage ; d'autre part, la connaissance des lois statistiques permet de mieux comprendre comment se comporte le phénomène dont la variable est un indicateur. Par exemple, la distribution de la taille dans la population française suit une « loi normale », qui vient notamment du fait qu'un grand nombre de causes indépendantes contribuent à la déterminer. Des facteurs sociaux, comme le revenu ou l'éducation, ou biologiques, comme l'âge ou le sexe, participent à cette détermination, mais ces facteurs jouent indépendamment les uns des autres, ce qui signifie notamment que lorsqu'un de ces facteurs est isolé, comme le sexe, les variations au sein d'une des modalités du facteur suivent également une loi normale. Chez les femmes comme chez les hommes, la taille se distribue normalement, même si la taille moyenne des hommes est supérieure à la taille moyenne des femmes. L'analyste doit être attentif à d'autres courbes de distributions. Ainsi, dans le domaine des opinions, une distribution en J indique une norme de groupe, alors qu'une distribution en U révèle l'existence d'une polarisation des jugements susceptible de devenir un clivage si elle est fortement structurée par des groupes sociaux antagonistes.

Les tests utilisés dans les analyses bivariées ont le plus souvent pour objet d'évaluer l'intensité de la relation entre les deux variables étudiées. Deux

logiques peuvent être suivies. La première consiste à rechercher s'il existe une relation significative entre les deux variables observées, et la deuxième si les deux varient ensemble, c'est-à-dire si, par exemple, l'augmentation de la valeur de l'une entraîne l'augmentation de la valeur de l'autre.

La première logique s'applique à l'étude bivariable des fréquences, portant sur des variables nominales (dont les modalités sont des codes) ou ordinales (les modalités sont des classes ordonnées), voire quantitatives. Elle passe par l'analyse des distributions observées dans un tableau à double entrée, une variable apparaissant en tête de colonne et l'autre en tête de ligne. Dans chaque case de ce tableau un effectif est observé. La question que se pose l'analyste, et à laquelle le test statistique cherchera à répondre, est de savoir si les variations observées dans les différentes cases de ce tableau ont une signification statistique. Ce qui revient à se demander comment juger si une variation est « significative ». Les statisticiens répondent à cette question en partant de ce que donnerait une distribution des réponses en l'absence de liaison entre les deux variables. Dans ce tableau théorique, appelé « tableau d'indépendance », les effectifs de chaque ligne ou colonne se distribuent en fonction des marges du tableau, c'est-à-dire en fonction de la répartition du phénomène sur l'ensemble de la population.

Par exemple, si le chercheur étudie la relation entre le sexe et les réponses à une question d'opinion, il va établir un tableau de la distribution des réponses pour les hommes, pour les femmes et pour l'ensemble de la population. Si les réponses sont distribuées indépendamment du sexe, cela signifie que les hommes répondent à peu près comme les femmes et à peu près comme la moyenne, c'est-à-dire que la distribution des réponses chez les hommes est très proche de celle de l'ensemble de la population telle qu'elle peut être observée dans la marge du tableau. À l'inverse, s'il existe une relation, le profil de réponses (lignes ou colonnes) des hommes est nettement différent de celui des femmes, l'un et l'autre s'écartant de la distribution d'ensemble. Le test statistique le plus couramment utilisé pour évaluer cet écart est le χ^2 . Plus les données réelles s'éloignent de la situation d'indépendance, plus la relation est jugée statistiquement significative. À partir de la valeur du χ^2 , plusieurs indices de liaison, dont le V de Cramer, sont calculés pour mesurer l'intensité de la relation entre les deux variables du tableau. Mais ce test n'apporte pas d'information sur la façon dont les deux phénomènes sont reliés entre eux. Cette limite n'est pas gênante lorsque les variables sont nominales car, leurs modalités étant

conventionnelles, il n'y a pas lieu de se demander si la relation est positive ou négative. Il suffit de savoir si elle existe et d'évaluer son intensité avec un indice comme le V de Cramer variant de 0 (*i.e.* absence de liaison) à 1 (*i.e.* liaison totale). En revanche, les variables numériques ou ordinales appellent un jugement plus précis sur la nature des relations observées.

La deuxième logique n'est adaptée qu'aux variables numériques adoptant des valeurs quantitatives, comme par exemple la taille et le poids : l'analyse porte sur des variations d'intensité de chaque variable, notion qui n'a de sens que si une intensité peut être mesurée. Dans le cas de variables nominales définies par l'appartenance à des catégories, comme le sexe, la profession ou la région, la variation d'intensité ne signifie rien : la catégorie « Homme » n'a pas une intensité supérieure ou inférieure à celle de « Femme », par exemple. L'âge, par contre, est bien une variable numérique : 38 ans est plus âgé que 23, et moins que 57. Outre leur caractère ordinal, les variables numériques sont également caractérisées par leur continuité, c'est-à-dire le fait que les écarts observés sont de même ordre, quelles que soient les valeurs prises par la variable. Par exemple un écart de deux ans d'âge représente la même durée temporelle, que l'on considère la différence entre 18 et 20 ans ou entre 53 et 55 ans. Le coefficient de corrélation linéaire (le R de Pearson) évalue l'intensité et le sens de la relation entre deux variables numériques. Il varie de -1 à $+1$, les fortes valeurs positives indiquant une variation de même sens des deux phénomènes observés, des valeurs proches de -1 , une variation de sens contraire (l'un augmente quand l'autre diminue) ; les valeurs voisines de 0 signalent l'absence de liaison. Les données d'enquêtes utilisent assez peu de variables numériques mais souvent des variables ordinales dont les modalités sont ordonnées. Un ordre (par exemple le rang d'unités géographiques classées selon l'intensité d'un phénomène, une classe de revenus et d'âge, etc.) est une information plus forte qu'un code conventionnel, mais beaucoup moins précise qu'une valeur numérique. Les échelles sémantiques utilisées dans les questions d'opinion (du type « beaucoup », « assez », « peu », « pas du tout ») sont assimilées à ce type de variable. Le coefficient de corrélation de rang de Spearman ou le Tau de Kendall permettent de calculer l'intensité et le sens de la relation entre deux variables de ce type. Ils varient aussi de $+1$ à -1 .

Avec des données nominales ou ordinales, un tableau croisé peut prendre en compte en même temps, trois variables et plus. Plus difficiles

à lire, toutefois, et donc à analyser, ces tableaux doivent être utilisés avec prudence car la multiplication des variables retenues réduit les effectifs présents dans chaque case, ce qui tend à invalider le tableau produit : même si les résultats restent statistiquement significatifs, l'existence d'effectifs très faibles, voire nuls, rend difficile l'élaboration d'un raisonnement pertinent. Il est possible de créer une situation plus favorable en réduisant la taille du tableau⁷. Pour cela, il faut diminuer le nombre de modalités des variables (par exemple en recodant le niveau de diplôme en deux positions). L'opération est délicate car elle doit être conduite en respectant les coupures « naturelles » de la variable, de façon à perdre le moins possible d'information, et ne pas gommer artificiellement une relation pertinente. On procède ainsi. Le recueil d'information est fait à un niveau très détaillé (par exemple l'âge, le dernier diplôme obtenu). Puis à l'aide de tris avec la variable d'origine, on recherche quelles valeurs entraînent de fortes variations de la variable dépendante. Ces valeurs guident ensuite les opérations de recodage, par exemple en trois classes d'âge, ou en deux classes de diplôme. Ainsi, dans le cas du diplôme, le baccalauréat introduit une segmentation forte de la population pour de très nombreux phénomènes d'opinion : une nouvelle variable est créée en tenant compte de cette coupure « naturelle », ses deux modalités distinguant entre ceux qui ont au moins le baccalauréat et ceux qui ne l'ont pas. On obtient alors des tableaux croisés de rang trois, voire plus, qui conservent des effectifs suffisants (à condition que l'échantillon de départ soit assez grand) et qui restent lisibles, du moins pour un lecteur un peu expérimenté. Reste alors à mettre en forme les résultats importants issus de ces tableaux complexes, sous une forme accessible par un public plus large.

Un troisième et dernier niveau de complexité consiste à analyser en même temps les relations entre plus de deux variables. Cette logique, dite analyse multivariée, va très au-delà de la décomposition des résultats de l'enquête en tableaux de rang trois et plus : son objectif est plutôt d'essayer de prendre en considération plusieurs résultats en même temps. L'analyse multivariée considère chaque individu en s'intéressant aux réponses qu'il

7. Rappelons que la taille d'un tableau est définie par le nombre de cases qu'il contient, lequel est égal au produit du nombre de modalités de chaque variable. L'analyse des relations entre le niveau d'études, en dix modalités, l'âge (cinq modalités) et le sexe (deux par définition) produit un tableau de cent cases.

a pu donner à la multiplicité des questions qui lui ont été posées. Par exemple, dans le cas de la construction d'une typologie, chaque individu est caractérisé par l'ensemble de réponses qu'il a donné à un ensemble de questions posées. L'analyse se donne pour objectif de rechercher dans l'ensemble du fichier d'enquête d'autres individus ayant le même profil de réponses ou un profil très proche.

Niveaux d'analyse et objectifs recherchés

Il est souvent reproché à l'analyse quantitative des faits sociaux de réduire leur complexité à un chiffre, une moyenne ou un pourcentage. Les méthodes quantitatives ne sauraient que compter, de façon plus ou moins arbitraire du fait des nomenclatures qu'elles utilisent, les phénomènes de société. Cette critique ignore que le dénombrement n'est que le degré zéro de l'analyse quantitative. Conduite par des praticiens expérimentés, celle-ci permet de rendre compte de processus sociaux complexes, non directement repérables, et selon des procédés rigoureux, reproductibles et contrôlables par la communauté scientifique. Tout dépend du niveau d'investigation et de présentation des résultats retenu pour explorer un fichier de données et rendre compte de son contenu.

Trois niveaux d'analyse

Les trois niveaux de complexité dans le traitement statistique des données correspondent à différents types d'exposition des résultats d'une enquête. Une première forme d'exposition consiste à présenter les résultats d'ensemble des questions, leur texte et le pourcentage de chaque réponse. C'est ainsi que les sondages sont le plus souvent commentés, par exemple, dans les journaux. Une autre version possible de ce type de présentation des résultats consiste à produire un graphique montrant la distribution de la variable étudiée, méthode qui est également assez largement utilisée dans les journaux, notamment en ayant recours aux possibilités de l'infographie. La lecture de ces résultats donne le sens premier du sondage, le plus évident et celui sur lequel la plupart des lecteurs pourront se mettre d'accord⁸. Ces résultats d'ensemble, pour intéressants qu'ils soient, ne

8. Pour présenter les réponses selon une échelle d'intensité allant par exemple de « très » à « pas du tout », un diagramme en bâton permet de visualiser

disent rien des variations de l'opinion au sein de la population interrogée, variations qui peuvent parfois être considérables et rendre caduque la simple lecture à plat.

La présentation des résultats aux mêmes questions en fonction de différents facteurs est un deuxième mode d'exposition des résultats d'un sondage. Dans ce deuxième cas, les résultats d'ensemble, s'ils peuvent faire l'objet d'une présentation et d'un commentaire, ne constituent plus l'objet principal du travail présenté : le chercheur se concentre sur l'idée que l'étude des variations sera plus riche que l'analyse de la distribution d'ensemble. Il opte pour une lecture « en relief » des résultats de l'enquête. La différence entre ces deux modes d'exposition correspond à celle qui existe entre une carte qui dessine les contours d'un pays ou d'une région, et une autre carte qui indique aussi le relief de cette région ou de ce pays, information non négligeable dans bien des cas : il suffit d'avoir fait quelques randonnées à pied pour comprendre toute la valeur de ce genre d'informations. L'analyse des variations offre en outre la possibilité d'élaborer de nouvelles hypothèses sur la signification des résultats. Une variation significative selon des facteurs tels que la profession ou le niveau d'instruction, peut suggérer, par exemple, une interprétation en terme de statut social ; une variation selon l'âge conduira à s'interroger sur la nature de ses effets : phénomène de cycle de vie ou de génération ? Une variation selon le sexe peut conduire à réfléchir sur les rôles sociaux ou sur l'identité sexuée des sujets. Il convient, toutefois, de souligner que ce type de déduction provenant de l'observation des variations en fonction de tel ou tel facteur explicatif ne doit pas être le fruit d'une démarche purement empirique, mais plutôt le résultat du test d'un certain nombre d'hypothèses élaborées en se reposant sur des travaux antérieurs. Le commentaire des résultats rappellera ces hypothèses initiales, mettra en évidence le fait qu'elles sont ou non validées et les conclusions qui peuvent en être tirées.

Le travail sur la structure des réponses à un ensemble de questions à l'aide de l'analyse multivariée relève d'un troisième niveau d'analyse.

et comparer rapidement les structures des réponses dans l'ensemble de la population étudiée, à condition de conserver la même échelle et la même origine pour tous les graphiques. Dans de tels cas, les graphiques sectoriels (dits « camemberts ») ne mettent pas en relief les phénomènes qui méritent d'être relevés.

L'objet n'est plus de se concentrer sur des aspects partiels de l'étude, mais sur des structures plus globales qui synthétisent une pluralité d'informations. L'exemple le plus caractéristique de cette démarche est celui qui consiste à mettre en évidence les attitudes latentes derrière différentes questions d'opinions ou de comportements déclarés recueillis lors d'un sondage. Les réponses à ces questions sont alors utilisées non pour elles-mêmes mais comme des manifestations observables d'un phénomène latent étudié. C'est à ce niveau qu'intervient pleinement le travail sur les indicateurs, les réponses aux questions étant considérées comme le symptôme d'une caractéristique plus fondamentale, et non directement perceptible, permettant elle-même de décrire les sujets sociaux dans leurs rapports avec certains objets. On cherche ainsi à mettre en correspondance, par exemple des comportements électoraux ou de consommation avec des attitudes ou des systèmes de valeurs des enquêtés. Ce type d'analyse des données permet de tenir un discours interprétatif qui relie observations, concepts et théories de moyenne portée, ce que ne permettent pas (ou de façon très approximative) les commentaires attachés à la lettre de chaque question. La dernière partie de ce chapitre illustre cette démarche.

En parcourant ces différents niveaux de complexité de l'analyse statistique, le chercheur tente de se libérer du cadre un peu contraint de son enquête pour monter en généralité, tout en s'appuyant sur des données d'observation. En bref, elle lui sert à faire un véritable travail de recherche qui part d'une problématique élaborée à partir d'un ensemble de connaissances, qui passe par la construction de résultats innovants et qui aboutit à des conclusions nouvelles sur l'objet étudié.

Les principaux objectifs

La statistique est un outil puissant permettant de révéler des relations entre des phénomènes, mais son utilisation doit toujours être orientée par des raisons à la fois scientifiques et méthodologiques. Scientifiquement, elle n'a de chance d'aboutir à un résultat intelligible et véritablement nouveau que si elle est produite en partant d'une problématique établie d'après les connaissances du chercheur et le travail de recherche préalable qu'il aura pu mener en se documentant sur l'objet de sa recherche. Il faut avoir une idée de la nature des phénomènes pertinents à découvrir. Comment ? Méthodologiquement, l'analyse statistique se doit aussi d'être précisément

orientée dans la mesure où les outils de traitement de données dont le chercheur peut disposer sont très nombreux et où leurs combinaisons avec les différentes informations qu'apporte l'enquête sont quasiment infinies : la définition d'un plan d'analyse est en fait aussi nécessaire que le choix d'un emplacement précis pour un pêcheur qui dépose son filet le soir et espère bien recueillir quelques poissons au cours de la nuit. D'une variété infinie, les objectifs qui peuvent être donnés à une analyse statistique ressortent souvent de quelques logiques essentielles qui sont exposées ci-dessous.

La première logique que suivent nombre d'analyses statistiques est de rechercher les régularités statistiques des résultats produits par l'enquête. Ce premier type d'analyse consiste notamment à caractériser les distributions de résultats et à étudier les variations des données selon des critères tenus pour potentiellement explicatifs. La nature sociologique de certains comportements qui semblent relever du libre arbitre ou du pur hasard est mise en évidence par de telles régularités : le suicide, le choix du conjoint, le vote. Travaux statistiques et sociologiques sont ici étroitement liés⁹, et restent soumis aux principes généraux de la démarche scientifique, notamment le processus élaboration-test d'hypothèses. Ce type d'analyse est souvent tenu pour descriptif, ce qui est assez juste mais ne doit pas faire oublier que l'observation des régularités demande à être interprétée et peut conduire à élaborer de nouvelles hypothèses de travail.

Une deuxième logique souvent mise en œuvre dans une analyse statistique consiste à rechercher, au-delà du discours apparent délivré par les personnes interrogées dans leurs réponses, les variables latentes c'est-à-dire celles qui s'expriment au travers du système de signes que constitue le questionnaire d'enquête. Les questions sont, en effet, considérées non en elles-mêmes mais en tant que symptômes d'une tendance essentielle non exprimée directement. Par exemple, un ensemble de questions pourra mettre en évidence le soutien plus ou moins fort d'une personne au libéralisme

-
9. La sociologie est fille de la philosophie et du droit, mais aussi de la statistique, comme le rappellent les œuvres fondatrices d'Émile Durkheim sur le suicide et d'André Siegfried sur les comportements électoraux. Rappelons que Durkheim ne découvre pas les régularités du suicide, mais change le paradigme de leur interprétation en démontrant que ce comportement a des causes sociales.

économique, et l'analyse statistique offrira la possibilité non seulement de démontrer qu'une telle tendance existe, mais aussi qu'elle peut avoir une traduction concrète dans une variable synthétique rendant compte non seulement de plusieurs indicateurs partiels mais également de cette tendance générale. Les méthodes factorielles (analyse des correspondances ou analyse en composantes principales) sont couramment utilisées dans ce sens, mais les méthodes de classification ressortent de la même logique dans la mesure où elles permettent de mettre à jour des classes latentes. L'analyse des corrélations entre les réponses permet de sélectionner les questions qui appartiennent au même groupe de sens.

La troisième logique est celle de l'analyse explicative et de la modélisation. Le propos est dans ce cas de rendre compte du pouvoir explicatif d'une ou plusieurs variables indépendantes sur une ou plusieurs variables dépendantes. L'analyse statistique, si elle permet d'évaluer la force des relations entre différentes variables, voire d'établir une hiérarchie dans le pouvoir explicatif des variables indépendantes, ne détermine en aucune manière l'existence et le sens des causalités. Faut-il rappeler qu'une corrélation n'est pas une causalité? Il y a des corrélations fortuites, d'autres induites par une variable ignorée, ou non contrôlée par l'analyste. Mais même lorsque la relation entre deux phénomènes est fermement établie toute chose égale par ailleurs, se pose presque toujours la question de son sens, c'est-à-dire de sa direction ($A \Rightarrow B$ ou $B \Rightarrow A$?) et de sa signification en terme de compréhension du monde social. La réponse ne peut pas venir des tests statistiques mais doit s'appuyer sur une théorie scientifique, en marketing, en science politique, en sociologie, en psychologie sociale, etc. Cette base scientifique est indispensable pour produire un véritable modèle explicatif qui sera testé dans le cadre de l'analyse statistique, produisant ainsi une véritable modélisation.

En arrière-plan de ces différentes logiques, l'analyse statistique des sondages d'opinion vise un objet principal qui est de comprendre les articulations complexes entre manières d'être, de penser et d'agir. Pour y parvenir, le chercheur doit souvent condenser l'information se rapportant à ces trois types de faits sociaux, de façon à avoir des mesures de statut social, d'attitudes et de comportements à la fois plus synthétiques et plus stables. Comme on l'a vu, à ce stade, il travaille sur des objets statistiques qui sont certes des abstractions (les facteurs de l'analyse factorielle par exemple) mais qui sont construits, non *a priori*, mais à partir des corrélations

entre les réponses des enquêtés. Ces abstractions empiriques issues de la statistique sont d'autant plus pertinentes qu'elles renvoient aux concepts propres au champ de l'étude. Ainsi, le facteur principal extrait par une analyse des correspondances d'une série de questions d'opinion est une traduction empirique du concept d'attitudes latentes utilisé en psychologie sociale, en marketing ou en science politique. Armé d'instruments de cette nature qui font sens à la fois dans le champ de la théorie et dans celui de l'observation, le chercheur dispose d'une représentation structurée de son objet qui n'est pas accessible à une observation superficielle. Il peut alors tester des hypothèses complexes sur les relations entre les trois grandes dimensions de la vie sociale.

Le déroulement d'une analyse statistique: l'opinion des Français sur la construction européenne en 1997

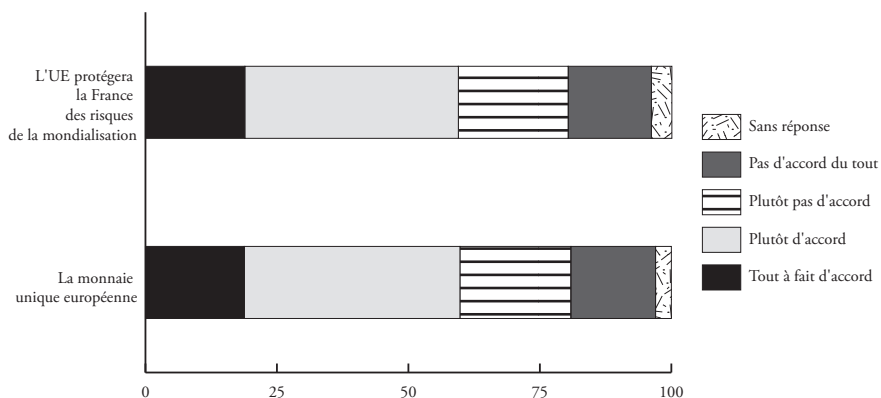
Dans une enquête nationale réalisée entre les deux tours des élections législatives de 1997 par la Sofres auprès d'un échantillon par quotas de plus de 3 000 électeurs¹⁰, neuf questions concernent l'opinion des Français sur la construction européenne. La richesse de l'information permet une exploration assez précise de ce que pensent, à cette époque, les Français sur ce sujet. La diversité du format des questions posées rend délicat la construction d'un seul tableau ou graphique permettant de comparer facilement l'ensemble des réponses. La restitution de l'information à ce premier niveau de l'analyse passe par la construction de deux graphiques et un tableau, mais en hiérarchisant toujours les réponses des plus favorables aux plus défavorables à l'Europe, afin d'en faciliter la lecture et le commentaire.

Le graphique 1 présente les réponses à deux questions portant sur la monnaie unique et le rôle protecteur de l'Union européenne face à la mondialisation, mesurées avec la même échelle d'accord. Il montre

10. Cette enquête a été conçue à l'initiative de trois laboratoires, le Cevipof de l'IEP de Paris, le Cidsp de l'IEP de Grenoble et le Craps de Lille, en partenariat avec le quotidien *Libération*. Ces données ont servi à la rédaction d'un ouvrage (Bréchon, Laurent, Perrineau, direction, 2000). Voir en particulier Bruno Cautrès et Bernard Denni, « Les attitudes des Français à l'égard de l'Union européenne: les logiques du refus ».

clairement que les réponses se distribuent de façon identique sur ces deux variables, et que l'opinion est majoritairement favorable au rôle de l'Union dans ces domaines. Toutefois, l'adhésion reste modérée, la modalité « plutôt d'accord » recueillant deux fois plus de réponses que l'item « tout à fait ». Il serait très hasardeux de conclure de ces deux réponses qu'une majorité absolue de Français adhèrent à l'idée de construction européenne. Le graphique 2 montre en effet que les opinions sur ce sujet varient sensiblement selon le thème précis de la question. Lorsque les enquêteurs demandent si « la poursuite de l'unification européenne aura des effets plutôt positifs, plutôt négatifs ou n'aura pas d'effet pour la France », les réponses varient fortement selon que l'opinion sollicitée concerne la croissance économique, les manières de vivre ou la protection sociale. La perception reste majoritairement favorable pour les effets sur la croissance, alors que les réponses négatives l'emportent sur les positives en ce qui concerne la protection sociale. Il se confirme que les Français associent positivement l'Europe au développement économique, mais il apparaît qu'ils redoutent ses conséquences en matière sociale. Ils ne perçoivent pas de menace, en revanche, sur leurs manières de vivre, une majorité relative (45 %) considérant que l'Europe n'aura pas d'effets en la matière.

Graphique 1. Deux opinions sur l'Union européenne.



Graphique 2. Les effets de la poursuite de l'Union européenne concernant...

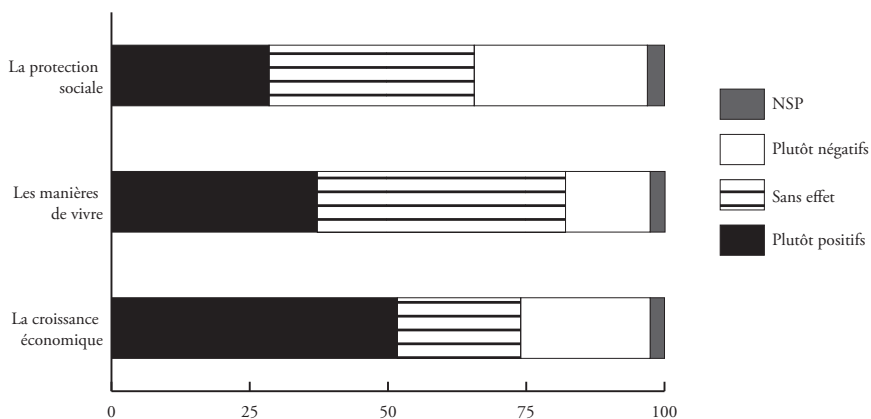


Tableau 1. Opinions sur l'Union européenne.

La France a-t-elle bénéficié ou non de son appartenance à l'UE?	
A bénéficié	49,6
N'a pas bénéficié	44,4
Sans réponse	6,0
Sentiment si l'UE était abandonnée	
Regrets	45,4
Indifférence	33,5
Soulagement	18,9
Sans réponse	2,2
À l'avenir, l'UE doit par rapport aux États membres...	
Renforcer son pouvoir	36,4
Maintenir son pouvoir	41,3
Diminuer son pouvoir	16,4
Sans réponse	5,9
Avez-vous le sentiment d'être...	
Autant Européen que Français	37,8
Plus Français qu'Européen	35,0
Seulement Français	25,9
Sans réponse	1,3

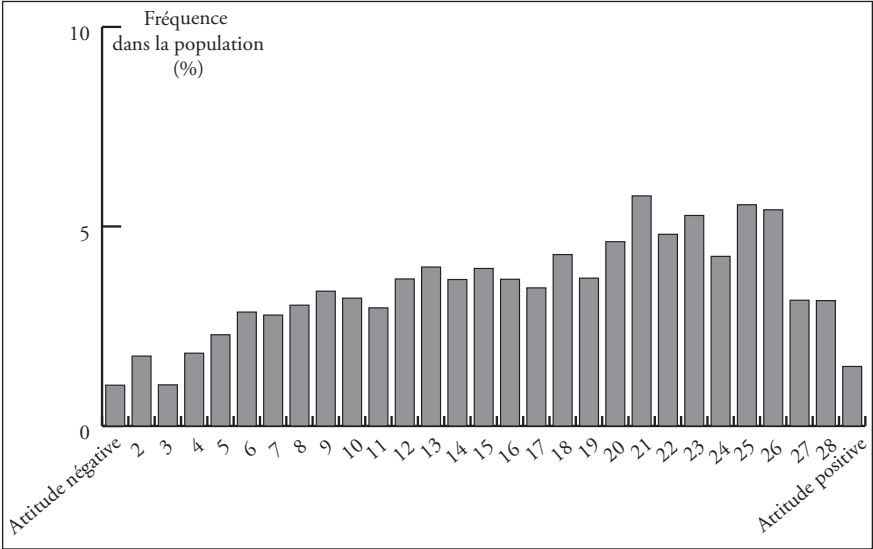
Le tableau 1 rassemble quatre autres questions qui abordent sous différents aspects la construction européenne. Les réponses favorables à la construction européenne ne sont jamais (ou à peine) majoritaires. Les réponses négatives varient de 44 % (qui estiment que la France n'a pas bénéficié de son appartenance à l'UE) à 16 % (favorables à la diminution des pouvoirs de l'UE par rapport à celui des États membres). On notera enfin, que sur des sujets réputés difficiles et souvent abstraits, les taux de sans réponse restent faibles, compris entre 1 % et 6 %.

Cette analyse de premier niveau met en évidence la variabilité des réponses, laquelle ne signifie nullement l'inconstance des personnes inter-

rogées. Elle nous révèle au contraire la complexité de l'opinion sur cet enjeu politique. Toutes les enquêtes le montrent (par exemple Céline Belot et Bruno Cautrès, 2006), les avis des personnes interrogées varient beaucoup selon le thème abordé par la question. Il n'est donc pas possible d'évaluer de façon fiable le degré d'adhésion des citoyens à l'Union Européenne à partir de leurs réponses à une seule question. Leur jugement varie au cas par cas. La question de savoir si les Français sont favorables ou non à la construction européenne n'a pas de réponse simple.

Le deuxième niveau d'analyse consiste à étudier le phénomène en fonction d'autres données sociologiques afin de savoir par exemple si l'opinion sur l'Europe varie en fonction de l'âge, du diplôme, de l'orientation politique, etc. Les résultats qui précèdent rendent cette étape délicate à mettre en œuvre. Avant de se lancer dans une analyse sociologique question par question, une autre voie doit être explorée. La variation des réponses avec les thèmes abordés n'exclut pas l'existence d'une logique dans les réponses données. Si elle existe, elle sera mise en évidence par l'analyse statistique, fondée sur la recherche des corrélations entre ces réponses. Une réponse favorable à l'Europe à une question augmente-t-elle la probabilité d'une réponse favorable à toutes les autres questions, et inversement? Une analyse des correspondances multiples effectuées sur les neuf questions étudiées permet de vérifier qu'il existe bien une telle structure complexe des corrélations. Toutes les questions sont corrélées entre elles, avec des intensités variables, mais toujours de façon significative. Le premier axe de l'analyse, qui résume ce que toutes ces réponses ont en commun, représente à lui seul 44 % de la variance globale des données; le deuxième axe, qui représente 23 %, agrège les personnes qui n'ont pas répondu. Ce résultat signifie que les différences d'appréciation sur l'Europe selon le thème de la question ne constituent pas autant de dimensions indépendantes les unes des autres sur ce sujet, mais s'organisent autour d'un même « facteur » qui résume une même tendance latente que l'on peut appeler attitude pro ou antieuropéenne. Ce sentiment pro ou antieuropéen connaît des variations de degré infinies d'un individu à l'autre. Le premier axe de l'analyse traduit empiriquement ces différences de sensibilité établies à partir du profil de réponses aux neuf questions posées. Le graphique 3 présente la répartition des répondants sur cet axe, en les hiérarchisant des plus défavorables à l'Union européenne (à la gauche du graphique) aux plus favorables, à droite.

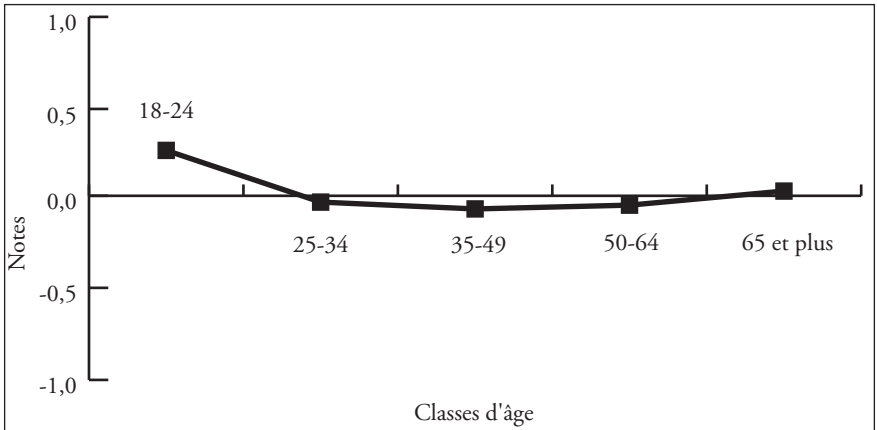
Graphique 3. L'attitude à l'égard de l'Europe : répartition des enquêtés sur le 1^{er} axe de l'analyse des correspondances.



On dispose ainsi d'un instrument précis et puissant pour analyser les relations entre l'attitude européenne et tout autre type de variables renvoyant aux manières d'être, de faire ou de penser. Cette nouvelle variable étant numérique, il est possible d'analyser les variations de sa moyenne (qui est par construction égale à 0 sur l'ensemble de l'échantillon, les notes négatives indiquant une attitude défavorable, les notes positives une attitude favorable) en fonction de différentes caractéristiques sociales ou politiques. Dans le cadre de cet exemple, retenons l'âge, le diplôme et la profession.

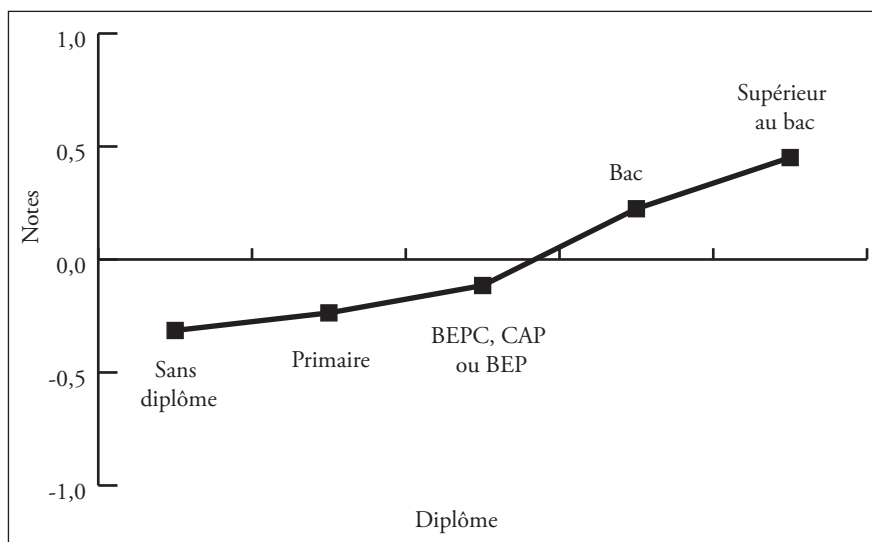
L'analyse en fonction de l'âge fait apparaître peu de variations, à l'exception des plus jeunes, dont l'opinion est nettement plus favorable, et des plus âgés (graphique 4).

Graphique 4. Note d'européanisme en fonction de l'âge.



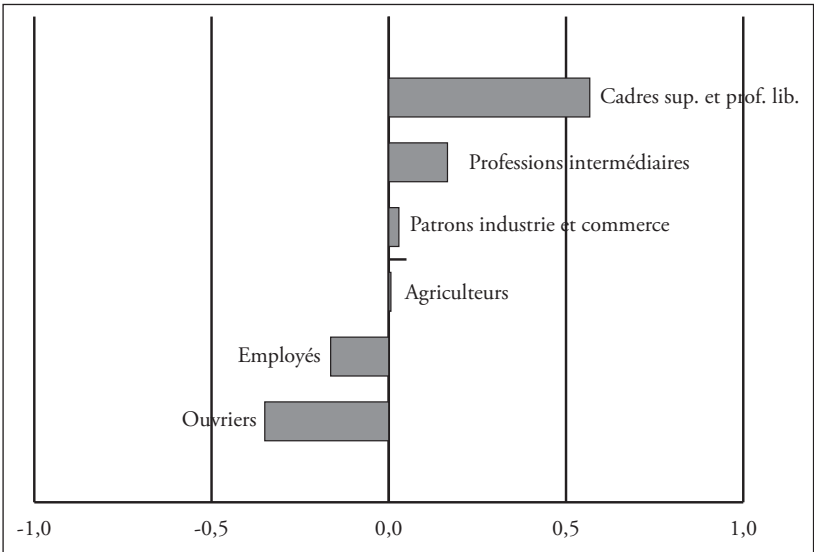
Cette courbe en U semble conforter une analyse souvent avancée : les plus jeunes adhèrent davantage à la construction européenne car « ils sont nés avec » et en ont davantage l'expérience, notamment à travers les échanges au cours de leurs études ; les plus âgés y sont attachés car ils sont les plus sensibles à ce principe fondateur de l'Europe : prévenir les risques de guerre sur le continent. Toutefois, les effets de l'âge sur un phénomène sont toujours délicats à interpréter. En effet, si cette variable est facile à objectiver (les personnes donnent aisément leur année de naissance), elle est beaucoup plus difficile à interpréter : traduit-elle un phénomène de génération comme le suggère l'interprétation proposée ? Un phénomène de cycle de vie : l'opinion change selon les statuts sociaux occupés au fil de l'existence ? Ou encore un effet de composition sociologique lié notamment aux transformations de la société et à la mortalité différentielle : les caractéristiques sociales des groupes d'âge ne sont pas identiques, les femmes, les pratiquants, les personnes possédant un patrimoine étant nettement plus nombreux parmi les plus âgés. Ainsi, l'âge et le diplôme sont fortement corrélés, les plus jeunes étant plus diplômés, ce que vérifient les données de l'enquête. Dès lors, il faut se demander s'il existe une relation entre le niveau de diplôme et l'attitude européenne. Le graphique 5 montre que les deux variables sont fortement liées, l'adhésion à l'idée européenne devenant de plus en plus forte au fur et à mesure que le niveau de diplôme augmente.

Graphique 5. Note d'européanisme en fonction du niveau diplôme.



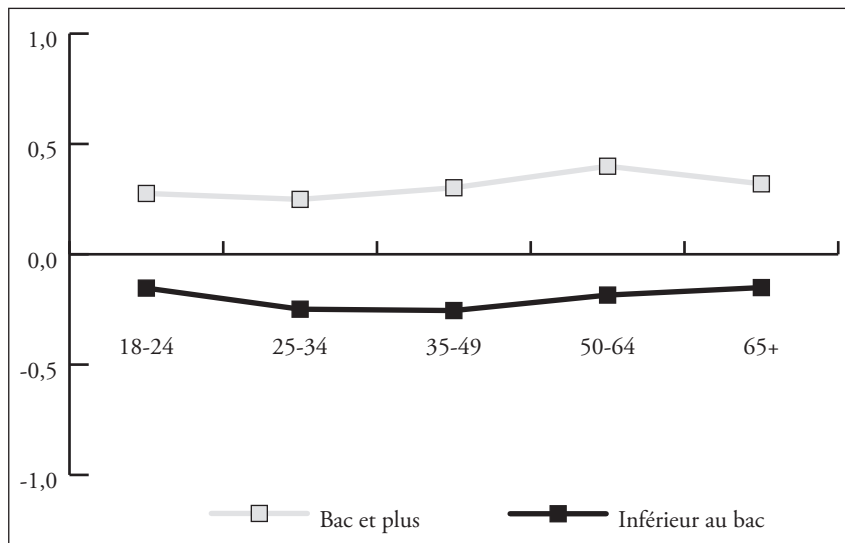
Ce constat empirique, confirmé par toutes les enquêtes et notamment les Eurobaromètres de la Commission européenne, appelle plusieurs interprétations. Selon une première hypothèse, il s'agit d'un effet direct de la durée de la période de formation scolaire qui favorise la diffusion et l'adoption de valeurs « universalistes » d'ouverture aux autres. De surcroît, plus les compétences acquises pendant les études sont importantes, plus l'individu a le sentiment de disposer des ressources nécessaires pour faire face au changement, voire souhaite se mesurer au changement. L'influence du diplôme peut aussi être indirecte : un diplôme élevé favorise l'accès aux emplois de cadres ou de professions libérales, donc aux catégories sociales « aisées » qui sont aussi les plus favorables à la construction européenne, par principe (adhésion à des valeurs universalistes) et par intérêt (occupation d'emplois très qualifiés dans des secteurs d'activité ouverts sur la mondialisation économique). Le graphique 6 montre très clairement le clivage entre les cadres proeuropéens et les catégories populaires (ouvriers et employés) bien moins favorables à l'Europe.

Graphique 6. Note d'européanisme en fonction du groupe socioprofessionnel de la personne interrogée (population active seulement).



Quoi qu'il en soit, le diplôme est une variable sociologique largement prédictive de l'opinion sur la construction européenne. Comment se combinent ses effets avec ceux de l'âge ? Une analyse à trois variables, après recodification du niveau de diplôme en deux catégories, le bac servant de frontière, permet de répondre à cette question. Le résultat est présenté au graphique 7.

Graphique 7. Note d'européanisme en fonction de l'âge contrôlé par le niveau de diplôme.



L'effet du diplôme sur l'opinion est beaucoup plus fort que celui de l'âge : pour toutes les classes d'âge, les plus diplômés sont nettement favorables à l'Europe et les moins diplômés, nettement défavorables. Le profil de la courbe du graphique 4 ne se retrouve, de façon très atténuée, que pour les moins diplômés. L'adhésion à l'idée européenne ou son rejet semble moins résulter d'un phénomène de génération, lié à l'histoire de la construction européenne, qu'à l'adhésion à un système de valeurs (largement transmis par l'école) et à des intérêts socio-économiques liés au statut professionnel. Ces deux facteurs (manière de penser et d'être) se combinent et jouent un rôle déterminant dans l'adhésion ou la méfiance à l'égard du projet européen.

L'analyse peut ainsi se poursuivre pour tester de nouvelles hypothèses, comme la relation entre l'attitude européenne et la tolérance, l'optimisme social, l'orientation politique, etc. On dresse ainsi un panorama assez complet de l'opinion des Français sur l'Europe, dans ces diverses dimensions et avec toute la complexité des mécanismes sociaux qui contribuent à construire cette opinion. L'analyse approfondie de ces données de 1997

montre l'existence, dès cette époque, d'une forte opposition à l'Europe libérale notamment au sein de l'électorat du parti socialiste. Le référendum constitutionnel de mai 2005 a donné l'occasion de renforcer ce courant d'opinion, de lui donner une consistance politique et de lui permettre de s'exprimer publiquement dans un comportement électoral.

Conclusion

L'analyse statistique des données ne se limite pas à la lecture de pourcentages dans des tableaux de fréquence à une ou deux dimensions. Pourtant, même à ce stade, les risques d'erreurs existent, ne serait-ce que la lecture erronée des pourcentages calculés en ligne ou en colonne. Mais dès qu'elle dépasse ce stade initial, cette analyse exige la maîtrise de multiples compétences, de l'intuition, de l'expérience accompagnée d'un solide savoir-faire, et la mobilisation de moyens humains, techniques et financiers importants. Pourtant, lorsque l'ordinateur produit les résultats des modélisations complexes imaginées par le chercheur, le véritable travail sociologique ne fait que commencer : quelle signification donner à la structuration des réponses mise à jour, à la hiérarchisation des « facteurs explicatifs » ? Il est fréquent qu'à cette phase ultime de l'interprétation, il soit nécessaire de retourner aux données pour vérifier la solidité d'une relation statistique ou tester une nouvelle hypothèse. L'exploitation des données d'enquête passe par un va-et-vient permanent entre la recherche d'une interprétation sociologique et la mobilisation d'outils et de méthodes, pour mettre en évidence des relations latentes entre les données qui soient compatibles avec cette interprétation.