Olivier Martin

L'ENQUÊTE ET SES MÉTHODES

L'ANALYSE DE DONNÉES QUANTITATIVES

2e édition

ARMAND COLIN

Olivier Martin

L'ENQUÊTE ET SES MÉTHODES

L'ANALYSE DE DONNÉES QUANTITATIVES

2e édition

sous la direction
de François de Singly

Dans la même collection

Série « L'enquête et ses méthodes »
Anne-Marie ARBORIO, Pierre FOURNIER, *L'Observation directe* (2e édition)
Daniel BERTAUX, *Le Récit de vie* (2e édition)
Alain BLANCHET, Anne GOTMAN, *L'Entretien* (2e édition)
Jean COPANS, *L'Enquête ethnologique de terrain* (2e édition)
Sophie DUCHESNE, Florence HAEGEL, *L'Entretien collectif* (2e édition)
Jean-Claude KAUFMANN, *L'Entretien compréhensif* (2e édition)
François DE SINGLY, *Le Questionnaire* (2e édition)

Série « Domaines et approches »
Philippe ADAM, Claudine HERZLICH, *Sociologie de la maladie et de la médecine*
Cyprien AVENEL, *Sociologie des « quartiers sensibles »* (2e édition)
Olivier BOBINEAU, Sébastien TANK-STORPER, *Sociologie des religions*
Michel BOZON, *Sociologie de la sexualité* (2e édition)
Maryse BRESSON, *Sociologie de la précarité*
Jean COPANS, *Introduction à l'ethnologie et à l'anthropologie* (2e édition)
Jean COPANS, *Développement mondial et mutations des sociétés contemporaines*
Philippe CORCUFF, *Les Grands Penseurs de la politique*
Pierre-Yves CUSSET, *Le Lien social*
Muriel DARMON, *La Socialisation*
Pascal DURET, Peggy ROUSSEL, *Le Corps et ses sociologies*
Emmanuel ETHIS, *Sociologie du cinéma et de ses publics* (2e édition)
Laurent FLEURY, *Sociologie de la culture et des pratiques culturelles*
Yves GRAFMEYER, *Sociologie urbaine* (2e édition)
Benoît HEILBRUNN, *La Consommation et ses sociologies*
Claudette LAFAYE, *Sociologie des organisations*
François LAPLANTINE, *La Description ethnographique*
Pierre LASCOUMES, Patrick LE GALÈS, *Sociologie de l'action publique*
Olivier MARTIN, *Sociologie des sciences*
Véronique MUNOZ-DARDÉ, *Rawls et la justice sociale*
Bruno PÉQUIGNOT, *Sociologie des arts*
Jean-Manuel DE QUEIROZ, *L'École et ses sociologies* (2e édition)
Catherine ROLLET, *Introduction à la démographie* (2e édition)
Martine SEGALEN, *Rites et rituels contemporains* (2e édition)
François DE SINGLY, *Sociologie de la famille contemporaine* (2e édition)
Marcelle STROOBANTS, *Sociologie du travail* (2e édition)

Série « Sociologies contemporaines »
Laurent BERGER, *Les Nouvelles Ethnologies*
Philippe CORCUFF, *Les Nouvelles Sociologies* (2e édition)
Pascal DURET, *Sociologie de la compétition*
Danilo MARTUCCELLI, François DE SINGLY, *Les Sociologies de l'individu*

ISBN : 978-2-200-24461-3

SOMMAIRE

PARTIE 2
ANALYSER LES RELATIONS ENTRE VARIABLES

Du même auteur

La Tentation du corps dans les sciences sociales françaises (avec Dominique Memmi et Dominique Guillo), Paris, Éditions de l'EHESS, 2009.
Savoirs et savants. Les recherches sur la science en France (avec Jean-Michel Berthelot et Cécile Collinet), Paris, PUF, 2005.
Internet en famille (dir. avec Benoît Lelong), Paris, Hermès-Lavoisier (*Réseaux*, n° 123), 2004.
Mathématiques et sciences sociales au XX^e siècle (dir.), Lille, Presses du Septentrion, 2002.
Sociologie des sciences, Paris, Nathan 2000.
La Mesure de l'esprit, Paris, L'Harmattan, 1997.

INTRODUCTION

Ce manuel présente les outils statistiques pour l'analyse des données quantitatives en sociologie. Il prend la suite, dans la série « L'enquête et ses méthodes », de l'ouvrage *Le Questionnaire,* de François de Singly[1] consacré à la présentation de la démarche quantitative en sociologie, à sa méthode emblématique (le questionnaire) ainsi qu'à son premier outil statistique (le tableau croisé).

Nous avons pris le parti de ne présenter que les aspects utiles en sociologie. Ce parti pris a deux conséquences. Premièrement, nous privilégions les outils statistiques les plus couramment utilisés, au détriment des techniques habituellement enseignées dans les cours et manuels de statistique, mais peu ou pas utilisées en sociologie. Deuxièmement, nous favorisons l'exposé des principes, des logiques, des intérêts et des limites des outils statistiques, au détriment des aspects mathématiques et calculatoires. Ainsi, par exemple, il semble préférable de bien assimiler la notion d'intervalle de confiance, les questions qu'elle permet de traiter, son intérêt et ses limites, plutôt que de savoir justifier mathématiquement la formule qui permet de calculer cet intervalle (c'est le travail du mathématicien), plutôt que de savoir calculer sans comprendre le sens du calcul (c'est le travail des logiciels statistiques).

Pour nous, les aspects mathématiques et calculatoires sont secondaires, non pas au sens où ils ne servent à rien, mais au sens où ils ne doivent venir qu'après la présentation et la justification, en langage ordinaire, des fondements et de la portée de chaque outil statistique. Lorsque ce manuel aborde leurs aspects mathématiques et calculatoires, c'est uniquement à la suite de l'exposé de leurs principes et de l'examen de leur portée.

Dès lors, deux usages du manuel sont possibles : soit comme outil de formation à la compréhension des fondements logiques des outils statistiques du sociologue (sans compétence mathématique particulière) ; soit comme outil de référence pour la mise en œuvre pratique des outils statistiques (comment les utiliser, interpréter et calculer ?).

La « seule » chose que le manuel ne fait vraiment pas, c'est fournir les justifications mathématiques exactes des outils statistiques. Ces justifications

1. Première édition Nathan, 1992 ; seconde édition revue, Armand Colin, 2005.

sont d'ailleurs impossibles sans de solides connaissances en théorie des probabilités et en statistique mathématique – nous renvoyons le lecteur intéressé aux ouvrages cités en bibliographie finale.

1. POURQUOI RECOURIR AUX OUTILS STATISTIQUES ?

La sociologie peut être vue comme une science étudiant les principes de variation sociale des caractéristiques individuelles, des comportements, des attitudes, des pratiques ou des opinions. Lorsqu'elle recourt aux données quantitatives, par exemple issues d'un questionnaire, la sociologie fait appel à la statistique puisque celle-ci fournit des outils destinés à analyser de grands ensembles de données. Face à de tels ensembles, la science statistique dispose d'outils théoriques et pratiques permettant d'identifier ces variations (par exemple dans les pratiques culturelles des Français), de comparer ces variations entre divers groupes (les hommes et les femmes ont-ils les mêmes pratiques ?), de saisir les liens pouvant unir ces variations (relation entre les pratiques de lecture et les pratiques d'écoute de la musique), d'identifier les groupes « typiques » ayant des pratiques plutôt homogènes, c'est-à-dire présentant peu de variations (les adolescents ont-ils des pratiques de lecture comparables ?), ou encore d'expliquer les principes de variations (la diversité des pratiques de lecture s'explique-elle par les différences d'âge ou de milieu social ?).

Ces notions de variation, de liens, de co-relation, d'explication, de typologie ou encore de comparaison, dont les exemples montrent bien toute l'importance en sociologie (comme dans toutes les sciences empiriques d'ailleurs), trouvent en statistique des expressions et des formalisations pratiques.

Ainsi, à titre d'illustration, l'idée de relation ou de lien s'exprime dans les notions statistiques de corrélation, de tableau croisé, de comparaison de pourcentages ou de test du khi^2 ; l'idée d'explication trouve une expression pratique dans la notion de régression ; l'idée de variation est bien incarnée dans la notion de variance ou dans le tri à plat... En somme, la statistique offre à la sociologie des instruments permettant d'opérationnaliser, c'est-à-dire de mettre en pratique, des questions que cette dernière se pose sur des faits sociaux.

2. LE PLAN DU MANUEL

Ce manuel est découpé en deux grandes parties. La première d'entre elles, divisée en deux chapitres, expose les différents aspects de la production et de la préparation des variables. Le chapitre 1 évoque les différentes sources de données statistiques en sociologie. Il aborde la question centrale de l'échantillonnage des individus ou situations analysées ainsi que la question de la validité des résultats établis sur un échantillon : dans quelle mesure une donnée issue d'une enquête auprès d'un échantillon d'individus renseigne-t-elle sur un phénomène dépassant le cadre du seul échantillon pour concerner l'ensemble de la population ? Le chapitre 2 est consacré aux aspects relevant de la préparation des données, du recodage, de la création de variables pertinentes pour le sociologue. Les contraintes pesant sur les protocoles de recueil et de constitution des données empiriques conduisent en effet à des variables primaires souvent imparfaitement adaptées au besoin de l'analyse : diverses opérations de préparation des variables sont nécessaires.

La seconde partie du manuel présente les outils d'analyse statistique proprement dits. Elle est également organisée en deux chapitres. Le premier d'entre eux (chapitre 3) expose les outils permettant d'analyser les relations en deux variables, quelle que soit leur nature. Le second (chapitre 4) aborde les outils destinés à l'analyse des relations entre plusieurs variables (trois voire beaucoup plus).

Entre les deux parties, un court « intermède » permet de présenter un point central pour les raisonnements et outils statistiques : la notion de variabilité et les principaux outils de sa mesure.

Partie 1

PRODUIRE ET PRÉPARER LES VARIABLES

PRODUIRE DES DONNÉES QUANTITATIVES

Pour étudier le social, le sociologue (quantitativiste) doit élaborer des outils (méthodes d'enquête, concepts, catégories, données) lui permettant de s'abstraire des cas particuliers, de se détacher des représentations individuelles (à commencer par la sienne). Ce travail d'objectivation est essentiel. Nous n'y revenons pas – le lecteur intéressé pourra consulter les ouvrages classiques[1] ou des manuels[2]. Nous nous attachons ici à préciser les aspects pratiques d'élaboration des données quantitatives en sociologie : les origines possibles de ces données, la construction des échantillons et enfin l'estimation de la fiabilité des résultats.

1. LES « DONNÉES » DU SOCIOLOGUE

Il est commun de dire que les sociologues, et plus généralement tous les scientifiques, travaillent sur des « données ». Le terme utilisé est très mal choisi car d'une part ces « données » sont construites et d'autre part elles sont coûteuses[3].

Elles sont construites au sens où elles résultent d'un travail d'élaboration théorique de la part du sociologue (et du statisticien) : celui-ci doit définir les dimensions du social qui semblent pertinentes (sa problématique), les concepts permettant de se représenter la réalité étudiée, les catégories servant

1. Émile Durkheim, *Les Règles de la méthode sociologique*, Paris, PUF, 1992 (première édition 1895) ; Pierre Bourdieu, Jean-Claude Chamboredon et Jean-Claude Passeron, *Le Métier de sociologue*, Paris, Mouton, 1968.
2. François de Singly, *op. cit.*, 2005, notamment le chapitre 1, § 2.
3. Olivier Martin, « Les statistiques parlent d'elles-mêmes ? Regards sur la construction sociale des statistiques », dans : *La Pensée confisquée*, Paris, La Découverte, 1997, p. 173-191 ; Alain Desrosières, « Entre réalisme métrologique et conventions d'équivalence : les ambiguïtés de la sociologie quantitative », *Genèses*, n° 43, 2001, p. 112-127 ; Alain Blum et Olivier Martin, *La Vérité des chiffres : une illusion ?*, Université Paris Descartes, 2009, vidéo téléchargeable sur <médiathèque.parisdescartes.fr/article.php3?id_article=3659> et sur iTunes U.

à coder les faits observés, ainsi que les modalités des protocoles d'interview ou d'observation... Les données ne s'offrent pas à lui : il doit les « conquérir ». Dire que les données sont construites ne signifie toutefois pas qu'elles sont inventées : affirmer que la conception d'un dispositif d'observation et d'enregistrement du réel est indispensable à l'étude de ce réel ne signifie pas que ce réel soit une invention, un artifice.

Elles sont coûteuses puisque la conception d'une enquête et sa réalisation nécessitent beaucoup de travail et donc de temps. Elles sont coûteuses parce qu'elles supposent la reproduction de questionnaires, leur diffusion puis leur saisie et parfois la rémunération des enquêteurs ou des personnes qui vont saisir les réponses. Le coût financier des enquêtes quantitatives constitue parfois un frein pour le sociologue dont les moyens peuvent être modestes, notamment dans l'espace universitaire.

2. Population, échantillon et individus

L'ensemble des situations qui intéressent le sociologue constitue la *population*. Les situations sur lesquelles il travaille réellement et qu'il va soumettre à son questionnaire ou à son protocole d'observation constituent son *échantillon* (qui est très souvent un petit sous-ensemble de la population, nous y reviendrons plus loin § 4). Enfin, chacune des situations étudiées est, selon une terminologie héritée des sciences statistiques, un *individu*. Ce terme possède en statistique, en sociologie quantitative et dans les logiciels d'analyse statistique un sens qui dépasse le sens habituel : un individu n'est pas nécessairement une personne, un homme ou une femme. C'est l'unité statistique élémentaire sur laquelle portent l'enquête et l'analyse sociologique. Ainsi, si le sociologue étudie les pratiques culturelles des Français, ses individus seront effectivement des individus au sens habituel, c'est-à-dire des personnes (françaises en l'occurrence). S'il travaille sur les usages des équipements électroménagers de ménages, ses individus sont des ménages. S'il travaille sur le rôle des politiques municipales dans la revitalisation des centres-villes, ses individus sont des communes. S'il étudie les systèmes de scolarisation des enfants et les compare entre différents pays, ses individus seront les divers systèmes identifiés ou les différents pays. S'il étudie le déroulement et le contenu d'émissions télévisuelles, ses unités sont les différents moments de la chronologie, voire les minutes de ces émissions... La notion d'individu est parfois remplacée par celle, moins ambiguë, d'*unité statistique*.

Il est fréquent que les enquêtes sociologiques portent à la fois sur les ménages et sur les individus qui composent ces ménages. C'est par exemple

le cas des enquêtes « conditions de vie » de l'INSEE : une partie du questionnement porte sur le ménage (composition, logement, patrimoine, revenus, équipements...) ; une autre porte sur les individus (état de santé, vie professionnelle et conditions de travail, habitudes de vie, relations sociales, loisirs...). Dans ce cas, l'analyse statistique devra soigneusement distinguer les données relatives aux individus et celles relatives aux ménages. Si l'enquête porte sur les animaux familiers, les taux de possessions de chats, de chiens et de poissons sont plutôt relatifs aux ménages (dans la mesure où les animaux familiers sont rarement la propriété exclusive d'un seul individu du ménage).

3. Les sources quantitatives en sociologie

L'analyse quantitative en sociologie suppose que les informations traitées soient standardisées, c'est-à-dire codées, transcrites, selon des critères communs à tous les individus. C'est à cette condition que le recours aux outils statistiques se justifie. Réaliser une enquête par questionnaire est l'une des manières d'obtenir des données standardisées. Mais ce n'est pas la seule.

3.1 Une source particulière mais courante : le questionnaire

Le questionnaire est sans conteste l'outil le plus fréquemment utilisé pour élaborer des données quantitatives en sociologie. Il ne s'agit pas, ici, de présenter la méthodologie du questionnaire[1]. Nous nous contentons de rappeler les grandes étapes de l'élaboration d'une enquête par questionnaire, et de souligner un aspect fondamental de la démarche quantitative : le recours à des indicateurs et à des questions-indices pour opérationnaliser les notions sociologiques.

Toute démarche quantitative en sociologie débute par une phase théorique : le sociologue doit expliciter sa problématique et les notions et concepts théoriques que celle-ci mobilise. Puis il doit rechercher des indicateurs empiriques opérationnalisant ses notions et concepts, les traduire en questions puis bâtir son questionnaire en organisant ces questions. De

1. François de Singly, *op. cit.*, 2005, chapitres 2 à 4.

manière conjointe, il doit définir les contours de la population qu'il souhaite enquêter. Une fois le questionnaire mis au point et la population cible définie, viennent les temps de la passation, de la saisie des réponses sur un logiciel dédié et enfin l'analyse.

L'opérationnalisation des notions et concepts est un point central, qui a des implications dans la fabrication du questionnaire mais aussi dans la phase d'analyse des données. Opérationnaliser un concept ou une notion revient à trouver des indicateurs empiriques de grandeurs trop abstraites ou trop complexes pour être mesurées par une seule question ou une seule observation : pour travailler sur « l'amour de la lecture », « la proximité sociale entre deux individus », « l'identité religieuse » ou encore « l'investissement scolaire », il est nécessaire de trouver des indicateurs de ces grandeurs. Ainsi, par exemple, la prise de cours particuliers, le temps passé avec les parents à discuter du travail scolaire et l'achat de manuels d'exercices sont des indicateurs d'investissement scolaire. Ces indicateurs doivent être transcrits en des questions. Ces questions ne sont pas des mesures directes de la notion ou du concept, mais des expressions partielles et sous-jacentes (latentes) : ces questions constituent des « indices ». Nous en verrons des exemples au chap. 2, § 4.

3.2 Les autres sources quantitatives

Le questionnaire permet de recueillir des données empiriques qui sont, par leurs conditions mêmes de récolte, standardisées : tel est bien l'objectif de la standardisation des questions et des modalités de réponse. Il est toutefois possible d'utiliser des données non standardisées *a priori* : c'est un traitement *a posteriori* de standardisation par codage qui rend possible l'analyse quantitative des informations empiriques. Cette situation n'est pas rare et ne doit pas être négligée : les matériaux susceptibles de faire l'objet d'une analyse quantitative ne s'arrêtent pas aux questionnaires. Il est par exemple possible de coder de façon standardisée puis d'analyser des textes, des lettres, des comportements, des sources médiatiques, des entretiens…

Le travail de Luc Boltanski dans son article « La dénonciation »[1], publié en 1984, fournit une bonne illustration de cette situation. Le matériau analysé est constitué par un corpus de 275 lettres adressées au journal *Le Monde*

1. Luc Boltanski (avec Yann Darré et Marie-Ange Schiltz), « La dénonciation », *Actes de la recherche en sciences sociales*, n° 51, mars 1984, p. 3-40.

entre 1979 et 1981 par des lecteurs dénonçant une ou des injustices et souhaitant voir leur lettre publiée. Boltanski analyse les modalités des dénonciations publiques exprimées dans ces lettres. Pour cela, il traite au « moyen des mêmes instruments » les lettres qui sont « remarquablement disparates » et en réduit « la diversité en leur appliquant uniformément un ensemble de codes », c'est-à-dire en les soumettant toutes aux mêmes interrogations » (p. 6). Sont ainsi codées : la description des affaires relatées dans les lettres (milieu, durée, nature, ressources...) ; la description du contenu des lettres (pièces jointes, photocopies d'actes de justice...) ; leurs propriétés graphiques (lisibilité, fautes, soulignements...) ; leurs propriétés stylistiques et rhétoriques (menaces, invectives, ironie, répétitions, discordances, titres utilisés par l'auteur pour se qualifier, genre littéraire...) ; les propriétés sociales de l'auteur (lieu de résidence, sexe, profession, âge...)...

Ces lettres, ainsi codées, font l'objet d'une analyse statistique fine et notamment d'une analyse factorielle des correspondances (voir chap. 4, § 2). Ce travail permet à Boltanski de conduire une analyse sociologique quantitative du discours et d'interroger la frontière et les liens entre l'action collective et l'action individuelle : il montre notamment comment sont mobilisés les arguments permettant de conduire d'un intérêt particulier à un intérêt général.

De la même manière, il est possible de coder des comportements et des paroles, y compris dans des situations *a priori* complexes comme une émission télévisuelle, le Téléthon, riche en séquences de nature très différentes (variétés, divertissements, reportages sur des manifestations se déroulant partout en France, témoignages de malades, interviews de parents d'enfants malades, interviews de médecins, de chercheurs comme de responsables publics...)[1].

3.3 L'analyse secondaire des données

Une dernière situation peut s'offrir au sociologue : disposer et analyser des données quantitatives dont il n'est pas le producteur. Dans ce cas, on parle communément d'*analyse secondaire*. C'est notamment le cas lorsqu'il utilise des données produites par des institutions ou sociétés (ministères, associa-

1. Dominique Cardon, Jean-Philippe Heurtin, Olivier Martin, Anne-Sylvie Pharabod et Sabine Rozier, « Les formats de la générosité : trois explorations du Téléthon », *Réseaux*, volume 17, n° 95, 1999, p. 16-105.

tions, entreprises, centres d'étude et de recherche, INSEE...) sans avoir pris part à la phase d'élaboration de ces données.

Les données utilisées n'ont pas nécessairement vocation à faire l'objet de traitements statistiques et d'analyse sociologique. Par exemple, le sociologue peut exploiter des annuaires ou des fichiers de personnes pour décrire la structure de la population des membres d'une association, des salariés d'une entreprise, des clients d'un commerçant... Il peut également utiliser, comme le font régulièrement les sociologues de l'éducation, les informations récoltées par les établissements scolaires à des fins administratives et informatives : les informations individuelles sur les élèves des écoles primaires, des collèges, des lycées ou des universités constituent un matériau utile pour l'étude statistique du système d'enseignement. Le sociologue peut aussi analyser les données récoltées par les opérateurs de téléphonie mobile pour étudier les pratiques téléphoniques, les réseaux sociaux d'échanges, les moments et les durées des appels.

Il arrive fréquemment que la sociologie utilise des données conçues par des instituts d'études statistiques (INSEE, INED ou encore CEREQ), par des directions statistiques des principaux ministères (Éducation nationale, Affaires sociales, Culture...) ou encore par des organismes d'études parapublics (Crédoc, CIDSP, OFDT...). Les enquêtes « Emplois », « Budget des familles », « Conditions de vie des ménages », « Logement », « Emploi du temps » ou encore « Handicaps, incapacités, dépendance » constituent des sources de données statistiques dans lesquelles le sociologue peut, en fonction de ses préoccupations et thématiques de recherche, venir puiser.

Dans tous les cas, le sociologue souhaitant conduire une analyse secondaire de données doit attentivement s'interroger sur les conditions de production, sur les modalités d'échantillonnage et sur les significations des questions posées. N'ayant pas participé au processus d'élaboration des données, il doit néanmoins parvenir à se familiariser avec l'esprit, les forces mais aussi les limites de l'enquête qu'il souhaite exploiter.

4. La notion d'échantillon

Tout sociologue dispose de deux stratégies pour conduire son étude : soit il réalise une enquête exhaustive auprès de tous les individus de la population qui l'intéresse ; soit il se contente d'examiner des « morceaux », « sous-ensembles » ou « fractions », appelés *échantillons*, de cette population. Dans

ce dernier cas, il réalise un *sondage*. Le terme « sondage » a un sens qui dépasse largement la signification qu'il prend à travers les pratiques des instituts de sondage ou les usages qu'en font les médias. Un sondage est une enquête sur une fraction de situations, choisies parmi toutes les situations possibles. L'enquête par sondage n'est pas propre au sociologue : c'est une méthode utilisée en marketing, en psychologie, en sciences politiques mais aussi dans les sciences médicales, pharmaceutiques ou biologiques (on ne peut pas tester un nouveau médicament sur l'ensemble d'une population), dans les sciences de la nature (le géologue ne peut pas sonder un sol en tout point) comme dans les sciences industrielles (un industriel se contente de tester la qualité de ses produits sur un échantillon). Au fond, tous les scientifiques sondent la réalité, chacun à leur manière.

4.1 Étudier toute une population ou un échantillon ?

Même lorsque c'est possible, il est souvent fastidieux et très coûteux de réaliser une étude exhaustive. Il ne faudrait d'ailleurs pas croire qu'une enquête exhaustive apporte une meilleure connaissance de la population : parce qu'une enquête exhaustive auprès d'une grande population suppose l'emploi d'un grand nombre d'enquêteurs, dont la formation doit être assurée, dont le travail doit souvent être contrôlé *a posteriori* et dont le coût est donc très élevé ; parce qu'un recensement nécessite d'opérer un maillage précis et systématique du territoire ou de l'espace à recenser ; enfin parce que la gestion de très grands ensembles de données présente de sérieuses difficultés (recoupement des informations, contrôle de la qualité des questionnaires, vérification du caractère réellement exhaustif...).

Les erreurs s'accumulant et les difficultés se multipliant avec la taille de la population à enquêter, une enquête exhaustive présente toujours des défauts : les erreurs, omissions ou doubles comptes, réponses inexactes ou omises, sont inévitables. Ainsi, le recensement de la population française par l'INSEE présente un taux d'erreur de 1 à 2 % : lors du recensement de 1990, l'INSEE estime avoir oublié environ 960 000 personnes et compter deux fois

400 000 autres individus[1]. D'ailleurs, l'INSEE a changé, en 2001, sa méthode de collecte des informations en abandonnant la recherche d'exhaustivité : chaque année, seule une partie des communes et seule une partie de la population des communes de grande taille seront interrogées. Ces sondages annuels sont plus simples à organiser tout en offrant une qualité d'information comparable voire supérieure à une enquête exhaustive.

Mieux vaut une enquête auprès d'un échantillon dont on connaît bien les conditions de recrutement et de passation qu'une enquête aspirant à être exhaustive ou très large. Ce principe a été illustré de manière éclatante par une affaire célèbre, qui a popularisé l'enquête par sondage : lors des élections présidentielles américaines de 1936, George Gallup a utilisé un échantillon de 5 000 personnes pour prévoir le vote tandis qu'un journal a sollicité 2 millions de personnes mais sans contrôler leur représentativité en espérant que la très grande taille de l'échantillon serait le garant de la qualité des résultats. La prédiction de Gallup (victoire de Roosevelt) s'est avérée exacte alors que le journal s'est trompé.

Ainsi, contrairement à une idée spontanée, l'enquête exhaustive d'une population n'est pas une solution idéale : travailler sur un échantillon bien conçu permet de mieux contrôler le choix des individus et les conditions de passation des questionnaires ainsi que de réduire les non-réponses, les biais de réponse et les erreurs de mesures. À l'exception des situations où les populations étudiées sont de petite taille (par exemple la population des élèves inscrits en terminale dans un lycée particulier, ou la population d'un immeuble), le sociologue travaille exclusivement sur des échantillons. Interroger un nombre restreint d'individus apporte autant d'informations, et des informations de meilleure qualité, qu'une enquête exhaustive. La condition est que ce nombre restreint d'individus – cet échantillon – soit « bien conçu » et respecte un certain nombre de principes. Il existe deux grandes catégories d'échantillon : les échantillons aléatoires ou « probabilistes » (les individus enquêtés sont choisis au hasard) ; et les échantillons empiriques ou « non probabilistes » (les individus enquêtés sont choisis selon des principes non aléatoires).

Construire un échantillon, c'est échantillonner une population (appelée population « mère », « cible » ou « de référence »).

1. Voir notamment François Héran et Laurent Toulemon, « Que faire quand la population recensée ne correspond pas à la population attendue ? », *Population et sociétés* (INED), n° 411, avril 2005.

4.2 Les échantillons aléatoires

Les individus composant un échantillon aléatoire sont choisis de manière probabiliste, c'est-à-dire au hasard, parmi les membres de la population de référence. Il ne faut pas se méprendre sur le sens des mots « hasard » ou « probabiliste » : ces termes signifient qu'aucun principe ou critère ne doit présider au choix des individus. Ainsi, interroger les 100 premières personnes qui franchissent les portes d'un musée ne constitue pas un échantillon aléatoire des visiteurs puisque le jour et l'heure de visite sont déterminés par la situation professionnelle et familiale des individus. Interroger les individus dont le prénom commence par une lettre choisie au hasard ne constitue pas plus un échantillon aléatoire : la distribution des prénoms est conditionnée par le sexe mais aussi par la langue et la culture. Il y a presque deux fois plus de prénoms masculins que féminins commençant par W ; et les plus fréquents sont William, Walid, Wassim, Wesley, Warren, Wilfried, Willy, Wilson, Wissan... prénoms dont les caractéristiques sociales des parents qui les choisissent ne sont pas neutres[1].

Pour construire un échantillon aléatoire, il faut que les individus soient choisis indépendamment de toutes leurs caractéristiques ou propriétés. Techniquement, il est nécessaire que tous les individus de la population de référence aient une probabilité (« chance ») connue et non nulle de faire partie de l'échantillon. Il existe trois manières de concevoir de tels échantillons.

– Le premier type d'échantillon aléatoire est l'*échantillon aléatoire simple*, qui correspond à la situation où tous les membres d'une population ont une probabilité identique de faire partie de l'échantillon. Ce type d'échantillon présente deux avantages essentiels. Premièrement, il ne présuppose aucune connaissance sur les principes structurant, sociologiquement, la population : il n'est pas nécessaire de connaître la répartition de la population-cible selon les âges, les sexes, les catégories sociales, etc., pour construire l'échantillon. L'enquête fournira, entre autres, ces informations. Deuxièmement, ce type d'échantillon fournit des indications fiables, c'est-à-dire représentatives (non biaisées), sur la population : si l'échantillon comporte 42 % de personnes s'étant rendu au cinéma au cours du dernier mois, on est presque certain qu'environ 42 % des personnes de la population-cible sont dans ce cas (*cf.* le § 1.5 consacré à l'évaluation de la fiabilité des chiffres).

1. Pour se convaincre de l'absence de neutralité sociale dans les choix des prénoms, voir : Philippe Besnard, Guy Desplanques, *Un prénom pour toujours. La cote des prénoms, hier, aujourd'hui et demain*, Paris, Balland, 1986 ; régulièrement actualisé.

Il est toutefois très difficile, voire impossible, de construire de tels échantillons : la condition selon laquelle tous les individus de la population ont des chances identiques de participer à l'échantillon est apparemment simple mais n'est en fait pas facile à respecter. Imaginons, par exemple, vouloir construire un échantillon aléatoire simple d'étudiants dans les universités de Lyon. Il est possible de se rendre dans les différents sites universitaires de la ville, puis d'interroger « au hasard » des étudiants sortant de l'enceinte des bâtiments universitaires. Une difficulté provient des principes qui vont guider ce choix « au hasard » : certains étudiants vont refuser de répondre ; l'enquêteur introduira certainement un biais de sélection, en fonction de la sympathie qu'il éprouve ou pas pour telle ou telle catégorie d'étudiants, pour tel ou tel « look » ou « sexe ». Une autre difficulté résulte du fait que tous les étudiants ne fréquentent les sites universitaires de manière identique : les étudiants en doctorat peuvent être en mission ou en déplacement pour réaliser une enquête à l'extérieur ; des étudiants malades ou souffrant de handicaps peuvent échapper à l'enquêteur ; des étudiants séchant les cours ont peu de chance d'être interrogés…

En fait, pour réaliser un échantillon aléatoire simple, il est nécessaire de connaître la liste exhaustive, complète et sans erreur, des individus composant la population. En l'occurrence, il faut posséder la liste nominative fiable (avec les coordonnées téléphoniques ou postales) des étudiants – avec toutes les difficultés que posent les étudiants ayant déménagé plusieurs fois ou logés chez divers amis, ceux ayant quitté la ville ou le pays en abandonnant leurs études. De manière générale, une telle liste, comportant l'ensemble des individus d'une population, est appelée *base de sondage*. C'est à partir de cette liste que le tirage aléatoire des individus peut être réalisé : en utilisant des méthodes permettant d'obtenir des nombres au hasard (« générateurs de nombres aléatoires »), le sociologue peut extraire de la base de sondage un échantillon de taille quelconque.

Cette méthode est notamment utilisée pour étudier les membres d'un groupe dont l'existence est instituée (objectivée par une « institution ») : les clients d'une banque ; les étudiants d'une école ; les adhérents d'une association ; les abonnés à un magazine… Mais, en dehors de ces quelques situations, cette méthode d'échantillonnage est peu utilisée en sociologie (et ailleurs) car disposer de la base de sondage est rarement possible, soir parce qu'elle n'existe pas, soit parce qu'elle est inaccessible au sociologue. Il n'existe pas (fort heureusement !) de liste des mères divorcées, des usagers des transports en commun parisiens, des couples homosexuels, des SDF (sans domicile fixe) ou encore des amateurs de musique celtique… Ces groupes

n'ont aucun caractère officiel permettant de penser qu'il existe une liste exhaustive de leurs membres. Et leurs frontières sont floues et incertaines : celui qui s'aviserait de constituer une telle liste ne pourrait qu'échouer... Par ailleurs, lorsqu'elles existent, les listes ne sont pas toujours facilement accessibles : s'il existe une liste des adhérents au Front national, il est peu probable que les responsables de ce parti fournissent cette liste aux sociologues pour leur permettre de réaliser une enquête !

Il faut par ailleurs se méfier des listes apparemment fiables mais qui présentent des biais indéniables : les annuaires téléphoniques (pages blanches) semblent constituer une base de sondage des ménages français, mais ce serait oublier que des ménages n'ont pas le téléphone, que d'autres sont sur liste rouge, que d'autres ont plusieurs lignes téléphoniques et apparaissent donc plusieurs fois dans les annuaires, que les personnes vivant en institutions (hôpitaux, asiles, maisons de retraites, casernes, prisons) échappent également à ces annuaires...

– Le deuxième type d'échantillon aléatoire est l'*échantillon stratifié*, qui consiste à découper la population en groupes (ou strates) et à réaliser un échantillon aléatoire au sein de chacun des groupes. Un sociologue de la famille souhaitant étudier la gestion des tâches ménagères au sein des ménages en fonction de leur composition (personne isolée, couple sans enfant, couple avec enfants, famille monoparentale) aura intérêt à stratifier son échantillon de ménages : pour cela, il choisira un échantillon aléatoire de ménages avec enfants à domicile, un échantillon aléatoire de ménages avec enfants ayant quitté le domicile, un échantillon aléatoire de ménages monoparentaux... Un échantillon stratifié assure au sociologue que son échantillon comprendra un nombre suffisant de chacune des situations qui l'intéresse en priorité : il pourra ainsi conduire des analyses fines de chacune de ces situations (même si elles sont relativement rares à l'échelle de la population globale) et comparer les situations entre elles. Un tel échantillon n'est en général pas représentatif de la population étudiée mais chacune des strates l'est – il est toutefois possible de redresser (*cf. infra*) l'échantillon pour le rendre représentatif.

On distingue l'échantillon stratifié proportionnel et l'échantillon stratifié non proportionnel. Dans le premier cas, le nombre d'individus enquêtés au sein de chaque strate (ou groupe) est proportionnel à l'importance du groupe par rapport à la population totale : dans notre exemple précédent, ce serait le cas si le sociologue interrogeait une part de familles monoparentales égale à la part de ce type de famille dans la population totale. Dans le second cas (échantillon stratifié non proportionnel), ce critère de proportionnalité n'est pas respecté. Le recours à ce type d'échantillon est utile pour étudier

finement des pratiques peu fréquentes à l'échelle de l'ensemble de la population. Si le sociologue veut par exemple avoir des résultats précis sur les pratiques musicales (jouer d'un instrument de musique), il aura intérêt à construire un échantillon stratifié pour lequel les classes sociales favorisées seront sur-représentées – puisqu'il sait que jouer d'un instrument de musique est une activité beaucoup plus fréquente dans ces catégories sociales.

– Il existe un troisième type d'échantillon aléatoire : les *échantillons en grappes* (ou « par grappe »). On suppose ici que les individus de la population sont naturellement regroupés en paquets relativement homogènes, appelés « grappes ». Réaliser un échantillon en grappes revient alors à constituer un échantillon aléatoire de ces grappes puis à enquêter l'ensemble des individus de chacune des grappes retenues. Cette méthode d'échantillonnage est par exemple utilisée pour réaliser les enquêtes « passagers/voyageurs », « visiteurs » ou « clients » : on constitue un échantillon aléatoire de trains (ou d'avions, de bus...), de moments de visite (jours et heures), de magasins ou de médecins puis on interroge tous les passagers, tous les visiteurs, tous les clients ou tous les patients de cet échantillon. C'est également une des méthodes utilisées pour fournir aux journalistes les estimations des résultats des élections à « 20 heures » : un échantillon de bureaux de vote est constitué et, au sein de chacun de ces bureaux, les bulletins sont dépouillés rapidement.

Notons enfin que les sondages par grappe sont appelés *sondages aréolaires* si le critère de découpage de la population est un critère géographique ou spatial. La constitution d'un tel échantillon est courante lorsqu'on souhaite enquêter les habitants d'une commune : on découpe la zone en « blocs » relativement homogènes (quartier, rue, immeuble,...), puis on tire aléatoirement un ensemble de « blocs » au sein desquels on interroge tous les habitants. Pour ses enquêtes « Emploi », l'INSEE utilise une méthode aréolaire qui facilite le repérage des logements « marginaux » (meublés, sous-locations, logements de domestiques) et permet d'éviter la sous-estimation de leurs occupants.

L'échantillonnage par grappe présente deux avantages majeurs. Premièrement, à la différence des deux méthodes précédentes, il n'est pas nécessaire de disposer d'une base de sondage complète et de bonne qualité : il suffit de choisir un principe permettant de découper les populations. Deuxièmement, un sondage par grappe est relativement moins coûteux en temps et en argent car il ne nécessite pas de parcourir toutes les grappes : les enquêteurs se contentent d'interroger les individus dans des lieux ou à des moments précis.

4.3 Les échantillons empiriques

La procédure d'échantillonnage est dite « empirique » lorsque les individus sont choisis en fonction de critères ne garantissant pas le caractère aléatoire de l'échantillon. Les échantillons par quotas sont les plus courants mais il existe également des échantillons volontaires et accidentels. Ils se distinguent essentiellement en fonction du caractère plus ou moins explicite et plus ou moins raisonné des critères de choix des enquêtés. On parle parfois d'échantillons à « choix raisonné » si ces critères de choix sont connus et bien choisis (au regard de la problématique). Notons que raisonner le choix de l'échantillon suppose de connaître quelques informations sur l'ensemble de la population.

– Les *échantillons par quotas* sont des échantillons respectant des critères de composition ou de structure : les individus ne sont pas choisis au hasard mais en fonction de leur capacité à respecter ces critères. Par exemple, constituer un échantillon par quota de sexe (50 % de femmes et 50 % d'hommes) revient à trouver autant d'hommes que de femmes ; constituer un échantillon par quotas de diplômes (avec ou sans baccalauréat) revient à imposer dans l'échantillon un nombre précis de titulaires du bac et un nombre précis de non-titulaires. Chaque enquêteur se voit attribuer des critères de recrutement : tant de femmes, tant d'hommes, tant de Parisiens, tant de provinciaux, tant d'ouvriers, tant de cadres... Libre à lui de trouver des individus permettant de satisfaire ces quotas. En général, les quotas sont définis à partir de critères sociodémographiques simples, comme le sexe, l'âge, la profession, la région de résidence : c'est notamment le cas des échantillons des enquêtes d'opinion ou d'intention de vote dont la presse fait souvent état... Rien n'interdit toutefois d'utiliser d'autres types de caractéristiques, sauf la difficulté à trouver des enquêtés adaptés si les caractéristiques ne sont pas faciles à connaître.

Le succès de l'enquête par quotas résulte de la facilité de sa mise en œuvre : il est inutile de disposer d'une base de sondage ; les enquêteurs sont libres, dans le cadre du respect des quotas, d'enquêter qui ils veulent. Cette médaille a des revers. D'une part, en dehors des critères de quotas, on ne sait pas précisément quels sont les principes qui ont guidé le choix des enquêtés : le recrutement peut être fortement biaisé. D'autre part, tout recours aux outils statistiques de mesure de la qualité ou de la fiabilité des résultats est, en toute rigueur, impossible – nous verrons pourquoi et comment ils peuvent néanmoins être utilisés.

– Beaucoup d'enquêtes recourent à des *échantillons « volontaires »* ou *« spontanés »*, c'est-à-dire des échantillons dont les membres ont eux-mêmes

décidé de se soumettre à l'enquête. C'est le cas de toutes les enquêtes utilisant des questionnaires mis librement à disposition d'un public, qu'ils soient empilés dans un lieu, ou qu'ils soient publiés dans la presse ou sur Internet : les individus sont entièrement libres d'y répondre. Si le recrutement des enquêtés ne coûte rien, il a l'inconvénient de ne pas permettre de savoir quels sont les critères ayant conduit certains à répondre et d'autres à ne pas répondre. Le problème de savoir à « qui on a affaire ». Ainsi l'échantillon de lecteurs constitué à partir d'une enquête par questionnaire glissé dans un magazine pourra-t-il nous renseigner sur une catégorie particulière de lecteurs, plutôt déterminés à faire valoir leur point de vue et à participer à la vie du magazine.

– Cette remarque s'applique également aux *échantillons accidentels*, c'est-à-dire aux échantillons constitués au gré des circonstances, sans réflexion sur les conditions de recrutement. Un sociologue travaillant sur les usages du téléphone portable dans les lieux publics peut par exemple observer les cent premières personnes utilisant leur portable dans une rue ou une place choisie. Faute d'avoir des idées précises sur les profils des individus possédant un téléphone portable et fréquentant cette rue ou cette place, le sociologue n'a aucune idée de la population d'où est issu cet échantillon. Faute d'être pleinement conscient des critères qui lui ont fait choisir cet individu plutôt qu'un autre, il ne peut rien dire sur la nature de son échantillon. Il ne pourra pas facilement généraliser les résultats, ni considérer que ses observations sont représentatives du comportement général des propriétaires de téléphones portables. Son enquête lui permettra néanmoins d'identifier certains comportements des utilisateurs de portables dans les lieux publics, de comprendre la logique et le sens de ces usages, et de saisir leur lien éventuel avec des traits sociaux généraux.

Afin de diminuer la place de la subjectivité de l'enquêteur dans le choix des enquêtés, il est possible d'adopter diverses méthodes imposant des contraintes plus ou moins fortes sur ce choix. Par exemple, la *méthode des itinéraires* impose à l'enquêteur un itinéraire et des arrêts en des points précis. En chacun de ces points, il doit interroger la première personne présente. Dans le même ordre d'idée, on peut imposer une *méthode de sélection* : souhaitant réaliser une étude du public d'un musée, l'enquêteur doit, après chaque enquêté, interroger la dixième (ou la vingtième...) personne qui se présente à l'entrée d'un musée.

De manière générale, toute enquête conduite sur un échantillon empirique doit être attentive aux biais introduits par les conditions de réalisation et par les critères de sélection utilisés. Par exemple, un sociologue souhaitant travailler sur les pratiques de lecture de la presse peut vouloir construire son

échantillon à partir de clients de kiosques de presse ou de boutiques de tabac-presse. Il doit simplement prendre conscience qu'il surestime probablement la part des « gros lecteurs » au détriment des « faibles lecteurs » : ces derniers se rendant moins souvent dans les kiosques, la probabilité que le sociologue les interroge est faible. Il doit également réfléchir aux populations qui échappent à son échantillon, en l'occurrence les abonnés, les non-lecteurs et ceux qui lisent la presse dans les bibliothèques publiques...

4.4 Qu'est-ce qu'un échantillon représentatif ?

Un échantillon est dit *représentatif* s'il possède la même « structure » que la population de référence. Cela signifie que les différents sous-groupes qui composent cet échantillon doivent représenter une part identique à la part qu'ils représentent dans la population : si la population comprend 52 % de femmes et 48 % d'hommes, un échantillon représentatif devra comprendre 52 % de femmes et 48 % d'hommes (sur un échantillon de 1 000 personnes, cela suppose d'enquêter 520 femmes et 480 hommes). En d'autres termes, dire qu'un échantillon est représentatif, c'est dire que les répartitions (ou distributions) des variables ou des caractères sont identiques dans l'échantillon et dans la population. La représentativité est assurée lorsque l'échantillon est probabiliste et que tous les individus ont une probabilité identique d'être inclus dans l'échantillon. Elle est également assurée pour les autres types de sondages aléatoires à condition de les redresser.

Par ailleurs, la représentativité est souvent recherchée dans les échantillons par quotas. En particulier, les échantillons par quotas utilisés dans les sondages d'opinion ou dans les enquêtes d'intention de vote sont représentatifs. Il est toutefois impossible de construire un échantillon d'individus dont toutes les variables/caractères ont des distributions identiques dans l'échantillon et dans la population. Seules les distributions d'un petit nombre de variables peuvent être respectées. La notion de représentativité n'a pas de sens si on ne précise pas la population à laquelle l'échantillon se réfère et les critères ou variables dont les distributions sont respectées. Un échantillon n'est pas « représentatif », il peut simplement être « représentatif d'une population au sens d'une série de critères ». Quand on entend ou qu'on dit qu'un échantillon est représentatif, il faut se demander : 1°) représentatif de quoi (de quelle population) ? 2°) au sens de quels critères (quelles sont les variables ou les caractères dont les distributions sont respectées) ? Par exemple, les échantillons représentatifs utilisés dans les enquêtes d'intention de vote ou d'opinion

sont souvent représentatifs de la population française en âge de voter (18 ans et plus) au sens des critères d'âge, de sexe, de diplôme et de catégorie sociale.

Lorsqu'un échantillon n'est pas représentatif, il est parfois possible de le redresser pour le rendre représentatif, c'est-à-dire de pondérer les individus de façon à respecter le poids de chacune des sous-populations. Un échantillon comportant 50 % d'hommes et 50 % de femmes alors que la représentativité impose respectivement 48 % et 52 % peut être redressé en affectant un poids de $48/50^e$ (soit 0,96) à chaque homme et un poids de $52/50^e$ (1,04) aux femmes de l'échantillon. Tous les calculs statistiques doivent ensuite se faire en tenant compte de ces pondérations – la moyenne doit par exemple être remplacée par la moyenne pondérée. Les pondérations, appelées « coefficients de redressement », constituent en fait le poids des réponses de chaque enquêté dans les résultats finaux. Les opérations de redressement n'ont de sens que si les coefficients de redressement ne sont pas trop élevés (c'est-à-dire si la structure de l'échantillon n'est pas trop différente de la structure de référence).

En général, la représentativité d'un échantillon ne garantit pas que les observations conduites sur cet échantillon soient nécessairement valables pour l'ensemble de la population. Dans les années 1960, les caractéristiques générales (âge, sexe, CSP) des lecteurs de *Paris Match* étaient proches des caractéristiques de l'ensemble de la population française (hors enfants). D'un point de vue strictement formel, cela faisait des lecteurs de *Paris Match* un échantillon représentatif de la population française. Il est pourtant évident que cet échantillon était biaisé et qu'une enquête auprès de ces lecteurs aurait fourni des estimations erronées des comportements généraux (à commencer par l'estimation des pratiques de lecture !). En toute rigueur, les variables étudiées par le sociologue doivent être fortement liées aux critères de représentativité. Si ce n'est pas le cas, rien ne dit que les résultats établis sur l'échantillon pourront être généralisés à l'ensemble de la population.

4.5 Remarques et conseils supplémentaires

Terminons cette présentation des méthodes d'échantillonnage par quelques remarques supplémentaires. Il est, premièrement, tout à fait possible de combiner les diverses méthodes d'échantillonnage : l'enquête sur les pratiques culturelles des Français réalisée en 1997 par le ministère de la Culture a par exemple été conduite sur un échantillon de 3 000 personnes : la population française (âgée de 15 ans et plus) a été stratifiée en fonction des régions et des

catégories d'agglomération puis, au sein des strates, des échantillons par quotas d'âge, de sexe, de CSP, de taille du foyer et d'activité de la femme ont été conçus[1].

Une situation fréquente mérite d'être signalée : le cas où le sociologue construit son échantillon de personnes en choisissant dans un premier temps des ménages, puis, dans un second temps, une personne au hasard parmi les membres de chaque ménage. La procédure de choix de cette personne au sein de chaque ménage doit respecter de stricts principes aléatoires : la personne ne doit pas être choisie parce qu'elle est plus disponible, parce qu'elle dit être le chef de ménage ou parce qu'elle est le seul adulte présent. Une des méthodes habituellement utilisée pour respecter le hasard dans le processus de choix est la méthode de « l'individu Kish[2] ». Cette situation ne doit pas être confondue avec le cas où l'on souhaite acquérir des informations sur l'ensemble du ménage en interrogeant une seule personne : dans ce cas, l'échantillon est un échantillon de ménages et la personne à interroger est la personne la plus compétente ou la mieux informée.

Deuxièmement, la publication des résultats issus d'une enquête par échantillon doit être accompagnée d'une fiche technique décrivant non seulement sa taille, mais aussi le mode de construction de l'échantillon (critères utilisés, base de sondage, type de tirage aléatoire, quotas...). C'est une obligation légale pour les sondages d'opinion publiés dans la presse et c'est un principe d'honnêteté et d'objectivité qui devrait être suivi dans tous les cas. Voici un exemple de fiche technique accompagnant la publication d'un sondage : « Échantillon de 1 006 personnes, représentatif de la population française âgée de 18 ans et plus, selon la méthode des quotas appliquée aux variables suivantes : sexe, âge, profession du chef de famille, après stratification par région et catégorie d'agglomération. » Il est également préférable d'indiquer la période et le mode de passation : « L'enquête a été réalisée par X du 15 au 20 février par téléphone. »

La troisième remarque est relative à une question fréquemment posée : « Qu'est-ce qu'un bon échantillon ? » Retenons, en premier lieu, qu'il n'existe pas de critère absolu permettant de savoir si un échantillon est « bon » ou « mauvais ». Il existe simplement des principes généraux de construction, plus ou moins faciles à mettre en œuvre et garantissant une « représentativité » plus ou moins grande. Et, surtout, il existe un principe

1. Olivier Donnat, *Les Pratiques culturelles des Français, enquête 1997*, Paris, La Documentation Française, 1998.
2. Pour une présentation détaillée : <www.cmh.acsdm2.ens.fr/glossair.php>.

affirmant qu'un échantillon n'est jamais « bon » ou « mauvais » de manière absolue : l'intérêt d'un échantillon ne se juge pas de manière intrinsèque, en fonction de ses seules propriétés statistiques mais en fonction de son adéquation à une problématique ou une série d'interrogations précises. Même le critère de représentativité n'est pas nécessairement un idéal.

Par exemple, souhaitant travailler sur les troupes de danse amateur (danse sportive, danse de salon...) et cherchant à identifier les caractéristiques générales des pratiquants, il est important de disposer d'un échantillon représentatif (en supposant que cela soit possible). Mais un échantillon représentatif comportera peu d'hommes (puisque la pratique est essentiellement féminine) et ne sera pas adapté à une recherche portant sur les motifs différenciés des femmes et des hommes à participer à de telles troupes de danse : dans ce cas, il est essentiel de concevoir un échantillon sur-représentant les hommes (avec par exemple un quota de 50 % d'hommes et de 50 % de femmes). Ainsi, si on souhaite pouvoir comparer des groupes, il est important que chacun de ces groupes soit bien représenté.

Cette considération débouche sur un conseil pratique pour aider à bien concevoir des échantillons : plus l'hétérogénéité et la complexité du phénomène étudié sont supposées grandes, plus l'échantillon (ou le sous-échantillon) devra être de grande taille. Notamment, si l'échantillon peut être stratifié et qu'au sein de certaines strates les comportements et pratiques des individus sont supposés être délicats à saisir car très divers et présentant peu de régularité, il est essentiel de sur-représenter ces strates. Considérons par exemple une enquête sur les pratiques de communication des ménages français via des « TIC » (technologies de l'information et de la communication : téléphones, téléphones portables, SMS, chat, e-mails...). Il est plus délicat d'étudier les pratiques des ménages multi-équipés (possédant plusieurs portables, des ordinateurs connectés à Internet et des lignes de web phone) que des ménages équipés d'un seul téléphone fixe. Bien saisir les pratiques des multi-équipés suppose de les sur-représenter et, inversement, de sous-représenter (alors qu'ils sont majoritaires en réalité) des non ou faiblement équipés. De manière générale, on peut dire qu'un groupe doit être d'autant mieux représenté dans un échantillon qu'il présente une grande hétérogénéité, une grande variabilité. Et, dans le cas particulier de la comparaison de plusieurs groupes sur lesquels le sociologue n'a pas d'hypothèse précise sur l'hétérogénéité ou au contraire leur homogénéité de comportement, le plus simple est d'égaliser la part de chacun des groupes dans l'échantillon – comme dans notre exemple précédent concernant les pratiquants de danse.

4.6 L'échantillon comme prisme

Le critère de représentativité n'est pas idéal. Il est, de plus, souvent illusoire de chercher à le respecter. D'une part, comme nous l'avons dit, cela suppose le tirage aléatoire d'enquêtés à partir d'une liste exhaustive précise, qui existe très rarement. D'autre part, la représentativité assurée par la méthode des quotas est, elle-même, difficile à respecter : cela suppose de connaître les caractéristiques générales de la population-mère. Et que sait-on de la répartition des sexes, des âges et des catégories sociales au sein de la population des danseurs ? Dispose-t-on d'éléments précis permettant de décrire la population des lecteurs de polars ou celle des SDF ? Sans ces éléments, il est impossible de concevoir des échantillons respectant la représentativité par la méthode des quotas.

Dès lors, quel sens et quelle confiance accorder à des échantillons non représentatifs ? Il faut peut-être commencer par rappeler que les sociologues ne sont pas les seuls à avoir recours à des échantillons dont la représentativité n'est pas garantie : les sciences du vivant (biologie, pharmacie et médecine) travaillent fréquemment sur des groupes d'individus dont on ne connaît pas la représentativité[1]. En fait, plutôt que de partir d'une population bien connue pour construire un échantillon représentatif, on part de l'échantillon obtenu empiriquement et raisonné « au mieux » et on considère que cet échantillon est représentatif d'une population aux contours ignorés *a priori*. Elle est mal connue mais bien réelle et il est possible d'en obtenir une meilleure connaissance grâce aux renseignements obtenus à travers l'échantillon (répartition des sexes, des âges ou de toute autre caractéristique connue sur l'échantillon). Selon cette perspective, l'échantillon est un prisme laissant entrevoir une population dont la description peut être faite *a posteriori*.

Cette inversion de perspective (partir de l'échantillon pour « construire » la population dont cet échantillon est représentatif) n'est pas un tour de passe-passe. Elle ne garantit pas que les résultats obtenus soient universaux et que les statistiques publiées vaillent pour l'ensemble de la population. Mais elle permet de s'assurer que ces résultats et ces statistiques ont un sens et une portée réelle : ils valent pour une population dont on est capable de décrire les principaux contours. Si, pour des raisons évidentes, il est impossible de construire un échantillon représentatif des SDF, il est néanmoins possible de concevoir un échantillon des utilisateurs de services d'hébergement gratuits ou à faible participation,

1. Daniel Schwartz, *Le Jeu de la science et du hasard. La Statistique et le vivant*, Paris, Flammarion, 1994, p. 33-39.

ou de distribution de repas chauds en échantillonnant convenablement ces services[1]. Cet échantillon ne peut pas prétendre représenter l'ensemble des SDF mais, d'une part, il semble difficile de faire mieux et, d'autre part, il permet d'approcher les conditions de vie d'une partie non négligeable des SDF.

Dernière remarque : tout comme les enquêtes mobilisant des matériaux qualitatifs (entretiens, récits de vie) qui ne sont jamais représentatifs, l'enquête quantitative sur un échantillon non représentatif permet néanmoins d'identifier des phénomènes, des mécanismes, des processus et des traits typiques... Il n'est pas nécessairement utile de montrer qu'un phénomène concerne une part précisément connue de la population générale pour estimer que ce phénomène est intéressant et pertinent à analyser.

4.7 À propos des « erreurs » et des « biais »

Toute enquête est entachée d'« erreurs ». Il n'existe pas de règles ou de principes permettant de s'en prémunir : il faut simplement être très attentif. Cette vigilance doit s'exercer durant la phase de préparation de l'enquête, durant sa réalisation comme durant son exploitation. Ces erreurs de mesure peuvent être accidentelles ou systématiques. Elles peuvent résulter de la conception du questionnaire (questions mal rédigées, oubli d'une modalité...) ou de problèmes survenus lors de la passation des questionnaires. Sur ce dernier point, il est particulièrement important d'être attentif aux biais d'auto-sélection, aux refus de répondre, aux absences... Les difficultés sont fréquentes chez les enquêtés à profil particulier : des individus ayant des horaires de travail décalés et donc absents de leur domicile lors du passage de l'enquêteur ; des individus s'estimant non concernés ou non aptes à remplir des questionnaires ; des individus jugeant superflu de perdre du temps à répondre à des questions... Et dans le cas des enquêtes par échantillon volontaire ou par questionnaires auto-administrés, les taux de réponses varient considérablement selon les caractéristiques des individus.

Il faut également ne pas sous-estimer les biais des bases de sondage. Un sociologue utilisant un fichier administratif pour constituer son échantillon doit s'interroger sur les profils des personnes dont les informations sont absentes de ce fichier ou incomplètes. L'exemple de l'annuaire téléphonique, déjà évoqué, illustre bien ce point.

1. Voir les résultats de l'enquête de l'INSEE : Cécile Brousse, Bernadette de la Rochère et Emmanuel Massé, « Hébergement et distribution de repas chauds », *Insee Première*, n° 823 et n° 824, janvier 2002. De manière plus générale, voir Cécile Brousse, Jean-Marie Firdion, Maryse Marpsat, *Les Sans-Domicile*, Paris, La Découverte, 2008.

Il ne faut toutefois pas envisager ces diverses sources d'erreurs ou de biais comme des « erreurs et biais de mesure » au sens classique de la métrologie. Cette science de la mesure est essentiellement fondée sur les questions soulevées dans les sciences physiques et notamment la mécanique. Elle s'intéresse aux difficultés rencontrées lors des mesures de longueur, des durées, des volumes, des fréquences ou encore des densités. La conception de la mesure est une conception réaliste : il existe de vraies, d'authentiques grandeurs que les dispositifs de mesure cherchent à identifier. En pratique, ces dispositifs commettent toujours des erreurs résultant de l'imprécision des instruments (du mètre ou de la balance), des conditions de mesure (la température ambiante peut modifier la longueur du mètre utilisé) ainsi que du comportement de l'opérateur (la précision de ses gestes, son calme...). Selon cette conception de la mesure, un biais ou une erreur est un écart par rapport à une valeur réelle. L'objectif du scientifique est alors de s'approcher au plus près de cette réalité.

En sciences sociales, et spécifiquement en sociologie, cette conception réaliste de la mesure n'est pas justifiable – ou seulement pour quelques rares variables. Prétendre le contraire serait supposer qu'il existe des réponses parfaitement exactes, sans ambiguïté, à des questions telles que, par exemple, « L'exécution de votre travail vous impose-t-elle de porter ou de déplacer de lourdes charges ? », « Avez-vous confiance dans notre médecin ? », « Combien de films avez-vous vus au cours du dernier mois ? », « Fumez-vous ? »... Même ces deux dernières questions, apparemment purement factuelles, n'appellent pas des réponses sans arbitraire ni équivoque : comment répondre si le visionnage d'un film a été interrompu, si on s'est endormi durant le film, si on est sorti de la salle avant la fin, si on regarde un film à la télévision tout en travaillant ? Comment répondre si on fume occasionnellement, si on vient de s'arrêter ou si on s'arrête momentanément en raison d'un traitement médical ?

Les réponses aux questions des sociologues renvoient à une réalité sociale complexe qui ne peut pas être facilement mise en ordre et catégorisée sans ambiguïté. Il existe un flou intrinsèque à toute catégorisation. Ce flou peut être plus ou moins important, en fonction du travail d'objectivation que le social (les institutions, l'économie et le politique) a réalisé. Les catégories les plus instituées (comme l'état matrimonial, la profession, le diplôme...) présentent moins d'équivoque que les catégories construites et utilisées par le seul sociologue pour les besoins de son enquête[1].

1. François Héran, « L'assise statistique de la sociologie », *Économie & statistique*, 1984, n° 169, p. 23-35.

Une réflexion similaire pourrait être conduite en ce qui concerne l'idée ou la notion de « biais d'échantillonnage ». L'idée selon laquelle il existe un échantillon parfait duquel il faudrait s'approcher au plus près doit être rejetée. Il peut exister des erreurs dans le choix et la sélection des enquêtés. Mais il ne faut pas voir ces erreurs ou biais comme des écarts à un idéal, à une réalité. Il est préférable de les penser comme heuristiques, d'en tirer profit. Il faut apprendre à faire avec « ce qu'on a » et à bien prendre conscience de « ce qu'on a ». Les erreurs apparentes ou les biais supposés d'une enquête peuvent être des outils heuristiques et nous renseigner comme le feraient d'hypothétiques données « idéales »[1].

À ces « erreurs de mesure » s'ajoutent les incertitudes liées à l'échantillonnage. Mais, comme nous allons le voir maintenant, celles-ci peuvent être estimées.

5. Que valent les informations issues d'un échantillon ?

Tout au long de la conception puis de la réalisation de son enquête, le sociologue doit s'interroger sur le statut des résultats qu'il compte obtenir. Deux cas se présentent. Soit il souhaite établir des résultats et commentaires qui n'auront de sens que pour son échantillon et qui n'ont pas vocation à le renseigner sur l'ensemble d'une population. Soit il espère que les résultats obtenus sur son échantillon valent aussi pour l'ensemble de la population de référence. Il sait que ses résultats sont entachés d'une marge d'« erreur » ou d'« incertitude » puisqu'il se contente d'interroger une petite partie de la population : s'il enquête deux échantillons, il est probable que les résultats obtenus seront différents d'un échantillon à l'autre. Il sent bien, toutefois, que les résultats obtenus sur ces deux échantillons devraient être proches et que s'il multipliait les échantillons les résultats tendraient à converger vers des valeurs proches, voire très proches.

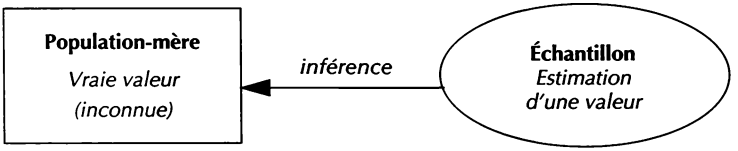
5.1 Passer de l'échantillon à la population

La statistique permet de préciser et de formaliser ce que le sociologue « sent ». Elle lui offre une série d'outils appelés « *tests statistiques* » fournissant les

1. Michel Gollac, « Des chiffres insensés ? Pourquoi et comment on donne un sens aux données statistiques », *Revue française de sociologie*, vol. 38, 1997, p. 5-36.

éléments de réponses à son problème : dans quelle mesure les résultats établis sur l'échantillon sont-ils valables pour décrire l'ensemble de la population ? Les tests statistiques permettent de saisir les effets des fluctuations d'échantillonnage.

Ce problème est une forme particulière du problème général de l'induction, c'est-à-dire du raisonnement consistant à remonter de données particulières (faits, expériences, énoncés) à des propositions plus générales. Les statisticiens utilisent le terme « inférence » pour désigner ce passage du particulier (l'échantillon) vers le général (la population). Pour indiquer qu'un résultat est établi sur un échantillon, il est courant de parler d'*estimation*. Le sociologue *estime* des informations sur sa population à partir de connaissances établies sur un seul échantillon.



5.2 Quelle est la « valeur » de l'estimation ?

Pour comprendre le raisonnement conduit par les statisticiens et probabilistes, et qu'utilisent les sociologues, considérons un cas simple. Supposons qu'au sein d'une population (de grande taille), 65 % des individus aient regardé la télévision au moins 10 minutes la veille de l'enquête. Et supposons que nous interrogions seulement un échantillon de 10 personnes par tirage aléatoire à partir cette population[1]. Nous demandons à chacune de ces 10 personnes si elle a, ou non, regardé la télévision au moins 10 minutes la veille.

Il est possible d'envisager quels sont les différents types d'échantillon qu'il est possible d'obtenir à partir d'une telle population : la première colonne du tableau 1.1 présente ces différents types d'échantillon. Notre échantillon correspond nécessairement à une des situations indiquées dans ce tableau, même si nous ne savons évidemment pas lequel *a priori*.

1. Il existe deux manières de concevoir un tirage au sort au sein d'une population : avec ou sans remise, c'est-à-dire en s'autorisant à tirer au sort et donc enquêter ou non deux fois le même individu. Le raisonnement est identique dans les deux cas, mais les calculs sont différents. Toutefois, les populations étant généralement de très grande taille en sociologie, les différences entre les calculs « avec remise » et les calculs « sans remise » sont négligeables.

Pour chacun de ces types d'échantillon, il est possible de calculer la part[1] des personnes ayant regardé la télévision : cette part est l'*estimation* de la part des individus ayant regardé la TV parmi toute la population (*cf.* deuxième colonne de notre tableau). Il est également possible de calculer la part des échantillons de ce type parmi tous les échantillons possibles c'est-à-dire la probabilité d'obtenir un échantillon de ce type (« d'être tombé sur un échantillon de ce type »). Considérons, par exemple, le type d'échantillon pour lequel tous les individus ont répondu « Oui » à notre question (première ligne du tableau 1.1). Pour obtenir un tel échantillon, il faut choisir, parmi les membres de la population, 10 fois un individu ayant regardé la TV. Puisqu'il y a 65 % de personnes ayant regardé la TV dans notre population, à chaque fois que nous choisissons un individu nous avons 65 chances sur 100 d'obtenir quelqu'un ayant regardé la TV (soit 0,65). La probabilité d'avoir 10 fois de suite cette chance est donc :

$$0{,}65 \times 0{,}65 \times 0{,}65 \times 0{,}65 \times 0{,}65 \times 0{,}65 \times 0{,}65 \times 0{,}65 \times 0{,}65 \times 0{,}65$$
$$= (0{,}65)^{10} = 0{,}013462743\ldots \approx 0{,}0135$$

À l'opposé, on peut calculer la probabilité d'obtenir un échantillon où personne n'a regardé la TV la veille. Il faut, lors de la désignation de chacun des 10 individus composant notre échantillon, que le hasard désigne une des personnes n'ayant pas regardé la TV : puisque 35 % de la population n'a pas regardé la TV, la probabilité d'obtenir une telle personne est de 0,35. La probabilité d'obtenir un échantillon de 10 personnes de ce type est donc :

$$0{,}35 \times 0{,}35 \times 0{,}35 \times 0{,}35 \times 0{,}35 \times 0{,}35 \times 0{,}35 \times 0{,}35 \times 0{,}35 \times 0{,}35$$
$$= (0{,}35)^{10} = 0{,}00002758547 \approx 0{,}000028$$

La probabilité d'obtenir 10 réponses « Non » est donc beaucoup plus faible que celle d'obtenir 10 réponses « Oui ». Cela résulte du fait qu'il y a beaucoup plus de personnes ayant regardé la TV que ne l'ayant pas fait dans notre population (65 % contre 35 %).

Pour chacun des types d'échantillon, il est possible de calculer cette probabilité – même si le calcul est un peu plus complexe car il nécessite le recours à des calculs de dénombrement.

1. Il existe deux manières courantes d'exprimer une « part » : soit sous la forme d'un pourcentage ; soit sous la forme d'une fraction (un nombre compris entre 0 et 1). Ainsi, 65 % s'expriment aussi 0,65 : « 65 personnes parmi 100 » ou « 0,65 personne pour une ». Cette remarque vaut également pour les probabilités : une probabilité de 0,33 peut s'exprimer comme une probabilité (ou chance) de 33 %.

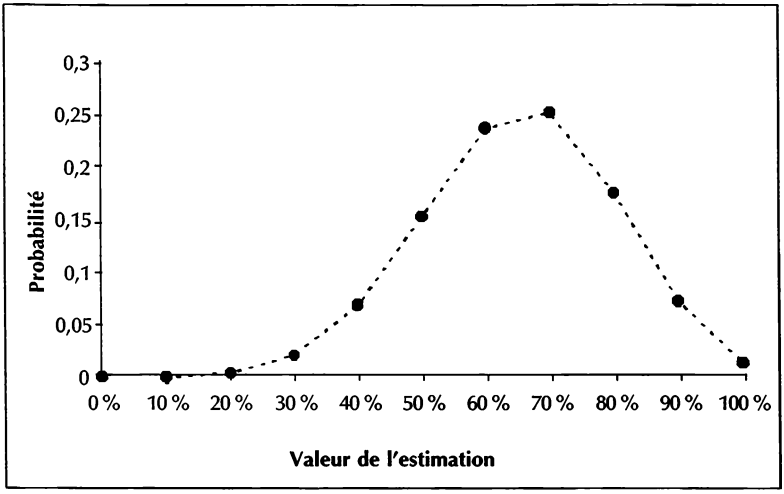
Tableau 1.1. Les types d'échantillon

Type d'échantillon	Part des personnes ayant regardé la TV pour ce type d'échantillon	Probabilité d'obtenir un échantillon de ce type
10 individus répondent « Oui » Personne ne répond « Non »	100 %	≈ 0,0135
9 « Oui » et 1 « Non »	90 %	≈ 0,0725
8 « Oui » et 2 « Non »	80 %	≈ 0,1756
7 « Oui » et 3 « Non »	70 %	≈ 0,2522
6 « Oui » et 4 « Non »	60 %	≈ 0,2377
5 « Oui » et 5 « Non »	50 %	≈ 0,1536
4 « Oui » et 6 « Non »	40 %	≈ 0,0689
3 « Oui » et 7 « Non »	30 %	≈ 0,0212
2 « Oui » et 8 « Non »	20 %	≈ 0,0043
1 « Oui » et 9 « Non »	10 %	≈ 0,00051
Personne ne répond « Oui » 10 individus répondent « Non »	0 %	≈ 0,000028

Considérons les deux dernières colonnes de ce tableau 1.1 : elles fournissent la probabilité associée à chaque estimation. Ainsi, nous avons par exemple 15,36 % de chance (0,1536) d'obtenir un échantillon estimant à 50 % la part des personnes ayant regardé la TV la veille. Il est possible de représenter graphiquement ces informations (graphique 1.1) et ainsi de prendre conscience que les types d'échantillon les plus probables sont ceux fournissant des estimations proches de la vraie valeur 65 %.

L'ensemble de ce raisonnement peut évidemment être généralisé et étendu à des situations où la taille de l'échantillon est plus élevée : 100, 200, 500, 1 000 personnes voire davantage. Le raisonnement n'est pas plus complexe : seuls les calculs sont un peu plus fastidieux. En suivant ce raisonnement, il est possible de représenter, pour chaque taille d'échantillon (par exemple 200, 500 et 1 000 personnes ici), les probabilités d'obtenir, par tirage au sort, un échantillon fournissant telle ou telle estimation de la part des individus ayant regardé la TV. Ces graphiques sont théoriques (ils sont construits par calcul) mais ils correspondent exactement à ce qu'un expérimentateur obtiendrait s'il avait la patience de construire puis d'interroger tous les échantillons possibles tirés au hasard à partir d'une population !

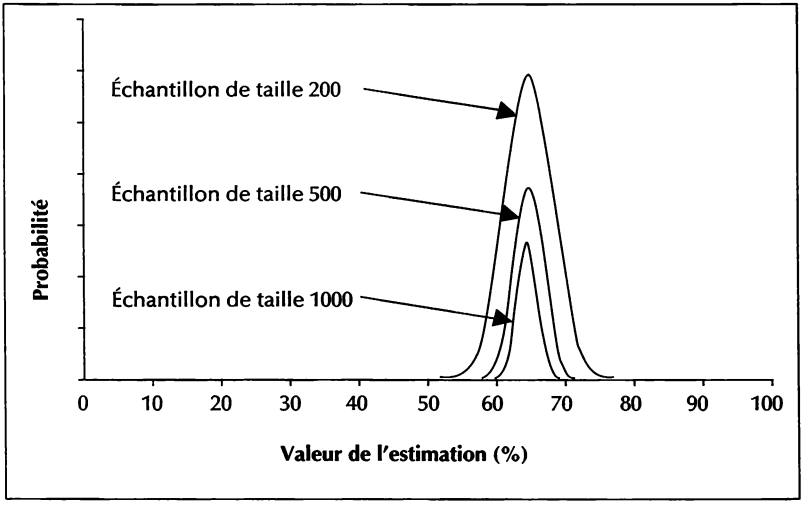
Graphique 1.1. Probabilité des diverses estimations pour un échantillon de taille 10



Qu'apprenons-nous grâce au graphique 1.2 ? Premièrement, la probabilité d'obtenir des échantillons fournissant des estimations éloignées de la vraie valeur (65 %) est faible : par exemple, nous avons seulement une chance sur 4×10^{45} d'obtenir une estimation de 0 % à partir d'un échantillon de 100 personnes[1]. Deuxièmement, cette probabilité est d'autant plus faible que la taille de l'échantillon est grande : ainsi la probabilité précédente passe à une chance sur 10^{226} pour un échantillon de 500 personnes. Troisièmement, cette probabilité est d'autant plus faible que l'estimation s'éloigne de la vraie valeur : à partir d'un échantillon de 100 personnes, la probabilité d'obtenir une estimation de 60 % est d'environ 0,05, la probabilité d'obtenir une estimation de 50 % est de 0,0007 et celle associée à l'estimation 40 % s'élève seulement à 10^{-7} environ (soit 0,0000007). Seul le quatrième et dernier constat peut être vu comme pessimiste : la probabilité que l'estimation soit

1. Ces probabilités sont très, très faibles. À titre de comparaison, un joueur de Loto a environ une chance sur $1,4 \times 10^9$ de trouver les six bons numéros (soit une chance sur 14 millions). Rappelons que la notation 10^n permet de représenter commodément des nombres comportant beaucoup de zéros. Ainsi 10^3 peut s'écrire 1 000 et 10^{10} s'écrit 10 000 000 000.

Graphique 1.2. Probabilités des diverses estimations (pour des échantillons de 200, 500 et 1000 individus)



exactement égale à la vraie valeur n'est pas très élevée et tend même à diminuer lorsque la taille de l'échantillon croît : pour un échantillon de 100 personnes, la probabilité d'avoir un échantillon fournissant une estimation d'exactement 65 % est de 0,084 ; elle vaut seulement 0,026 pour un échantillon de 1 000 personnes.

Globalement, ces constats conduisent à affirmer qu'il est très peu probable que la part des individus ayant regardé la TV la veille au sein de notre échantillon soit très différente de 65 % – même si la probabilité de « tomber » sur la vraie valeur est finalement assez faible. En d'autres termes, « nous avons très peu de chances (voire presque aucune chance) d'obtenir une estimation éloignée (voire très éloignée) de la vraie valeur » ou, inversement, « il y a de très grandes chances que l'estimation fournie par l'échantillon soit proche de la vraie valeur ».

Non seulement « il y a de très grandes chances que l'estimation fournie par l'échantillon soit proche de la vraie valeur » mais on peut préciser la valeur de cette « très grande chance » et cette notion de « proximité avec la vraie valeur ». Nous pouvons calculer la probabilité d'obtenir un échantillon de taille 100 fournissant une estimation comprise entre, par exemple, 60 % et 70 % : il

suffit de calculer la somme des probabilités d'obtenir une estimation de 60 %, de 61 %, de 62 %... jusqu'à 70 %. En l'occurrence, cette somme vaut 75 %. Pour un échantillon de 1 000 personnes, cette probabilité vaut plus de 99,9 %. Autrement dit, il y a 99,9 % de chances que l'estimation obtenue sur l'échantillon de taille 1 000 soit comprise entre 60 % et 70 %. C'est donc quasi certain !

Avant de tirer des conclusions pratiques de l'ensemble de ce raisonnement, résumons-le. Nous avons supposé connaître la vraie valeur d'un pourcentage au sein d'une population (en l'occurrence la part des individus ayant regardé la TV la veille parmi l'ensemble des individus de la population). Puis nous avons calculé l'estimation de ce pourcentage sur chacun des échantillons possibles – plus exactement sur chacun des types d'échantillon possible. Nous avons enfin montré que cette estimation a de grandes chances, voire de très grandes chances d'être proche de la vraie valeur. Ces « grandes chances » et cette « proximité » augmentent l'une comme l'autre au fur et à mesure que la taille de l'échantillon croît. Plus précisément, cette estimation se situe dans un intervalle de valeurs autour de la vraie valeur avec une probabilité calculable. Cette probabilité correspond à la part des échantillons pour lesquels l'estimation est dans l'intervalle.

Un peu plus formellement, nous pouvons écrire que l'affirmation *(A)* :

vraie valeur – marge < estimation < vraie valeur + marge (A)

est vraie avec une probabilité p, c'est-à-dire pour une part p des échantillons.

En somme, notre raisonnement a permis de saisir l'effet de l'aléa, du hasard, dans le processus d'estimation.

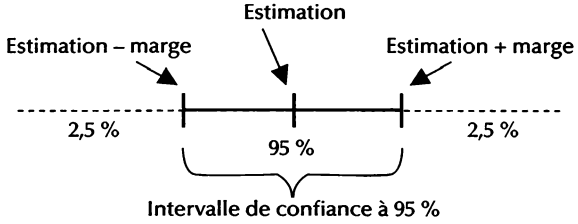
5.3 L'intervalle et le niveau de confiance

En apparence, un seul problème subsiste : dans tout ce qui précède, nous avons supposé connaître la vraie valeur. Ce n'est évidemment pas le cas en pratique. Il suffit toutefois de constater que si nous sommes en mesure d'apprécier l'écart séparant notre estimation de la vraie valeur, nous sommes évidemment capables d'apprécier l'écart entre la vraie valeur et notre estimation : pour obtenir un encadrement de la vraie valeur, il suffit de réécrire l'affirmation *(A)* précédente de la façon strictement équivalente suivante[1] :

estimation – marge < vraie valeur < estimation + marge (B)

1. Cette affirmation est obtenue par une transformation exacte : si $a - b < c < a + b$ alors $a < c + b$ et $c - b < a$ donc $c - b < a < c + b$.

Comme précédemment, cette affirmation est vraie avec une probabilité p, c'est-à-dire pour une part p des échantillons. Cette écriture peut être transposée en une représentation graphique simple :



Selon la terminologie utilisée par les statisticiens et probabilistes et largement adoptée par les sociologues, cet encadrement est un intervalle appelé *intervalle de confiance*. Et la probabilité est appelée le *niveau de confiance* de cet intervalle. Cette probabilité p est souvent exprimée en pourcent. Son complément à 100 % est le niveau (ou seuil) de risque, qui exprime le risque de se tromper en affirmant que la vraie valeur est située dans l'intervalle. Ainsi, si le niveau de confiance est de 95 %, le seuil de risque s'élève à 5 % : il y a 2,5 % de chance que la vraie valeur se situe en dessous de l'intervalle et 2,5 % au-dessus.

Les raisonnements et calculs précédents montrent qu'il est possible de déterminer la probabilité (niveau de confiance) associée à un intervalle de confiance. Inversement, il est possible de déterminer l'intervalle associé à un niveau de confiance. Il n'est pas du ressort de ce manuel de fournir la démonstration et l'expression mathématique de ces affirmations[1].

Avant d'en venir aux aspects pratiques de l'utilisation de l'intervalle de confiance, ajoutons une remarque. Pour une taille d'échantillon donnée, l'intervalle de confiance sera d'autant plus grand que le niveau de confiance est élevé et, inversement, plus le niveau de risque sera grand, plus l'intervalle de confiance sera étroit. Tout utilisateur de l'intervalle de confiance est donc face à un choix cornélien : soit accepter un risque élevé, mais avoir un encadrement précis ; soit refuser tout risque élevé, mais disposer d'un encadrement grossier de la vraie valeur. À la limite, le sociologue souhaitant ne prendre aucun risque

1. Nous renvoyons le lecteur aux ouvrages de Philippe Tassi ou de Gilbert Saporta indiqués en bibliographie finale.

(niveau de confiance égale à 100 %) ne pourra fournir comme intervalle de confiance de ses pourcentages que l'intervalle [0 % ; 100 %] !

5.4 L'usage pratique

De manière pratique, les choses se passent de la façon suivante : le sociologue réalise une enquête par sondage auprès d'un échantillon de *n* personnes. Sur cet échantillon, il calcule la part des individus qui ont telle ou telle caractéristique. Afin de fixer les idées, quittons notre exemple précédent et supposons que le sociologue étudie les situations et parcours professionnels et qu'il calcule la part des enquêtés ayant connu une période de chômage d'au moins 6 mois au cours de leur vie. La part qu'il calcule est une estimation de la part des individus ayant chômé au moins six mois dans la population générale. Le sociologue sait qu'il ne peut pas déterminer quelle est la vraie valeur, mais il sait qu'à chaque intervalle correspond une probabilité que la vraie valeur soit contenue dans cet intervalle. Précisons l'alternative qui s'offre à lui.

Premier choix : le sociologue souhaite connaître quelle est la probabilité pour que la vraie valeur soit contenue dans un intervalle autour de son estimation. Il lui incombe de choisir l'ampleur, la taille, de cet intervalle. Plus l'intervalle qu'il se fixe est ample, plus la probabilité (le niveau de confiance) sera élevée mais plus l'idée qu'il aura de la vraie valeur sera floue, approximative (l'estimation est « grossière »). Supposons qu'au sein de son échantillon 20 % des enquêtés aient chômé au moins 6 mois. Le tableau 1.2 fournit les résultats du calcul de l'intervalle de confiance pour diverses tailles d'échantillon et pour divers intervalles de confiance. Par exemple, pour un échantillon de 500 personnes, la probabilité que la vraie valeur se situe autour de 20 % avec une tolérance de plus ou moins 2 % s'élève à 0,736. En d'autres termes, le sociologue a 73,6 % de chance de ne pas se tromper en affirmant que la vraie valeur est comprise entre 18 % et 22 %.

Si nous suivons tout ce qui précède, cette probabilité peut s'interpréter comme la part des échantillons fournissant un intervalle de confiance contenant la vraie valeur. Cette probabilité ne vaut jamais 1 : il y a donc toujours un risque pour que le sociologue se trompe. Ce risque est d'autant plus faible que le niveau de confiance est élevé : le risque est par exemple de 5 % pour un niveau de confiance de 95 %. Le sociologue ne peut pas se débarrasser de ce risque : il peut simplement le réduire en augmentant la taille de ses échantillons ou en utilisant des encadrements (intervalles de confiance) plus grands.

Tableau 1.2. Niveau de confiance pour divers échantillons et intervalles
(pourcentage estimé = 20 %)

Taille de l'échantillon (n)	Intervalle de confiance [a, b]	Niveau de confiance *p*
100	15 % < vraie valeur < 25 %	0,789
	18 % < vraie valeur < 22 %	0,383
500	15 % < vraie valeur < 25 %	0,995
	18 % < vraie valeur < 22 %	0,736
1000	15 % < vraie valeur < 25 %	≈1
	18 % < vraie valeur < 22 %	0,886
	19 % < vraie valeur < 21 %	0,571
2000	18 % < vraie valeur < 22 %	0,975
	19 % < vraie valeur < 21 %	0,736
5000	18 % < vraie valeur < 22 %	≈1
	19 % < vraie valeur < 21 %	0,923

Lecture : pour un échantillon de taille n, la probabilité que la vraie valeur soit comprise entre a et b est de *p*.

En sociologie, l'usage est de considérer qu'un encadrement est satisfaisant si le niveau de confiance qui lui est associé dépasse 90 % ou 95 %. En dessous de ces seuils, l'encadrement risque d'être trop incertain. Et au-dessus, l'encadrement sera plus certain mais peut-être trop ample (« grossier »).

Le second choix possible consiste à déterminer quel est l'intervalle de confiance correspondant à un niveau de confiance fixé. Le sociologue fixe un niveau de confiance, par exemple 90 %, 95 % ou encore 98 %. Il définit ainsi un seuil en dessous duquel il considéra les résultats comme non fiables, comme peu dignes de confiance ou d'intérêt. S'il veut être presque sûr de son encadrement, il choisira un niveau de confiance élevé mais obtiendra alors un intervalle de grande taille.

À partir de ce niveau ou seuil de confiance, il peut calculer les intervalles de confiance associés à tous les pourcentages estimés. Le tableau 1.3 fournit quelques exemples d'intervalle de confiance correspondant à trois exemples de niveau de confiance, pour des échantillons de taille 100, 500, 1 000 ou 2 000 personnes.

Tableau 1.3. Intervalles de confiance selon la taille de l'échantillon et le niveau de confiance (pourcentage estimé = 20 %)

Taille de l'échantillon (n)	Niveau de confiance *p*	Intervalle de confiance [a, b]
100	90 %	13,4 % < vraie valeur < 26,6 %
	95 %	12,2 % < vraie valeur < 27,8 %
500	90 %	17,1 % < vraie valeur < 22,9 %
	95 %	16,5 % < vraie valeur < 23,5 %
1000	90 %	17,9 % < vraie valeur < 22,1 %
	95 %	17,5 % < vraie valeur < 22,5 %
	98 %	17,1 % < vraie valeur < 22,9 %
2000	95 %	18,2 % < vraie valeur < 21,8 %
	98 %	17,9 % < vraie valeur < 22,1 %
5000	98 %	18,7 % < vraie valeur < 21,3 %
	99 %	18,5 % < vraie valeur < 21,5 %

Lecture : pour un échantillon de taille n et un niveau de confiance *p*, l'intervalle de confiance est [a, b].

L'intervalle est d'autant plus étroit, et donc la précision avec laquelle nous estimons les pourcentages est d'autant plus grande, que le niveau de confiance est faible. Le constat est similaire lorsque la taille de l'échantillon croît.

5.5 Les outils de calcul de l'intervalle de confiance

Le niveau de confiance est très souvent fixé à 95 % (le seuil de risque étant alors de 5 %). Dans ce cas, l'intervalle de confiance de niveau de confiance 95 % d'un pourcentage *p* (exprimé sous forme de fraction, c'est-à-dire par un nombre compris entre 0 et 1) pour un échantillon de taille *n* est fourni par la formule $IC_1^{95\,\%}$ suivante :

$$p - 1{,}96 \times \sqrt{\frac{p \times (1-p)}{n}} \leq \textit{vraie valeur} \leq p + 1{,}96 \times \sqrt{\frac{p \times (1-p)}{n}}$$

À titre d'exemple d'application de cette formule, calculons l'intervalle de confiance de 20 % (0,20) pour un échantillon de taille 1 000 :

$$0{,}20 - 1{,}96 \times \sqrt{\frac{0{,}2 \times (1 - 0{,}2)}{1\,000}} \leq \textit{vraie valeur} \leq 0{,}2 + 1{,}96 \times \sqrt{\frac{0{,}2 \times (1 - 0{,}2)}{1\,000}}$$

$$\text{soit : } 0{,}17520 \leq \textit{vraie valeur} \leq 0{,}22479$$

$$\text{soit : } 17{,}52\ \% \leq \textit{vraie valeur} \leq 22{,}5\ \%$$

La vraie valeur est donc très probablement comprise entre 17,5 % et 22,5 %. Il n'y a que 5 % de risque que cette affirmation soit fausse (si le hasard a voulu que l'échantillon soit un des 5 % d'échantillons fournissant une image relativement déformée de la population).

Sous réserve d'une petite approximation couramment admise, cette formule peut être simplifiée et devient (toujours pour un niveau de confiance de 95 %) la formule $IC_2^{95\,\%}$ suivante :

$$p - \frac{1}{\sqrt{n}} \leq \textit{vraie valeur} \leq p + \frac{1}{\sqrt{n}}$$

Les deux formules suivantes indiquent les intervalles de confiance pour les niveaux de confiance 90 % et 98 %, qui sont, après le niveau de confiance 95 %, les seuils les plus couramment utilisés en sociologie :

$$\text{A 90 \% : } p - 1{,}64 \times \sqrt{\frac{p \times (1 - p)}{n}} \leq \textit{vraie valeur} \leq p + 1{,}64 \times \sqrt{\frac{p \times (1 - p)}{n}}$$

$$\text{A 98 \% : } p - 2{,}34 \times \sqrt{\frac{p \times (1 - p)}{n}} \leq \textit{vraie valeur} \leq p + 2{,}34 \times \sqrt{\frac{p \times (1 - p)}{n}}$$

Une propriété apparemment surprenante mérite d'être soulignée : l'intervalle de confiance ne dépend que du pourcentage estimé et de la taille de l'échantillon : la taille de la population de référence n'entre pas en ligne de compte. Même si cela peut paraître contre-intuitif, cette propriété résulte du fait que les individus sont choisis au hasard : la précision d'une estimation pour la population française ou pour la population d'une commune ne dépend que de la taille de l'échantillon[1].

Signalons enfin que cette notion d'intervalle de confiance peut être utilisée pour beaucoup d'autres indicateurs statistiques, et pas seulement dans le cas d'un pourcentage comme nous venons de le faire[2]. Et même si les sociologues

1. Sauf si la taille de l'échantillon est du même ordre de grandeur que la taille de la population, mais ce cas est rarissime en sociologie.
2. Voir, par exemple, les calculs d'intervalle de confiance des probabilités d'accès aux grandes écoles en fonction des origines sociales dans : Valérie Albouy et Thomas Wanecq, « Les inégalités sociales d'accès aux grandes écoles », *Économie et Statistique*, 2003, n° 361, p. 27-52.

ne calculent pas systématiquement les intervalles associés à toutes les grandeurs qu'ils calculent sur leurs échantillons, il est indispensable que chacun ait à l'esprit l'incertitude qui entoure tout calcul effectué sur un échantillon.

5.6 Quelques généralités supplémentaires sur les tests statistiques

Une enquête auprès d'un échantillon ne permet pas d'obtenir une image exacte et sans erreur de la réalité. Mais nous avons montré qu'un échantillon permet d'obtenir, de manière d'autant plus vraisemblable que l'échantillon est de grande taille, un encadrement de la vraie valeur. Cet encadrement est l'intervalle de confiance. Le caractère « vraisemblable » de cet encadrement est mesuré par une probabilité appelée « niveau de confiance ». Ainsi, sous réserve d'accepter un peu d'incertitude (mesurable), il est possible de travailler sur des échantillons. Ce constat est essentiel pour toute activité scientifique !

Le cheminement du raisonnement est donc le suivant : 1) le sociologue estime la valeur d'un pourcentage sur un échantillon ; 2) il envisage un ou plusieurs encadrements de ce pourcentage, qui constituent donc des hypothèses sur la gamme des valeurs possibles de la vraie valeur du pourcentage ; 3) il calcule un niveau de confiance pour chacun de ces encadrements qui sont, en fait, des mesures de la plausibilité ou de la crédibilité de chacune de ses hypothèses ; 4) il arbitre pour choisir un encadrement qui soit à la fois suffisamment précis et suffisamment plausible/crédible.

Ce raisonnement est un exemple d'un raisonnement plus général que les statisticiens et probabilistes appellent « test statistique ». Un « test statistique » consiste à faire une l'hypothèse[1] et à la supposer vraie tant qu'on n'a pas une forte présomption du contraire. On suppose alors que les résultats empiriques sont dus au hasard puis on calcule la probabilité que l'hypothèse soit vérifiée. L'hypothèse sera rejetée si son improbabilité est démontrée, c'est-à-dire si l'hypothèse formulée est fortement improbable à la vue des résultats obtenus empiriquement.

Un test consiste donc : 1) à calculer une caractéristique sur un échantillon (un pourcentage, une moyenne, une corrélation...) ; 2) à faire une hypothèse sur les propriétés de la population ; 3) à examiner le caractère plausible ou

1. Cette hypothèse est appelée « hypothèse zéro » ou « nulle » par les statisticiens. Cette terminologie ne doit pas surprendre le sociologue : il ne s'agit pas d'un jugement de valeur, mais simplement d'un usage signifiant « hypothèse de départ » ou « hypothèse initiale ».

non (en fait le niveau de plausibilité) de cette hypothèse au regard de la caractéristique calculée sur l'échantillon ; 4) à conclure sur le caractère plausible ou non, acceptable ou non, des hypothèses.

D'un point de vue général, un test statistique permet de saisir l'effet de l'aléa, du hasard, dans le processus d'estimation sur un échantillon. Il existe une gamme très étendue de tests statistiques, utilisés aussi bien en sciences de la vie et de la nature qu'en sciences humaines et sociales[1]. Il est par exemple possible de concevoir un test pour savoir si un jeu de hasard est biaisé. Considérons, par exemple, le jeu de pièce « pile ou face ». Imaginons avoir lancé une pièce 100 fois et que le côté « pile » soit tombé 62 fois. Ces 100 lancers constituent notre échantillon. Faisons ensuite l'hypothèse que la pièce n'est pas biaisée, c'est-à-dire qu'elle a la même probabilité de tomber du côté pile (0,5 soit 50 %) et du côté face (idem). Cette hypothèse est-elle plausible au regard de notre échantillon de lancers ? Quel est son niveau de plausibilité ? Sans faire de calculs, notre intuition tend à nous faire penser que tomber 62 fois sur « pile » alors que la pièce est censée tomber seulement environ 50 fois sur « pile » est « bizarre ». Le test statistique vise simplement à formaliser cette intuition et ce sentiment de « bizarrerie ». En l'occurrence le niveau de crédibilité de l'hypothèse de non-biais est très faible (moins de 1 %).

Nous présenterons dans la suite de ce manuel d'autres tests particulièrement utiles en sociologie (notamment le test du khi-deux).

5.7 Usages et limites des tests statistiques en sociologie

S'il est bon de faire confiance aux tests, il ne saurait être question de leur accorder un pouvoir qu'ils n'ont pas : les tests statistiques ne fournissent jamais une certitude ; ils expriment simplement des présomptions. Il faut les concevoir comme des indices, comme des outils aidant le sociologue dans son travail d'interprétation.

D'un côté, parce qu'ils peuvent se tromper et aboutir à une conclusion inverse de la réalité : la conclusion d'un test statistique n'est pas « vrai ou faux », « oui ou non », mais une probabilité mesurant le caractère plausible ou non d'une hypothèse. Ainsi, par exemple, si [35 % ; 45 %] est l'intervalle de confiance d'un pourcentage à un seuil de 95 %, il reste 5 % de chance que

1. La mise au point et la justification théorique de ces tests sont l'affaire d'une branche de la statistique appelée « statistique inférentielle », par opposition à la « statistique descriptive » qui cherche les meilleurs outils pour décrire de grands ensembles de données.

le « vrai » pourcentage se situe en dehors de cet intervalle... Les tests ne sont pas des outils divinatoires et ne constituent pas des preuves irréfutables.

D'un autre, parce qu'ils reposent sur des hypothèses dont nous ne pouvons pas toujours vérifier l'adéquation avec la réalité empirique. L'ensemble de notre raisonnement repose notamment sur l'hypothèse que l'échantillon est effectivement construit par hasard : notre calcul de l'intervalle de confiance n'est possible que si les individus de l'échantillon sont choisis de manière probabiliste. Plus généralement, tous les tests statistiques supposent que les échantillons sont probabilistes. De manière apparemment paradoxale, il est possible d'estimer les effets du hasard à condition qu'on le laisse faire.

Or, comme nous l'avons déjà écrit (chap. 1, § 4), les échantillons utilisés en sociologie sont rarement aléatoires : ils résultent de tirages « empiriques », notamment de la méthode des quotas. Si le choix des individus composant l'échantillon n'est pas le fruit du hasard mais de l'action du sociologue, s'il est le produit de sa volonté ou de sa subjectivité, il est impossible d'affirmer que l'échantillon va fournir une image imparfaite mais proche de la population : un échantillon non aléatoire peut offrir une image très décalée, fortement biaisée, de la population.

Dès lors, quelle est la validité des tests statistiques ? En toute rigueur, les tests statistiques ne peuvent pas être utilisés. Il est néanmoins possible d'y avoir recours en considérant que les tests fournissent les seuils minimaux d'incertitude. Par exemple, lorsqu'un échantillon de taille 1 000 est parfaitement aléatoire, l'incertitude qui entoure tout pourcentage est d'environ ± 3 % (selon la formule $IC_2^{95\,\%}$ page 41). Il s'agit donc d'une erreur « mécanique » presque incompressible, à laquelle tout scientifique doit s'attendre lorsqu'il travaille sur un échantillon de taille 1 000. Lorsque l'échantillon ne respecte pas les critères probabilistes, il est vraisemblable que l'erreur est au moins aussi élevée : il est alors vain d'être trop précis dans la publication des résultats. En d'autres termes, le calcul de l'intervalle de confiance fournit un ordre de grandeur permettant d'apprécier le niveau des erreurs attendues.

Un autre argument justifie l'utilisation des tests pour les échantillons empiriques : l'expérience accumulée par les sociologues et statisticiens depuis plusieurs décennies tend à suggérer que les résultats obtenus grâce aux échantillons par quotas sont d'une précision égale, sinon comparable, aux résultats établis grâce aux échantillons probabilistes. La répétition des enquêtes par quotas et leur comparaison avec les enquêtes probabilistes semblent assurer que les estimations ne sont pas davantage dispersées ou instables. Il ne s'agit pas d'une justification théorique mais d'un constat empirique qui conforte les sociologues dans leurs usages des tests statistiques.

Enfin, un dernier argument plaide en faveur des tests, même dans le cas d'échantillons non aléatoire ou dont la représentativité est assurée par des quotas : considérer, comme nous l'avons exposé précédemment (§1.4), que l'échantillon est représentatif d'une population dont les contours sont donnés par les résultats obtenus sur l'échantillon. Celui-ci est alors conçu comme un prisme laissant entrevoir une population inconnue *a priori* mais éclairée par ce prisme *a posteriori*.

2 |

CONCEVOIR ET PRÉPARER LES VARIABLES NÉCESSAIRES À L'ANALYSE

1. QUESTIONS, VARIABLES ET MODALITÉS

Il est indispensable de distinguer deux niveaux d'information, même s'ils se recoupent largement. Le premier niveau est celui des réponses fournies par les enquêtés aux questions ou les codages de matériaux qualitatifs. Il s'agit d'informations primaires, liées à la logique et aux exigences spécifiques de l'enquête ou du codage : les informations sont réduites, limitées à un aspect précis mais ponctuel. Il s'agit par exemple des titres des livres lus durant le dernier mois, du nombre d'enfants, de la situation professionnelle, de la destination des dernières vacances, de la durée de la dernière période de chômage...

Ce premier niveau d'informations regroupe les variables qualifiées de « primaires » (au sens de « premières ») et qui sont évidemment indispensables : elles constituent le matériau empirique. Pourtant, ces variables ne sont pas suffisantes ou satisfaisantes dans bien des cas : elles répondent à des exigences empiriques et aux impératifs méthodologiques de réalisation de l'enquête ; elles ne renseignent que partiellement le sociologue sur ce qui l'intéresse en priorité.

Le second niveau est celui des variables « dérivées » (ou « secondaires ») élaborées pour mieux correspondre aux exigences techniques du traitement statistique ainsi qu'aux exigences théoriques de la problématique. Ces variables sont dérivées au sens où elles résultent des variables primaires par recodage ou agrégation de plusieurs informations primaires. Ce sont les « vraies » variables sociologiques, celles directement liées à la problématique ou au questionnement théorique du sociologue.

2. VARIABLES QUALITATIVES ET VARIABLES QUANTITATIVES

Deux grands types de variables peuvent être distingués : les variables quantitatives, qui expriment des grandeurs quantifiables, et les variables qualitatives, qui reflètent des grandeurs non quantitatives, des « qualités ». En sociologie les secondes sont plus fréquentes que les premières : l'essentiel des informations est de nature qualitative. Ceci résulte de la nature des phénomènes analysés par le sociologue : les pratiques, les opinions, les représentations, les caractéristiques sociales, ou encore les attitudes s'expriment rarement à l'aide de variables quantitatives. Et il n'est pas rare que les quelques variables quantitatives soient recodées en variables qualitatives afin d'harmoniser le statut des variables et d'écarter l'illusion de précision que peuvent incarner les variables quantitatives.

La distinction entre variables quantitatives et qualitatives n'est pas anodine. Elle ne résulte pas d'un raffinement conceptuel inutile mais d'une contrainte technique forte : la nature des variables conditionne le type de méthodes d'analyse utilisables. Il est par exemple impossible de calculer un statut matrimonial moyen ou un diplôme moyen.

2.1 Variables quantitatives

Une variable quantitative permet d'exprimer une grandeur quantifiable c'est-à-dire une grandeur mesurable à l'aide d'une unité. C'est par exemple le cas de l'âge (exprimables en « années » ou en « mois »), du revenu (en euros ou en Yens) ou encore du nombre d'enfants. Une variable quantitative s'exprime à l'aide de nombres et ses diverses valeurs peuvent être numériquement comparées.

De manière générale, les sociologues utilisent des variables quantitatives dans deux grands types de situations. Premièrement, lorsqu'ils veulent exprimer des durées (âge, ancienneté d'une pratique, durée d'une expérience professionnelle, temps consacré à une activité, nombre d'années d'études, durée entre deux événements...), des valeurs monétaires (revenus, patrimoine, salaires, montant de l'argent de poche, dépenses, consommation, épargne...), des indicateurs de « volume » (nombre de livres lus, nombre d'enfants, taille du réseau amical...) ou des indicateurs d'« intensité » (fréquence d'une pratique culturelle...). Les variables synthétiques, que nous définirons plus loin et qui jouent un rôle central en sociologie quantitative,

relèvent également de cette catégorie : elles expriment grâce à un indicateur quantitatif la position d'un individu selon une grandeur sociologique – par exemple, son niveau de participation aux tâches ménagères, son niveau d'investissement sociale, son degré de « religiosité »...

Le second cas d'utilisation de variables quantitatives en sociologie est relatif aux situations où les sociologues ne travaillent pas sur des personnes, mais sur des entités collectives (par exemple des familles, ménages, associations, communes, entreprises...). Dans ce cas, ces collectifs peuvent être caractérisés par des variables quantitatives exprimant des parts ou des taux : part des individus de sexe masculin ; taux de redoublement ; part des plus de 65 ans ; part de ceux déclarant aimer la musique Rap ou RnB ; probabilité des enfants des différents groupes sociaux d'accéder à une grande école... Dans ce cas, on parle parfois de *données agrégées* car pour obtenir des caractéristiques relatives à des groupes, il est souvent nécessaire d'agréger les réponses individuelles.

2.2 Variables qualitatives

Les grandeurs non quantifiables sont celles qui ne peuvent pas s'exprimer en unités : ces modalités marquent des différences qui ne sont pas des différences numériques mais des différences de nature. Le diplôme, le sexe, la catégorie sociale, les sympathies politiques, le titre du dernier ouvrage lu, le statut matrimonial ou encore la couleur des yeux sont non quantifiables : elles s'expriment grâce à des variables qualitatives. Les modalités de ces variables ne sont pas comparables quantitativement : il n'existe aucune mesure commune de la modalité « marié » et de la modalité « divorcé » de la variable « statut matrimonial ».

Sont également considérées comme qualitatives les variables qui sont fondamentalement quantitatives mais que le sociologue utilise sous une forme recodée, avec des modalités qui correspondent à des classes. L'âge biologique est une variable quantitative mais elle est presque exclusivement utilisée sous la forme d'une variable qualitative définie à partir de classes d'âge : par exemple [18-25 ans] ; [26-30 ans] ; [31-40 ans] ; [41-55 ans] ; [56 ans et plus].

Parmi les variables qualitatives, il est possible de distinguer les variables à modalités ordonnables et celles à modalités non ordonnables. Comme leur nom l'indique, les modalités ordonnables peuvent être classées, hiérarchisées : c'est notamment le cas de toutes les variables dont les modalités sont semblables à « Tout à fait, assez, peu, pas du tout » ou « Très souvent, assez souvent, de temps en temps, rarement, jamais ». C'est aussi le

cas de toutes les variables fondamentalement quantitatives mais qui sont codées selon une échelle comme dans l'exemple suivant :

« Au cours de la dernière année, combien de livres avez-vous acheté ?

1. Aucun
2. Un ou deux livres
3. Entre 3 et 10 livres
4. Entre 11 et 30 livres
5. Plus de 30 livres »

Il est également possible de considérer que les variables « diplôme » voire « opinion politique » sont ordonnables : les diplômes peuvent être classés selon un principe de hiérarchie scolaire et de nombre d'années d'études ; les opinions politiques peuvent être classées en fonction de l'axe gauche-droite (à condition d'ignorer les difficultés concernant les apolitiques ou les écologistes). La catégorie sociale donne également lieu à un classement dans beaucoup de travaux sociologiques : catégories sociales supérieures, intermédiaires ou populaires...

Une variable qualitative peut être simple (lorsqu'elle reflète une seule information), multiple (lorsqu'elle reflète plusieurs informations en même temps) ou ordonnées (lorsqu'elle reflète plusieurs informations classées par ordre). La question « Quelles sont vos trois stations de radio préférées ? » constitue une variable multiple. S'il est, en plus, demandé de classer ces trois stations de radio préférées, elle devient une variable multiple ordonnée.

3. De la nécessité de recoder les variables

Le travail de recodage résulte de deux nécessités. L'une d'entre elles correspond à des contraintes statistiques et techniques : 1) certaines réponses, notamment les réponses aux questions ouvertes, doivent être recodées de manière à être exploitables dans une perspective quantitative ; 2) certaines modalités de réponses sont rarement choisies et doivent donc être regroupées car les effectifs ne permettent pas de les analyser en tant que telles ; 3) enfin, il est parfois nécessaire, pour pouvoir utiliser certaines méthodes statistiques, de diminuer le nombre de modalités des variables (c'est le cas dans les analyses factorielles).

La seconde nécessité correspond aux exigences et choix théoriques : elle résulte de la problématique sociologique choisie. Recoder une variable, c'est préparer les données de façon à les rendre adéquates à la problématique. Cette

dernière affirmation est essentielle : en dehors des contraintes techniques signalées ci-dessus, le recodage d'une variable doit être réalisé en fonction d'un questionnement et non de présupposés extérieurs à la problématique.

Il est donc faux de croire que le recodage est une simple opération technique. Il s'agit d'une opération théorique, visant à rendre les variables les plus adéquates possibles à la problématique et aux notions en œuvre dans celle-ci. Bien recoder les variables est un impératif pour conduire une bonne analyse sociologique.

3.1 Techniques de recodage 1 : regrouper des modalités

Considérons la question suivante, adressée à des titulaires du baccalauréat :

Quelles études avez-vous poursuivies après votre baccalauréat ?

a) Aucune, arrêt des études	f) Faculté de droit
b) Classes préparatoires	g) Autre filière universitaire
c) IUT	h) École d'infirmières
d) BTS	i) École d'architecture
e) Faculté de médecine ou de pharmacie	j)...

Il y a au moins trois manières de recoder cette variable, selon qu'on s'intéresse à l'opposition entre ceux qui ont poursuivi des études post-bac et ceux qui ont arrêté ; à l'opposition entre ceux qui ont engagé des études courtes (IUT, BTS...) et ceux ayant débuté des cursus longs (médecine, classes préparatoires) ; ou à l'opposition entre les filières sélectives (classes préparatoires, IUT, médecine, pharmacie...) et filières moins sélectives (filière universitaire hors médecine, pharmacie et droit...). C'est la problématique et la question théorique posée au traitement statistique (par exemple un tableau croisé) utilisant la variable qui vont déterminer la nature du recodage, en l'occurrence du regroupement de modalités.

3.2 Techniques de recodage 2 : simplifier les variables multiples

L'analyse des variables multiples et ordonnées est parfois plus facile si elles sont transformées en variables qualitatives simples. Il est par exemple possible de

transformer une variable ordonnée en une variable multiple en ne retenant que les modalités choisies par les enquêtés et en écartant l'ordre indiqué. Et il est possible de transformer une variable multiple en une série de variables dites indicatrices : à chaque modalité *M* de la variable multiple est associée une variable indicatrice dont les modalités sont « a choisi » et « n'a pas choisi » la modalité *M*.

Il est également parfois utile de transformer une variable multiple en une simple variable quantitative comptant le nombre de modalités choisies par chaque enquêté.

3.3 Techniques de recodage 3 : simplifier les variables quantitatives

Le recodage des variables quantitatives est souvent indispensable. Il y a au moins deux raisons à cela. Il est, d'une part, commode voire impératif de disposer de variables ayant toutes un statut identique : la plupart des variables manipulées par les sociologues étant des variables qualitatives, il est commode de recoder les quelques variables quantitatives en variables qualitatives. Cette remarque ne s'applique évidemment pas aux quelques situations où l'essentiel des variables sont quantitatives, notamment dans les travaux de socio-démographie, de socio-économie, ou lorsque le sociologue travaille sur des collectifs (voir le chap. 2, § 2).

Recoder une variable quantitative revient à définir les bornes (ou frontières) des diverses catégories (appelées « classes »).

Il existe trois principes généraux de recodage d'une variable quantitative. Le premier principe est un principe « esthétique » ou « mathématique » : les diverses valeurs de la variable sont regroupées en tranches d'égale amplitude et dont les bornes sont « naturelles ». Selon ce principe, la variable « âge » sera recodée en tranches de 5 ou 10 ans, avec des frontières « rondes » : [10-20 ans] ; [21-30 ans] ; [31-40 ans]... Ce principe semble être le plus naturel et est d'usage très fréquent (notamment en démographie et dans les enquêtes très générales) mais il n'est pas nécessairement le plus pertinent ni toujours le plus adéquat aux données dont dispose le sociologue. Les deux autres modes de recodage répondent davantage, de ce point de vue, aux exigences du travail sociologique.

Le deuxième principe de codage est de nature « statistique » et vise à assurer que les catégories créées regroupent un nombre suffisant d'individus. Le sociologue essaie de trouver un compromis entre des catégories (ou classes) regroupant un trop grand nombre d'individus (et donc trop grossières

et tentant à « écraser » les éventuelles différences entre individus) et des catégories regroupant un trop petit nombre d'individus (rendant ainsi impossible ou illusoire leur analyse statistique). Une solution « optimale » consiste à créer des classes équilibrées, c'est-à-dire regroupant un nombre d'individus proche d'une classe à l'autre. Certains logiciels permettent de déterminer automatiquement les classes statistiquement équilibrées. Sinon, il faut procéder par tâtonnement, en essayant plusieurs configurations.

Le troisième principe de recodage est de nature plus sociologique et vise à assurer que les catégories créées correspondent à des situations sociologiques homogènes, similaires. Ainsi, un sociologue travaillant sur les transformations induites par l'arrivée d'un premier enfant dans une famille devrait concevoir les différentes classes de la variable « âge » en fonction de son objet : si la taille de l'échantillon le permet, il devra concevoir des classes d'âge fines autour de l'âge moyen d'arrivée du premier enfant (entre 28 et 30 ans), quitte à concevoir des classes plus vastes pour les âges éloignés de cet âge moyen.

En pratique, c'est au sociologue de trouver un compromis raisonnable et acceptable du point de vue statistique et sociologique : le recodage d'une variable quantitative doit respecter le principe statistique, sans pour autant sacrifier l'exigence du sens sociologique de la variable. Le critère esthétique ou mathématique est plus superflu mais peut malgré tout entrer en ligne de compte pour rendre les résultats plus pédagogiques (puisque plus familiers et plus simples en apparence).

3.4 Techniques de recodage 4 : coder les matériaux qualitatifs

Devant un matériau de nature qualitative (des lettres, les images, des textes... voire des entretiens), le sociologue doit commencer par déterminer quelles sont les informations à retenir : quelles sont les données pertinentes pour sa problématique ? Une fois ces choix opérés, il doit coder les données selon une grille standardisée. Nous avons déjà présenté (chap. 1, § 3) le cas de codage de lettres ou de sources vidéos. Considérons ici un autre exemple : le codage de petites annonces matrimoniales parues dans le *Chasseur français*[1]. Ces petites annonces présentent des formes trop hétérogènes pour être analysables sans codage préalable. À côté du sexe et de l'âge de l'annonceur, il existe bien

1. François de Singly, « Les manœuvres de séduction : une analyse des petites annonces matrimoniales », *Revue française de sociologie*, 1984, XXV, 4, p. 523-559.

d'autres caractéristiques méritant d'être analysées et recodées : le nombre de mots de l'annonce, la présence d'enfants, le verbe formant jonction entre l'offre et la demande (« rencontrerait, épouserait, cherche... »), le nombre d'éléments corporels cités, la présence de référence économique, les références morales ou culturelles, la présence de qualificatifs d'excellence physique (« bien physiquement, joli, beau... ») ou encore la présence de qualificatifs d'excellence sociale (« belle situation, grande propriété... »). En tout, l'auteur a repéré et codé 78 traits dans son corpus d'annonces – certains qualitatifs, d'autres quantitatifs. Ce travail lui permet d'appréhender les processus par lesquels « un individu tente de faire reconnaître sa valeur sociale en mettant en scène ses richesses les plus propres à séduire ». Cet exemple illustre bien un principe : il faut faire feu de tout bois et longuement réfléchir aux informations méritant d'être recodées. Même les matériaux apparemment pauvres (ici des annonces de quelques lignes) peuvent faire l'objet de codages précis et nombreux (ici 78 critères distinctifs ont été repérés).

Les réponses aux questions ouvertes (par exemple « Quels sont les titres des films que vous avez vus au cinéma au cours du dernier mois ? ») constituent un cas fréquent de variables nécessitant ce type de travail de codage[1].

3.5 Techniques de recodage 5 : combiner les variables

Afin de simplifier le travail d'analyse et de croisement, il est souvent utile de concevoir des variables combinant deux variables primaires. Les modalités de la nouvelle variable sont obtenues par combinaison des modalités des deux variables primaires. Cette technique est particulièrement utile lorsque l'analyse conduit à tenir compte de deux variables contextuelles ou explicatives en même temps. Il est par exemple fréquent de recourir à une variable combinant à la fois une information sur le sexe et une information sur l'âge[2] :

Variable âge × sexe

1. Homme de 18 à 34 ans

1. Il est parfois possible de recourir à des outils d'analyse textuelle : voir Pascal Marchand, *L'Analyse du discours assistée par ordinateur*, Paris, Armand Colin, 1998 ; Ludovic Lebart, André Salem, *Statistique textuelle*, Paris, Dunod, 1994 (épuisé mais consultable sur le site <http://ses.telecom-paristech.fr/lebart/>).
2. On pourra prendre soin de réfléchir à l'ordre avec lequel on croise les variables : dans l'exemple, le sexe vient avant l'âge et la variable reflète des groupes de sexe découpés selon l'âge. L'inversion des rôles fournit une variable davantage structurée par l'âge.

2. Homme de 35 à 59 ans
3. Homme de plus de 60 ans
4. Femme de 18 à 34 ans
5. Femme de 35 à 59 ans
6. Femme de plus de 60 ans

Cette technique est également utile pour rassembler deux informations qui vont naturellement ensemble mais qui font l'objet de deux questions différentes dans le questionnaire. Les questions « Quelle est votre religion ? » et « Êtes-vous pratiquant(e) ? » peuvent être assemblées de la manière suivante :

1. Sans religion
2. Catholique non pratiquant
3. Catholique pratiquant
4. Protestant non pratiquant
5. Protestant pratiquant
6. Musulman non pratiquant
7. Musulman pratiquant
8. *etc.*

Le nombre de modalités de la nouvelle variable est égal au produit du nombre de modalités de chacune des questions : il peut donc être élevé et rendre nécessaire un nouveau recodage pour regrouper des modalités (notamment celles qui sont rares).

4. Passer des variables aux indicateurs théoriques : les variables synthétiques

Nous avons souligné la nécessité de recoder les informations recueillies pour les ajuster à la problématique et au questionnement théorique. Mais ce premier travail sur les variables ne suffit pas : il est souvent nécessaire de concevoir, à partir des réponses aux questions, de nouvelles variables incarnant les concepts et notions utilisés en les opérationnalisant. Ces variables sont des *variables* ou *indicateurs synthétiques* : elles rassemblent (« synthétisent ») les informations issues de diverses questions liées à un concept ou une notion.

Les notions ainsi opérationnalisées peuvent être abstraites et être issues de la théorie sociologique : c'est par exemple le cas des notions d'autonomie, d'individualisation, d'investissement scolaire, de proximité sociale ou encore d'intégration qui ne s'observent pas directement. Mais il peut également s'agir de notions moins abstraites dont l'objectivation passe nécessairement par plusieurs questions. Ainsi, plutôt que de demander « Lisez-vous beaucoup ? », il est préférable de poser plusieurs questions plus précises comme « Au cours du dernier mois, combien de romans avez-vous lus ? », « Combien de BD ? », « Combien d'essais ? », « Lisez-vous régulièrement un magazine ? », « Lisez-vous régulièrement un quotidien ? »... Ces diverses questions ne nous intéressent peut-être pas en tant que telles. Elles prennent sens dans la mesure où, prises ensemble, elles renseignent sur la pratique de lecture de l'enquêté. Mais, étant nombreuses, elles ne sont pas aisément utilisables dans les traitements statistiques. Il est dès lors utile de les rassembler pour constituer un indicateur synthétique d'intensité de la pratique de lecture.

Le nombre d'informations primaires intervenant dans la définition de la variable synthétique peut être très différent (de deux ou trois à quelques dizaines). Nous allons présenter les principales techniques permettant de construire et mettre au point de tels indicateurs synthétiques.

4.1 Créer des variables synthétiques par combinaison

La première technique, déjà entrevue précédemment, consiste à fusionner deux ou trois variables primaires en combinant leurs modalités. Imaginons travailler sur les pratiques de lecture et de « consommation » de livres et considérons par exemple les trois questions suivantes :

Q1. Au cours du dernier mois, avez-vous acheté des livres ?

1. Oui
2. Non

Q2. Au cours du dernier mois, avez-vous emprunté des livres en bibliothèque ?

1. Oui
2. Non

Q3. Au cours du dernier mois, avez-vous emprunté des livres à des proches, des amis, des connaissances... ?

1. Oui
2. Non

Il est possible de combiner ces trois questions pour construire la variable synthétique « Pratiques de l'achat et de l'emprunt de livres au cours du dernier mois » qui est un assez bon indicateur de la pratique livresque et qui peut être utile pour estimer la circulation et la manipulation des livres de manière indépendante de leur lecture :

1. N'a ni acheté ni emprunté
2. A acheté mais n'a pas emprunté
3. A acheté et a emprunté à des proches
4. A acheté et a emprunté en bibliothèque
5. A acheté et a emprunté à des proches et en bibliothèque
6. N'a pas acheté mais a emprunté en bibliothèque
7. N'a pas acheté mais a emprunté à des proches
8. N'a pas acheté mais a emprunté à des proches et en bibliothèque

Cette technique est notamment utilisée par Bernard Lahire dans son travail sur la *Culture des individus*[1] pour identifier le caractère consonant ou dissonant des pratiques et goûts culturels des individus. Ils sont dits dissonants s'ils mêlent à la fois des goûts et des pratiques très légitimes et peu légitimes ; ils sont consonants s'ils mêlent uniquement des aspects très légitimes (consonants légitimes) ou uniquement des aspects peu légitimes (consonants peu légitimes). Pour cela, la première étape est de classer, par simple recodage (regroupement de modalités), les différentes modalités de chaque variable décrivant les goûts et les pratiques (TV, musique, livres, visites, spectacles, cinéma...) selon leur degré de légitimité. Ainsi, par exemple, les préférences télévisuelles sont classées en trois modalités : peu légitimes (*Le Juste Prix*, *Tout est possible*, *Perdu de vue*...), très légitimes (*Bouillon de culture*, *Faut pas rêver*, *Les Mercredis de l'histoire*...) et mixtes. Par combinaison de ces variables, il est possible de qualifier chaque individu selon son profil : ses goûts et pratiques son dissonants en matière télévisuelle, livresque et cinématographique s'il combine, en ces matières, à la fois des modalités « peu légitimes » et « très légitimes ».

Cette technique de construction de variables synthétiques est seulement utilisable si le nombre de variables à combiner n'est pas trop élevé (deux ou trois, quatre au plus) et que le nombre de modalités de chacune de ces variables n'est également pas trop grand. Dans le cas contraire, la variable synthétique obtenue n'est pas commode d'utilisation puisque son nombre

1. Bernard Lahire, *La Culture des individus. Dissonances culturelles et distinction de soi*, Paris, La Découverte, 2004, chapitres 3 et 6.

de modalités est très élevé – et donc le nombre d'individus par modalité faible.

4.2 Créer des variables synthétiques par calcul de scores

Une seconde technique, qui est certainement la plus utilisée et la plus facile à mettre en œuvre, est de calculer des variables-scores. Le principe est le suivant : après avoir identifié la liste de toutes les variables utiles, on combine les variables ou certaines de leurs modalités. Cette combinaison de variables est différente selon que nous avons à faire à des variables qualitatives ou des variables quantitatives.

S'il s'agit de variables qualitatives, on affecte des notes (généralement des nombres entiers 0, 1 ou 2) à chacune des modalités des variables de cette liste puis, pour chaque individu, on compte son score ou sa note finale. Cette variable est par définition quantitative : elle doit être recodée car il est bien difficile de donner un sens à chacune des valeurs de cette variable. Il est usuel de créer un nombre réduit de classes (3, 4 ou 5 environ) : un échelonnement des comportements ou des situations individuelles en 3, 4 ou 5 niveaux suffit en général. Les modalités correspondent alors à des intensités : « très faible », « faible », « assez élevé », « très élevé »... Le principe n'est pas difficile à mettre en œuvre. Et il est très efficace pour créer des variables objectivant des notions plus ou moins abstraites.

Un cas simple et très fréquent de variable-score est celui où le sociologue souhaite calculer un nombre indiquant une intensité ou une diversité. Considérons par exemple la question « Parmi les activités physiques et sportives suivantes, indiquez celles que vous pratiquez au moins une fois par an ? Natation ; Jogging ; Vélo ; Randonnée ; Danse ; Gymnastique... ». Il peut être intéressant de calculer le nombre d'activités pratiquées, qui peut être interprété comme l'intensité de pratiques physiques et sportives d'un individu. Pour cela, il suffit d'attribuer la note 1 à toutes les activités puis de compter le score de chaque individu. Il peut également être intéressant de calculer la diversité des pratiques en distinguant les pratiques sportives individuelles, collectives et en comptant le nombre d'activités de nature différente pratiquées par chaque enquêté...

Une autre situation classique de construction d'une variable-score est le cas de l'opérationnalisation d'un concept ou d'une notion abstraite. Considérons par exemple notre recherche portant sur les usages du téléphone portable au

sein des couples, où nous nous intéressions notamment à la question de l'individualisation du portable[1]. Le portable peut en effet être utilisé par un individu à titre purement privé et individuel, sans que son conjoint n'intervienne. Cette notion s'oppose à celle de partage ou de collectivisation du portable.

Afin de préciser et d'opérationnaliser cette notion d'« individualisation du portable », nous avons utilisé les réponses à quatre questions du questionnaire : « Arrive-t-il que le conjoint réponde à votre place avec votre portable ? », « Au cours de la dernière semaine, votre conjoint a-t-il emprunté votre portable ? », « Au cours de la dernière semaine, avez-vous reçu des appels sur votre portable pour votre conjoint ? », « Votre conjoint connaît-il le code PIN de votre portable ? ». Notre indice général d'individualisation synthétise les réponses à ces quatre questions : pour chaque question où il a répondu négativement, un individu se voit attribuer un point. Ainsi, chaque individu est caractérisé par un « score » résumant le degré d'individualisme de son portable. La variable ainsi créée constitue une échelle d'individualisation du portable. De nature quantitative, elle varie de 0 à 4 : les scores 0 et 1 correspondant à un « très faible » ou « faible individualisme » du portable, le score 2 à un « individualisme moyen », le score 3 à un « individualisme assez fort » et le score 4 à un « très fort individualisme ».

Un autre exemple est celui mis en œuvre par Alain Girard dans son travail sur *Le Choix du conjoint* (1974). Pour étudier la « distance » ou inversement la « proximité » entre deux conjoints, c'est-à-dire leur dissemblance ou leur ressemblance sociale, culturelle et géographique, Alain Girard construit un indice global en retenant douze variables caractérisant les deux conjoints[2] :

- leur nationalité ;
- la taille de leur commune de naissance ;
- la situation géographique de leur commune de naissance ;
- leur niveau d'études ;
- leur religion ;
- la taille de leur commune de résidence au moment du mariage ;
- la situation géographique de leur commune de résidence lors du mariage ;

1. Olivier Martin, François de Singly, « Le téléphone portable dans la vie conjugale : retrouver un territoire pour soi ou maintenir le lien conjugal ? », *Réseaux*, vol. 20, 2002, n° 112-113, p. 211-248.
2. Alain Girard, *Le Choix du conjoint. Une enquête psycho-sociologique en France*, Paris, PUF-INED (nouvelle édition), 1974, p. 87-94. Le choix des variables participant à la définition de l'indicateur est critiquable, mais ce n'est pas l'essentiel ici.

– le nombre de localités habitées depuis leur naissance ;

– leurs professions ;

– la nationalité du père de chaque conjoint ;

– la profession du père de chaque conjoint ;

– la profession actuelle du mari et celle de son beau-père.

Selon la plus ou moins grande différence au sein de chacune des variables, Girard affecte une note variant de 1 à 7 : 1 si les deux conjoints sont très différents ; 7 s'ils sont semblables[1]. Ainsi, pour la variable « Niveau d'étude » divisée en sept modalités hiérarchisées en « degrés » (pas d'études, études primaires sans CEP, études primaires avec CEP, études techniques, études primaires, études secondaires, études supérieures), les conjoints sont affectés de la note 7 s'ils ont exactement le même niveau d'étude, de la note 6 si leurs niveaux d'étude diffèrent d'un degré..., de la note 1 si leurs niveaux d'étude diffèrent de six degrés.

Au final, les individus obtiennent une note (ou score) variant du minimum 12 au maximum 84 : d'une faible proximité à une proximité très forte. Il s'agit d'une variable quantitative que Girard regroupe en classes.

D'un point de vue technique, deux questions peuvent se poser lors de la création d'une variable-score. La première question est celle du choix des variables rentrant dans la composition de la variable synthétique : comment sélectionner les variables ? La principale réponse est d'ordre théorique : il faut inclure les variables issues des questions conçues et utilisées comme des indicateurs de la notion théorique au moment de la mise en point du questionnaire. Il est possible, mais pas indispensable, de compléter cette réponse par un second critère : il faut étudier les relations (par calcul des corrélations ou test de l'indépendance) entre les variables et inclure celles qui sont « positivement liées », c'est-à-dire qui « vont dans le même sens ».

La seconde question est celle du choix des coefficients de pondération affectés à chacune des variables ou des modalités. Il n'existe aucun critère indiscutable pour justifier le choix de la valeur de coefficients de pondération. Les sociologues recourent simplement à leur bon sens, c'est-à-dire choisissent les valeurs en fonction de la pertinence de la variable ou de la modalité et de sa capacité à exprimer une notion théorique. L'expérience

1. Girard interprète son indicateur comme une distance. Mais il s'agit plutôt d'un indicateur de proximité : une valeur élevée de l'indicateur étant synonyme d'une forte proximité entre les conjoints ; une valeur faible étant synonyme de fortes différences sociales entre les individus.

tend de toute façon à montrer que le choix des coefficients n'est pas crucial. Le plus souvent, on se contente d'affecter des notes simples : 0 et 1 ; voire 0, 1 et 2...

La qualité de la variable synthétique dépend bien davantage de la pertinence des variables qui rentrent dans sa composition. Cette qualité dépend aussi du nombre de ces « variables pertinentes » : *a priori*, plus leur nombre est élevé, plus la variable synthétique aura un sens incontestable et robuste.

4.3 Créer des variables synthétiques à partir de variables quantitatives

Dans le cas du calcul de variables-scores, des points sont attribués aux modalités puis additionnés car, les variables étant qualitatives, il n'est pas possible de combiner numériquement leurs modalités. Dans le cas où les variables sont quantitatives, il est possible de combiner directement leurs valeurs par addition, soustraction, division... ou toute autre opération mathématique[1]. Le seul critère pour juger du bien-fondé de l'indicateur calculé est la signification ou le sens qu'il est possible de lui attribuer.

Par exemple, pour étudier le « lien de germanité à l'âge adulte » et notamment l'homophilie de sexe (à l'âge adulte, a-t-on davantage tendance à fréquenter les individus de même sexe que soi parmi les membres de sa fratrie ?), des sociologues[2] ont eu recours à des indicateurs d'homophilie définis comme la différence entre le nombre de rencontres avec des germains de même sexe au cours des 12 derniers mois et le nombre de rencontres avec des germains de sexe différent au cours des 12 derniers mois. Cet indicateur, de définition et d'interprétation simples, permet d'objectiver cette notion d'homophilie de sexe : les valeurs élevées de l'indicateur sont des signes d'homophilie prononcée ; les valeurs faibles (négatives) sont des signes d'absence d'homophilie, voire d'hétérophilie de sexe.

De manière comparable, étudiant les activités et les loisirs et notamment les disparités entre les hommes et les femmes, Alain Chenu et Nicolas Herpin

1. Quitte à pondérer ou normaliser les variables initiales si leurs valeurs et leurs variabilités sont trop hétérogènes (en d'autres termes, si elles prennent des gammes de valeurs très différentes : l'une variant de 0 à 1 ; l'autre variant de 1 000 à 10 000 par exemple).
2. Emmanuelle Crener, Jean-Hugues Déchaux, Nicolas Herpin, « Le lien de germanité à l'âge adulte. Une approche par l'étude des fréquentations », *Revue française de sociologie*, vol. 41, n° 2, 2000, pp. 211-239.

ont défini un indicateur du caractère plutôt masculin ou plutôt féminin d'une activité[1]. Pour cela, ils ont considéré, pour chaque activité, le temps moyen D_h passé par les hommes dans cette activité et le temps moyen D_f passé par les femmes. Leur indicateur est alors défini ainsi :

$$I = 200 \times \frac{D_f}{D_f + D_h} - 100$$

Cet indicateur peut varier entre -100 et 100 : il prend la valeur 100 pour une activité exclusivement féminine, la valeur -100 pour une activité exclusivement masculine, la valeur 0 pour une activité indifféremment masculine et féminine. Cet indicateur leur permet d'identifier facilement les dominantes plutôt féminines ou masculines de diverses activités : la lecture, la promenade, le sport, le bricolage, les courses, le soin des enfants... Par exemple, l'indicateur vaut environ 50 en ce qui concerne la cuisine, le linge et le ménage : ces activités sont nettement féminines (les femmes y passent trois fois plus de temps que les hommes) ; et l'indicateur vaut presque 100 en ce qui concerne la couture (cette activité est quasi exclusivement féminine). Le calcul de ces indicateurs à deux dates différentes permet par ailleurs de déterminer si une activité se féminise ou non au cours du temps, ou si elle perd progressivement de sa dominante féminine...

4.4 Créer des variables synthétiques par analyse factorielle

Afin de diminuer l'arbitraire dans le choix des variables et de leur poids (coefficient de pondération) dans la définition d'un indicateur synthétique, il est possible de recourir à des méthodes statistiques dites multidimensionnelles c'est-à-dire destinées à analyser un grand nombre de variables en même temps. Nous les présenterons en détail au chapitre 4. Retenons pour l'instant que certaines d'entre elles – notamment les méthodes factorielles – permettent de construire de nouvelles variables qui soient des combinaisons de variables et qui restituent au mieux les différents liens entre ces variables. Ces nouvelles variables sont appelées des « axes », des « facteurs » ou des « dimensions ».

1. Alain Chenu et Nicolas Herpin, « Une pause dans la marche vers la civilisation des loisirs », *Économie et statistique*, n° 352-353, 2002, p. 15-37.

Dans sa recherche sur la carrière scolaire des enfants issus de l'immigration[1], Philippe Cibois formule l'hypothèse que si ces enfants réussissent mieux (à caractéristiques sociales identiques) que les enfants de parents français, c'est en raison de la force de leur projet migratoire et de leurs attentes vis-à-vis du système scolaire. Les familles immigrées ont, selon l'auteur, des attentes qui se caractérisent par une bonne volonté scolaire, c'est-à-dire « par un ensemble de comportements de respect des consignes données par l'école dans le comportement scolaire et hors école des enfants ». Pour construire son indicateur de « bonne volonté scolaire », Philippe Cibois utilise des modalités à 15 questions en ne retenant que les modalités considérées comme des indices de bonne volonté scolaire. Par exemple : « avoir préparé son cartable la veille avant de se coucher », « avoir préparé son cartable la veille avant le repas », « n'oublie jamais ou rarement un livre ou un cahier à la maison », « estime qu'arriver en retard à l'école est grave »... Mais, plutôt que de construire un indicateur « à la main » en choisissant une pondération de manière arbitraire, l'auteur réalise une analyse des correspondances qui lui permet d'obtenir une nouvelle variable (un facteur) opposant, de manière synthétique et cohérente, la bonne volonté scolaire à une attitude qu'il qualifie de « décontractée » vis-à-vis de l'école[2].

1. Philippe Cibois, « La bonne volonté scolaire. Expliquer la carrière scolaire d'élèves issus de l'immigration », *in* Philippe Blanchard et Thomas Ribémont (dir.), *Méthodes et outils des sciences sociales. Innovation et renouvellement*, L'Harmattan, 2002, p. 111-126. Pour un autre exemple, voir Olivier Galland, Yannick Lemel et Jean-François Tchernia, « Les valeurs en France », *Données sociales*, Paris, INSEE, 2002, p. 559-564.
2. En fait, cette analyse a une fonction exploratoire (voir cette notion dans notre conclusion) : elle permet à l'auteur de construire son indicateur en toute connaissance de cause.

UN INTERMÈDE : SAISIR LA DIVERSITÉ DES SITUATIONS

Après avoir présenté les diverses logiques et méthodes de préparation des données (Partie I) et avant d'aborder les logiques et méthodes d'analyse des données (Partie II), arrêtons-vous sur une notion fondamentale de la statistique : la notion de variabilité.

La notion de variabilité renvoie à l'idée de variation, de diversité, d'hétérogénéité. Un phénomène de grande variabilité est un phénomène présentant une grande disparité, de fortes différences entre les individus. Inversement, un phénomène présentant une faible variabilité est un phénomène plutôt homogène, montrant de grandes similitudes : les individus présentent des caractéristiques proches. Par exemple l'âge de la population française est très variable : il y a des individus de 4 ans, de 7 ans, de 38 ans, de 58 ans comme de 98 ans... En revanche, l'âge des individus d'une classe de terminale est peu variable : la plupart des élèves de terminale ont 17 ou 18 ans, quelques-uns ont 19 ou 20 ans et quelques autres ont 15 ou 16 ans, mais ils sont finalement assez peu nombreux.

Cette notion de variabilité peut être objectivée, opérationnalisée, par des indicateurs statistiques. L'exemple précédent montre que ces indicateurs de mesure de variabilité doivent à la fois rendre compte de l'étendue des valeurs (quelles sont les diverses valeurs prises par la variable ?) et de la fréquence avec laquelle ces valeurs surviennent. Il apparaît en effet essentiel que la mesure de la variabilité soit capable de distinguer un groupe où tous les individus ont le même âge (30 ans) sauf un (40 ans) d'un groupe où tous les individus ont des âges compris entre 30 et 40 ans.

Lorsque la variable est de nature qualitative, il est artificiel de concevoir un indicateur numérique de variabilité. Il est toutefois possible de saisir, de manière plus subjective, cette variabilité en examinant le tri à plat de la variable, c'est-à-dire la répartition des individus selon les diverses modalités. Le tri à plat présenté ci-dessous permet de constater que certaines modalités (DAEU, bac STT, autre) sont rarement choisies et, qu'en revanche, les individus sont bien répartis sur les autres modalités. La variabilité est donc relativement grande : les individus ne sont pas tous regroupés ; leur répartition est assez forte.

Tri à plat de la variable « Type de baccalauréat »

Modalité	L	ES	S	STT	DAEU	Autre
Effectif	325	290	245	40	23	10
%	34, 8 %	31,1 %	26,3 %	4,3 %	2,5 %	1 %

Lorsque la variable est quantitative, il existe de multiples manières pour définir un indicateur de variabilité[1]. Les deux indicateurs les plus courants et les plus pertinents d'un point de vue théorique sont la variance et l'écart-type : ils expriment la dispersion des valeurs d'une variable autour de la moyenne de cette variable. Plus les valeurs seront concentrées et proches de la moyenne, plus la variance et l'écart-type seront faibles ; plus les valeurs seront dispersées et différentes de la moyenne, plus la variance et l'écart-type seront élevés.

Mathématiquement, si on connaît les valeurs $(X_1, X_2, \ldots, X_n)$ de la variable X pour n individus, et si $\overline{X}$ est la moyenne de ces valeurs, la variance est définie par la relation suivante :

$$\text{variance} = \frac{1}{n} \times \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Cette formule est l'expression en langage formel de la définition de la variance comme « moyenne des carrés des écarts des valeurs à leur moyenne ».

Le second indicateur courant est l'écart-type, simplement défini comme la racine de la variance. Le calcul de la racine permet d'obtenir un indicateur de variabilité s'exprimant dans la même grandeur (la même unité) que la variable elle-même : si la variable s'exprime en minutes ou en euros, l'écart-type s'exprime également en minutes ou en euros.

$$\textit{écart-type} = \sqrt{\text{variance}} = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Indépendamment de leurs définitions mathématiques, il est essentiel de retenir que la variance et l'écart-type sont toujours positifs et prennent des valeurs d'autant plus grandes que la variabilité, l'hétérogénéité d'une variable

1. La plupart des ouvrages de statistiques offrent une présentation de ces divers indicateurs et de leurs intérêts respectifs. Voir par exemple : Thomas Wannacott et Ronald Wannacott, *Statistiques*, Paris, Economica, 1991, p. 45-50.

(c'est-à-dire des valeurs de cette variable chez les différents individus de l'enquête) est grande.

Dans la grande majorité des cas pratiques rencontrés par le sociologue analysant des données quantitatives, le calcul de la variance et de l'écart-type n'a pas d'intérêt en tant que tel : il est rare que ces indicateurs soient calculés et commentés pour eux-mêmes. Mais ils interviennent dans de nombreux raisonnements et outils statistiques – et c'est ce qui justifie cette brève présentation dans ce manuel.

La notion de variabilité est en effet centrale dans toute démarche scientifique empirique puisque le chercheur doit faire face à des situations présentant une certaine diversité/hétérogénéité, doit essayer de déchiffrer les principes structurant cette diversité, voire en trouver les causes : cela vaut pour le biologiste qui cherche à comprendre la diversité des réactions à un même agent infectieux, pour le psychologue qui souhaite saisir les différentes attitudes face à un même stimulus, ou encore le sociologue qui s'interroge sur les différences d'opinion ou de comportement des individus d'une même culture et d'une même société. D'une certaine manière, sans la variabilité intrinsèque à la plupart des phénomènes naturels, une grande partie des scientifiques seraient au chômage (et la vie serait sans doute bien triste !)

La statistique peut être vue comme la science permettant aux scientifiques, en l'occurrence aux sociologues, de saisir la variabilité des situations, de catégoriser cette diversité, d'en identifier les principes, d'en chercher les facteurs explicatifs... Par exemple, dans son travail sur *Le Suicide*, Émile Durkheim s'est attaché à identifier la diversité des taux de suicide, les facteurs expliquant cette diversité, pour finalement proposer une catégorisation des suicides.

Partie 2

ANALYSER LES RELATIONS ENTRE VARIABLES

ANALYSER LES RELATIONS ENTRE DEUX VARIABLES

Le sociologue ne se contente pas de saisir les comportements majoritaires, ni même d'étudier la diversité des situations. Son ambition est d'étudier les « variations concomitantes » (selon le vocabulaire utilisé par Durkheim dans *Les Règles de la méthode sociologique*[1]), c'est-à-dire les relations, les dépendances ou les corrélations entre variables. Les variables étant fréquemment de nature qualitative, l'essentiel de ce chapitre est consacré aux techniques d'analyse destinées à ce type de variable. Nous aborderons le cas des variables quantitatives en fin de chapitre (§ 3 et § 4).

L'outil principal pour étudier les relations entre variables qualitatives est le *tableau croisé* (parfois appelé *tri croisé*). Il s'agit d'un tableau indiquant la distribution des individus selon deux variables simultanément[2]. De tels tableaux ont vocation à mettre en évidence l'influence d'une variable sur une autre (afin d'identifier les déterminants sociaux) ou, plus simplement, la dépendance d'une variable vis-à-vis d'une autre (afin de montrer l'existence d'interdépendances entre des phénomènes).

La question à laquelle ce type de tableau répond est : « Dans quelle mesure tel phénomène ou telle caractéristique sociale dépend-t-elle de tel autre phénomène ou caractéristique ? » En termes plus techniques, la question s'exprime ainsi : « Dans quelle mesure une variable dépend-t-elle d'une autre ? » Notons que la notion de dépendance en jeu dans ces questions ne renvoie pas nécessairement à une idée déterministe ou causale. Notons également, de manière fondamentale, que les deux variables croisées jouent des rôles distincts, dissymétriques : on s'interroge sur la dépendance de l'une vis-à-vis de l'autre. L'une est la variable dépendante ou « à expliquer » ; l'autre est la variable indépendante ou « explicative ».

1. *Op. cit.*, chapitre 6, pages 129 et suivantes.
2. Sur les principes de construction et de lecture de tels tableaux, nous renvoyons le lecteur à François de Singly, *op. cit.*, 2005, p. 93-101.

1. JUGER LA DIFFÉRENCE ENTRE DEUX POURCENTAGES

Le tableau 3.1 croise la variable « sexe » avec la variable « sorties au cinéma au cours des douze derniers mois » sur un échantillon de 1 001 personnes. La question soulevée par ce tableau est « la fréquentation des cinémas dépend-t-elle du sexe ? ». Pour répondre à cette question, il est possible de comparer la répartition des femmes selon les différentes modalités en colonne avec la répartition des hommes. Concrètement, cela revient à calculer les pourcentages en ligne (figurant dans ce même tableau 3.1). En d'autres termes, selon une présentation plus technique, la variable « explicative » étant en ligne et la variable « expliquée » en colonne, l'étude de ce tableau suppose le calcul de pourcentages en ligne.

Tableau 3.1. Sorties au cinéma au cours des 12 derniers mois selon le sexe

	Jamais	Moins d'une fois par mois en moyenne	Plus d'une fois par mois en moyenne	Total
Femme	261 *51 %*	180 *35 %*	69 *14 %*	510 *100 %*
Homme	225 *46 %*	150 *30 %*	116 *24 %*	491 *100 %*
Total	486 *49 %*	330 *33 %*	185 *18 %*	1001 *100 %*

Source : Effectifs fictifs mais inspirés de l'enquête sur les conditions de vie de l'INSEE, 2003.

L'intérêt de tels tableaux est de mettre en évidence d'éventuelles différences de pratique, d'attitude ou d'opinion représentées en colonne entre les divers groupes représentés par les modalités en ligne. En l'occurrence, le tableau 3.1 permet de constater que hommes et les femmes ont des pratiques différentes en ce qui concerne la fréquentation des cinémas : 24 % des hommes contre seulement 14 % des femmes fréquentent plus d'une fois par mois les salles obscures.

Il est en effet incontestable que parmi les 1 001 individus de notre échantillon, les hommes vont plus souvent au cinéma que les femmes : l'écart entre le pourcentage relatif aux hommes et celui relatif aux femmes s'élève à 10 %. Mais qu'en est-il réellement parmi l'ensemble de la population dont notre échantillon est extrait ? Peut-on faire confiance aux résultats établis sur

l'échantillon ? De manière plus précise, peut-on faire confiance à cet écart de 10 points entre les pratiques masculines et les pratiques féminines ?

1.1 L'intervalle de confiance d'une différence

Pour répondre à cette question il faut conduire un raisonnement similaire à celui mené précédemment (chap. 1, § 5) : les pourcentages calculés sur l'échantillon ne sont que des estimations des « vrais » pourcentages, ceux relatifs à l'ensemble de la population de référence ; et dans notre cas, la différence entre les deux pourcentages n'est qu'une estimation de la « vraie » différence, celle existant éventuellement au sein de la population de référence.

Nous savons maintenant qu'une estimation peut donner lieu au calcul d'un intervalle de confiance (qui devrait même systématiquement accompagner l'estimation) : cet intervalle est un encadrement probable de la vraie valeur, c'est-à-dire un encadrement contenant la vraie valeur (inconnue) avec une forte probabilité (au moins 90 ou 95 % en fonction des exigences du sociologue).

Dans la situation qui nous intéresse ici, nous avons affaire à deux pourcentages ou, plus exactement à la différence entre deux pourcentages : ce qui nous intéresse, ce ne sont pas tant les deux pourcentages que leur différence apparente. Cette différence n'est-elle qu'apparente, c'est-à-dire produite par le hasard, ou bien traduit-elle une différence existant fondamentalement au sein de la population ?

Pour répondre de manière précise à cette question, il faut recourir à la notion d'intervalle de confiance appliquée à notre estimation de la différence entre les deux pourcentages[1]. Nous avons estimé à 10 points l'écart entre le comportement des hommes et celui des femmes, mais au sein de la population de référence cet écart est-il de 11, 12, 13, voire 20 points ou, au contraire, de 9, 8, 7,..., 2, 1 voire 0, –1 ou –2 points ? Dans le cas d'un écart nul au sein de la population de référence, l'écart constaté sur l'échantillon résulte des aléas d'échantillonnage et non d'une différence fondamentale entre le comportement des hommes et celui des femmes.

1. Un autre raisonnement consisterait simplement à comparer les intervalles de confiance de chacun des pourcentages intervenant dans le calcul de la différence. La comparaison de deux intervalles de confiance n'est logiquement et statistiquement pas possible. Les statisticiens disent, de manière simple : « Il faut calculer l'intervalle de confiance des différences et non la différence des intervalles de confiance. »

Un cas particulier mérite toute notre attention. Lorsque l'intervalle de confiance de la différence, qui présente les valeurs probables de cette différence, comprend la valeur 0 (zéro), celle-ci est une valeur possible et même probable de la différence : nous devons alors conclure que l'écart constaté entre nos deux pourcentages n'est pas significatif et donc que le comportement des hommes et celui des femmes ne sont pas significativement différents.

Inversement, si l'intervalle de confiance de la différence ne comprend pas la valeur 0, celle-ci n'est pas une valeur probable de la différence et donc la différence constatée au sein de l'échantillon est probablement le signe d'une différence existant au sein de l'échantillon.

1.2 Les formules de calcul et leur usage

Les principes et raisonnements conduisant à l'établissement des formules de calcul des intervalles de confiance d'une différence sont identiques à ceux ayant présidé à l'établissement des intervalles de confiance simples (chap. 1, § 5). Nous faisons l'hypothèse que la différence constatée sur l'échantillon est valable sur l'ensemble de la population. Puis, sous cette hypothèse, nous estimons grâce à un calcul probabiliste exact quel est le domaine (l'intervalle) de variation probable de cette différence si nous l'estimons sur un échantillon de taille n. Cela revient à envisager tous les échantillons possibles issus de cette population et, pour chacun de ces échantillons, à calculer la différence observée. Cela revient à déterminer quelle est la probabilité d'apparition de chacune des différences possibles entre deux pourcentages.

Selon l'usage, on écarte les valeurs des différences ayant une trop faible probabilité d'apparition (5 ou 10 % par exemple) pour ne retenir que les valeurs « probables ». L'ensemble de ces valeurs probables constitue, par définition, l'intervalle de confiance.

Les formules de l'intervalle de confiance d'une différence sont, en apparence, plus compliquées que celles relatives à l'intervalle de confiance d'un pourcentage. Mais leur forme générale et leur interprétation sont identiques.

Supposons que nous voulions comparer les réponses ou le comportement de deux groupes A et B (ici les hommes et les femmes) : nous avons estimé sur notre échantillon, la part p_A des n_A individus de type A et la part p_B des n_B individus de type B qui ont fourni telle réponse ou qui ont tel comportement. L'intervalle de confiance des différences à un niveau de confiance 95 % est fourni par la formule $ICD_1^{95\ \%}$ suivante :

$$(p_A - p_B) - 1,96\sqrt{\frac{p_A \times (1-p_A)}{n_A} + \frac{p_B \times (1-p_B)}{n_B}} \leq vraie\ valeur \leq \ldots$$

$$\ldots (p_A - p_B) + 1,96 \times \sqrt{\frac{p_A \times (1-p_A)}{n_A} + \frac{p_B \times (1-p_B)}{n_B}}$$

Dans le cas de notre exemple sur le comportement cinématographique des hommes et des femmes, $p_{hommes} = 0,24$, $p_{femmes} = 0,14$, $n_{hommes} = 491$ et $n_{femmes} = 510$. L'intervalle de confiance à 95 % est donc :

$$(0,24 - 0,14) - 1,96\sqrt{\frac{0,24 \times 0,76}{491} + \frac{0,14 \times 0,86}{510}} \leq vraie\ valeur \leq \ldots$$

$$\ldots (0,24 - 0,14) + 1,96\sqrt{\frac{0,24 \times 0,76}{491} + \frac{0,14 \times 0,86}{510}}$$

soit : $0,05 \leq$ vraie valeur $\leq 0,149$

Selon toute probabilité (95 %), la différence entre le comportement des hommes et celui des femmes est comprise entre 5 % et 14,9 %. Zéro n'étant pas une valeur probable, la différence entre hommes et femmes est significative : le comportement des hommes peut être jugé significativement différent de celui des femmes.

Comme précédemment, il est possible de simplifier la formule précédente, de façon à la rendre plus commode (bien qu'un peu moins précise).

$$(p_A - p_B) - \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \leq vraie\ valeur$$

$$\leq (p_A - p_B) - 1,96\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

Il est également possible d'établir les formules associées à d'autres niveaux de confiance. Si, pour simplifier l'écriture, nous posons par convention, $S = \sqrt{\frac{p_A \times (1-p_A)}{n_A} + \frac{p_B \times (1-p_B)}{n_B}}$, les formules d'intervalle de confiance à 90 % ($ICD^{90\,\%}$) ou de 98 % ($ICD^{98\,\%}$) s'expriment ainsi :

$$(p_A - p_B) - 1,64 \times S \leq vraie\ valeur \leq (p_A - p_B) + 1,64 \times S$$
$$(p_A - p_B) - 2,34 \times S \leq vraie\ valeur \leq (p_A - p_B) + 2,34 \times S$$

Notons à nouveau que la taille de la population générale n'intervient pas dans ces formules. Et notons également, de manière plus spécifique, que la taille totale de l'échantillon n'intervient pas non plus ; seules les tailles des groupes comparés interviennent (n_A et n_B). Plus les tailles de ces groupes sont importantes, plus un écart constaté sur l'échantillon aura des chances d'être significatif.

Il n'est pas toujours commode ni même utile de calculer les intervalles de confiance à chaque fois qu'on compare deux pourcentages. Il est cependant indispensable de retenir quel est l'écart minimal garantissant, pour un échantillon de taille donné, la significativité d'une différence (pour un niveau de confiance fixé, souvent 95 %). Le tableau 3.2 indique les valeurs de cet écart minimal pour diverses tailles des deux groupes comparés[1]. En dessous de ce seuil, les pourcentages sont jugés équivalents et nous ne pouvons pas conclure à l'existence d'une différence entre les deux groupes.

Tableau 3.2. Écarts minimaux nécessaires pour juger significative une différence entre deux pourcentages (aux niveaux de confiance de 95 et *90 %*)

Tailles des sous-groupes	100	200	300	500	750	1000	2000
100	13,5 % *(11,5 %)*	12,0 % *(10 %)*	11,0 % *(9,5 %)*	10,5 % *(9,0 %)*	10,0 % *(8,5 %)*	10,0 % *(8,5 %)*	10,0 % *(8,5 %)*
200		9,5 % *(8,0 %)*	9,0 % *(7,5 %)*	8,0 % *(7,0 %)*	7,5 % *(6,5 %)*	7,5 % *(6,5 %)*	7,0 % *(6,0 %)*
300			8,0 % *(6,5 %)*	7,0 % *(6,0 %)*	6,5 % *(5,5 %)*	6,5 % *(5,5 %)*	6,0 % *(5,0 %)*
500				6,0 % *(5,0 %)*	5,5 % *(4,5 %)*	5,5 % *(4,5 %)*	5,0 % *(4,0 %)*
750					5,0 % *(4,0 %)*	4,5 % *(4,0 %)*	4,0 % *(3,5 %)*
1000						4,5 % *(3,5 %)*	3,5 % *(3,0 %)*
2000							3,0 % *(2,5 %)*

Lecture : Si deux pourcentages calculés sur deux sous-groupes de taille 500 et 750 diffèrent d'au moins 5,5 points alors il est probable (à un seuil de 95 %) qu'il existe une différence de comportement entre ces deux sous-groupes à l'échelle de la population. Au niveau de confiance de 90 % ce seuil de significativité minimale s'élève à 4,5 points.

1. En toute rigueur, comme le montrent les formules précédentes, la significativité d'une différence dépend de la taille des sous-groupes ainsi que, plus marginalement, de la valeur des pourcentages. Nous avons, dans ce tableau, négligé ce dernier aspect.

Dans son manuel exposant les principes de la démarche d'enquête par questionnaire, François de Singly précise que deux pourcentages sont significativement différents si leur différence est d'au moins 5 points[1]. Cette règle trouve ici sa justification, en même temps qu'elle devient plus précise : pour deux sous-groupes d'environ 500 personnes chacun, une différence de 5 points est significative au niveau de confiance de 90 %.

2. Le test du khi-deux (khi^2)

Devant un tableau croisé, le sociologue peut s'intéresser à l'existence d'une relation entre des modalités en ligne et des modalités en colonne. Il peut aussi, plus globalement, s'interroger sur la relation pouvant exister entre l'ensemble des modalités en ligne et l'ensemble des modalités en colonne, autrement dit entre la variable en ligne et la variable en colonne : ces deux variables entretiennent-elles une relation ? Ou bien sont-elles sans lien ?

2.1 La notion d'indépendance

Pour répondre à cette question, il est possible de faire appel à la notion de dépendance ou, plutôt, à celle, complémentaire, d'*indépendance*. Deux variables sont indépendantes s'il n'existe pas de relation entre les modalités en ligne et les modalités en colonne, ce qui revient à dire que les modalités en ligne ne conditionnent pas les modalités en colonne (et réciproquement). Autrement dit, les variables sont indépendantes si les distributions des modalités selon les diverses colonnes ne dépendent pas des modalités en ligne (et réciproquement), c'est-à-dire si les distributions des pourcentages en ligne sont identiques sur toutes les lignes (ou les pourcentages en colonnes sont identiques dans toutes les colonnes).

Considérons à nouveau le tableau 3.1 (page 68) représentant la fréquence des sorties au cinéma selon le sexe. Quelle serait l'allure de ce tableau si les deux variables étaient parfaitement indépendantes ? Il suffit de déterminer les effectifs dits « effectifs théoriques » qui devraient composer chacune des cases du tableau si les deux variables étaient parfaitement indépendantes : cela revient

1. François de Singly, *op. cit.*, 95-96.

à déterminer les effectifs des cases de telle manière que les individus soient répartis de la même manière sur chacune des lignes du tableau (ou, si on préfère, sur chacune des colonnes du tableau). Les pourcentages calculés en ligne (ou en colonne) doivent être identiques sur chacune des lignes (ou colonnes).

À la vue du tableau observé, 49 % de l'ensemble des individus ne sont jamais allés au cinéma au cours des douze derniers mois, 33 % y sont allés moins d'une fois par mois en moyenne et 18 % y sont allés au moins une fois par mois en moyenne. Si les deux variables étaient indépendantes, le nombre de femmes n'étant jamais allées au cinéma serait donc : 49 % de 510 soit $^{486}/_{1001} \times 510$ soit 248 femmes. De manière plus générale, l'effectif de chaque case du tableau d'indépendance peut être calculé de la façon suivante :

$$\textit{Effectif d'indépendance} = \frac{\textit{Effectif total de la colonne}}{\textit{Effectif total de l'échantillon}} \times \textit{Effectif total de la ligne}$$

Cet effectif est souvent appelé « effectif théorique » puisqu'il ne correspond pas à une situation empirique authentique mais seulement à un effectif abstrait si le tableau exprimait une parfaite indépendance entre les deux variables. Notons que cet effectif théorique, qui résulte d'un calcul numérique, n'est pas nécessairement un nombre entier : pour obtenir un effectif plausible d'un point de vue empirique, il est possible de l'arrondir à l'entier de plus proche.

Le calcul de ces effectifs « théoriques », auxquels nous pourrions nous attendre si les deux variables étaient indépendantes et si le hasard n'intervenait pas, permet d'établir le tableau T_{ind} dit « tableau d'indépendance » (repéré 3.3).

Tableau 3.3. Tableau d'indépendance (T_{Ind})
(les sorties au cinéma sont supposées être indépendantes du sexe)

	Jamais	Moins d'une fois par mois en moyenne	Plus d'une fois par mois en moyenne	Total
Femme	248 *49 %*	168 *33 %*	94 *18 %*	510 *100 %*
Homme	238 *49 %*	162 *33 %*	91 *18 %*	491 *100 %*
Total	486 *49 %*	330 *33 %*	185 *18 %*	1001 *100 %*

2.2 Évaluer l'effet du hasard

Dans l'immense majorité des cas, nous avons affaire à un échantillon mais ce qui nous intéresse est la relation que les variables peuvent entretenir entre elles à l'échelle de la population. Se pose alors la question, déjà soulevée au § 1.5, de l'inférence ou de l'induction d'un constat établi à l'échelle d'un échantillon à un constat dépassant la particularité de ce seul échantillon. Formulée simplement, cette question est la suivante : est-il raisonnable de conclure à l'indépendance des deux variables à la vue de la répartition de ces variables dans l'échantillon enquêté ?

Pour répondre à cette question, il faut, une fois encore, prendre conscience que l'échantillon sur lequel nous travaillons, appelé « échantillon observé », n'est qu'un échantillon parmi les très nombreux échantillons possibles. Et il faut évaluer les effets du hasard sur l'image que l'échantillon nous fournit de la population. Comme précédemment (§ 1.5), il faut faire une hypothèse, en l'occurrence une hypothèse sur la relation liant les deux variables à l'échelle de l'ensemble de la population ; puis voir si cette hypothèse est compatible avec les données de nôtre échantillon ; et enfin conclure sur la validité ou, plus exactement, la plausibilité (crédibilité) de cette hypothèse. Ce raisonnement constitue un nouveau test statistique : le *test du khi-deux* (parfois noté khi^2, chi^2, chi-carré, ou avec la lettre grecque χ^2). Ce test consiste à tester l'hypothèse d'indépendance entre les deux variables ou, dit autrement, à mesurer le niveau de plausibilité de cette hypothèse[1].

On suppose donc qu'à l'échelle de l'ensemble de la population, les deux variables n'entretiennent aucune relation de dépendance. Dès lors, quelle est la forme probable d'un échantillon d'individus choisis au sein d'une telle population ? Bien entendu, il y a relativement peu de chances que cet échantillon nous fournisse une image de parfaite indépendance entre les deux variables : le hasard du choix des enquêtés va introduire un « biais », une « distorsion » par rapport à la situation d'indépendance. De manière plus pratique, il n'est, par exemple, pas certain que l'effectif de la première cellule du tableau soit exactement égal à 248 : en raison des hasards de l'échantillonnage, l'effectif s'élèvera peut-être à 249, 250, 251, ou inversement à 247, 246, 245...

1. Le test du khi^2 peut être utilisé dans d'autres situations que le croisement de deux variables. Mais c'est cet usage qui domine en sociologie.

Mais il y a également peu de chance que cette image s'écarte considérablement de la situation d'indépendance et nous suggère l'existence d'une forte relation entre les deux variables. Il est par exemple très peu probable que l'effectif de la première cellule soit égal à 200, 150 ou *a fortiori* 100, 50...

L'hypothèse d'indépendance entre les deux variables permet de déterminer (par simulation ou, plus sûrement, par calcul) à quoi doivent ressembler les divers échantillons possibles : quels types d'échantillon peut-on obtenir et avec quelle probabilité peut-on les obtenir ? Cela revient à identifier les différents types de tableaux et leurs probabilités d'apparition. Il suffit alors de déterminer dans quel type d'échantillon se situe notre échantillon observé (celui que nous avons effectivement obtenu et sur lequel nous travaillons). Cela revient à déterminer quel est le type du tableau dit « tableau observé » résultant de l'échantillon. Deux cas se présentent :

– Si notre échantillon observé (ou tableau observé) est semblable à un échantillon (ou tableau) qu'il est relativement probable d'obtenir lorsque les deux variables sont indépendantes, alors il est raisonnable de penser que l'hypothèse d'indépendance doit être acceptée (les statisticiens disent qu'elle ne peut pas être rejetée).

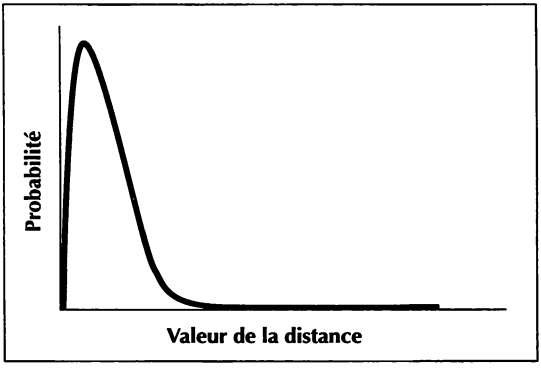
– Inversement, si notre échantillon observé (tableau observé) s'apparente à un type d'échantillon (tableau) peu ou très peu probable lorsque les deux variables sont indépendantes, alors il est raisonnable de penser que l'hypothèse d'indépendance doit être écartée (les statisticiens parlent de « rejet »).

2.3 Une distance pour juger de la proximité entre tableaux

D'un point de vue pratique, il suffit d'expliciter la manière dont on va comparer les échantillons – ou plus exactement les tableaux. Pour cela, on fait appel à la notion de distance : une distance est un indicateur qui prend des petites valeurs lorsque les deux situations comparées sont proches et donc se ressemblent, et qui prend de grandes valeurs lorsque deux situations sont éloignées et donc différentes. L'indicateur de distance utilisé ici est la « distance du khi^2 » : il permet d'apprécier la ressemblance ou la dissemblance de deux tableaux. En l'occurrence, il permet d'apprécier la distance, et donc la ressemblance ou la dissemblance, entre le tableau d'indépendance et le tableau issu de tout autre échantillon.

Le raisonnement précédent peut être reformulé en faisant appel à cette notion de distance du khi^2. En supposant que les deux variables sont indépendantes, il est possible (une fois de plus par simulation ou, plus sûrement, par calcul) de déterminer à quoi doivent ressembler les divers échantillons possibles. Nous avions dit que cela revenait à identifier les différents types de tableaux et leurs probabilités d'apparition. Nous pouvons maintenant reformuler cette dernière affirmation en disant que « cela revient à identifier les diverses valeurs possibles de la distance du khi^2 entre le tableau de parfaite indépendance et chacun des types de tableaux possibles ». Autrement dit, il est possible de déterminer quelles sont les valeurs de la distance du khi^2 si les deux variables sont indépendantes et avec quelle probabilité chacune de ces valeurs peut apparaître. D'un point de vue formel, cela revient à construire une courbe où, pour chaque valeur de la distance du khi^2, on indique sa probabilité d'apparition. Cette courbe est ici représentée figure 3.1.

Figure 3.1. La courbe de la distance du khi^2



Cette courbe est une courbe abstraite dans la mesure où elle n'est pas établie à partir de notre seul échantillon : elle est le reflet de la façon dont la distance du khi^2 se comporte, c'est-à-dire le reflet de ce qui pourrait se passer si on recommençait l'échantillonnage un très grand nombre de fois et qu'on calculait la distance à chaque fois. Elle est abstraite mais exacte : il ne s'agit en aucune manière d'une approximation ou d'une extrapolation mais bien du calcul probabiliste (à peine plus compliqué que le calcul des probabilités de gagner au Loto ou à tout autre jeu de hasard). Rappelons que cette courbe est

construite en supposant (c'est notre hypothèse) que les deux variables sont indépendantes.

Cette courbe exprime bien des résultats intuitifs. Si les deux variables sont effectivement indépendantes, il est très peu probable d'obtenir un échantillon qui s'éloigne beaucoup de la situation d'indépendance ; il est donc peu probable d'obtenir une distance du khi^2 très élevée – ce que la partie droite de la courbe restitue bien. Inversement, il est très probable d'obtenir un échantillon qui s'éloigne peu de la situation d'indépendance ; il est donc plutôt probable d'obtenir une distance du khi^2 faible – ce que la partie gauche de la courbe restitue (même s'il est peu probable d'obtenir un échantillon ne présentant aucun écart par rapport à l'échantillon de parfaite indépendance – distance nulle).

Pour tirer les conséquences pratiques de ce calcul, il suffit de calculer la distance entre le tableau d'indépendance et notre propre tableau empirique (celui que nous avons effectivement obtenu par enquête) puis de déterminer où se situe cette valeur de la distance parmi l'ensemble des valeurs possibles. Les deux cas envisagés précédemment deviennent maintenant :

– Si notre distance du khi^2 est faible : notre valeur fait partie des valeurs qu'il est vraisemblable d'obtenir si les deux variables sont indépendantes. Il est donc raisonnable de penser que l'hypothèse d'indépendance ne peut pas être écartée. On conclut alors à une indépendance entre les deux variables.

– Inversement, si notre distance du khi^2 est élevée : notre valeur fait partie des valeurs qu'il est peu vraisemblable d'obtenir si les deux variables sont indépendantes. Il est donc raisonnable de penser que l'hypothèse d'indépendance n'est pas plausible. On écarte donc cette hypothèse pour conclure à une dépendance entre les deux variables. Notons qu'en toute rigueur (épistémo-) logique, rejeter une hypothèse ce n'est pas accepter une hypothèse apparemment inverse ou contraire. C'est pourtant ce que font la plupart des utilisateurs du test du khi^2. Et c'est un usage courant et accepté en sociologie.

Reste une dernière question, sous-tendue par ce raisonnement : à partir de quelle valeur considère-t-on que la distance est « grande » ou « petite », « faible » ou « élevée »…

Tout en ne faisant pas encore appel aux formules exactes (que nous verrons plus loin), il est nécessaire de savoir que la répartition de la distance du khi^2 dépend non seulement de la dépendance ou de l'indépendance des deux variables entre elles mais aussi de la taille de l'échantillon et de la taille du tableau (nombre de lignes et nombre de colonnes). Ainsi, plus le tableau sera de grande taille plus la distance du khi^2 sera, structurellement, élevée (puisque le nombre

de termes intervenant dans la somme sera grand comme nous le verrons plus loin). Et plus l'effectif de l'échantillon sera élevé, plus la distance du khi^2 sera grande (puisque la distance s'exprime à partir des effectifs comme nous le verrons également plus loin). D'un point de vue technique, cela signifie que la courbe présentant l'ensemble des valeurs possibles de la distance du khi^2 prend des valeurs différentes tout en conservant la même forme générale.

Pour surmonter cette difficulté, il ne faut pas s'attacher à la valeur de la distance mais à la valeur de la probabilité associée à cette distance : « Plutôt que de juger à la vue de la distance on juge à la vue de la probabilité. » En effet, la graphique 3.1 permet de déterminer, pour chaque valeur de la distance, une probabilité qui s'interprète comme la probabilité d'obtenir un échantillon dont la distance à la situation de parfaite indépendance est au moins aussi élevée que notre distance (celle calculée sur notre échantillon observé). En d'autres termes, cette probabilité est la part des échantillons dont la distance à la situation de parfaite indépendance est au moins aussi grande que la distance que nous avons obtenue à partir de notre échantillon observé. Elle s'interprète comme la probabilité d'obtenir un échantillon au moins aussi éloigné de la situation d'indépendance que celui que nous avons effectivement obtenu empiriquement.

Si cette probabilité (dite « probabilité associée à la distance du khi^2 ») est faible et si les deux variables sont réellement indépendantes, il est peu probable d'obtenir un échantillon au moins aussi éloigné de la situation d'indépendance que le nôtre. L'alternative est donc la suivante : soit considérer que notre échantillon est issu d'une situation où les deux variables sont réellement indépendantes mais que nous n'avons vraiment pas eu de chance ; soit considérer que l'hypothèse d'indépendance n'est pas plausible. En pratique, il est usuel de considérer que la « chance » ne nous a pas été trop défavorable et donc que c'est l'hypothèse d'indépendance qui n'est pas plausible.

Inversement, si la probabilité associée à la distance du khi^2 est élevée et si les deux variables sont réellement indépendantes, il était tout à fait probable d'obtenir un échantillon au moins aussi éloigné de la situation d'indépendance que le nôtre. Dès lors, l'hypothèse d'indépendance est tout à fait plausible et nous l'acceptons.

Reste simplement à savoir ce qui peut être considéré comme une « probabilité » faible. L'usage, qui reflète à la fois une norme collective et des contraintes techniques de taille d'échantillon, est de considérer qu'une probabilité est faible si elle est inférieure à 5 % (0,05). On tolère parfois 10 % (0,10) et, inversement, les sociologues travaillant sur de gros échantillons et souhaitant être très prudents dans leurs affirmations, considèrent parfois qu'un seuil

de 1 ou 2 % est nécessaire. Ces seuils sont des seuils jugés acceptables par la communauté des sociologues. Ils sont différents dans l'espace de la recherche médicale ou biologique, comme dans l'industrie : un industriel construisant des avions ne peut pas se permettre de tolérer des niveaux de risque si élevés lorsqu'il teste la résistance ou le comportement des pièces mécaniques !

2.4 Le calcul de la distance et de l'effet du hasard

Nous avons, jusqu'à présent, exposé le principe et la logique du test du khi^2 sans recourir aux formalismes mathématiques. Sans que cela change quoi que ce soit à ce principe et à cette logique, le recours à ces formalismes permet de préciser certains aspects techniques du test, notamment le calcul de la distance qui doit permettre de juger de la ressemblance ou de la dissemblance entre la situation observée (T_{Obs}) et la situation d'indépendance (T_{Ind}, tableau que nous pourrions obtenir si les variables étaient réellement indépendantes et si le hasard n'intervenait pas).

Tableau 3.4. Effectifs observés (T_{Obs}) et effectifs d'indépendance (T_{Ind})

Obs / *Ind*	Jamais	Moins d'une fois par mois en moyenne	Plus d'une fois par mois en moyenne	Total
Femme	261 / *248*	180 / *168*	69 / *94*	510 / *510*
Homme	225 / *238*	150 / *162*	116 / *91*	491 / *491*
Total	486 / *486*	330 / *330*	185 / *185*	1001 / *1001*

Pour cette ressemblance ou au contraire cette dissemblance entre T_{ind} et T_{obs}, il faut comparer les effectifs : à cette fin, on calcule les écarts (dit absolus) entre les effectifs observés et les effectifs théoriques selon la formule suivante :

écart absolu = effectif observé – effectif théorique

Le tableau 3.5 présente les valeurs de ces écarts sur notre exemple. Les valeurs positives figurant dans ce tableau s'interprètent comme des attractions entre modalités. Par exemple, la modalité « jamais » est davantage représentée chez les femmes qu'elle ne devrait l'être si les deux variables étaient indépendantes : il y

a 13 femmes « en trop » par rapport à la situation d'indépendance. Inversement, les valeurs négatives s'interprètent comme des « répulsions » entre modalités : les modalités « homme » et « jamais » se « répulsent ».

Tableau 3.5. Tableau des écarts absolus ($T_{Obs} - T_{Ind}$)

	Jamais	Moins d'une fois par mois en moyenne	Plus d'une fois par mois en moyenne
Femme	+ 13 (= 261 – 248)	+ 12	– 25
Homme	– 13	– 12	+ 25

Toutefois, il est délicat de juger de l'importance c'est-à-dire de la force de l'attraction ou de la répulsion entre modalités à la seule vue de ces écarts : un écart de 10 individus entre 200 et 210 n'a pas la même signification d'un écart de 10 individus entre 50 et 60. Il est nécessaire de tenir compte des effectifs sur lesquels les écarts sont calculés. On définit donc l'écart relatif en divisant l'écart absolu par l'effectif (en l'occurrence l'effectif théorique) :

$$\textit{écart relatif} = \frac{\textit{effectif observé} - \textit{effectif théorique}}{\textit{effectif théorique}}$$

Les écarts relatifs expriment bien l'intensité de la répulsion ou de l'attraction entre modalités des deux variables. Ainsi, à la vue du tableau de ces écarts relatifs (tableau 3.6), nous pouvons constater que la répulsion entre la modalité « femme » et la modalité « plus d'une fois par mois en moyenne » est forte ; plus forte en tout cas que l'attraction de la modalité « femme » et les deux autres modalités. Et le constat étant semblable pour les hommes, nous pouvons penser que la différence entre le comportement des hommes et celui des femmes résulte surtout de leur opposition dans la dernière colonne. Les écarts relatifs peuvent utilement aider le sociologue dans son interprétation des tableaux croisés, même s'il est très rare de les utiliser dans des publications.

Tableau 3.6. Tableau des écarts relatifs

	Jamais	Moins d'une fois par mois en moyenne	Plus d'une fois par mois en moyenne
Femme	+ 0,052	+ 0,071	– 0,266
Homme	– 0,055	– 0,074	+ 0,275

En somme, nous venons de voir comment l'écart entre deux cases pouvait être mesuré : soit par l'écart absolu qui exprime une distance « absolue » ou « brute » et qui représente un nombre d'individus (donc un « poids ») ; soit l'écart relatif qui exprime un lien de répulsion ou attraction et qui représente l'intensité de ce lien[1]. Une bonne façon de faire la synthèse de ces deux aspects est de tenir compte à la fois de l'écart absolu (c'est-à-dire des effectifs) et de l'écart relatif (c'est-à-dire de l'intensité du lien) en les multipliant. Cette multiplication de nos deux indicateurs d'écart permet de supprimer les signes négatifs : que les écarts soient positifs ou négatifs, ils expriment une différence et c'est le niveau de cette différence qui nous intéresse, quel que soit son signe (positif ou négatif).

Cette multiplication a également le mérite de tenir compte de l'intensité d'un lien d'attraction ou de répulsion tout en pondérant l'intensité de ce lien par les effectifs sur lesquels elle porte afin de ne pas accorder trop d'importance à des liens portant sur des effectifs trop faibles et, inversement, d'accorder toute son importance aux intensités de liens portant sur de nombreux individus.

Le résultat de cette multiplication, qui est peut-être interprété comme une distance, s'exprime formellement de la manière suivante.

$$\begin{aligned} &\textit{distance entre deux cases} \\ &= \textit{écart absolu} \times \textit{écart relatif} \\ &= (\textit{effectif observé} - \textit{effectif théorique}) \times \left(\frac{\textit{effectif observé} - \textit{effectif théorique}}{\textit{effectif théorique}} \right) \\ &= \frac{(\textit{effectif observé} - \textit{effectif théorique})^2}{\textit{effectif théorique}} \end{aligned}$$

Cette formule vérifie bien les propriétés que nous pouvions attendre d'une distance : elle est toujours positive et, surtout, elle est d'autant plus élevée que les écarts (absolu et relatif) entre une cellule observée et la cellule théorique correspondante sont grands.

Revenons au khi^2 et à la définition de la distance du khi^2, qui doit nous permettre d'estimer l'écart global entre les tableaux T_{Obs} et T_{Ind}. La distance

1. Philippe Cibois propose, avec son indicateur PEM (Pourcentage de l'écart maximum), un autre outil pour estimer l'attraction ou la répulsion de deux modalités. Son emploi est malheureusement peu répandu. Voir : Philippe Cibois, « Le PEM, pourcentage de l'écart maximum : un indice de liaison entre modalités d'un tableau de contingence », *Bulletin de méthodologie sociologique*, n° 40, 1993, p. 43-63 (article consultable sur le site Web de l'auteur).

du khi^2 doit exprimer une distance entre l'ensemble des cases du tableau observé et l'ensemble des cases du tableau théorique. Elle peut simplement être définie comme la somme des distances entre cases c'est-à-dire comme la somme des contributions de chaque cellule. Ces distances entre cases sont d'ailleurs appelées « contributions au khi^2 ». Le tableau 3.7 présente les valeurs de ces contributions pour notre exemple.

Tableau 3.7. Tableau des contributions de chaque cellule au khi^2

	Jamais	Moins d'une fois par mois en moyenne	Plus d'une fois par mois en moyenne
Femme	0,68	0,86	6,65
Homme	0,71	0,89	6,87

Formellement, la formule de la distance du khi^2 entre le tableau observé et le tableau théorique s'exprime ainsi[1] :

$$distance\ du\ khi^2 = \sum_{ensemble\ des\ cellules} \frac{(effectif\ observé - effectif\ théorique)^2}{effectif\ théorique}$$

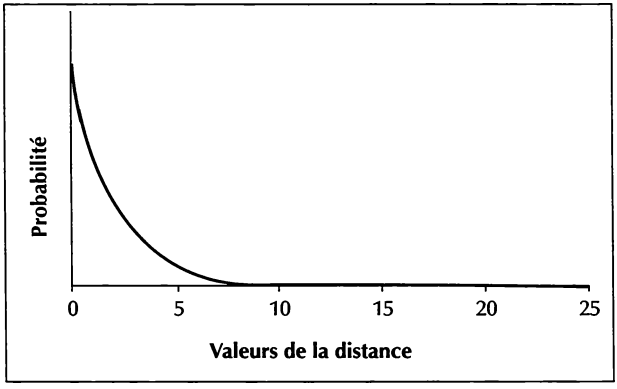
Conformément à ce que nous avions annoncé, cette distance dépend de la taille du tableau (c'est-à-dire du nombre de lignes et de colonnes) : plus le tableau est grand, plus le nombre de contributions est grand et donc, mécaniquement, plus la distance du khi^2 sera grande[2]. Cette distance dépend également de la taille de l'échantillon : plus l'échantillon est de grande taille, plus la distance est, structurellement, élevée. Pour s'en convaincre, il suffit de calculer la distance sur un tableau où tous les effectifs ont été multipliés par 10 : la distance du khi^2 augmente elle-même d'un facteur 10.

Ces deux constats nous avaient précédemment conduits à ne pas accorder trop d'importance à la valeur de cette distance en tant que telle mais à privilégier la probabilité associée à cette distance. Dans le cas de notre exemple, la distance du khi^2 vaut 0,68 + 0,86 + 6,65 + 0,71 + 0,89 + 6,87 soit 16,66.

1. Le symbole Σ est une façon commode d'écrire « somme de toutes les valeurs sur l'ensemble des cellules ».
2. Les statisticiens et les logiciels statistiques utilisent la notion de « degré de liberté » (*ddl*) pour exprimer cette idée de taille du tableau. Le *ddl* est égal à (nombre de colonnes – 1) × (nombre de lignes – 1). Plus de tableau est de grande taille (nombreuses lignes, nombreuses colonnes) plus le *ddl* est élevé.

La probabilité associée est la probabilité d'obtenir un échantillon (observé) dont la distance au tableau d'indépendance est au moins aussi grande.

Figure 3.2. La distribution des valeurs de la distance du khi2 si les deux variables sont indépendantes



La valeur de la distance (16,66) se situe clairement du côté des valeurs qu'il est peu, voire très peu probable d'obtenir si les variables sont indépendantes. En l'occurrence, la probabilité d'obtenir un tableau au moins aussi éloigné de la situation d'indépendance que le nôtre s'élève environ à 0,0002. Autant dire que ce n'est probablement pas le hasard qui explique la distribution des effectifs dans le tableau : une autre explication s'impose. Cette explication est évidemment que l'hypothèse d'indépendance n'est pas acceptable et donc que les variables sont dépendantes l'une de l'autre : il existe un lien entre le sexe et la fréquentation des cinémas.

2.5 Une illustration pratique

En pratique, le sociologue ne parcourt pas tout le cheminement que nous avons suivi pour justifier et exposer la logique du test du khi^2. Tout logiciel statistique réalise les calculs nécessaires : écarts ; écarts relatifs ; contributions aux khi^2 ; distance ; probabilité associée à cette distance... La seule information intéressant le sociologue est la probabilité : plus elle est faible (inférieure à 10 % ou

5 % selon l'usage) moins l'hypothèse d'indépendance entre les deux variables est crédible et donc plus l'hypothèse de l'existence d'un lien entre ces variables est acceptable.

Les commentaires d'un test du khi^2 peuvent prendre diverses formes, en fonction du contexte et du public visé.

– *Voici une formulation résumant l'ensemble du raisonnement :* si les variables étaient indépendantes, la probabilité d'obtenir un échantillon conduisant à une distance du khi^2 au moins égale à la distance obtenue sur notre échantillon (16,66) serait de 0,0002. Cette probabilité est inférieure au seuil de risque (5 %) donc l'hypothèse d'indépendance n'est pas « raisonnable » et on la rejette.

– *Voici une formulation rigoureuse mais plus allusive* : la distance du Khi^2 étant de 16,66 et la probabilité associée de 0,0002, on rejette l'hypothèse d'indépendance.

– *Voici enfin une formulation minimale* : à la vue du test du khi^2, les deux variables semblent dépendre l'une de l'autre.

Dans les articles ou ouvrages de sociologie, il est de plus en plus fréquent de ne pas commenter chaque test statistique réalisé mais d'affecter à chaque résultat testé des étoiles représentant, en fonction d'une convention fixée en début d'article, la significativité du test associé. Il est par exemple usuel d'attribuer trois étoiles (***) aux résultats testés avec une probabilité inférieure à 0,01 (soit 1 %) ; deux étoiles (**) aux résultats ayant une probabilité associée inférieure à 0,05 (soit 5 %) ; et enfin une seule étoile (*) pour les probabilités inférieures à 0,1 (soit 10 %). Tous les résultats moins fiables, c'est-à-dire pour lesquels la probabilité associée est supérieure à 0,1, ne sont pas jugés significatifs. Ces étoiles accompagnent les données statistiques dans les tableaux, graphiques ou commentaires.

2.6 Intérêts et limites

Le test du khi^2 est très utile pour indiquer l'existence d'une relation de dépendance entre deux variables. Mais il ne constitue pas un indicateur de l'intensité de cette dépendance. La probabilité associée à l'hypothèse d'indépendance ne permet pas de hiérarchiser les relations entre variables en identifiant celles qui sont fortement liées et celles qui le sont un peu moins : la probabilité indique la confiance qu'il est possible d'accorder à l'hypothèse d'indépendance et non l'intensité de leur éventuelle dépendance.

Le test du khi^2 ne constitue pas non plus un indicateur du sens de la relation : la conclusion issue de notre exemple précédent est qu'il existe un lien entre le sexe et la fréquentation des cinémas. Mais nous ne savons rien de ce lien : les femmes y vont-elles davantage que les hommes ? Ou bien est-ce le contraire ? Ce n'est pas le test du khi^2 qui permet de répondre à cette question, mais la lecture du tableau des pourcentages ou du tableau des écarts à l'indépendance.

Enfin, quatre limites théoriques et pratiques à l'utilisation du test du khi^2 méritent d'être signalées. Premièrement, le tableau croisé doit être un tableau de contingence : un individu doit être présent dans une et une seule cellule du tableau. Cela interdit d'avoir recours au test du khi^2 sur des tableaux croisant une ou deux variables multiples.

Deuxièmement, les effectifs doivent être suffisants pour que nous puissions juger des effets du hasard et, donc, distinguer ce qui relève du hasard et ce qui n'en relève pas. Imaginons que l'effectif théorique d'une cellule du tableau soit seulement de 4 personnes. Si l'effectif observé de cette même cellule est de 2 ou de 6 personnes, l'écart relatif est important mais peut-on juger sur de si petits effectifs ? De toutes petites fluctuations dans l'échantillonnage vont considérablement changer la distance du khi^2 et donc les conclusions du test. Pour cette raison, il est préférable de ne pas utiliser le test du khi^2 dès que le tableau est trop « creux ». Les recettes pour déterminer si un tableau est trop creux sont nombreuses – presque aussi nombreuses que les manuels ou les statisticiens. La plus prudente est certainement de suivre le principe proposé par Philippe Cibois[1] : examiner les contributions des diverses cases du tableau à la distance totale du khi^2 afin d'identifier les cases qui, à elles seules, expliqueraient l'essentiel de la distance du khi^2 ; et si seules une ou deux cases expliquent la valeur de la distance, s'interroger sur la nature de la relation unissant les modalités correspondantes.

Troisièmement, comme tout test statistique, le test du khi^2 n'est pas une preuve absolue de la présence ou de l'absence d'une dépendance entre deux variables. Il ne fournit que des présomptions de relations… qui devront être étayées par d'autres analyses, d'autres croisements…

Enfin, quatrièmement, l'existence d'une relation statistique entre deux variables ne signifie pas que cette relation ait un sens empirique ou sociologique. Son interprétation reste à faire.

1. Philippe Cibois, « Faire comprendre le khideux », *Bulletin de méthodologie sociologique*, 2001, p. 37-45.

3. LE COEFFICIENT DE CORRÉLATION LINÉAIRE

Lorsque deux variables sont de nature quantitative, il est possible de les recoder en créant des classes de valeur afin de se situer dans le cas précédent : celui de deux variables qualitatives, de leur tableau croisé et du test du khi^2 associé. Mais il est également possible de conserver le statut quantitatif des variables en recourant au coefficient de corrélation : ne pas recoder permet de conserver toute l'information (et toute sa précision, si on la juge utile)[1]. Le coefficient de corrélation (dit de Bravais-Pearson) est un indicateur tellement classique en statistique que le terme de « corrélation », initialement forgé pour désigner la co-relation de deux variables quantitatives telle qu'elle est exprimée par ce seul coefficient, est régulièrement utilisé pour désigner l'idée que deux phénomènes (ou deux variables) sont reliés l'un à l'autre : il est fréquent de rencontrer des expressions telles que « la corrélation entre le développement économique et l'épanouissement personnel » ou « la corrélation entre la création artistique et les structures matérielles et culturelles de la société », sans que ces expressions fassent réellement référence au calcul du coefficient de corrélation. Ce coefficient est particulièrement utile lorsque les variables analysées expriment des quantités de temps (âge, durée entre deux événements...), des quantités monétaires (revenus, patrimoine, dépenses, consommation...), des fréquences ou des indicateurs synthétiques construits pour les besoins de l'enquête.

Il importe de bien comprendre quel est le type de relation que le coefficient de corrélation permet de mettre en évidence car, à la différence du test du khi^2 qui permet de savoir si deux variables sont indépendantes ou pas sans faire d'hypothèse sur la nature de leur relation (dépendance), le coefficient de corrélation cherche à identifier un type tout à fait précis de relation : la relation linéaire. En toute rigueur, il serait préférable de parler du coefficient de corrélation linéaire et non, simplement, du coefficient de corrélation – puisque cette dernière expression laisse entendre que la relation mesurée par le coefficient peut être de nature quelconque alors qu'elle est exclusivement de nature linéaire.

---

1. Il existe un coefficient de corrélation dit « des rangs » ou « de Spearman » permettant de traiter le cas intermédiaire : celui où les variables sont ordonnables. La corrélation des rangs revient à calculer le coefficient de corrélation sur les numéros de classement (1er, 2^e, 3^e...).

3.1 La notion de relation linéaire

Deux variables entretiennent une relation linéaire entre elles si la variation relative de l'une d'entre elles entraîne immanquablement une variation relative constante de l'autre. Autrement dit, les variables X et Y sont linéairement liées si une variation de p % de X entraîne toujours une variation constante de q % de Y. Par exemple la consommation de livres (estimée par le nombre de livres achetés par an) est liée linéairement à leur prix si une variation de 10 % du prix des livres entraîne immanquablement une variation de q % de leur consommation : la valeur de q peut être négative (dans ce cas, la consommation baisse si le prix augmente) ou positive (dans ce cas, la consommation croît si le prix augmente) ; q peut valoir – 15 %, – 9 %, – 3 %, 5 %, 11 %, 17 %... L'important est ici que cette variation soit toujours identique si le prix des livres augmente de 10 % pour passer de 10 à 11 €, ou de 20 à 22 €, ou de 50 à 55 €... Mathématiquement, la propriété de liaison linéaire entre deux variables s'écrit : $Y = a \times X + b$ (où a et b sont des nombres constants). Notons que cette relation n'a pas de symétrique (X et Y jouent des rôles identiques) : si X et Y sont liées linéairement, Y et X le sont également. Ainsi si $Y = a \times X + b$ alors : $X = {}^{Y}/_{a} - {}^{b}/_{a}$: c'est une autre relation linéaire ($X = a' \times Y + b'$ avec $a' = 1/a$ et $b' = -b/a$).

3.2 La notion de co-variation

Pour exprimer la variation (variabilité) d'une variable, nous avons vu qu'il était possible de recourir à l'indicateur de variance ou d'écart-type. Il est possible de concevoir un indicateur comparable pour exprimer la co-variation de deux variables c'est-à-dire pour rendre compte de leurs variations simultanées. Si ces variations simultanées sont élevées, les variables sont probablement liées (puisqu'une variation de l'une est associée, presque systématiquement, à une variation de l'autre – comme par exemple dans le cas du poids et de la taille des individus). Si ces variations simultanées sont faibles, les variables sont probablement indépendantes (puisque l'une peut varier sans que l'autre le fasse).

L'indicateur de co-variation est appelé covariance et est défini comme la moyenne des produits des écarts à la moyenne de chaque variable. D'un point de vue formel, si nous avons observé les variables X et Y pour n individus, si

pour chaque individu repéré i les valeurs observées sont notées X_i et Y_i, et si $\overline{X}$ et $\overline{Y}$ sont les moyennes calculées sur l'ensemble des individus, alors la covariance est définie par la relation suivante :

$$covariance = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}) \times (Y_i - \overline{Y})$$

La covariance prend des valeurs positives d'autant plus élevées que les deux variables varient simultanément dans le même sens (une hausse de l'une est associée à une hausse de l'autre ; une baisse de l'une est associée à une baisse de l'autre). Elle prend des valeurs négatives d'autant plus petites (c'est-à-dire éloignées de zéro) que les variables varient simultanément dans des sens contraires (une baisse de l'une est associée à une hausse de l'autre).

3.3 La définition du coefficient de corrélation linéaire

Le coefficient de corrélation (noté r_{XY}) est défini comme le rapport entre la covariance des deux variables et les écarts-types de ces mêmes variables. En somme, le coefficient de corrélation est une mesure de la co-variation mutuelle de X et de Y, compte tenu de la variation propre de X et de la variation propre de Y.

$$r_{XY} = \frac{covariance\ de\ X\ et\ de\ Y}{(écart\text{-}type\ de\ X) \times (écart\text{-}type\ de\ Y)}$$

En tenant compte des définitions présentées précédemment, la définition r_{XY} peut être exprimée de la manière suivante (à condition de simplifier par n en haut et en bas) :

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X}) \times (Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2} \times \sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

Le coefficient de corrélation linéaire r_{XY} permet de mesurer jusqu'à quel point deux variables entretiennent une telle relation linéaire entre elles. Ses valeurs sont toujours comprises entre –1 et 1.

– Le coefficient calculé entre deux variables X et Y vaut –1 si X et Y sont liées par une relation linéaire parfaite mais qu'une croissance de X est associée

à une décroissance de Y (et inversement : une décroissance de X est associée à une croissance de Y) ;

– Le coefficient vaut 1 si X et Y entretiennent une relation linéaire parfaite et une croissance de X est associée à une croissance de Y. Il est possible de parler de relation linéaire croissante ou positive ;

– Le coefficient vaut 0 si X et Y n'entretiennent pas de relation linéaire : X et Y sont dites linéairement indépendantes.

– En dehors de ces situations idéales-typiques, le coefficient de corrélation exprime une plus ou moins grande proximité avec chacune de ces situations. Si, par exemple, le coefficient vaut 0,9, il est aisé de considérer que les deux variables entretiennent une relation presque parfaitement linéaire ; si le coefficient vaut 0,1 ou – 0,1, les deux variables ne sont quasiment pas liées de manière linéaire ; si le coefficient vaut – 0,9, les deux variables sont liées par une relation presque parfaitement linéaire mais décroissante.

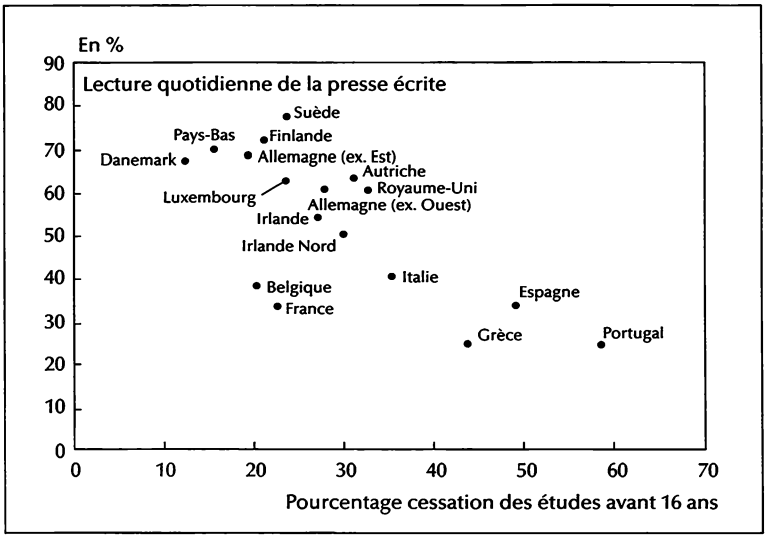
Il est également possible d'interpréter r_{XY}, ou plus exactement son carré r^2_{XY}, comme une mesure de la variabilité de Y expliquée par X (et réciproquement), c'est-à-dire la part des variations de Y qui peut être expliquée par les variations de X (et réciproquement). Plus cette part est élevée, plus Y est expliquée par X (et réciproquement). Ce coefficient r^2_{XY} est appelé « coefficient de détermination » : il est toujours compris entre 0 et 1 : plus il est proche de 1, plus X permet d'expliquer les variations de Y ; plus il est proche de 0, moins X permet d'expliquer Y (et réciproquement).

3.4 Une illustration pratique

Considérons par exemple, pour divers pays européens, deux indicateurs quantitatifs : le pourcentage d'individus cessant leurs études avant 16 ans (qui est un indice du niveau scolaire du pays) et la part de ceux lisant quotidiennement la presse écrite[1]. Le graphique 3.3 fournit une représentation de ces données.

1. Louis Chauvel, « Les Européens et l'information », *Revue de l'OFCE*, n° 69, avril 1999, p. 277-285.

Graphique 3.3. Niveau scolaire et lecture de la presse



Source : Revue de l'OFCE, n° 69, 1999, p. 282.

Les deux variables considérées ne sont pas linéairement liées : pour cela, il aurait fallu que les points représentant les différents pays soient parfaitement alignés. La répartition de ces points ne semble toutefois pas aléatoire : ils sont disposés selon une direction partant du haut gauche pour aller vers le bas droit. Cette représentation semble suggérer l'existence d'une relation entre le niveau scolaire d'un pays et les pratiques de lecture de la presse. Le calcul du coefficient de corrélation, qui vaut ici – 0,57, confirme cette impression : « Moins un [pays] compte de personnes faiblement scolarisées, plus les habitudes de lecture dont importantes[1]. »

Ici, le coefficient de détermination vaut 0,50 : la moitié de la variation des pratiques de lecture peut être expliquée de la variation de niveau scolaire. Le reste des variations n'est pas imputable au niveau scolaire et devra, si besoin, être expliqué par d'autres facteurs. Le coefficient de détermination permet d'apprécier, quantitativement, la force d'une explication et d'une relation et donc, inversement, la part de la variation qui reste à expliquer.

1. Louis Chauvel, *op. cit.*, p. 282.

3.5 Intérêts, usages et limites

Le principal intérêt du coefficient de corrélation est de fournir une indication de l'intensité de la relation (linéaire) qu'entretiennent deux variables. À la différence du khi^2, qui livre simplement une indication sur la plausibilité de l'hypothèse d'indépendance entre deux variables sans apporter de renseignement sur la force de leur éventuelle dépendance, les valeurs des coefficients de corrélation peuvent être interprétées comme des intensités : elles permettent donc de hiérarchiser les relations entre variables.

Le coefficient de corrélation est très fréquemment utilisé dès que les variables sont de nature quantitatives : d'une part, parce qu'il est simple à calculer et à interpréter ; d'autre part, parce que dans le cas de petites variations, il est toujours possible de considérer, par approximation, que les variations sont linéaires – car, à l'échelle des faibles variations de valeur, toute courbe reliant deux variables X et Y peut être considérée comme un morceau de droite. Une autre raison de son succès est qu'il est connu par tous : il fait partie des tout premiers outils statistiques présentés dans tous les cours de statistiques. Cette familiarité n'a toutefois pas que des avantages : elle conduit à oublier les conditions d'utilisation et le sens réel de ce coefficient pour ne voir en lui que « la » méthode de mesure des co-relations entre variables quantitatives.

La principale limite du coefficient de corrélation a déjà été signalée – elle est contenue dans sa définition : il ne permet d'identifier que les relations de nature linéaire entre deux variables et signale la présence ou l'absence d'une telle relation, sans fournir la moindre indication sur la présence d'une relation d'un tout autre type entre deux variables. Par exemple, le coefficient de corrélation linéaire entre l'âge des enquêtés et le nombre de fois où ils ne sont rendus au cinéma au cours des douze derniers mois peut être nul sans pour autant signifier qu'il n'existe aucun lien entre l'âge et la fréquentation des cinémas : le nombre de films vus au cinéma au cours d'une année peut croître progressivement de 12 à 25 ans pour baisser de 25 à 45 ans et croître à nouveau au-delà de 45 ans... Le coefficient de corrélation linéaire sera incapable de restituer une telle relation.

Dans tous les cas, il est bon voire indispensable de se faire une idée de la relation qu'entretiennent deux variables en construisant une représentation graphique des valeurs de ces deux variables – comme dans l'exemple précédent. Un graphique permet d'identifier les éventuelles relations entre deux variables, même si ces relations ne sont pas linéaires.

Une autre limite classiquement attribuée au coefficient de corrélation est que s'il permet de renseigner le sociologue sur l'éventuelle présence d'une

relation linéaire entre deux variables, il ne dit rien sur la nature causale ou non de cette relation. Cette limite n'est pas propre à ce coefficient : la critique vaut également pour le test du khi^2. De manière générale, la recherche de relation causale est hors de portée des seuls outils statistiques simples.

Notons enfin que le coefficient de corrélation peut, comme tout indicateur statistique, faire l'objet d'un test destiné à déterminer si la valeur de ce coefficient calculé sur un échantillon fournit une information fiable, c'est-à-dire valant, de manière probable, pour l'ensemble de la population.

4. Une variable qualitative et une variable quantitative

Après le cas où les deux variables sont qualitatives puis celui où elles sont quantitatives, un dernier cas de figure peut survenir : une des deux variables est quantitative et l'autre est qualitative. Comment apprécier la relation qu'une variable quantitative et une variable qualitative entretiennent ? En d'autres termes, et en considérant que les diverses modalités de la variable qualitative correspondent à des groupes d'individus, comment comparer les valeurs prises par une variable quantitative au sein de différents groupes ?

Ce cas de figure se rencontre lorsqu'on cherche à identifier le lien entre le niveau de diplôme (variable qualitative) et le revenu mensuel (variable quantitative), le milieu social (qualitatif) et la consommation d'alcool (quantitatif), le type de lycée (public ou privé) et la note moyenne obtenue au baccalauréat... Pour les besoins de notre présentation, nous allons étudier la relation que peuvent entretenir, chez les adolescents, le sexe et le nombre d'activités sportives pratiquées de manière régulière[1]. Cela revient à comparer le groupe des garçons avec celui des filles.

La méthode habituellement utilisée pour estimer la présence ou l'absence d'un tel lien est appelée l'analyse de la variance, souvent abrégée en ANOVA (*ANalysis Of VAriance*). Il s'agit d'un test, permettant d'aboutir à l'acception ou au rejet d'une hypothèse, en l'occurrence l'hypothèse qu'il n'y a pas de lien entre la variable qualitative et la variable quantitative. On parle parfois du « test ANOVA ».

1. Nous empruntons les données à une enquête réalisée en 2002 dans le cadre du CERLIS (Centre de Recherche sur les Liens Sociaux) et de l'enseignement de méthodologie de la Faculté des Sciences Humaines et Sociales de l'Université Paris Descartes.

4.1 Le principe de l'ANOVA

Nous savons que la variance est une mesure de la variabilité : elle permet d'estimer l'hétérogénéité ou, au contraire, l'homogénéité d'une série de valeurs. Supposons que le nombre de sports pratiqués soit parfaitement déterminé par le sexe, par exemple que les adolescentes pratiquent un seul sport tandis que les adolescents (masculins) pratiquent deux sports. Dans ce cas, l'homogénéité de comportement des filles (comme celle des garçons) est totale : la variance du nombre de sports pratiqués est nulle parmi les filles (comme parmi les garçons). En revanche, entre le groupe des garçons et celui des filles, il existe une hétérogénéité : la variabilité entre les deux groupes n'est pas nulle. Ainsi la variabilité entre le comportement des garçons et celui des filles est plus importante que la variance des comportements au sein du groupe des filles et au sein du groupe des garçons. Et ceci est associé à une situation où la variable qualitative est liée à la variable quantitative.

Imaginons maintenant que le sexe et la pratique sportive n'aient aucun lien entre eux. Dans ce cas, les comportements des garçons diffèrent peu ou pas du tout de ceux des filles : les garçons peuvent avoir des pratiques très différentes ; les filles peuvent également avoir des pratiques sportives très variables ; mais ce qui importe ici est que les garçons et les filles aient des attitudes proches voire identiques. En termes techniques, cela signifie que la variabilité entre les garçons et les filles est faible ou nulle comparativement à la variabilité des comportements masculins d'une part et féminins d'autre part.

Cet exemple suggère que pour se faire une idée de l'existence d'un lien éventuel entre la pratique sportive et le sexe des adolescents, il suffit de comparer les variabilités au sein des groupes et la variabilité entre les groupes. Le principe général de l'ANOVA réside dans cette comparaison des variabilités.

4.2 Mise en œuvre technique de l'ANOVA

Bien entendu, cette comparaison s'opère numériquement et il faut préciser les indicateurs quantitatifs utilisés pour mesurer les variabilités ainsi que les seuils permettant de comparer et donc de juger si une variabilité est plus grande qu'une autre. Il s'agit donc de présenter ces aspects techniques de mesure de la variabilité autant que de préciser le vocabulaire usuellement employé. La mise en œuvre technique de l'analyse de la variance est simplement la transposition mathématique du principe général présenté précédemment : l'esprit et la logique d'ensemble ne changent pas.

Les statisticiens ont l'habitude de parler de « variabilité intraclasse (ou intragroupe) » pour désigner la variabilité au sein des groupes et de « variabilité interclasse (ou intergroupe) » pour désigner la variabilité entre les groupes. Si la variabilité intragroupe est faible par rapport à la variabilité intergroupes, les différences de comportement au sein des groupes sont faibles comparativement aux différences de comportement entre les groupes : dans ce cas, il est vraisemblable que le critère définissant les groupes (la variable qualitative) soit lié à la variable quantitative étudiée. Inversement, si la variabilité intragroupe est élevée par rapport à la variabilité intergroupe, les différences de comportement au sein des groupes sont importantes comparativement aux différences de comportement entre les groupes : dans ce cas, il est vraisemblable que le critère définissant les groupes ne soit pas une bonne explication de l'hétérogénéité de la variable quantitative.

Pour estimer numériquement le niveau de variabilité, il est possible d'avoir recours à la variance, mais il est plus fréquent d'utiliser une grandeur directement liée à la variance : la somme des carrés des écarts à la moyenne, notée *SC*.

$$SC = \sum_{i=1}^{N} (X_i - \overline{X})^2 = N \times variance$$

Cet indicateur *SC* doit être calculé pour chaque groupe (définis par les modalités de la variable qualitative) ainsi qu'entre les groupes. Par convention, supposons que la variable qualitative comporte m modalités, que l'effectif de chaque groupe j soir N_j, que la moyenne au sein de chaque groupe défini par ces modalités soit notée $\overline{M}_j$ ($j = 1...m$) et que la moyenne générale soit notée $\overline{M}$. La somme des carrés intergroupes s'écrit alors :

$$SC_{inter} = \sum_{j=1}^{m} N_j \times (\overline{M}_j - \overline{M})^2$$

La variabilité intragroupe est la somme des variabilités internes à chaque groupe ce qui, en termes formels, s'écrit :

$$SC_{intra} = SC^1_{intra} + SC^2_{intra} + SC^3_{intra} + \ldots + SC^m_{intra}$$

Chacun des termes SC^j_{intra} est la somme des carrés au sein du groupe ou de la modalité j. Si N_j est la taille de ce groupe j, cette somme s'écrit simplement :

$$SC^{j}_{intra} = \sum_{i=1}^{N_j} (X_i - \overline{M}_j)^2$$

Pour finaliser le test, il suffit de comparer SC_{inter} et SC_{intra}, à condition de tenir compte de la taille de l'échantillon et du nombre de modalités de la variable qualitative puisqu'une variable ayant un grand nombre de modalités va, structurellement, augmenter la variabilité interclasse et que la taille de l'échantillon aura un effet identique. Pour tenir compte de ces deux aspects, on calcule les « carrés moyens », c'est-à-dire la somme des carrés divisés par des coefficients (communément appelés « degrés de liberté ») : la somme des carrés intergroupes est divisée par *m–1* (nombre de modalités moins un) ; la somme des carrés intragroupes est divisée par *N–m* (Taille de l'échantillon moins le nombre de modalités).

La comparaison numérique de la variabilité intergroupe et de la variabilité intragroupe s'opère en divisant le carré moyen intergroupe par le carré moyen intragroupe. Le résultat de cette division est noté *F* (en raison du nom du statisticien ayant conçu cet indicateur, Ronald Fisher). Plus le résultat de cette division est élevé, plus la variabilité intergroupe est importante par rapport à la variabilité intragroupe : comme nous l'avons vu, c'est plutôt le signe de l'existence d'un lien entre la variable qualitative et la variable quantitative. Si la valeur de cette division est faible, les deux variables sont vraisemblablement indépendantes.

Comme dans le cas du khi^2, toute la question est de trouver un critère permettant de décider quand la valeur *F* peut être jugée faible ou, au contraire, élevée. La réponse à cette question est toujours la même en statistique : il faut comparer cette valeur à la distribution théorique de cette valeur si les deux variables sont effectivement indépendantes. En d'autres termes, il faut faire appel à un test statistique : on fait l'hypothèse que les deux variables sont indépendantes ; on détermine le comportement théorique de la valeur si cette hypothèse est supposée vraie ; on compare la valeur *F* obtenue empiriquement avec son comportement théorique[1]. Cette comparaison permet de conclure : si la valeur *F* empirique (ou observée) est une valeur probable sous l'hypothèse d'absence de lien entre les deux variables, il est raisonnable de conclure à l'absence de lien ; si la valeur empirique *F* est une valeur peu voire très peu probable sous l'hypothèse d'absence de lien entre les deux variables, il est raisonnable de conclure à la présence d'un lien. En pratique, comme dans les

1. En toute rigueur, ceci suppose que la variable quantitative vérifie certaines propriétés empiriques (normalité).

autres tests, le jugement s'opère à la vue de la probabilité. Le seuil à partir duquel on juge la valeur *F* probable ou peu probable est toujours le même : 5 % ou 10 %, en fonction du niveau d'exigence ou de la prudence souhaitée.

Il est usuel de présenter ces différentes informations dans un tableau synthétique (voir le tableau 3.8). Ce type de tableau est très fréquent en statistique : c'est notamment la présentation adoptée par la plupart des logiciels d'analyse statistique. Il est donc important de bien comprendre sa logique d'ensemble (même si les aspects techniques et les détails calculatoires peuvent être davantage oubliés) car il résume, à lui seul, le raisonnement et les aspects techniques du test ANOVA.

Tableau 3.8. Tableau de l'analyse de la variance

	Somme des carrés	Degré de libertés	Carré moyen	*F* (Fisher)	Probabilité
Intergroupe	SCi_{nter}	$m-1$	$\frac{SCi_{nter}}{m-1}$	$\frac{\frac{SCi_{nter}}{m-1}}{\frac{SCi_{intra}}{N-m}}$	p
Intragroupe	SCi_{intra}	$N-m$	$\frac{SCi_{nter}}{m-1}$		

4.3 Une illustration pratique

Revenons à notre exemple du lien entre le sexe (variable qualitative à deux modalités) et la pratique sportive (nombre de sports régulièrement pratiqués) chez les adolescents. En empruntant les données empiriques à une enquête auprès de 552 individus, il est possible de présenter les valeurs numériques des diverses grandeurs utiles dans le tableau 3.9.

Tableau 3.9. Un exemple d'analyse de la variance

	Somme des carrés	Degré de libertés	Carré moyen	*F* (Fisher)	Probabilité
Intergroupes	9,38	$2-1=1$	9,38	12,85	*0,001*
Intragroupes	401,76	$552-2=550$	0,73		

Source des données : Enquêtes CERLIS, Faculté SHS de l'Université Paris-Descartes, 2002.

La valeur de *F* est élevée. Elle semble indiquer que la variabilité intergroupe est bien plus grande que la variabilité intragroupe : il semble exister une différence entre le comportement sportif des garçons et celui filles. Cette

impression est confirmée par la probabilité associée au test du Fisher : il est très peu probable d'obtenir une telle valeur de F si le sexe et la pratique sportive n'ont pas de lien. On conclut donc, de manière raisonnable, à l'existence d'une relation entre ces deux variables.

À la simple vue des nombres moyens de sports pratiqués par les garçons et les filles, il n'était pas évident qu'une telle relation existe : les garçons pratiquent 1,3 sport en moyenne, tandis que les filles pratiquent 1,04 sport. Le test ANOVA était indispensable pour se faire une opinion fondée sur la relation qu'entretiennent les deux variables.

4.4 Compléments

Il existe d'ailleurs une autre façon de concevoir le rôle de l'analyse de la variance : elle permet de déterminer l'existence de différences significatives entre les moyennes de plusieurs groupes. Cela revient à tester l'hypothèse selon laquelle les valeurs au sein des sous-groupes sont distribuées selon la même logique. En termes plus techniques, cela revient à tester l'hypothèse selon laquelle les moyennes de chaque groupe proviennent de la même distribution statistique : les moyennes diffèrent-elles à cause des différences entre les groupes ou bien à cause de simples fluctuations d'échantillonnage ? Cette conception est formellement équivalente à la présentation que nous avons suivie précédemment. Seuls l'esprit et l'interprétation sont différents.

Le cadre général ainsi que l'exemple que nous venons d'exposer constitue le cas d'une analyse de la variance à un seul facteur, c'est-à-dire une seule variable qualitative. Il est possible de généraliser le raisonnement suivi pour tenir compte non pas d'une seule variable qualitative mais de deux, trois, voire davantage de variables qualitatives : on parle d'analyse de la variance à plusieurs facteurs (MANOVA : Multiple ANalysis Of VAriance). La difficulté n'est pas tant d'ordre théorique que pratique : pour tenir compte de plusieurs variables qualitatives et espérer identifier des effets significatifs, il est indispensable de disposer d'un échantillon de taille suffisante : il est en effet possible de constater, dans notre exposé précédent, que la taille de l'échantillon joue un rôle majeur dans la « puissance » du test statistique, c'est-à-dire dans la capacité de l'analyse de la variance de nous renseigner de manière fiable sur le lien entre deux variables.

L'analyse de la variance n'est pas la seule démarche statistique à pouvoir être utilisée pour analyser plus de deux variables. C'est justement l'objectif du chapitre suivant : présenter les principales méthodes d'analyses multidimensionnelles c'est-à-dire destinées à analyser plusieurs variables en même temps.

ANALYSER LES RELATIONS ENTRE PLUSIEURS VARIABLES

Toutes les techniques présentées précédemment permettent d'analyser les relations entre deux variables. Or, habituellement, le nombre de variables dont il faut étudier les interrelations est bien supérieur à deux : il est fréquent que les individus de l'échantillon soient caractérisés par 50 ou 100 variables, et il n'est pas rare que ce nombre soit bien supérieur (200 à 500 informations). Au-delà, il est souvent nécessaire d'avoir recours à plusieurs variables pour décrire et bien comprendre un phénomène. Il existe des méthodes adaptées à ce type de situation : les méthodes statistiques multivariées ou multidimensionnelles.

Chacune de ces méthodes répond toutefois à des questions différentes. Recourir à ces méthodes dépend donc du type de question posée et, comme toujours, il est essentiel d'expliciter sa demande et son questionnement avant de recourir à tel ou tel outil statistique :

– Quelles sont les variables qui entretiennent de fortes relations entre elles (§ 1) ? Par exemple, quelles sont les pratiques culturelles fortement associées ?

– Quels sont les grands principes structurant les relations entre variables (§ 2) ? Par exemple, est-il possible de découvrir la logique générale, les grandes lignes directrices des relations qu'entretiennent les diverses activités culturelles ?

– Quels sont les groupes ou les profils d'individus ayant d'importantes caractéristiques communes (§ 3) ? Est-il possible de classer les individus en fonction de leurs pratiques culturelles afin de constituer des profils typiques de comportement ?

– En quoi une variable dépend-t-elle et peut-elle être expliquée par une série d'autres variables (§ 4) ? Par exemple, la fréquentation des musées est-elle explicable à partir du sexe, du diplôme, de la CSP et de la situation de famille ? Et si oui, quel est le rôle de chacun de ces facteurs explicatifs ?

Nous allons successivement présenter les méthodes permettant de répondre à ces grands types de question. Seuls les principes généraux ainsi que des illustrations pratiques seront exposés – la complexité mathématique de ces méthodes rend impossible, ici, la présentation des formalismes et des détails techniques.

1. AUTOMATISER LE CROISEMENT DE VARIABLES

Une des premières stratégies pour analyser un grand ensemble de variables est d'automatiser le calcul des tableaux croisés et, surtout, des tests du khi^2 associés. Les logiciels statistiques permettent, en général, d'identifier les paires de variables qui sont fortement liées entre elles (en fixant, à volonté, le seuil à partir duquel la relation est jugée forte, seuil de significativité du test du khi^2).

La méthode peut paraître séduisante et elle permet, en effet, d'explorer les variables, pour se familiariser avec des résultats empiriques ou pour s'aider à trouver des principes gouvernant les liens entre les variables. Mais elle présente deux faiblesses, l'une technique, l'autre épistémologique.

Sa faiblesse technique vient du fait que la réalisation, à l'aveugle, des tableaux croisés est rarement une opération fertile puisque la plupart des tableaux croisés ne deviennent intéressants et pertinents qu'après suppression des modalités inutiles ou marginales (les non-réponses notamment ; les modalités trop rares) ou regroupement des modalités selon des principes statistiques (avoir des effectifs suffisants) ou sociologiques (faire surgir des significations cachées ou noyées par des modalités trop nombreuses ou trop diverses). Il est rare qu'un tableau croisant deux variables soit immédiatement intéressant : il est souvent nécessaire de « nettoyer » le tableau en regroupant des modalités, en supprimant les moins pertinentes...

La faiblesse épistémologique de la démarche croisant systématiquement les variables entre elles provient de son absence de problématique précise, de l'absence de hiérarchisation des variables. Ne pas savoir ce qu'on cherche à comprendre, ne pas avoir explicité quelles sont les variables plutôt explicatives et les variables plutôt expliquées, et espérer que l'analyse exhaustive des variables va nous aider est une illusion : une question, une interrogation, une problématique, doivent toujours guider les pas du sociologue, même si cette question, interrogation ou problématique n'est pas encore totalement limpide et définitive. Il faut se débarrasser de la croyance selon laquelle l'analyse systématique des données peut, à elle seule, faire surgir une pensée, une interprétation ou une théorie.

Un cas particulier de l'analyse automatique et systématique des tableaux croisés est cependant fort utile : il s'agit du cas où on cherche à identifier toutes les variables fortement liées à une variable ou un indicateur précis et choisi. Imaginons que l'on dispose d'une enquête sur les pratiques culturelles et que l'on s'interroge sur les différences entre niveaux de diplômes (classés en trois catégories : aucun diplôme ou diplôme inférieur au bac ; baccalauréat seul ; diplôme de l'enseignement supérieur) : une première chose à faire est de

connaître les variables décrivant les pratiques culturelles fortement dépendantes du diplôme. Cela permet d'obtenir rapidement quelques résultats généraux, d'identifier les grandes différences de pratiques selon le diplôme... Même si cela ne suffit évidemment pas pour élaborer une analyse sociologique fine et construire une interprétation générale.

D'un point de vue technique, deux stratégies de croisement peuvent être adoptées : croiser les variables telles qu'elles se présentent dans l'enquête ; croiser des variables transformées en variables binaires. La transformation d'une variable en variables binaires consiste à décomposer une variable à *m* modalités en *m* variables indicatrices « oui/non » comme dans l'exemple suivant :

Question initiale :

« Quel est le genre du dernier film vu : a) un film comique ; b) un film d'action ; c) un film historique ; d) un film policier ou d'espionnage ; e) un film d'aventure ; f) une comédie dramatique ? »

Questions binaires ou indicatrices :

« Le dernier film vu est un film comique : a) Oui ; b) Non. »

« Le dernier film vu est un film d'action : a) Oui ; b) Non. »

« Le dernier film vu est un film historique : a) Oui ; b) Non. »

...

Ainsi, à la place d'un tableau ordinaire croisant toutes les modalités de la variable diplôme avec toutes les modalités de la variable « Genre du dernier film vu », il est possible d'établir une série de tableaux croisant chacune des modalités de la variable « diplôme » avec chacune des modalités de la variable « Genre du dernier film vu », comme dans l'exemple présenté tableau 4.1.

Tableau 4.1. Un exemple de croisement de variables indicatrices

	Le dernier film vu est une comédie dramatique	Autre cas	Total
Diplôme de l'enseignement supérieur	40 % *57 %*	60 % *33 %*	100 % *40 %*
Autre cas	20 % *43 %*	80 % *67 %*	100 % *60 %*
Total	28 % *100 %*	72 % *30 %*	100 % *100 %*

Données fictives — Lecture : 40 % des diplômés de l'enseignement supérieur ont vu une comédie dramatique ; et parmi ceux ayant vu une comédie, 57 % sont des diplômés de l'enseignement supérieur.

Cette transformation a le mérite de faire surgir chaque modalité en la dégageant des autres modalités qui peuvent compliquer son interprétation et faire disparaître, mécaniquement, la relation qu'elle peut entretenir avec la variable principale. Elle s'impose lorsque les modalités ne sont pas ordonnées et renvoient à des significations très différentes. Cette transformation n'est pas indispensable lorsque les variables sont ordonnables ou, *a fortiori*, lorsqu'il s'agit de variables quantitatives recodées en classes.

Il est commode d'exposer les résultats en adoptant une présentation semblable à celle du tableau 4.2.

Tableau 4.2. Résultats de croisements systématiques

Modalités liées à la modalité « Diplôme de l'enseignement supérieur »			
Modalité	**% de la modalité**	**% du groupe**	**Probabilité du test du khi²**
Dernier film vu : comédie dramatique	57 %	40 %	< 0,001
Dernier spectacle vu : danse	50 %	30 %	< 0,001
Dernier spectacle vu : opéra	57 %	20 %	< 0,001
Dernier spectacle vu : théâtre classique	50 %	60 %	< 0,001
Dernier lieu visité : galerie d'art	53 %	15 %	0,003
Dernier lieu visité : Musée d'histoire	45 %	32 %	0,04
...	...	...	...

Données fictives – Lecture : 50 % des individus ayant vu un spectacle de danse la dernière fois sont des diplômés de l'enseignement supérieur ; 30 % des diplômés de l'enseignement supérieur ont vu un spectacle de danse la dernière fois ; la probabilité du test du khi² associé au croisement est inférieure à 0,001 (les deux modalités sont donc très probablement dépendantes). Le test du khi² est relatif au tableau croisant les variables indicatrices « diplômé du supérieur : oui/non » et « dernier spectacle danse : oui/non ».

Cette technique d'analyse est parfois appelée « analyse des profils ». Elle est notamment utilisée pour décrire les principales caractéristiques de groupes d'individus déterminés automatiquement (voir le § 3) ou définis *a priori*. Il est, par exemple, possible de définir des groupes à partir de critères jugés pertinents, puis de qualifier ces groupes à l'aide de l'analyse de leurs profils.

C'est la méthode suivie dans notre recherche sur les usages des ordinateurs et d'Internet par les 10-20 ans[1]. À partir de deux indicateurs, un indicateur de sociabilité (estimant la taille du réseau amical et l'intensité des sorties avec les copains) et un indicateur de contrôle parental (estimant le niveau de contrôle que les parents exercent sur les faits et gestes de l'enfant), nous avons défini quatre groupes : les « casaniers » (faible sociabilité et faible contrôle parental), les « contrôlés » (faible sociabilité et fort contrôle), les « indépendants » (forte sociabilité et faible contrôle) et enfin les « en liberté surveillée » (forte sociabilité et fort contrôle). Nous avons ensuite caractérisé les pratiques informatiques et les usages d'Internet de ces quatre groupes en examinant systématiquement les modalités ou variables liées à l'un ou l'autre de ces groupes : la problématique est donc de voir en quoi les usages et pratiques de l'ordinateur font écho et correspondent à des caractéristiques générales de la vie des enfants, en l'occurrence leur sociabilité amicale et le contrôle que leurs parents exercent[2]. En surgit une qualification de ces groupes montrant, par exemple, que le groupe jouissant d'une liberté surveillée a des usages importants et très variés de l'ordinateur, que celui-ci est omniprésent dans les activités (loisirs, sorties, hobbies), que les enfants de ce groupe en ont un usage communicationnel important mais que tout cela se fait sous la surveillance parentale et avec leur accord. Au final, l'ensemble de ces indices permet de qualifier et d'interpréter les statuts de l'ordinateur et d'Internet parmi les individus en « liberté surveillée » : l'ordinateur est une sorte de fenêtre ouverte vers l'extérieur, et leur confère de l'autonomie dans la gestion de leurs relations et de leurs loisirs.

2. Repérer et synthétiser les relations : l'analyse factorielle

La méthode de croisement automatique que nous venons de présenter facilite l'analyse des variables dès que celles-ci sont nombreuses. Elle peut aider la sociologie à faire la synthèse de ses données, c'est-à-dire à repérer les principes généraux, à élaborer une vue d'ensemble, à avoir une idée générale de

1. Olivier Martin, « L'Internet des 10-20 ans : une ressource pour une communication autonome », *Réseaux*, vol. 22, n° 123, 2004, p. 25-58.
2. Les deux indicateurs synthétisent des réponses à des questions ne portant pas sur l'ordinateur ni sur les usages d'Internet.

la façon dont les diverses variables sont globalement reliées. Il existe cependant une gamme de méthodes statistiques qui facilite encore davantage ce travail de synthèse : il s'agit des méthodes dites d'analyse factorielle.

2.1 Principes et intérêts généraux de l'analyse factorielle

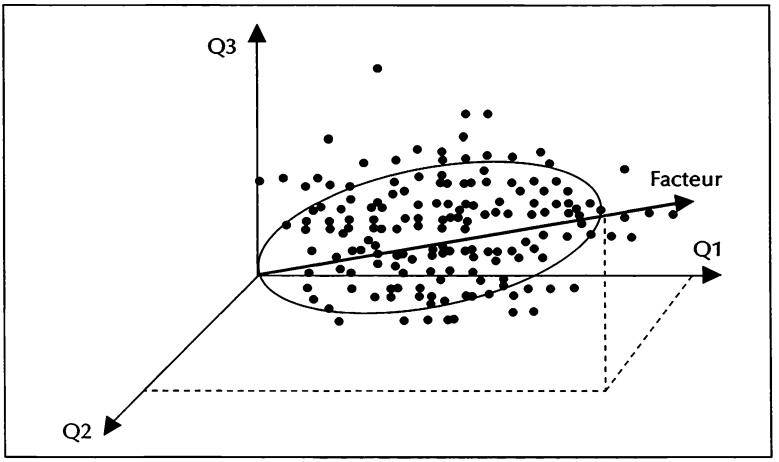
Il existe au moins trois manières de concevoir c'est-à-dire de penser l'analyse factorielle. Elles sont mathématiquement équivalentes. Nous choisissons toutefois de les présenter toutes les trois afin de bien faire comprendre l'esprit de la démarche. Nous laissons « simplement » de côté les justifications mathématiques complètes qui sortent du cadre de ce manuel – le lecteur intéressé pourra se reporter à la bibliographie finale.

La première manière de concevoir l'objectif de l'analyse factorielle est de considérer que l'enquête apporte de l'« information » sur les individus. La quantité d'information apportée est grande, trop grande pour être spontanément intelligible. Elle est trop riche pour être immédiatement compréhensible. L'analyse factorielle cherche à faire surgir les traits saillants, dominants, de cette information tout en la trahissant le moins possible, tout en étant le plus fidèle possible. Elle résume cette information en un petit nombre de caractéristiques plus facilement intelligibles – un peu à la manière d'un résumé d'une page condensant l'essentiel des propos d'un livre de 200 pages.

La deuxième façon de concevoir l'analyse factorielle est de recourir à une métaphore spatiale – en fait, il ne s'agit pas simplement d'une métaphore puisqu'il est possible de justifier de manière rigoureuse ces méthodes en recourant à la géométrie. Chaque individu de l'enquête étant caractérisé par v variables (questions), il est possible de le positionner dans un espace à v dimensions, où chaque dimension représente une variable. Les différents individus ne se répartissent ni uniformément ni au hasard dans cet espace à v dimensions : ils sont regroupés selon certaines directions ; ils sont rares dans certaines zones et, au contraire, nombreux dans d'autres. L'analyse factorielle consiste à essayer de trouver une représentation économique de ce nuage de point, une représentation qui permette de le voir de manière commode sans trop le déformer. Pour illustrer ce principe, considérons le cas simple d'un nuage de points dans un espace à trois dimensions.

Imaginons avoir enquêté les individus sur leurs pratiques culturelles et leurs loisirs. Supposons connaître l'intensité (fréquence) de leurs sorties au cinéma (Q1), au théâtre (Q2) ainsi que dans les musées (Q3).

Figure 4.1. Le nuage des points-individus



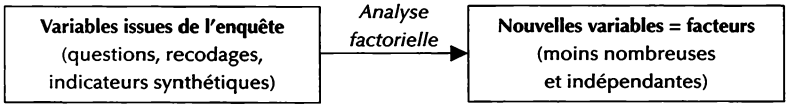
Chaque individu enquêté est repéré par ses réponses aux questions Q1, Q2 et Q3 qui sont les coordonnées de cet individu (représenté par un point) dans l'espace à trois dimensions. L'ensemble des individus de l'échantillon constitue un nuage de points. L'objectif de l'analyse factorielle est de trouver une représentation simple et juste de ce nuage : elle doit être simple dans la mesure où l'on souhaite avoir un nombre plus réduit de critères pour repérer les individus (deux voire une seule variable à la place de trois) ; elle doit être juste dans la mesure où l'on souhaite que cette représentation ne trahisse pas trop la forme du nuage de points.

Dans le cas de la figure 4.1, le nuage de point est très concentré autour d'un axe : cet axe est le facteur. Il répond assez bien à nos deux exigences : il est simple et plutôt fidèle. Ainsi, au lieu de caractériser les individus par les trois variables Q1, Q2 et Q3, il est plus commode et « pas trop faux » de les caractériser par une seule donnée : leur position sur cet axe.

Reste à donner un sens à ce qui est pour l'instant une pure construction géométrique et mathématique sans valeur sociologique. Pour trouver le sens de cet axe, il suffit de voir ce qu'il partage avec Q1, Q2 et Q3. La situation est assez simple : les grandes valeurs de Q1, Q2 et Q3 sont associées aux grandes valeurs sur l'axe ; les petites valeurs de Q1, Q2 et Q3 sont associées

aux petites valeurs sur l'axe. Cet axe peut être interprété comme un axe d'intensité, mesurant l'intensité des pratiques des individus (toutes pratiques confondues). Il a surgi de l'analyse parce que, empiriquement, les individus pratiquant intensément l'une des activités pratiquent également intensément les deux autres activités. Cette situation est en tout cas fréquente dans l'enquête et elle explique que les points représentant les individus ne soient pas répartis n'importe où mais soient, au contraire, concentrés le long d'une direction précise.

La présentation de la troisième manière de concevoir l'analyse factorielle va nous permettre de préciser encore davantage les choses. On cherche à exprimer les variables (questions) issues de l'enquête grâce à un nombre minimal de nouvelles variables en partant de l'hypothèse que les variables empiriques entretiennent des relations entre elles (elles sont corrélées) et qu'elles sont donc redondantes. On cherche à réduire le nombre de variables nécessaires pour décrire de manière fiable les individus. Ces nouvelles variables sont les *facteurs* : pour être pertinents, ces facteurs doivent être moins nombreux (sinon on ne gagne rien et on ne simplifie pas la situation) et indépendants les uns des autres (de façon à exprimer des aspects indépendants de la réalité enquêtée et à ne pas être redondants). Ces facteurs s'expriment à l'aide des seules variables initialement présentes dans l'enquête (sans apport de données extérieures). Mathématiquement, ils s'obtiennent par combinaison linéaire des variables initiales.



Considérons à nouveau notre exemple d'enquête sur les pratiques culturelles. Imaginons posséder non seulement les questions Q1, Q2 et Q3 précédentes, mais aussi des questions sur l'intensité de la lecture (Q4), de l'écoute de la musique (Q5), des visites de galeries et d'exposition d'art (Q6), d'écoute de la radio (Q7) ou de la télévision (Q8). À nouveau, il est raisonnable de supposer que ces variables ne sont pas indépendantes les unes des autres : visiter fréquemment des musées est par exemple souvent associé à d'assez fortes pratiques de lecture... Leur analyse factorielle permettra peut-être de montrer que les relations entretenues par ces variables peuvent être décomposées en plusieurs effets : un effet de volume et un effet de spécialisation. L'effet de volume restitue le fait que les pratiques intenses de chacune

de ces activités sont souvent associées. L'effet de spécialisation restitue le fait que les pratiques ont leurs spécificités et que les pratiques intérieures (TV, radio, lecture) s'opposent aux pratiques extérieures (cinéma, musée...). Dès lors, plutôt que de décrire un individu par huit variables, il est possible de le décrire à l'aide des deux facteurs : le volume global de ses pratiques culturelles et son profil (ayant de préférence plutôt des activités d'intérieur ou d'extérieur). En tout cas, si les données empiriques sont plutôt conformes à cette répartition et cette structuration des relations, l'analyse factorielle devrait l'identifier.

2.2 La pratique de l'analyse factorielle

Ainsi, l'analyse factorielle des données permet d'identifier un nombre « raisonnable » d'axes (facteurs) capables de rendre compte des répartitions des individus. Dit autrement, l'analyse factorielle permet de trouver les nouvelles variables (facteurs) qui résument plutôt bien les informations obtenues par enquête (les variables initiales) et qui, en même temps, soient moins nombreuses que ces variables initiales.

La capacité de l'analyse factorielle à trouver de nouveaux axes/variables ne serait rien si elle ne permettait pas de juger de l'intérêt de ces axes/variables ainsi que du sens de ces axes/variables. Heureusement, elle fournit des indicateurs sur la qualité de ces résumés ainsi que des éléments pour interpréter ces nouvelles variables.

La qualité d'une analyse factorielle se juge grâce à une mesure de la part d'information initiale qu'elle permet de restituer. Plus précisément, à chaque facteur est associée une mesure de la part d'information qu'il restitue par rapport à l'information totale initiale. Si on considère que les données initiales représentent une certaine quantité d'information, la part de l'information restituée par chaque facteur s'exprime en pourcentage de cette part initiale : plus ce pourcentage est important, plus le facteur est « fidèle » et « riche » en informations. Le sociologue aura donc intérêt à interpréter les axes les plus riches. Selon le vocabulaire mathématique utilisé par les statisticiens et les logiciels statistiques, cette part d'information est directement liée à la « valeur propre » de l'axe. Peu importe, pour nous, le sens technique exact de ces termes : l'essentiel est de savoir que plus cette valeur est grande plus le facteur (l'axe ou la variable) a de l'intérêt.

Il n'est pas possible de décider, *a priori*, un seuil en dessous duquel il ne serait pas souhaitable d'utiliser les facteurs. La part d'information apportée

par un facteur dépend, mécaniquement, du nombre des variables initialement présentes dans l'analyse ainsi que de la nature de leurs relations. Si le nombre de variables est très grand, il y a peu d'espoir, par nature, de parvenir à résumer fidèlement cette grande quantité d'informations à un seul ou même un petit nombre d'axes. Il n'est pas rare d'obtenir, à partir d'une analyse de grands ensembles de données, des parts d'information restituées égales à 10 %. Il ne faut pas se désespérer de cette situation. Pas plus qu'il n'est possible de résumer parfaitement un ouvrage de 1 000 pages en une seule phrase, l'analyse factorielle ne peut pas résumer fidèlement des centaines de variables en une seule variable.

En pratique, il s'avère que les sociologues se contentent d'utiliser seulement les deux, trois, voire au maximum quatre facteurs ayant les plus grandes valeurs propres. Au-delà, l'interprétation devient laborieuse et souvent délicate. Un des critères pour choisir ce nombre (1, 2, 3 voire 4 ?) est d'interpréter les facteurs tant que leur sens semble pertinent et identifiable.

Comment, justement, déterminer la signification d'un facteur ? Pour cela, l'analyse factorielle nous fournit ce qu'il est communément appelé des « contributions » : les contributions expriment l'importance du poids de chaque variable initiale dans les facteurs. Plus la contribution est importante, plus la variable pèse c'est-à-dire participe à la définition du facteur et donc plus le facteur hérite (absorbe) la signification de la variable. Pour donner un sens aux facteurs, il suffit de trouver les variables qui pèsent le plus, c'est-à-dire les variables qui ont les plus fortes contributions sur ce facteur.

Après avoir identifié les nouveaux facteurs, les avoir interprétés, il reste à utiliser ces résultats pour analyser les données initiales. Pour cela il est commode d'avoir recours à une représentation graphique. On construit une carte où les axes (horizontal et vertical) sont des facteurs et où les modalités ou variables sont représentées. Cette carte est appelée le « plan factoriel ». Elle permet de faire figurer les convergences entre modalités ou variables (c'est-à-dire les modalités ou variables fréquemment associées) ainsi que les divergences (c'est-à-dire les modalités ou variables antagonistes, incompatibles entre elles). Cette carte factorielle (dont l'interprétation doit néanmoins être prudente) fournit une image presque spontanément intelligible des relations que les variables ou modalités entretiennent entre elles.

2.3 Rôles des variables

Bien qu'il soit techniquement possible de traiter ensemble toutes les variables d'une enquête, on a rarement intérêt à procéder ainsi : d'une part, parce que cette attitude reviendrait à mélanger des variables qui peuvent avoir des statuts différents dans la perspective théorique adoptée (variables explicatives/expliquées, opinions/comportements...) ; d'autre part, parce qu'en procédant ainsi, on attribue un statut identique à toutes les variables, on risque de donner plus de poids à des variables nombreuses mais mesurant des phénomènes jugés marginaux, au détriment de variables mesurant des phénomènes fondamentaux mais moins nombreuses. On a donc toujours intérêt à distinguer les variables qui nous intéressent en premier ressort et celles qui pourront compléter nos analyses ou venir illustrer les résultats. Les premières, qui vont directement participer à l'analyse statistique et contribuer à la forme que prendront les résultats, sont appelées « actives ». Les secondes, qui ne sont pas déterminantes pour l'élaboration des résultats, sont appelées « variables supplémentaires ou illustratives » : elles permettent d'éclairer la nature de la structure obtenue par une analyse factorielle sans intervenir dans son élaboration. Ainsi, cherchant à comprendre comment se répartissent les diverses pratiques culturelles selon les caractéristiques socio-démographiques des individus, le sociologue aura intérêt à choisir comme variables actives les informations socio-démographiques et comme variables supplémentaires les pratiques. Ou, cherchant à analyser le lien entre les usages des nouveaux outils de communication et les formes de sociabilité, le sociologue pourra choisir les indicateurs de sociabilité comme variables actives et les indicateurs d'usages du téléphone, du portable, des SMS ou du courriel comme variables supplémentaires.

2.4 Usages

Les méthodes factorielles les plus couramment utilisées sont l'analyse en composantes principales (ACP, destinée aux variables quantitatives[1]) et l'analyse des correspondances multiples (ACM, destinée aux variables qualitatives). L'esprit de ces méthodes ne diffère pas réellement : nous

1. Pour un exemple : Anne-Carole Rivière, « Le téléphone : un facteur d'intégration sociale », *Économie et statistique*, n° 345, 2001, p. 3-32.

venons d'en présenter les principaux traits. D'un point de vue technique, les choses sont toutefois bien différentes car les outils mathématiques pour trouver les facteurs ne sont pas absolument identiques – mais ce n'est pas essentiel ici.

L'analyse factorielle, et notamment l'analyse des correspondances multiples, a connu un grand succès dans les années 1970 et 1980. Pierre Bourdieu l'utilise souvent dans ses ouvrages (*La Distinction*, *Homo Academicus*, *La Noblesse d'État*...). Aujourd'hui, elle est un peu moins présente dans les publications en sociologie mais continue toujours à être utilisée de manière exploratoire (voir le § 1.4 de la conclusion).

Travaillant sur les jeunes de 16-24 ans sans domicile et en situation précaire (enquête INED 1998), Maryse Marspat et Jean-Marie Firdion[1] cherchent à saisir la logique d'ensemble du recours aux services d'aides (hébergement, repas, accueil, soutien...). Pour cela, ils réalisent une ACM en prenant comme variables actives les différentes modalités d'accès aux services, le recours à la solidarité des proches, et les autres ressources permettant d'organiser la vie quotidienne (appel au Samu social ou à une mission locale, fréquentation d'un centre d'accueil, type d'hébergement, lieu habituel de prise des repas de midi et du soir, personnes ayant aidé, origine des ressources financières...).

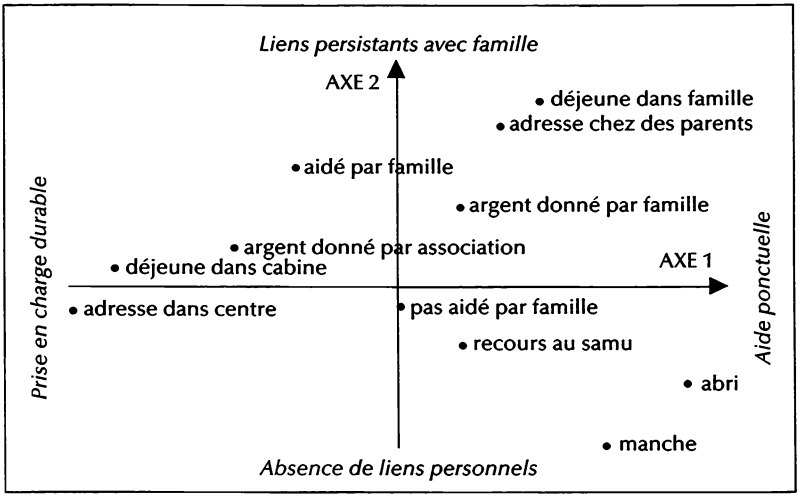
Cette analyse leur permet d'identifier deux facteurs structurants.

Le premier facteur/axe oppose les « pratiques en lien avec les services qui prennent en charge les jeunes sur la durée, souvent avec une aide, pour accéder à l'emploi ou à la formation et au logement, qui s'accompagne d'un engagement du jeune » et les « pratiques de ceux qui utilisent les services ponctuels (accueils de jour, centres d'urgence, distributions de repas) ainsi que la manche et l'hébergement dans des abris précaires ».

Le second facteur oppose les « pratiques reposant sur la persistance de liens avec la famille » aux situations marquées par « l'absence de liens personnels et le faible recours aux services d'aide (dormir dans un abri, faire la manche) ». La figure 4.2 permet de présenter graphiquement ces résultats.

1. « Les situations des jeunes sans domicile et en situation précaire », *Recherches et prévisions*, n° 65, 2001, p. 91-112.

Figure 4.2. Plan factoriel



Ce graphique représente quelques modalités des variables actives. Il est possible d'en établir un autre pour connaître la distribution des modalités illustratives dans cet espace structuré autour de la double opposition « Aide ponctuelle/Prise en charge durable », et « Absence de lien/Persistance des liens ».

3. CLASSER LES INDIVIDUS POUR DÉFINIR DES TYPES

Une autre manière de traiter et de tenir simultanément compte de plusieurs variables est de chercher à définir des groupes (ou classes) d'individus en fonction de l'homogénéité de leurs comportements, selon la proximité ou la ressemblance de leurs réponses aux questions. Pour cela, il est possible d'avoir recours à des « méthodes de classification (automatique) ». Elles permettent d'obtenir automatiquement (c'est-à-dire sans intervention du sociologue) des découpages de la population en groupes homogènes dans un ensemble hétérogène : l'usage veut que ces groupes soient appelés « classes ».

3.1 Principes généraux

L'opération consistant à découper un ensemble d'individus en groupes homogènes est une attitude courante, parfois même spontanée, dans l'activité scientifique : c'est l'opération que réalise par exemple le naturaliste lorsqu'il essaie d'établir des familles d'animaux ou de plantes, ou le sociologue lorsqu'il classe des individus en diverses catégories censées représenter des professions différentes mais proches. Dans tous les cas, il s'agit de regrouper dans une même classe les individus qui se ressemblent et de placer dans des classes différentes les individus trop dissemblables.

Afin de pouvoir juger de la ressemblance entre deux individus, il faut avoir recours à un indicateur quantitatif : une distance. Si les variables caractérisant les individus sont des variables quantitatives, la distance utilisée est la distance « classique » (euclidienne). Si les variables sont de nature qualitative, il est possible d'avoir recours à la distance du khi^2 (voir le chap. 3, § 2) ou encore à un indicateur comptant simplement le nombre de points différents (c'est-à-dire de modalités différentes) entre deux individus. Si les modalités choisies par deux individus à un ensemble de questions diffèrent en de nombreux points, ces individus sont très différents et devront donc être classés dans des groupes distincts ; si les modalités choisies sont presque toujours les mêmes, ces individus sont très proches et pourront être classés dans le même groupe.

En plus du choix d'une distance, il est nécessaire de déterminer le critère selon lequel on agrège un individu à un groupe : pour un individu *I* donné, comment choisir le groupe auquel il sera affecté ? Dans le cas de variables quantitatives, il est possible de calculer les distances entre l'individu *I* et l'individu moyen de chaque groupe, puis de choisir le groupe pour lequel la distance est la plus faible. Dans le cas de variables qualitatives, la notion d'individu moyen n'a pas de sens : il faut adopter d'autres critères, par exemple le critère « Max » (choisir parmi les groupes celui dont l'individu le plus différent de *I* est malgré tout le plus proche de *I*) ou « Min » (trouver le groupe ayant un individu le plus proche possible de *I*). Une autre solution, souvent adoptée, consiste à réaliser une analyse des correspondances multiples puis une classification sur les variables quantitatives (facteurs) issues de cette ACM.

Il existe deux stratégies principales pour réaliser une classification. La première stratégie consiste à agréger, petit à petit, les individus en créant des classes de plus en plus grosses. Cette méthode est dite ascendante hiérarchique : le terme « ascendant » renvoie à l'idée que la situation de

départ est la situation où tous les individus sont distincts et séparés, et qu'ils sont regroupés progressivement dans des classes de plus en plus importantes ; le terme « hiérarchique » renvoie à l'idée qu'à chaque étape de la classification les groupes d'individus sont obtenus par assemblage des classes obtenues à l'étape précédente. Cette méthode, dont le sigle courant est CAH (Classification Ascendante Hiérarchique) présente l'avantage de fournir des critères permettant de déterminer le nombre optimal de classes. En revanche, elle est plutôt réservée aux échantillons ne dépassant pas 1 000 individus (en raison des très nombreux calculs nécessaires).

La seconde stratégie, dite méthode des « nuées dynamiques » (ou des centres mobiles), repose sur le principe suivant : on définit plusieurs petits groupes appelés « noyaux » et constitués d'individus choisis de façon plus ou moins aléatoire dans la population ; on fait grossir ces noyaux en agrégeant à chacun d'eux les individus qui ont un profil proche de ce noyau et en excluant ceux qui ont un profil éloigné ; on cherche alors au sein de chacun des groupes ainsi obtenus ce qui en constitue le noyau (c'est-à-dire ce qui domine, ce qui est central, dans le groupe) ; les nouveaux noyaux ainsi obtenus permettent par agrégation, comme précédemment, de constituer de nouveaux groupes qui, à leur tour, permettent de définir de nouveaux noyaux... Au terme d'un grand nombre d'itérations, les groupes et leurs noyaux se stabilisent : la classification est achevée. Cette seconde méthode suppose de choisir, au préalable, le nombre de classes qu'on souhaite obtenir. Elle suppose donc d'avoir une idée précise de la logique des regroupements. Elle a toutefois l'avantage d'être utilisable sur des échantillons de très grande taille.

Comme dans le cas des analyses factorielles, il est nécessaire de distinguer les variables actives qui vont constituer les critères pour classer les individus, et les variables supplémentaires qui vont servir, une fois les classes identifiées, à les qualifier, à en trouver les traits caractéristiques.

3.2 Usages

Les méthodes de classification ont l'inconvénient d'être un peu instables (les résultats sont trop sensibles aux données et aux critères techniques utilisés). Pour cette raison, elles sont rarement publiées. Mais elles sont souvent utilisées de manière exploratoire, afin d'identifier les grandes lignes de division d'un échantillon, de repérer les grands types de comportements ou d'attitudes par les individus, quitte à utiliser les résultats obtenus pour définir des critères permettant de classer les individus « à la main ».

À l'issue d'une classification, chaque individu est caractérisé par une information supplémentaire : le numéro de la classe à laquelle il appartient. Cette information est représentée par une variable qualitative dont les modalités correspondent aux différentes classes obtenues. Il est alors possible de déterminer le sens, la nature, de chacune des classes : il suffit de croiser cette variable avec toutes les variables actives ayant participé à la classification. Il est ensuite possible de croiser cette même variable à toute autre variable supplémentaire.

Revenons à l'enquête sur les jeunes sans domicile et en situation précaire. Il est possible de classer les individus à partir des variables déjà utilisées dans l'ACM. La classification permet d'identifier trois principaux groupes dont on peut connaître non seulement les traits « actifs » mais également les principales caractéristiques sociodémographiques. L'un d'eux rassemble les jeunes bénéficiant d'une prise en charge dans une structure de long terme : en majorité, ils dorment dans des centres d'accueil de longue durée, dînent dans ce type de centre et disposent d'une adresse fixe : ils n'ont pas recours aux travailleurs sociaux ni aux aides d'urgence. En croisant la variable issue de la classification avec des variables telles le sexe, l'âge ou encore la situation d'emploi et le parcours biographique, il est possible de constater que ce groupe rassemble plutôt des femmes, plutôt jeunes, ayant séjourné à la DDASS et ayant fait une fugue. Les individus de ce groupe ont, plus souvent que les autres groupes, un emploi régulier.

Les deux autres classes rassemblent les individus faisant appel de manière ponctuelle à des structures d'aide et à des services d'urgence d'une part, et les individus vivant dans l'espace public (squat, abri de fortune, rue), faisant appel aux services d'aide et d'urgence sauf pour l'hébergement. Ces deux groupes sont plus masculins et plus âgés.

4. Décomposer les effets de chaque variable

Les méthodes présentées jusqu'ici ont vocation à saisir de manière simple les diverses relations entretenues par plusieurs variables. Elles contribuent à rendre intelligible de grands ensembles de variables. Mais elles ne permettent pas de bien identifier le rôle de chaque variable, de l'isoler des autres variables. Les méthodes de régression répondent à cette exigence : connaître le rôle « exact » de chaque variable. Ce rôle « exact » est appelé l'effet pur ou net d'une variable, par opposition à l'« effet brut ».

4.1 L'effet brut, l'effet pur et le raisonnement « toutes choses étant égales par ailleurs »

Considérons le tableau 4.3, croisant le devenir scolaire au collège (admission ou non en seconde générale ou technologique) en fonction de la nationalité. Ces données sont issues d'une analyse de Louis-André Vallet et Jean-Paul Caille menée sur un panel d'environ 18 500 élèves français[1]. Les résultats semblent sans ambiguïté : les enfants étrangers rencontrent plus de difficultés que les élèves français à poursuivre leur scolarité au lycée.

Tableau 4.3. Carrière scolaire en fonction de la nationalité de l'élève

Nationalité	Admis en seconde générale ou technologique	Non admis	Total
Française	74,1 %	25,9 %	100 %
Étrangère	63,9 %	36,1 %	100 %

Une question mérite toutefois d'être posée : en tenant compte simultanément du milieu social et culturel, de la taille de la famille, de la structure familiale ainsi que des situations professionnelles des parents, la différence entre les élèves français et étrangers est-elle aussi importante ? Ce tableau 4.3 permet d'identifier un « effet brut » de la nationalité, mais qu'en est-il de son « effet pur » tenant compte du fait que la nationalité, l'origine sociale et culturelle, la situation familiale et le devenir scolaire sont probablement liés ? Est-ce réellement la nationalité qui explique la réussite ou l'échec scolaire ? Si la réponse est affirmative, l'effet brut identifié par ce tableau est un effet pur. Si la réponse est négative l'effet brut est un artefact, lié au fait que la nationalité est le révélateur de situations sociales, culturelles et familiales particulières.

L'effet pur permet d'estimer le rôle d'une variable ou d'une modalité compte tenu de toutes les autres modalités c'est-à-dire une fois contrôlés et neutralisés les effets des autres modalités. L'effet pur exprime l'effet d'une variable/modalité sur une autre, « toutes choses étant égales par ailleurs ».

Pour connaître les effets purs, il est possible de recourir aux méthodes de régression : ces méthodes cherchent à décomposer les effets d'une série de variables sur une autre variable. Plus précisément et selon le vocabulaire

1. « Les carrières scolaires au collège des élèves étrangers ou issus de l'immigration », *Éducation et formations*, n° 40, 1995, p. 5-14.

usuel, les régressions permettent de déterminer les effets des variables explicatives sur une variable dite expliquée : les régressions sont des « méthodes explicatives ». Dans l'exemple, la variable à expliquer est le devenir scolaire (l'admission en seconde ou non) et les variables explicatives sont les indicateurs de situation culturelle, sociale et familiale...

Comme toujours, les méthodes doivent être distinguées en fonction de la nature des variables à analyser : les régressions linéaires sont adaptées aux variables quantitatives ; les régressions logistiques sont adaptées aux variables qualitatives (comme d'autres méthodes très proches : probit ; polytomiques[1]). Elles ne diffèrent pas tant par leur esprit que par les formalismes mathématiques et les techniques statistiques nécessaires à leur justification et à leur mise en œuvre.

Toute analyse par régression suit un cheminement en trois étapes : 1°) l'explication du modèle explicatif ; 2°) l'estimation du modèle et de sa pertinence (qualité globale ; qualité explicative de chaque variable) ; 3°) révision et changement du modèle (ajout ou suppression de variables). Ces trois étapes doivent être enchaînées et recommencées.

4.2 Expliciter un modèle

Pour recourir à ces méthodes, il est indispensable d'expliciter le modèle qu'on veut tester : quelles sont les variables explicatives et quelle est la variable expliquée ? Ces précisions sont indispensables pour pouvoir mettre en œuvre les méthodes de régression. Mais elles sont également indispensables d'un point de vue logique ou épistémologique : l'effet pur d'une variable ne peut pas être estimé dans l'absolu mais par rapport à un ensemble explicite de variables. Il n'existe pas un effet pur, compte tenu de l'ensemble des facteurs possibles, compte tenu de toutes les variables possibles et imaginables. Malgré son nom, l'effet pur est toujours relatif à un ensemble précis de variables.

Dans l'exemple, Louis-André Vallet et Jean-Paul Caille ont choisi de tester un modèle où la variable expliquée est l'admission ou non en seconde (générale comme technologique) et où les variables explicatives sont : la nationalité de l'enfant, la PCS du chef de famille, le diplôme du père, le diplôme de la mère, l'activité de la mère, le sexe de l'enfant, la taille de la famille, le rang de l'enfant dans la fratrie, la présence de frères ou sœurs dans l'enseignement

1. Les régressions logistiques permettent de traiter le cas de variables expliquées à seulement deux modalités. Les régressions polytomiques peuvent traiter le cas de variables expliquées à trois modalités ordonnées (ou plus).

supérieur, la structure de la famille (bi- ou monoparentale). Les variables étant qualitatives, la méthode utilisée est une régression logistique.

4.3 Évaluer la pertinence d'un modèle

Mesurer l'effet d'une variable ou plutôt de chacune de ses modalités suppose la définition d'une situation de référence. Cette situation de référence est la situation à partir de laquelle on juge de l'effet de chacune des modalités. Son choix est relativement arbitraire et sans conséquence : il consiste à désigner une modalité pour chaque variable explicative (le plus simple et le plus prudent est de choisir la situation globale la plus fréquente, en tout cas une situation assez fréquente). En l'occurrence, dans l'exemple, la situation de « référence » est la situation d'un élève français de sexe masculin, aîné de sa fratrie, dont le père de famille est ouvrier qualifié, dont les parents possèdent tous les deux un diplôme du type CAP-BEP-BEPC...

Les effets de chacune de ces variables, ou plus exactement de chacune des modalités de ces variables, sont appréciés à l'aide d'un coefficient dit « coefficient de régression » : il mesure l'ampleur du rôle que joue la modalité sur la variable à expliquer. Un coefficient positif associé à une modalité (qui exprime une caractéristique) indique que posséder la caractéristique augmente les chances d'être admis en seconde (par rapport à la situation de référence). Un coefficient négatif est synonyme d'une chance plus faible (par rapport à la situation de référence).

Par exemple (tableau 4.4), le coefficient 0,67 associé à la modalité « fille » signifie que les filles ont davantage de chance d'être admises en seconde que les individus de la situation de référence (en l'occurrence les garçons). Inversement, le coefficient (-0,31) associé à la modalité « famille monoparentale » signifie qu'un collégien dans cette situation a moins de chances que les enfants issus de famille biparentale (situation de référence) d'être admis en seconde.

Tableau 4.3. Variables expliquant l'admission en seconde

Variable	Modalités	Coefficient	Test
Nationalité de l'enfant	Française Étrangère	(ref) 0,30	(ref) $p < 0,001$
Sexe de l'enfant	Garçon Fille	(ref) 0,67	(ref) $p < 0,001$

Variable	Modalités	Coefficient	Test
Rang dans la fratrie	Rang 1	(ref)	(ref)
	Rang 2	– 0,23	p < 0,001
	Rang 3	– 0,08	ns
	Rang 4 et plus	– 0,19	p < 0,02
Structure familiale	Biparentale	(ref)	(ref)
	Monoparentale	– 0,31	p < 0,001
	Autre situation	– 0,58	p < 0,001
Taille de la famille	Un enfant	– 0,01	ns
	Deux enfants	(ref)	(ref)
	Trois enfants	– 0,19	p < 0,001
	Quatre enfants	– 0,43	p < 0,001
	Cinq à sept enfants	– 0,43	p < 0,001
	Huit enfants et plus	– 0,44	p < 0,01
Frère ou sœur dans l'enseignement supérieur	Non	(ref)	(ref)
	Oui	0,40	p < 0,001
Diplôme du père	Sans diplôme	– 0,40	p < 0,001
	Cap, bep, bepc	(ref)	(ref)
	Cep	– 0,15	p < 0,01
	Baccalauréat ou plus	0,37	p < 0,001
	inconnu	– 0,23	p < 0,001
Diplôme de la mère	Sans diplôme	– 0,53	p < 0,001
	Cap, bep, bepc	(ref)	(ref)
	Cep	– 0,39	p < 0,001
	Baccalauréat ou plus	0,64	p < 0,001
	inconnu	– 0,42	p < 0,001
PCS du chef de famille	Agriculteur	0,52	p < 0,001
	Artisan, commerçant	0,03	ns
	Cadre, chef d'entreprise	0,68	p < 0,001
	Profession intermédiaire	0,46	p < 0,001
	Employé	0,12	p < 0,05
	Ouvrier qualifié	(ref)	(ref)
	Ouvrier non qualifié	0,00	ns
	Inactif	– 0,10	ns

Source : Louis-André Vallet et Jean-Paul Caille, *op. cit.*, p. 10.
Lecture : « (ref) » signifie qu'il s'agit d'une modalité de référence ; « ns » signifie que le test de non-nullité du coefficient n'est pas significatif et que la modalité n'a pas d'effet.

Il est courant de transformer ces coefficients en des indicateurs appelés « odds-ratio » ou « rapports logistiques de chances » et définis comme l'exponentielle du coefficient. L'odds-ratio des filles s'élève à $e^{0,67} = 1,95$. Sous réserve d'un léger abus de langage, il est usuel d'interpréter ce coefficient de la manière suivante : « Les filles ont 1,95 fois plus de chances d'être admises en seconde que les garçons, toutes choses étant égales par ailleurs. » Si on ne veut pas risquer l'abus de langage, il est simplement possible de dire que les filles ont davantage de chances que les garçons d'être admises en seconde (toutes choses égales par ailleurs) et que l'effet du sexe est plus important que la présence d'un frère ou d'une sœur dans l'enseignement supérieur (odds-ratio = $1,49 = e^{0,40}$).

Un test statistique est associé à chacune des modalités (en fait à chacun des coefficients). Ce test est un test de « non-nullité » : il teste l'hypothèse de nullité du coefficient c'est-à-dire l'hypothèse d'absence d'effet de la modalité sur la variable à expliquer. Plus la probabilité associée à ce test est faible, moins l'hypothèse de nullité est acceptable et donc plus la modalité a « probablement » un effet.

À côté de ces tests permettant de déterminer la pertinence de la présence de telle ou telle modalité dans le modèle, il existe des tests indiquant le niveau de pertinence globale du modèle. Ils sont nombreux et de présentation souvent délicate : l'essentiel est ici de retenir qu'il est possible de disposer de critères permettant de comparer la pertinence de chaque modèle et de choisir le meilleur des modèles parmi tous les modèles testés.

Un indicateur de pertinence globale d'un modèle est cependant facile à interpréter et utiliser : il consiste simplement à comparer ce que le modèle prédit avec la réalité c'est-à-dire les données issues de l'enquête. Plus le nombre de situations correctement prédites par le modèle sera élevé, plus le modèle sera pertinent (au moins du point de vue statistique).

4.4 Remarques finales

Après avoir évalué un modèle, après avoir jugé sa qualité globale et la pertinence de chacune des variables, il ne faut pas hésiter à le réviser : ajouter ou supprimer des modalités, regrouper ou recoder des variables... Il ne faut pas hésiter à faire évoluer le modèle jusqu'à ce qu'il semble pertinent, à la fois du point de vue statistique (mesure de ses qualités) et du point de vue sociologique.

Le recours aux méthodes de régression, notamment logistiques, rencontre un grand succès en sociologie depuis plusieurs années. Il ne faut toutefois pas oublier que la recherche d'effets purs et le raisonnement « toutes choses étant égales par ailleurs » peuvent masquer des effets de structure majeurs.

Dans l'exemple présent, il est tout à fait intéressant de pouvoir montrer que les enfants de nationalité étrangère ont plus de chances que les enfants français d'être admis en seconde (le coefficient associé à la modalité « étranger » est positif et l'odds-ratio s'élève à 1,35). Mais il ne faut pas oublier que pour ces enfants, l'accès en seconde est, la plupart du temps, difficile. Ils sont « victimes » d'effets de structure : ils sont souvent issus de milieux peu diplômés, avec un chef de famille exerçant une profession située en bas de l'échelle sociale... À trop décomposer les effets de chaque variable, il ne faut pas oublier que les variables sont, dans la réalité sociale, fortement imbriquées. Cette critique du raisonnement « toutes choses étant égales par ailleurs » est ancienne[1]. Elle ne doit pas nous interdire d'y avoir recours ; elle doit nous aider à rester vigilants sur les forces et faiblesses de cette méthode – comme de toute autre.

1. François Simiand et Maurice Halbwachs avaient déjà formulé de telles critiques. Voir Olivier Martin, « Raison statistique et raison sociologique chez Maurice Halbwachs », *Revue d'histoire des sciences humaines*, n° 1, 1999, p. 69-101.

QUELQUES CONSEILS POUR CONCLURE

1. COMMENT UTILISER INTELLIGEMMENT LES OUTILS STATISTIQUES ?

1.1 Ne pas oublier la réflexion sociologique

La statistique peut être vue comme une boîte à outils que le sociologue doit apprendre à utiliser. Il doit prendre conscience de son intérêt mais aussi de ses limites. Il doit saisir ses effets sur sa conception et sa représentation des phénomènes étudiés. Ces outils aident le sociologue dans son travail d'interprétation des données d'enquête mais ne s'y substituent pas : il faut se défaire de l'idée selon laquelle analyser des données quantitatives se résumerait à appuyer sur des boutons-poussoirs, à enchaîner méthodiquement et presque aveuglement des techniques statistiques, puis à commenter automatiquement les résultats chiffrés qui sortent de la machine. Le recours aux méthodes et outils informatiques ne doit pas faire oublier l'essentiel : analyser une enquête sociologique (que celle-ci soit « qualitative » ou « quantitative ») suppose de faire preuve d'« imagination sociologique », de conduire une réflexion autour d'une problématique, d'élaborer progressivement une interprétation des faits... La réflexion théorique commence avec la formulation d'une problématique et se poursuit sans discontinuer jusqu'à la production du texte (article, livre ou rapport) restituant les conclusions de l'enquête. Les logiciels informatiques, aussi puissants soient-ils, ne remplacent pas ce travail intellectuel : ce qui est vrai pour le simple traitement de texte qui nous aide simplement à mettre en page les mémoires et les textes est vrai des logiciels statistiques qui nous aident simplement à manipuler aisément de grands ensembles de données. En conséquence, il est délicat de sous-traiter aveuglément l'analyse statistique des données auprès d'un service statistique ou informatique : le sociologue comme le statisticien doivent être présents et actifs à toutes les étapes du processus de construction et d'analyse de l'enquête.

1.2 Ne pas croire en une recette unique

L'analyse de données statistiques ne consiste pas à suivre méthodiquement une suite préétablie d'opérations. Elle ne s'apparente pas à l'application d'une recette toute faite garantissant la qualité du produit fini. Le travail du sociologue s'apparente davantage au travail d'un artisan qu'au travail d'un technicien appliquant des recettes éprouvées, au travail de l'artiste qui maîtrise des gestes et des instruments qu'au travail d'un opérateur répétant, en toutes circonstances, les mêmes gestes et les mêmes techniques éprouvées. Plus généralement, ceux qui ont observé le travail des scientifiques dans leurs laboratoires en ont bien montré le caractère artisanal, non algorithmique : la recherche suppose des compétences, mais aussi pas mal d'improvisation et de tâtonnements[1]. Il ne faut donc pas hésiter à « remettre l'ouvrage sur le métier »...

Il existe toujours une part d'inconnu et donc d'improvisation durant l'analyse, d'une part parce que les données ne sont pas toujours de même nature (nombre de questions, taille de l'échantillon, profondeur des questions...) et d'autre part parce que la réalité saisie par ces données ne laisse pas voir des mécanismes identiques et des régularités semblables d'un terrain à un autre.

1.3 Accepter de tâtonner

Le travail d'analyse nécessite du temps, en raison de la nécessaire familiarisation avec les données mais aussi du processus d'interprétation. Analyser des données quantitatives revient en effet à leur donner du sens, à trouver un ordre et une logique derrière l'apparent désordre. Les « données ne parlent pas d'elles mêmes » et parvenir à leur attribuer une signification est un travail parfois de longue haleine.

Ce temps de l'analyse ne peut pas être découpé en une série de phases consacrées à un type de traitement particulier : il n'y a pas la phase de « préparation des données » (codages et recodages notamment), puis la phase de « traitements statistiques simples », puis la phase de « traitements statistiques sophistiqués » et enfin la phase de « rédaction des résultats ».

1. Nous renvoyons le lecteur aux travaux des anthropologues des sciences (pour une présentation générale, voir Olivier Martin, *Sociologie des sciences*, Paris, Nathan, 2000, chapitre 5).

S'il est clair qu'il existe une progression qui va de la réalisation de l'enquête vers la rédaction et la présentation des résultats, cette progression n'est pas unidirectionnelle. Il est fréquent que le sociologue soit amené à recoder ses variables lors d'une analyse statistique sophistiquée, qu'il doive revenir à des traitements simples pour comprendre ou mieux interpréter des résultats issus d'analyses multidimensionnelles, qu'il redéfinisse ses indicateurs synthétiques après avoir conduit une analyse multidimensionnelle ou encore qu'il revienne sur l'analyse de deux variables au moment de la rédaction finale...

Au moins dans un premier temps, il faut élaborer des hypothèses et les multiplier, il faut construire des modèles et des interprétations, même sans y croire. La croyance dans les hypothèses, les modèles et les interprétations vient peu à peu, au fur et à mesure de leurs évolutions, de leurs transformations, de leur raffinement. Il ne faut en tout cas pas attendre d'avoir l'intuition géniale, la révélation ou l'interprétation parfaite avant de se lancer dans l'analyse statistique des données.

1.4 Distinguer l'exploration et la « démonstration »

Les analyses statistiques jouent au moins deux rôles distincts. Plus précisément, elles interviennent à deux moments de nature différente du processus d'analyse et de présentation des résultats.

Premièrement, elles permettent d'explorer les données, de se familiariser avec leur réalité, de trouver progressivement des cohérences, d'identifier les relations qu'elles entretiennent... Il s'agit d'une fonction exploratoire qui aide le sociologue dans son travail de reformulation des hypothèses et d'ajustement des interprétations.

Deuxièmement, elles contribuent à la validation de ces hypothèses, à la justification de ces interprétations, à la confirmation des relations suspectées entre les variables. Il s'agit de leur fonction confirmatoire : elles permettent au sociologue de prendre confiance dans ses résultats ; elles lui permettent également de justifier son point de vue dans une publication ou une communication orale.

Toute technique statistique peut jouer ces deux rôles, exploratoire et confirmatoire, même si certaines méthodes ont une vocation davantage exploratoire (analyse factorielle, ACM, classification) et d'autres une fonction plutôt confirmatoire (régression, tests).

1.5 Ne pas chercher la preuve absolue : préférer l'accumulation d'indices

Enfin, il est inutile d'espérer trouver une preuve absolue et unique. Le schéma selon lequel on part d'hypothèses, on conçoit des expériences et on valide ou non ces hypothèses à la vue des résultats des expériences est un schéma trop simplificateur et trop vague pour correspondre à la réalité des pratiques quotidiennes de la recherche. L'idée selon laquelle une expérience (dite « expérience cruciale ») permet, à elle seule, d'accepter ou de rejeter une hypothèse est une idée généralement naïve. Elle a d'ailleurs été radicalement remise en cause par les développements récents en épistémologie et sociologie des sciences[1]. En tout cas, elle ne correspond pas à la réalité de la « preuve » en sociologie, y compris en sociologie quantitative.

La « démonstration », la validation des résultats en sociologie reposent plutôt sur des faisceaux d'arguments, sur des assemblages d'indices concordants. Il s'agit en effet plutôt d'« éléments de réponse » que de réponses absolues et incontestables. Les données et analyses constituent des indices qu'il s'agit d'agencer de façon à trouver une interprétation et à convaincre les lecteurs de sa pertinence. Les preuves irréfutables existent rarement. Il est beaucoup plus fréquent de devoir multiplier les indices et les signes partiels. La preuve en sociologie repose sur un paradigme dit indiciaire[2].

2. Comment rédiger un rapport ou un article quantitatif ?

Il n'existe pas de modèle unique pour élaborer une argumentation et pour construire un texte en sociologie quantitative. Dès lors, les seuls conseils possibles sont nécessairement généraux :

– Présenter les données (le questionnaire et l'échantillonnage, au moins dans leurs grandes lignes) ; qualifier l'échantillon en en donnant les principales caractéristiques (répartition des caractéristiques classiques ou

1. Voir Olivier Martin, *op. cit*, 2000, chapitre 3.
2. Alain Desrosières, « Du singulier au général. L'argument statistique entre la science et l'État », in *Cognition et information en société*, *Raisons pratiques*, n° 8, 1997, p. 267-282.

centrales : sexe, âge, CSP...) ; cette présentation peut se faire dès l'introduction, dans une partie méthodologique ou en annexe.

– Adopter un fil conducteur et s'y tenir : le lecteur doit comprendre le sens de la démarche dès l'introduction du texte et ne doit pas perdre pied le long de l'article. Cela signifie en particulier que les aspects les plus statistiques ne doivent pas faire oublier la finalité sociologique du texte.

– Justifier les choix méthodologiques (pour quelle raison a-t-on recours à une analyse factorielle ? pourquoi réalise-t-on une telle régression ?).

– Présenter et justifier les indicateurs synthétiques construits (à partir de quelles variables et selon quels critères d'agrégation ?), voire les principaux recodages.

– Ne pas chercher à restituer toutes les analyses et tous les traitements réalisés : se contenter d'exposer les résultats les plus probants ou ceux qui sont stratégiques pour défendre la thèse avancée.

– Ne pas submerger le lecteur par des dizaines de tableaux et de nombreuses analyses sophistiquées : mieux vaut quelques tableaux bien choisis, quelques variables synthétiques bien construites et justifiées, et quelques graphiques pédagogiques qu'une avalanche de pourcentages, de tests et de statistiques sans lien apparent les uns avec les autres.

– Choisir un critère d'arrondi pour les données publiées, en fonction de la taille de l'échantillon. Pour un échantillon de 1 000 personnes, un chiffre après la virgule est un maximum.

– Ne pas hésiter à adopter un mode de présentation des résultats qui soit bien différent du mode de découverte : il est par exemple possible d'utiliser une méthode factorielle pour identifier les variables saillantes et fortement corrélées puis d'utiliser ces résultats pour construire des variables synthétiques dont la pertinence se justifie grâce à de simples tableaux croisés.

– Enfin, le dernier conseil est certainement de lire des articles de sociologie quantitative, pour en identifier la structure et l'enchaînement des arguments et des analyses statistiques au service d'un raisonnement sociologique.

POUR EN SAVOIR PLUS

En complément des références citées dans le texte, nous indiquons ici quelques ouvrages ou articles prolongeant ce manuel : soit ils en approfondissent certains aspects ; soit ils permettent au lecteur de trouver les justifications mathématiques des méthodes et outils présentés ici.

• Sur la démarche générale de l'enquête par questionnaire, nous renvoyons au manuel dont nous prenons la suite :

SINGLY (DE) François, *L'Enquête et ses méthodes : le questionnaire*, Paris, Armand Colin, 2005 (1re éd. : Nathan, 1992).

• Sur la statistique, son histoire, les questions dont elle se saisit et les grands types de réponse qu'elle fournit :

BLUM Alain, MARTIN Olivier, *La Vérité des chiffres : une illusion ?*, Université Paris Descartes, 2009, vidéo téléchargeable sur <mediatheque.paris-descartes.fr/article.php3?id_article=3659> et sur iTunes U).

DESROSIÈRES Alain, *Gouverner par les nombres* (2 volumes), Paris, Presses de l'École des Mines de Paris, 2008.

DESROSIÈRES Alain, *La Politique des grands nombres. Histoire de la raison statistique,* Paris, La Découverte, 2000 (1re éd. 1993).

• Pour un exposé très pédagogique (sans recours aux mathématiques) des principes des tests statistiques, de l'échantillonnage et de l'inférence statistique :

SCHWARTZ Daniel, *Le Jeu de la science et du hasard. La Statistique et le vivant*, Paris, Flammarion, 1994.

La lecture de ce beau livre est conseillée à tous. Les exemples sont empruntés à la médecine et à la biologie mais ils sont éclairants pour les sciences humaines et sociales ; et ils montrent incidemment que les SHS partagent beaucoup plus de choses avec les sciences de la vie qu'on ne le croit habituellement.

• Pour un exposé de l'ensemble des techniques statistiques simples ne nécessitant pas de grandes compétences en mathématiques :

FOX William, *Statistique sociales*, Paris, De Boeck, 1999.

WANNACOTT Thomas H., WANNACOTT Ronald J., *Statistique*, Paris, Economica, 1991.

• Sur l'ensemble des méthodes statistiques et de leur justification théorique complète (supposant des compétences en mathématiques) :

SAPORTA Gilbert, *Probabilités, analyse des données et statistique*, Paris, éditions Technip, 1990 (ouvrage très complet)

TASSI Philippe, *Méthodes statistiques*, Paris, Economica, 2004.

• Sur les méthodes de sondages et les principes de construction des échantillons (avec des justifications mathématiques, mais aussi des conseils et exemples pratiques) :

ARDILLY Pascal, *Les Techniques de sondage*, Paris, Technip, 1994.

GROSBRAS Jean-Marie, *Méthodes statistiques des sondages*, Paris, Economica, 1987.

• Sur les méthodes multi-variables (multidimensionnelles) :

CIBOIS Philippe, *Les Méthodes d'analyse d'enquêtes*, Paris, PUF, 2007.

DIDIER Busca, TOUTAIN Stéphanie, *Analyse factorielle simple*, Bruxelles, De Boeck, 2009.

HOWELL David C., *Méthodes statistiques en sciences humaines*, Bruxelles, DeBoeck, 1997.

SAPORTA Gilbert, *op. cit.*

VOLLE Michel, *L'Analyse des données*, Paris, Economica, 1997.

• Ressources électroniques

Olivier Godechot et Pierre Mercklé ont créé un site Web dédié à la promotion des méthodes quantitatives en sciences sociales et à l'assistance à leur utilisation : <quanti.hypotheses.org/>.

Mon propre site propose des ressources et conseils, par exemple pour mettre facilement en ligne des questionnaires : <www.olivier-martin.fr/ressources/websurvey/index.html>.

Le site de Philippe Cibois propose des textes et des exemples : <pagesperso-orange.fr/cibois/SitePhCibois.htm>.

INDEX

11009496 - (I) - (2) - OSB 80° - C2000 - BTT

Imprimerie Nouvelle
45800 Saint-Jean de Braye
N° d'Imprimeur : 429243N
Dépôt légal : Août 2009

Imprimé en France

128

La collection universitaire de poche

Cinéma Image

Communication

Droit Science politique

Éducation

Géographie Géopolitique

Histoire

Langues

Lettres Linguistique

Philosophie Spiritualités

Psychologie Psychanalyse

Science économique Gestion

Sociologie Anthropologie

L'ANALYSE DE DONNÉES QUANTITATIVES

Ce livre répond avec clarté et rigueur aux questions majeures que se pose tout concepteur d'enquête par questionnaire (comment concevoir un échantillon ? qu'est-ce qu'un sondage représentatif, un chiffre « significatif » ou une « bonne estimation » ?) ou toute personne souhaitant analyser des données quantitatives (comment recoder des variables et concevoir des indicateurs ? comment étudier leurs relations ?).

Sans recours inutile au formalisme mathématique, il expose les principes des raisonnements statistiques et des arguments probabilistes en s'appuyant sur des exemples issus d'enquêtes récentes ou classiques. Que ce soit en sociologie ou dans les domaines des études, du marketing, des sondages d'opinion ou des enquêtes de comportement, il répond aux besoins bien identifiés des étudiants, enseignants et intervenants en sciences sociales.

***Olivier Martin**, sociologue et statisticien, est professeur à l'Université Paris Descartes, chercheur au Cerlis (Centre de recherche sur les liens sociaux, CNRS) et directeur de l'École doctorale.*

6691059
ISBN : 978-2-200-24461-3
9 782200 244613

ARMAND COLIN

Atelier Didier Thimonier