

Introduction

Les données

Ce chapitre introductif précise quelles données utiliser et, sans viser l'exhaustivité, fournit quelques pistes de recherche pour qui souhaite fonder son étude sur des données existantes. Dans la mesure où l'ouvrage s'appuie en grande partie (mais pas seulement cependant) sur des exemples scientifiques issus d'enquêtes par sondage (c'est-à-dire d'une interrogation d'un échantillon d'individus par le moyen d'un questionnaire en vue de saisir leurs comportements, pratiques, attitudes, etc.), nous avons souhaité en outre, dans cette introduction, préciser succinctement quelques spécificités liées à ce mode de recueil de l'information sociale et les précautions indispensables à prendre lors de son emploi. Il présente en particulier trois notions fondamentales liées à l'échantillonnage : la marge d'erreur, les principaux biais de la mesure par ce moyen et les méthodes de redressement¹.

I. Quel type de données ?

I.1. Définitions : recensement et sondage, données individuelles et agrégées

Pour répondre à une question sociale, on peut penser de prime abord que l'idéal est de disposer de données exhaustives, c'est-à-dire d'une information portant sur l'ensemble de la population composant son objet d'étude ou sa population d'intérêt (les lycéens français, les électeurs marseillais, etc.). C'est le cas lorsque l'on procède par ce que l'on appelle un *recensement*. Si le plus connu d'entre eux en France est le *Recensement général de la population*, toute enquête

¹ Nous n'évoquerons pas ici la question de la protection des données personnelles et de l'anonymisation, très importante dans le cadre de l'utilisation de données. Nous renvoyons néanmoins le lecteur au site Internet de la Commission Nationale de l'Informatique et des Libertés (CNIL) <<http://www.cnil.fr/>> (consulté le 5 juin 2014), ainsi qu'au chapitre écrit par Roxane Silberman : « Chapitre 12 : La protection des données individuelles en France et la recherche en sciences sociales », Alain Chenu, Laurent Lesnard (dir.), *La France dans les comparaisons internationales : guide d'accès aux grandes enquêtes statistiques en sciences sociales*, Paris, Presses de Sciences Po, 2011, p. 183-204.

De la même manière, ce manuel de méthodologie n'est pas un ouvrage de sociologie de la mesure de l'opinion et du débat qui entoure cette question. Des éléments de réflexion à cet égard sont disponibles dans les articles suivants : Loïc Blondiaux, « Ce que les sondages font à l'opinion publique », *Politix*, vol. 10, n°37, 1997, p. 117-136 ; Pierre Bourdieu, « L'opinion publique n'existe pas », *Questions de sociologie*, Paris, Les Éditions de Minuit, 1980, p. 222-235 ; Claude Dargent, *Sociologie des opinions*, Paris, Armand Colin, 2011, 240 p. ; Gérard Grunberg, Nonna Mayer, Paul Sniderman (dir.), *La démocratie à l'épreuve : une nouvelle approche de l'opinion des Français*, Paris, Presses de Sciences po, 2002, 349 p. ; Nonna Mayer, *Sociologie des comportements politiques*, Paris, Armand Colin, 2010, 288 p. ; Paul M. Sniderman, « Les nouvelles perspectives de la recherche sur l'opinion publique », *Politix*, n°41, 1998, p. 123-175.

par questionnaire qui vise à interroger l'entièreté d'une population d'intérêt se nomme en réalité un recensement¹.

À moins que l'objet d'enquête porte sur une population de petite taille et facile d'accès, les coûts de collecte associés à ce mode de recueil sont cependant considérables, en termes financiers et humains, mais aussi en temps. Dans la plupart des cas, on ne dispose pas de l'ensemble de ces moyens : on préfère alors procéder par sondage, ce qui signifie que l'on s'appuiera sur une extraction, un échantillon de la population-mère. Les données ainsi obtenues sont donc partielles et ne représentent qu'une partie de la population d'intérêt. On parle dans ce cas de *données d'échantillon* et la qualité de l'échantillonnage mis en œuvre (c'est-à-dire des techniques déployées pour procéder à l'extraction²) est alors essentielle pour obtenir des données fiables et analysables ensuite du point de vue scientifique. Il s'agit en effet cette fois d'obtenir un miroir en modèle réduit de la population étudiée de telle sorte qu'il lui soit le plus fidèle, et donc le plus représentatif possible. Il existe bien entendu une palette d'outils et de précautions à respecter pour atteindre cet objectif indépassable.

Par ailleurs, les données peuvent être classées en deux grands types :

- les données individuelles (micro) : elles portent sur les unités statistiques de base et sont issues la plupart du temps d'enquêtes par sondage. On pense en tout premier lieu aux personnes physiques interrogées par questionnaire ;
- les données agrégées (macro) : elles sont issues de l'agrégation de données individuelles et proviennent généralement d'organismes institutionnels producteurs de données (voir ci-dessous). Il peut s'agir de quartiers, d'établissements scolaires, de divisions administratives – communes, départements, régions –, etc.

S'il est aisé de transformer des données individuelles en données agrégées, l'inverse n'est pas vrai. Prenons un exemple fictif : un fichier constitué d'individus d'une commune renseignant leur statut professionnel et leur quartier d'habitation. Le tableau correspondant serait le suivant (tableau 1).

¹ Ainsi, le recensement de la population, assuré par l'Institut national de la statistique et des études économiques (INSEE) en France, se fonde sur deux questionnaires, l'un portant sur le logement (feuille de logement), l'autre sur les individus qui le composent (bulletin individuel). Il consiste non seulement à compter les personnes qui vivent sur le territoire, mais également à connaître leurs principales caractéristiques sociodémographiques (âge, sexe, composition des familles, niveau de formation, etc.) et conditions de vie (conditions de logement en particulier, ainsi que conditions d'emploi, etc.). Notons néanmoins que les techniques de recueil ont été modifiées à partir de 2004 (annualisation).

² Se reporter à la partie II de ce chapitre pour une présentation générale des principales méthodes d'échantillonnage existantes.

Tableau 1 : Exemple de données individuelles, statut professionnel en trois catégories et quartier d'habitation (extraction)

Individus	Statut professionnel	Quartier d'habitation
1	Chômeur	A
2	Actif non chômeur	B
3	Inactif	B
4	Actif non chômeur	B
5	Actif non chômeur	B
6	Actif non chômeur	C
7	Actif non chômeur	A
8	Chômeur	C
9	Inactif	C
10	Inactif	C
11	Inactif	A
12	Actif non chômeur	A
13	Inactif	A
14	Actif non chômeur	A
15	Chômeur	B
16	Chômeur	B
17	Inactif	C
18	Chômeur	A
19	Actif non chômeur	A
20	Actif non chômeur	C
21	Actif non chômeur	A
22	Chômeur	B
23	Chômeur	A
24	Chômeur	C
25	Actif non chômeur	C
26	Inactif	A
27	Actif non chômeur	B
28	Actif non chômeur	C
29	Inactif	B
30	Actif non chômeur	A
etc.	etc.	etc.

Sources : données fictives

Il est possible d'agréger facilement la variable statut professionnel par quartier d'habitation afin d'obtenir les taux de chômage dans chaque quartier (tableau 2).

Tableau 2 : Exemple de données agrégées, taux de chômage selon le quartier d'habitation, %

Quartier	Taux de chômage
A	14
B	11
C	8

Sources : données fictives

En revanche, si l'on disposait comme unique source d'informations du tableau 2, il serait impossible de retrouver les détails individuels (ici le statut professionnel de chaque habitant).

I.2. Où trouver des données ?

Il existe de nombreuses sources, en particulier en ligne, donnant accès à des données sociales portant sur des objets d'étude infinis¹. En voici certaines d'entre elles, couramment utilisées par les auteurs du fait de leurs spécialités thématiques. Il va sans dire qu'au vu de l'ampleur des données accessibles, le catalogue pourrait être très largement étoffé. En outre, le mouvement des *open data* met à disposition de plus en plus de données et rendra certainement assez vite caduque ce premier inventaire. Celui-ci reste néanmoins susceptible de guider en première instance le lecteur peu familier des méandres de la statistique sociale disponible en ligne.

Où trouver des données agrégées ?²

- pour les données concernant principalement **la France**, plusieurs organismes généralistes utiles existent, en particulier l'Institut national de la statistique et des études économiques (INSEE³), l'Institut national d'études démographiques (INED⁴), le Portail de la statistique publique⁵, la Plateforme ouverte des données publiques françaises (DATAGOUV⁶) ou encore le Centre de recherche pour l'étude et l'observation des conditions de vie (CREDOC⁷);
- pour les **données européennes ou internationales**, les équivalents étrangers de l'INSEE sont indiqués sur son site Internet⁸, et d'autres organismes sont également importants à connaître, comme l'Organisation de coopération et de développement économiques (OCDE⁹), EUROSTAT¹⁰, la division statistique de l'Organisation des Nations Unies (ONU¹¹) ou encore la Data world bank¹²;
- pour les **données cartographiques**, de multiples sources, généralistes ou spécifiques, existent également. Citons par exemple l'atelier de cartographie de Sciences Po¹³, ainsi que deux outils de cartographie électorale, CARTELEC¹⁴ et VIZLAB¹⁵.

¹ Les listes qui suivent n'ont pas la prétention d'être exhaustives. Elles permettent néanmoins d'avoir une vision globale des principales sources de données existantes. Notons par ailleurs que les droits d'accès sont plus ou moins stricts selon les sources citées.

² Nous indiquons ici les liens vers la page d'accueil de chacun des organismes ou banques de données. Pour les premiers, un onglet « bases de données » ou « statistiques » est généralement disponible dès l'ouverture de cette page d'accueil.

³ <<http://www.insee.fr/fr/>>, consulté le 19 mai 2014.

⁴ <<http://www.ined.fr/>>, consulté le 19 mai 2014.

⁵ <<http://www.statistique-publique.fr/>>, consulté le 19 mai 2014.

⁶ <<http://www.data.gouv.fr/>>, consulté le 19 mai 2014.

⁷ <<http://www.credoc.fr/>>, consulté le 19 mai 2014.

⁸ <<http://www.insee.fr/fr/Insee-statistique-publique/default.asp?page=sites-statistiques/instituts-nationaux-statistiques.htm>>, consulté le 19 mai 2014.

⁹ <<http://www.oecd.org/fr/>>, consulté le 19 mai 2014.

¹⁰ <<http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/>>, consulté le 19 mai 2014.

¹¹ <<http://www.un.org/fr/databases/index.shtml>>, consulté le 19 mai 2014.

¹² <<http://data.worldbank.org/>>, consulté le 19 mai 2014.

¹³ <<http://cartographie.sciences-po.fr/>>, consulté le 19 mai 2014.

¹⁴ <<http://www.cartelec.net/>>, consulté le 19 mai 2014.

¹⁵ <<http://cdsp.sciences-po.fr/vizlab/?locale=fr>>, consulté le 19 mai 2014.

Où trouver des données individuelles ?

- en France, le réseau Quetelet¹ et le Centre de données socio-politiques de Sciences Po (CDSP²) archivent et diffusent les données d'enquête par questionnaire en sciences humaines et sociales ;
- le consortium interuniversitaire pour la recherche en science politique et en sociologie (ICPSR³) ainsi que l'institut de Leibniz pour les sciences sociales (GESIS⁴) archivent et diffusent les données à l'international. Chaque grande enquête comparative possède également son propre site Internet⁵, en particulier les enquêtes Eurobaromètres (EB⁶), celles de l'*International social survey programme* (ISSP⁷), les *European value studies* (EVS⁸) et *World value studies* (WVS⁹), les enquêtes *European social survey* (ESS¹⁰), ainsi que les autres baromètres régionaux (Latino barometer¹¹, African barometer¹², Asian barometer¹³, Arab barometer¹⁴). Enfin le Pew research center¹⁵ donne également accès à des données utiles.

Voici enfin les principaux instituts de sondage en France : BVA¹⁶, CSA¹⁷, GFK¹⁸, IFOP¹⁹, IPSOS²⁰, LH2²¹, OPINION WAY²², TNS-SOFRES²³.

¹ <<http://www.reseau-quetelet.cnrs.fr/>>, consulté le 19 mai 2014. Il est intéressant de noter que le réseau Quetelet dispose également d'un outil spécifique, la Base De Questions (BDQ), permettant de rechercher une question spécifique d'un des questionnaires archivés à l'adresse suivante <<http://bdq.reseau-quetelet.cnrs.fr/fr/Accueil>>, consulté le 19 mai 2014.

² <<http://cdsp.sciences-po.fr/>>, consulté le 19 mai 2014.

³ <<http://www.icpsr.umich.edu>>, consulté le 19 mai 2014.

⁴ <<http://zacat.gesis.org>>, consulté le 12 juin 2014.

⁵ Pour une présentation détaillée des trois principales enquêtes européennes (outre l'Enquête sociale européenne) : Pierre Bréchon, « Les Grandes enquêtes internationales (Eurobaromètres, Valeurs, ISSP) : apports et limites », *Année sociologique*, vol. 52, n°1, 2002, p. 105-130.

⁶ <http://ec.europa.eu/public_opinion/index_en.htm>, consulté le 19 mai 2014

⁷ <<http://www.issp.org/>>, consulté le 19 mai 2014.

⁸ <<http://www.europeanvaluesstudy.eu/>>, consulté le 19 mai 2014.

⁹ <<http://www.worldvaluessurvey.org/>>, consulté le 19 mai 2014.

¹⁰ <<http://www.europeansocialsurvey.org/>>, consulté le 19 mai 2014.

¹¹ <<http://www.latinobarometro.org>>, consulté le 19 mai 2014.

¹² <<http://www.afrobarometer.org>>, consulté le 19 mai 2014.

¹³ <<http://www.asianbarometer.org>>, consulté le 19 mai 2014.

¹⁴ <<http://www.arabbarometer.org>>, consulté le 19 mai 2014.

¹⁵ <<http://www.pewresearch.org>>, consulté le 19 mai 2014.

¹⁶ <<http://www.bva.fr>>, consulté le 19 mai 2014.

¹⁷ <<http://www.csa.eu>>, consulté le 19 mai 2014.

¹⁸ <<http://www.gfk.com/fr>>, consulté le 19 mai 2014.

¹⁹ <<http://www.ifop.com>>, consulté le 19 mai 2014.

²⁰ <<http://www.ipsos.fr>>, consulté le 19 mai 2014.

²¹ <<http://www.lh2.fr>>, consulté le 19 mai 2014.

²² <<http://www.opinion-way.com>>, consulté le 19 mai 2014.

²³ <<http://www.tns-sofres.com>>, consulté le 19 mai 2014.

II. Les données issues d'enquêtes par sondage

L'élaboration d'une enquête par questionnaire (d'un sondage) nécessite la plupart du temps de procéder à un échantillonnage de la population d'intérêt pour n'en interroger qu'un sous-ensemble que l'on souhaite le plus fidèle possible (il n'est généralement pas envisageable d'interroger la totalité des sujets qui composent son objet d'étude). Cela consiste à en extraire une partie appelée échantillon. Celui-ci est généralement de petite taille comparée à celle de la population-mère. Pour pouvoir conclure du point de vue de la population d'intérêt à partir de cet échantillon, il importe que ce dernier soit constitué avec le plus grand soin : de la qualité de cette phase technique dépend en effet la fiabilité de l'échantillon et donc les résultats du sondage. On parlera ici de la *représentativité* de l'échantillon, qui conditionne en grande partie les possibilités d'analyse. Différentes méthodes d'échantillonnage, aux caractéristiques techniques spécifiques et adaptées à des contextes d'utilisation particuliers, peuvent être employées. Nous en esquissons un rapide portrait ici, afin que le lecteur puisse saisir la complexité de cette étape, qui requiert une grande maîtrise et un long travail de réflexion avant toute mise en œuvre concrète.

Ce qu'il doit absolument retenir, c'est qu'un échantillon, par construction et quels que soient les efforts déployés pour l'élaborer avec la plus grande rigueur, n'en demeure pas moins un échantillon. Puisque la population d'intérêt dans son ensemble n'aura pas été touchée, il s'accompagne de fait d'une incertitude dans sa construction (un échantillon n'est qu'une approximation) et d'une erreur de mesure contre laquelle on ne peut rien, si ce n'est ne pas l'ignorer.

En outre, il ne s'agit pas de l'unique erreur de mesure de ce type d'outil de collecte de l'information. Nous présenterons également deux biais supplémentaires importants de l'enquête par questionnaire, qu'il convient d'avoir en tête avant toute analyse de ce type de matériau : une telle connaissance permet en effet d'éviter les surinterprétations et les mésusages malheureux.

Précisons néanmoins d'ores et déjà que parler d'incertitude (et de biais) n'invalidé en aucun cas la démarche, à partir du moment où l'on s'efforce d'une part de la minimiser, d'autre part d'en tenir compte dans son approche et de faire avec.

II.1. Les principales méthodes d'échantillonnage¹ et leur biais intrinsèque : nous avons affaire à des échantillons !

Il existe deux grandes familles d'échantillonnage :

- l'échantillonnage *aléatoire*, aussi appelé *probabiliste* : ici, nous sommes dans l'univers du hasard. Il consiste à tirer au sort, au sein de la population

¹ Le lecteur souhaitant en apprendre davantage sur les méthodes d'échantillonnage et leurs implications mathématiques peut se référer à Pascal Ardilly, *Les techniques de sondage*, Paris, Technip, 2006, 675 p. Une présentation plus accessible au lecteur débutant est consultable dans : Jean-Paul Bozonnet, Pierre Bréchon, « Établir un échantillon représentatif », Pierre Bréchon (dir.), *Enquêtes qualitatives, enquêtes quantitatives*, Grenoble, Presses universitaires de Grenoble, 2011, p. 123-143. Enfin, nous invitons le lecteur à consulter les sites Internet des grandes enquêtes internationales et les sites de diffusion d'enquête (précédemment indiqués) qui fournissent un vivier considérable d'exemples d'échantillonnage (chaque enquête présente en détail sa méthode de collecte).

- d'intérêt et selon des techniques plus ou moins originales, novatrices et complexes, les personnes que l'on va effectivement contacter pour être interrogées. La probabilité pour un individu de figurer dans l'échantillon est appelée *probabilité d'inclusion* et découle de la loi des grands nombres. Celle-ci dépend du type d'échantillonnage aléatoire utilisé mais présente l'intérêt d'être connue. C'est ainsi que l'on peut estimer l'incertitude, c'est-à-dire l'imprécision qui accompagne le sous-ensemble ainsi constitué. Dans ses formules les plus simples qui consistent à s'appuyer sur des listes d'individus au sein desquelles on tire au sort (chaque personne figurant dans la liste ayant la même chance ou probabilité d'être choisie), cette technique nécessite une connaissance fine de la population-mère et implique la possibilité d'accéder à des listes nominatives, ce qui n'est pas toujours possible, en France en particulier. C'est alors qu'il faut user d'inventivité dans la méthode de prélèvement aléatoire (mais c'est généralement très coûteux financièrement) ou procéder de façon non probabiliste ;
- l'échantillonnage *non aléatoire*, aussi appelé *empirique* : les méthodes aléatoires présentant un coût de mise en œuvre élevé, la méthode empirique des quotas est souvent utilisée en France. Lorsque l'on procède de la sorte, il s'agit de reproduire dans l'échantillon les proportions observées, dans la population d'intérêt, de certaines caractéristiques jugées déterminantes. Par exemple, la population française comporte environ 52% de femmes. Un échantillon de 1000 personnes (souhaitant reproduire la population française) utilisant un quota de sexe devra comporter 520 femmes et 480 hommes. De la sorte, ces catégories de population ne seront ni sur-, ni sous-représentées dans l'échantillon. Plus l'on mobilisera de quotas différents, plus il s'avèrera délicat de tirer un échantillon correspondant exactement aux distributions de chacun d'entre eux. L'objectif sera donc de s'en rapprocher au plus près pour aboutir à un échantillon le moins déformé possible. Il conviendra également de ne pas multiplier à l'excès les critères retenus. Classiquement et parce qu'ils sont connus dans la population générale grâce au recensement de la population effectué par l'Insee, ce sont les quotas d'âge, de sexe et de profession, assortis d'un contrôle par la région et la catégorie d'agglomération d'appartenance, auxquels les sondages d'opinion ont recours. L'univers académique a coutume, quand cela est possible, d'y ajouter un critère de niveau de diplôme (la question du coût entre en ligne de compte pour tout rajout de quota). Cependant, par cette méthode d'échantillonnage, la probabilité d'inclusion est inconnue et ne peut donc être, tout au moins en théorie, calculée. La conséquence la plus forte pour ce qui nous concerne est qu'il serait alors impossible d'estimer l'incertitude (l'erreur) qui accompagne notre échantillon. Ceci explique le rejet et la méfiance généralement suscités par cette méthode aux fondements non scientifiques dans l'univers des statisticiens.

Revenons un instant sur la question de la mesure de l'incertitude en fonction de ces deux grandes familles d'échantillonnage. Comme nous l'avons souligné, les fondements mathématiques et probabilistes sur lesquels s'appuient les démonstrations liées à l'inférence statistique (c'est-à-dire l'extrapolation des tendances observées dans son échantillon à l'ensemble de la population mère), à la marge d'erreur et au calcul d'intervalles de confiance (*cf. chapitre 4*) impliquent un échantillonnage aléatoire et s'appliquent donc, en théorie, uniquement dans ce cadre. En d'autres termes, un échantillon par quota ne pourrait faire l'objet de traitements statistiques ayant recours à ces méthodes et aucune mesure de l'incertitude ne serait permise.

Nous souhaitons néanmoins nuancer cette première conclusion et défendons au contraire une thèse inverse. Une longue tradition de sondage par quota nous indique même la pertinence de ce type de pratiques. En cela, nous rejoignons Jean-Paul Bozonnet et Pierre Bréchon : « Malgré l'absence de fondement scientifique, cette technique "marche" aussi bien que la méthode aléatoire. Sa fiabilité est démontrée par la constance des résultats obtenus lors de sondages successifs sur des questions qui ne sont pas soumises aux aléas de l'opinion. (...) Il est vrai qu'en toute rigueur la méthode aléatoire est la seule à autoriser le calcul d'une marge d'erreur. Mais, si on considère que la procédure des quotas est en fait une reconstitution de l'aléa à moindre coût, on peut en déduire que les marges d'erreur valables pour la méthodologie aléatoire le sont aussi pour les quotas »¹.

En suivant la ligne de ces auteurs, nous recommandons donc au lecteur :

- 1. d'admettre que quelle que soit la façon dont il a constitué son échantillon il ne s'agit que d'un échantillon et que par conséquent, il ne peut pas être une copie à l'exact identique de sa population-mère. Même s'il s'en rapproche, il existe par conséquent une incertitude, même très faible, appelée marge d'erreur, que l'on peut plus ou moins bien reconstituer et qui varie en fonction de trois grands critères (l'encart 1 développe la notion de marge d'erreur) ;
- 2. à défaut de pouvoir les calculer rigoureusement quel que soit son mode d'échantillonnage et plutôt que de ne rien faire, mieux vaut appliquer les marges d'erreur que l'on connaît dans le cas de la procédure aléatoire à d'autres types d'échantillons². Ces derniers devront toutefois être élaborés en fonction de critères de rigueur déjà éprouvés. Un échantillon élaboré « au petit bonheur la chance », c'est-à-dire sans réflexion approfondie préalable sur les procédures à mettre en œuvre et sans travail de documentation fouillée sur les méthodes existantes, n'entre pas dans cette catégorie. Nous excluons également ici tout échantillon au sein duquel la personne interrogée n'est pas préalablement et rigoureusement sélectionnée par l'analyste. Nous pouvons citer ici l'exemple des expériences de type *vote de paille* : poser une question à plusieurs millions d'individus sans prendre en compte le moindre critère statistique dans leur

¹ Jean-Paul Bozonnet et Pierre Bréchon, *op. cit.*, p. 135 et 140.

² Des études ont été effectuées pour montrer qu'il est possible d'approximer la marge d'erreur dans le cas d'échantillons construits de manière non aléatoire, notamment Jean-Claude Deville, « Une théorie des enquêtes par quota », *Techniques d'enquête*, vol. 17, n°2, 1991, p. 163-181.