Yves Tillé

Professeur à l'Université de Neuchâtel (Suisse)

Théorie des sondages

Échantillonage et estimation en populations finies

2e édition

DUNOD

Illustration de couverture : © Vijay kumar – istock.com

11, rue Paul Bert, 92240 Malakoff
www.dunod.com

ISBN 978-2-10-079355-6

Préface

La première version de ce livre a été publiée en 2001, l'année où j'ai quitté l'Ecole Nationale de la Statistique et de l'Analyse de l'Information à Rennes pour enseigner à l'Université de Neuchâtel. Cette version était issue de plusieurs supports des cours de théorie des sondages que j'avais enseignés à Rennes. A l'ENSAI, la collaboration avec Jean-Claude Deville a été particulièrement stimulante.

La rédaction de cette nouvelle édition a été laborieuse et s'est faite par à-coups. Je remercie toutes les personnes qui ont examiné les versions préliminaires et qui m'ont fait part de leurs remarques. Merci particulièrement à Monique Graf et Alexandre Oettli pour leurs relectures de certains chapitres et à Alexandrine Améziane pour sa correction méticuleuse de l'orthographe des épreuves.

Les presque vingt années que j'ai passées à Neuchâtel ont été émaillées de multiples péripéties. Je suis particulièrement reconnaissant à Philippe Eichenberger et Jean-Pierre Renfer qui ont successivement dirigé la Section des Méthodes Statistiques de l'Office Fédéral de la Statistique. Leur confiance et leur professionnalisme ont contribué à nouer un échange fécond entre l'Institut de Statistique de l'Université et l'Office Fédéral de la Statistique.

Je suis aussi très redevable aux doctorants que j'ai eu le plaisir d'encadrer jusqu'ici. Chaque thèse est une aventure qui apprend autant au superviseur qu'au doctorant. Merci donc à Alina Matei, Lionel Qualité, Desislava Nedyalkova, Erika Antal, Matti Langel, Toky Randrianasolo, Éric Graf, Caren Hasler, Matthieu Wilhelm, Mihaela Guinand-Anastasiade et Audrey-Anne Vallée qui m'ont fait confiance et que j'ai eu le plaisir d'encadrer pendant quelques années.

Yves Tillé, Neuchâtel 2019

Préface de la première édition

Cet ouvrage reprend un matériel pédagogique que j'ai commencé à mettre au point en 1994. Tous les chapitres ont en effet servi de support à un enseignement, un cours, une formation, un atelier ou un séminaire. En regroupant ce matériel, j'espère présenter un ensemble de résultats cohérents et modernes sur l'échantillonnage, l'estimation et le traitement des non-réponses, autrement dit sur la totalité des opérations statistiques d'une enquête par sondage classique.

En réalisant ce livre, mon but n'est pas de fournir un aperçu exhaustif de la théorie des sondages, mais plutôt de montrer que la théorie de l'échantillonnage est une discipline vivante, ayant un champ d'application très large. Si, dans plusieurs chapitres, des démonstrations ont été écartées, j'ai toujours veillé à renvoyer le lecteur vers des références bibliographiques. L'abondance de publications très récentes atteste d'ailleurs de la fécondité des années 90 en la matière. Tous les développements présentés dans ce livre se basent sur l'approche dite « basée sur le plan de sondage ». Il existe en théorie des sondages un autre point de vue fondé sur une modélisation de la population. J'ai laissé intentionnellement cette approche de côté, non par désintérêt, mais pour proposer une démarche que je juge cohérente et déontologiquement acceptable pour le statisticien public.

Je voudrais remercier toutes les personnes qui, d'une manière ou d'une autre, m'ont aidé à réaliser ce livre : Laurence Broze qui m'a confié mon premier cours de sondage à l'Université Lille 3, Carl Särndal qui m'a encouragé à plusieurs reprises, Yves Berger avec qui j'ai partagé un bureau à l'Université libre de Bruxelles (ULB) pendant plusieurs années et qui m'a fait part d'une multitude de remarques pertinentes. Mes remerciements vont aussi à Antonio Canedo qui m'a appris à utiliser LaTeX, à Lydia Zaïd qui a corrigé à plusieurs reprises le manuscrit et à Jean Dumais pour ses nombreux commentaires constructifs.

J'ai rédigé l'essentiel de ce livre à l'École Nationale de la Statistique et de l'Analyse de l'Information (ENSAI). La chaleureuse ambiance qui a régné au sein du Département de Statistique m'a procuré un soutien important. Je remercie particulièrement mes collègues Fabienne Gaude, Camelia Goga et Sylvie Rousseau qui ont méticuleusement relu le manuscrit et Germaine Razé qui a assuré le travail de reproduction des épreuves. Plusieurs exercices sont dus à Pascal Ardilly, Jean-Claude Deville et Laurent Wilms. Je tiens à les remercier de m'avoir autorisé à les reproduire. Ma gratitude va tout particulièrement à Jean-Claude Deville pour notre fructueuse collaboration au sein du Laboratoire de Statistique d'Enquête (LSE) du Centre de Recherche en

Économie et en Statistique (CREST). Les chapitres concernant l'échantillonnage par scission et les plans équilibrés reprennent par ailleurs des travaux de recherche que nous avons réalisés ensemble.

Yves Tillé, Bruz 2001

Table des matières

Chapitre 1

Une histoire des idées en théorie des sondages

1.1 Introduction

Avec le recul, les débats qui ont animé une discipline scientifique apparaissent souvent futiles. L'histoire de la théorie des sondages se révèle pourtant particulièrement instructive. Cette théorie est une des spécialisations de la statistique qui a elle-même une position un peu particulière, car elle est utilisée dans presque toutes les disciplines scientifiques. La statistique est indissociable de ses champs d'application puisqu'elle détermine comment des données doivent être traitées. La statistique est la clé de voûte des méthodes scientifiques quantitatives. En effet, il n'est pas possible de déterminer la pertinence de l'application d'une technique statistique sans se référer aux méthodes scientifiques des disciplines où elle est appliquée.

La vérité scientifique est souvent présentée comme le consensus ponctuel d'une communauté scientifique. L'histoire d'une discipline scientifique est donc l'histoire de ces consensus et surtout de leurs changements. Depuis les travaux de Thomas Samuel Kuhn, on considère que les sciences se développent autour de paradigmes qui sont « des modèles qui donnent naissance à des traditions particulières et cohérentes de recherche » (Kuhn, 1983, p. 30). Ces modèles présentent deux caractéristiques : « leurs accomplissements étaient suffisamment remarquables pour soustraire un groupe cohérent d'adeptes à d'autres formes d'activités scientifiques concurrentes ; d'autre part, ils ouvraient des perspectives suffisamment vastes pour fournir à ce nouveau groupe de chercheurs toutes sortes de problèmes à résoudre » (voir Kuhn, 1983, p. 31).

De nombreux auteurs ont proposé une chronologie des découvertes en théorie des sondages qui rend compte des importantes polémiques qui ont jalonné son développement (voir entre autres Hansen & Madow, 1974, Hansen *et al.*, 1983, Owen & Cochran, 1976, Sheynin, 1986, Stigler, 1986). Bellhouse (1988a) lit cette chronologie comme une histoire des grandes idées qui ont contribué au développement de la théorie des sondages. La statistique est une science particulière. Avec les mathématiques

pour outil, elle permet de finaliser la méthodologie des autres disciplines. En raison de la corrélation étroite entre la méthode et la multiplicité de ses champs d'action, la statistique repose sur une multitude d'idées différentes issues des diverses disciplines où elle est appliquée.

La théorie des sondages a joué un rôle prépondérant dans le développement de la statistique. Cependant, l'utilisation de techniques d'échantillonnage n'a été acceptée que très récemment. Parmi les polémiques qui ont animé cette théorie, on retrouve certains des débats classiques de la statistique mathématique, comme la place de la modélisation et la discussion sur des techniques d'estimation. La théorie des sondages a été tiraillée entre les grands courants de la statistique et a donné naissance à de multiples approches : basée sur le plan, basée sur un modèle, assistée par un modèle, prédictive ou bayésienne.

1.2 La statistique énumérative du 19e siècle

On trouve dans Droesbeke *et al*. (1987) plusieurs exemples de tentatives d'extrapolation de données partielles à la totalité de la population dès le Moyen Âge. En 1783, en France, Pierre Simon de Laplace (1847) présente à l'Académie des Sciences une méthode permettant de déterminer le nombre d'habitants à partir des registres de naissances en utilisant un échantillon de régions. Celui-ci propose de calculer, à partir de cet échantillon de régions, le ratio du nombre d'habitants sur le nombre de naissances et ensuite de le multiplier par le nombre total de naissances qui peut être obtenu avec précision pour toute la population. Laplace suggère même d'estimer « l'erreur à craindre » en faisant référence au théorème central limite. De plus, il préconise l'usage d'un estimateur par le quotient, en utilisant le nombre total de naissances comme information auxiliaire. La méthodologie des sondages ainsi que les outils probabilistes sont donc déjà connus avant le 19e siècle. Cependant, jamais au cours de cette période, un consensus n'a existé au sujet de sa validité.

Le développement de la statistique (étymologiquement, de l'allemand : science de l'État) est indissociable de l'émergence des États modernes au 19e siècle. Une des personnalités les plus marquantes de la statistique officielle du 19e siècle est le Belge Adolphe Quételet (1796–1874). Celui-ci connaissait la méthode de Laplace avec lequel il avait entretenu une correspondance. Selon Stigler (1986, pp. 164–165), Quételet fut d'abord attiré par l'idée d'utiliser des données partielles. Il tenta même d'appliquer la méthode de Laplace pour estimer la population des Pays-Bas en 1824 (dont la Belgique a fait partie jusqu'en 1830). Toutefois, il semble qu'il s'est alors rallié à une note de Keverberg (1827) qui critique sévèrement l'utilisation de données partielles au nom de la précision et de l'exactitude :

> À mon avis, il n'existe qu'un seul moyen de parvenir à une connaissance exacte de la population et des élémens dont elle se compose : c'est celle d'un dénombrement effectif et détaillé ; c'est-à-dire de la formation d'états nominatifs de tous les habitans, avec indication de leur âge et de leur profession. Ce n'est que par ce mode d'opérer, qu'on peut obtenir des documens dignes de confiance sur le nombre réel d'habitans d'un pays, et en même temps sur la statistique des âges dont la population se compose, et des branches d'industrie dans lesquelles elle trouve des moyens d'aisance et de prospérité.

Dans une de ses Lettres au Duc de Saxe-Cobourg Gotha, Quételet (1846, p. 293) plaide également en faveur du relevé exhaustif :

> La Place avait proposé de substituer au recensement d'un grand pays, tel que la France, quelques recensements particuliers dans des départements choisis, où ce genre d'opération pouvait avoir plus de chances de succès, puis d'y déterminer avec soin le rapport de la population soit aux naissances soit aux décès. Au moyen de ces rapports des naissances et des décès de tous les autres départements, chiffres qu'on peut constater avec assez d'exactitude, il devient facile ensuite de déterminer la population de tout le royaume. Cette manière d'opérer est très expéditive, mais elle suppose un rapport invariable en passant d'un département à un autre. [···] Cette méthode indirecte doit être évitée autant que possible, bien qu'elle puisse être utile dans certains cas, où l'administration aurait à procéder avec rapidité ; on peut aussi l'employer avec avantage comme moyen de contrôle.

Il est intéressant d'examiner l'argument utilisé par Quételet (1846, p. 293), pour justifier sa position.

> Ne pas se procurer la faculté de vérifier les documents que l'on réunit, c'est manquer à l'une des principales règles de la science. La statistique n'a de valeur que par son exactitude ; sans cette qualité essentielle, elle devient nulle, dangereuse même puisqu'elle conduit à l'erreur.

De nouveau, l'exactitude est considérée comme un principe de base de la science statistique. Malgré l'existence des outils probabilistes, malgré diverses applications de techniques de sondage, l'utilisation de données partielles fut perçue comme une méthode douteuse et peu scientifique. Quételet eut une grande influence sur le développement de la statistique officielle. Il participa à la création d'une section de Statistique au sein de la *British Association of Advancement of Sciences* en 1833 avec Thomas Malthus et Charles Babbage (voir à ce sujet Horvàth, 1974). Un de ses objectifs visait l'harmonisation de la production des statistiques officielles. Il organisa le premier Congrès international de la Statistique à Bruxelles en 1853. Quételet connaissait bien les systèmes administratifs de la France, du Royaume-Uni, des Pays-Bas et de la Belgique. Il a vraisemblablement contribué à faire admettre l'idée que l'utilisation de données partielles est peu scientifique.

Certaines personnalités, comme Malthus et Babbage en Grande-Bretagne et Quételet en Belgique, ont grandement concouru au développement de la méthodologie statistique. Par ailleurs, l'établissement d'un appareil statistique fut une nécessité dans l'édification des États modernes et ce n'est sans doute pas un hasard si ces personnalités proviennent des deux pays les plus rapidement touchés par la révolution industrielle. À cette époque, l'objectif du statisticien était surtout de réaliser des énumérations. La préoccupation majeure était d'inventorier les ressources des nations. Dans ce contexte, le recours à l'échantillonnage fut unanimement rejeté comme une procédure inexacte et donc foncièrement anti-scientifique. Tout au long du 19^{e} siècle, les discussions des statisticiens portent essentiellement sur la méthode à appliquer pour obtenir des données fiables et sur la présentation, l'interprétation et éventuellement la modélisation (par un ajustement) de ces données.

1.3 Polémiques sur l'utilisation de données partielles

En 1895, le Norvégien Anders Nicolai Kiær, directeur du Bureau central de la Statistique de Norvège, présente au Congrès de l'Institut international de Statistique

(IIS) à Berne un travail intitulé *Observations et expériences concernant des dénombrements représentatifs* relatif à un sondage réalisé en Norvège. Kiær (1896) sélectionne d'abord un échantillon de villes et de communes. Ensuite, dans chacune de ces communes, il ne sélectionne qu'une partie des individus à partir de la première lettre de leur patronyme. Il applique donc un plan à deux degrés, mais le choix des unités n'est pas aléatoire. Kiær défend l'intérêt de l'utilisation de données partielles pour peu qu'elles soient produites au moyen d'une « méthode représentative ». Selon cette méthode, l'échantillon doit être une représentation à taille réduite de la population. La notion de représentativité de Kiær est donc liée à la méthode des quotas. Son intervention est suivie d'un débat houleux, les actes du congrès de l'IIS rendent compte d'une longue polémique. Examinons de plus près l'argumentation de deux opposants à la méthode de Kiær (voir procès-verbal de l'Assemblée Générale de l'IIS, 1896).

> Georg von Mayr (Prusse) [· · ·] C'est surtout dangereux de se déclarer pour ce système des investigations représentatives au sein d'une assemblée de statisticiens. On comprend que pour des buts législatifs ou administratifs un tel dénombrement restreint peut être utile – mais alors il ne faut pas oublier qu'il ne peut jamais remplacer l'observation statistique complète. Il est d'autant plus nécessaire d'appuyer là-dessus, qu'il y a parmi nous dans ces jours un courant au sein des mathématiciens qui, dans de nombreuses directions, voudraient plutôt calculer qu'observer. Mais il faut rester ferme et dire : pas de calcul là où l'observation peut être faite.

> Guillaume Milliet (Suisse). Je crois qu'il n'est pas juste de donner par un vœu du congrès à la méthode représentative (qui enfin ne peut être qu'un expédient) une importance que la statistique sérieuse ne reconnaîtra jamais. Sans doute, la statistique faite avec cette méthode ou, comme je pourrais l'appeler, la statistique, *pars pro toto*, nous a donné çà et là des renseignements intéressants ; mais son principe est tellement en contradiction avec les exigences que doit avoir la méthode statistique, que, comme statisticiens, nous ne devons pas accorder aux choses imparfaites le même droit de bourgeoisie, pour ainsi dire, que nous accordons à l'idéal que scientifiquement nous nous proposons d'atteindre.

Le contenu de ces réactions peut à nouveau se résumer ainsi : comme la statistique est par définition exhaustive, renoncer au dénombrement complet c'est nier la mission même de la science statistique. La discussion ne porte donc pas sur la méthode proposée par Kiær mais sur la définition de la science statistique. Kiær ne désarme pourtant pas et continue à défendre la méthode représentative en 1897 au congrès de l'IIS à Saint-Petersbourg (voir Kiær, 1899), en 1901 à Budapest et en 1903 à Berlin (voir Kiær, 1903, 1905). Après cette date, la question ne sera plus mentionnée au congrès de l'IIS. Kiær obtient cependant l'appui d'Arthur Bowley (1869-1957) qui jouera ensuite un rôle déterminant dans le développement des sondages. Bowley (1906) présente une vérification empirique de l'application du théorème central limite à l'échantillonnage. Celui-ci fut le véritable promoteur des techniques de sondage aléatoire, il développe les plans stratifiés avec allocations proportionnelles et utilise le théorème de la variance totale. Il faudra attendre la fin de la Première Guerre mondiale et l'émergence d'une nouvelle génération de statisticiens pour que le problème soit rediscuté au sein de l'IIS. À ce sujet, on ne peut s'empêcher de citer la réflexion de Max Plank concernant l'apparition de nouvelles vérités scientifiques : « une nouvelle vérité scientifique ne triomphe pas en convainquant les opposants et en leur faisant entrevoir la lumière, mais plutôt parce que ses opposants mourront un jour et qu'une nouvelle génération, familiarisée avec elle, paraîtra » (cité par Kuhn, 1983, p. 208).

En 1924, une commission (composée de Arthur Bowley, Corrado Gini, Adolphe Jensen, Lucien March, Verrijn Stuart et Frantz Zizek) est créée afin d'évaluer la pertinence de l'utilisation de la méthode représentative. Les résultats de cette commission intitulés *Reports on the representative method in statistics* sont présentés au congrès de l'IIS de 1925 à Rome. La commission accepte le principe du sondage pour autant que la méthodologie soit respectée. Trente ans après la communication de Kiær, l'idée de l'échantillonnage est donc officiellement acceptée. La commission jettera les bases des recherches futures. Deux méthodes sont clairement distinguées : « la sélection aléatoire » et la « sélection raisonnée ». Ces deux méthodes correspondent à deux démarches scientifiques fondamentalement différentes. D'une part, la validation des méthodes aléatoires est basée sur le calcul des probabilités qui permet de construire des intervalles de confiance pour certains paramètres. D'autre part, la validation des méthodes par sélection raisonnée ne peut être obtenue que par l'expérimentation en comparant les estimations obtenues à des résultats de recensement. Les méthodes aléatoires sont donc validées par un argument strictement mathématique tandis que les méthodes par choix raisonné sont validées par une démarche expérimentale.

1.4 Développement d'une théorie des sondages

Le rapport de la commission présenté au congrès de l'IIS en 1925 marque la reconnaissance officielle de l'utilisation des sondages. La plupart des problèmes de base sont déjà posés comme l'utilisation d'échantillons aléatoires et le calcul de la variance des estimateurs pour les plans simples et stratifiés. L'acceptation de l'utilisation de données partielles et surtout la recommandation de recourir aux plans aléatoires vont aboutir à une mathématisation rapide de cette théorie. À cette époque, le calcul des probabilités était déjà connu. De plus, les statisticiens avaient déjà développé une théorie pour la statistique expérimentale. Tout était en place pour que progresse rapidement un champ de recherche fécond : la construction d'une théorie statistique des sondages.

Jerzy Neyman (1894-1981) développe une grande partie des fondements de la théorie probabiliste des sondages, plans simples, stratifiés et par grappes. Il détermine également l'allocation optimale d'un plan stratifié. La méthode de l'allocation optimale remet en cause l'idée de base de la méthode des quotas qui est la « représentativité » . En effet, selon la stratification optimale, l'échantillon ne doit pas être une miniature de la population puisque certaines strates doivent être surreprésentées. L'article publié par Neyman (1934) dans le *Journal of the Royal Statistical Society* est actuellement considéré comme un des textes fondateurs de la théorie des sondages. Neyman a balisé les principaux champs de recherche. Son travail aura un impact très important dans les années ultérieures. On sait maintenant que Tschuprow (1923) avait déjà obtenu certains des résultats qui furent attribués à Neyman. Ce dernier semble les avoir trouvés indépendamment de Tschuprow. Il n'est pas étonnant qu'une telle découverte ait été faite simultanément à plusieurs endroits. Dès le moment où l'utilisation de sondages probabilistes sera considérée comme une méthode valable, la théorie surgira directement de l'application du calcul des probabilités.

1.5 Les élections américaines de 1936

Au cours de cette même période, la mise en œuvre de la méthode des quotas a bien plus contribué au développement de l'utilisation des méthodes de sondage que les études théoriques. L'épisode des élections américaines de 1936 marque un tournant important dans la réalisation d'enquêtes par questionnaires. Les faits peuvent être résumés ainsi : les grands journaux américains avaient coutume de publier avant les élections des résultats d'enquêtes empiriques produites à partir de grands échantillons (deux millions de personnes sondées pour le *Literary Digest*) mais sans méthode pour sélectionner les individus. Alors que la plupart des sondages prédisaient la victoire de Landon, Roosevelt fut élu. Des sondages réalisés par Crossley, Roper et Gallup sur des échantillons plus réduits mais au moyen de la méthode des quotas donnèrent une prévision correcte. Cet événement a contribué à faire admettre la validité des données fournies par les sondages d'opinion.

Cet épisode, qui favorisa l'extension de la pratique des enquêtes par sondage, s'est fait sans référence à la théorie probabiliste qui était déjà développée. La méthode de Crossley, Roper et Gallup n'est en effet pas probabiliste mais empirique. La validation de l'adéquation de la méthode est donc donnée de manière expérimentale et absolument pas mathématique.

1.6 La théorie statistique des sondages

La mise en place d'un nouveau consensus scientifique en 1925 et la détermination des grandes voies de recherche dans les années suivantes va déboucher sur un développement très rapide de la théorie des sondages. Durant la Seconde Guerre mondiale, les recherches se poursuivent aux États-Unis. Une importante contribution est due à Deming & Stephan (1940), Stephan (1942, 1945, 1948) et Deming (1948, 1950, 1960) notamment au sujet de la question de l'ajustement de tableaux statistiques sur des données de recensement. Cornfield (1944) propose d'utiliser des variables indicatrices de la présence des unités dans l'échantillon. Cochran (1939, 1942, 1946, 1961) et Hansen & Hurwitz (1943, 1949) montrent l'intérêt de l'échantillonnage à probabilités inégales avec remise. Madow (1949) préconise le tirage systématique à probabilités inégales (voir également Hansen *et al.*, 1953a,b). Il est rapidement établi que réaliser un sondage à probabilités inégales de taille fixe sans remise est un problème complexe. Narain (1951), Horvitz & Thompson (1952), Sen (1953) et Yates & Grundy (1953) présentent plusieurs méthodes à probabilités inégales dans deux articles qui sont certainement parmi les plus cités dans ce domaine. Consacrés à l'examen de plusieurs plans à probabilités inégales, ces textes sont mentionnés pour l'estimateur général (estimateur par expansion) du total qui y est également proposé et discuté. L'estimateur par expansion est, en effet, un estimateur général sans biais applicable à tout plan de sondage sans remise. L'estimateur de variance proposé possède cependant un défaut. Yates & Grundy (1953) montrent que l'estimateur de la variance proposé par Horvitz et Thompson peut être négatif. Ils proposent une variante valide quand l'échantillon est de taille fixe et donnent des conditions suffisantes pour qu'il soit positif. Dès les

années cinquante, le problème de l'échantillonnage à probabilités inégales va susciter un engouement important qui se concrétisera par la publication de plus de deux cents articles. Avant de se tourner vers la statistique de rangs, Hájek (1981) traite du problème en détail. Un livre de synthèse de Brewer & Hanif (1983) a été consacré entièrement à ce sujet qui semble d'ailleurs loin d'être épuisé, comme en témoignent les publications dont il fait l'objet régulièrement.

La théorie des sondages, qui fait dès lors abondamment usage du calcul des probabilités, va retenir l'attention de statisticiens universitaires et, très rapidement, ceux-ci vont passer en revue tous les aspects de cette théorie qui possèdent un intérêt mathématique. Une théorie mathématique cohérente des sondages va être construite. Les statisticiens vont très vite se heurter à une difficulté de taille : dans les sondages en population finie, le modèle proposé postule l'identifiabilité des unités. Cette composante du modèle rend non pertinente l'application de la technique de réduction par exhaustivité et de la méthode de maximum de vraisemblance. Godambe (1955) établit qu'il n'existe pas d'estimateur linéaire optimal. Ce résultat est l'une des nombreuses preuves montrant l'impossibilité de définir des procédures optimales d'estimation pour des plans de sondage généraux dans les populations finies. Basu & Ghosh (1967) puis Basu (1969) démontrent ensuite que la réduction par exhaustivité se limite à la suppression des informations concernant la multiplicité des unités et donc de la non-opérationnalité de celle-ci. Plusieurs approches sont examinées dont celle issue de la théorie de la décision. Des propriétés nouvelles, comme l'hyperadmissibilité (voir Hanurav, 1968), sont définies pour des estimateurs applicables en populations finies.

Une école purement théorique de sondage se développe donc rapidement. Cette théorie va éveiller l'attention de chercheurs spécialisés en statistique mathématique, comme Debabrata Basu, qui vont s'intéresser aux spécificités de la théorie des sondages. Beaucoup de résultats proposés sont cependant des théorèmes de non-existence de solutions optimales. Les recherches sur la question des fondements de l'inférence en théorie des sondages deviennent à ce point importantes qu'elles font l'objet d'un symposium à Waterloo (Canada) en 1971. Lors de ce symposium, l'intervention de Calyampudi Radhakrishna Rao (1971, p. 178) débute par un constat bien pessimiste :

> Je peux mentionner que, dans la méthodologie statistique, l'existence de procédures uniformément optimales (comme l'estimateur sans biais uniformément à variance minimale, la région critique uniformément la plus puissante pour tester une hypothèse) est une exception rare plutôt qu'une règle. C'est la raison pour laquelle les critères ad hoc sont introduits pour restreindre la classe de procédures dans laquelle un optimum peut être recherché. Il n'est pas surprenant que la même situation soit obtenue dans l'échantillonnage pour une situation finie. Cependant, il présente d'autres difficultés qui ne semblent pas exister pour l'échantillonnage dans des populations infinies. [1]

Cette introduction annonçait les orientations des recherches actuelles.

1. Traduit de l'anglais « I may mention that in statistical methodology, the existence of uniformly optimum procedures (such as UMV unbiased estimator, uniformly most powerful critical region for testing a hypothesis) is rare exception rather than a rule. That is the reason why adhoc criteria are introduced to restrict the class of procedures in which an optimum may be sought. It is not surprising that the same situation obtains in sampling for finite situation. However, it presents some further complications which do not seem to exist for sampling from infinite populations. »

En théorie des sondages, il n'existe aucun théorème montrant l'optimalité d'une procédure d'estimation pour des plans de sondage généraux. On ne peut trouver des méthodes d'estimation optimales qu'en se restreignant à des classes particulières de procédures. Même en se limitant à une classe d'estimateurs particuliers (comme la classe des estimateurs linéaires ou sans biais), il n'est pas possible d'obtenir des résultats intéressants. Une option possible pour sortir de cette impasse consiste à modifier la formalisation du problème, par exemple en supposant que la population elle-même est aléatoire.

1.7 Modélisation de la population

C'est certainement l'absence de résultats généraux tangibles concernant certaines classes d'estimateurs qui a conduit à l'essor d'une modélisation de la population au moyen d'un modèle dit de « superpopulation ». Dans l'approche modèle, on suppose que les valeurs prises par la variable d'intérêt sur les unités d'observation de la population sont les réalisations de variables aléatoires. Le modèle de superpopulation définit une classe de distributions à laquelle ces variables aléatoires sont supposées appartenir. L'échantillon est alors issu d'une double expérience aléatoire : une réalisation du modèle qui engendre la population et ensuite le choix de l'échantillon. L'idée de modéliser la population était déjà présente dans Brewer (1963a), mais elle fut surtout développée par Royall (1970b, 1971, 1976b) (voir aussi Valliant *et al*., 2000, Chambers & Clark, 2012).

Tirant argument du fait que l'échantillon aléatoire est une statistique « ancillaire », Royall propose de travailler conditionnellement à celui-ci. Autrement dit, il considère qu'une fois l'échantillon sélectionné le choix des unités n'est plus aléatoire. Cette nouvelle modélisation a permis le développement d'une école particulière de recherche. Le modèle doit exprimer une relation connue et admise préalablement. Selon Royall, si le modèle de superpopulation décrit de manière « adéquate » la population, l'inférence peut être menée uniquement en fonction du modèle, conditionnellement au tirage de l'échantillon. L'utilisation du modèle permet alors de déterminer un estimateur optimal.

On peut objecter qu'un modèle est toujours une représentation approximative de la population. Toutefois, le modèle n'est pas construit pour être mis à l'épreuve des données mais pour « assister » l'estimation. Si le modèle est correct, alors la méthode de Royall fournira un estimateur performant. Si le modèle est faux, le biais peut être à ce point important que les intervalles de confiance construits pour le paramètre ne sont pas valides. C'est essentiellement la critique énoncée par Hansen *et al*. (1983).

Le débat est intéressant, car les arguments en présence ne sont pas du ressort de la statistique mathématique. Mathématiquement, ces deux théories sont évidemment correctes. L'argumentation porte sur l'adéquation de la formalisation à la réalité et est donc forcément étrangère à l'aspect mathématique du développement statistique. De plus, la modélisation proposée par Royall est particulière. Elle permet surtout de sortir d'une impasse théorique et fournit dès lors des estimateurs optimaux. Cependant, la pertinence d'une modélisation est discutable et sera envisagée de manière

complètement différente selon que l'on prend les arguments de la sociologie, de la démographie ou de l'économétrie, trois disciplines qui sont intimement liées à la méthodologie de la statistique officielle. Un commentaire de Dalenius (voir Hansen *et al.*, 1983, p. 800) met ce problème en évidence :

> Cela ne veut pas dire que les arguments pour ou contre l'inférence paramétrique dans la théorie statistique habituelle ne sont pas intéressants dans le contexte de la théorie des sondages par échantillonnage. Dans notre évaluation de ces arguments, cependant, nous devons prêter attention aux spécificités pertinentes des applications. [2]

Selon Dalenius, c'est donc bien de la discipline où est appliquée la théorie des sondages qu'il faudrait tirer les conclusions utiles concernant l'adéquation d'un modèle de superpopulation.

La théorie statistique des sondages s'applique essentiellement dans les instituts officiels de statistique. Ces instituts ne développent pas une science mais assurent une mission auprès de leurs États. À juste titre, on peut comprendre un argument assez classique des responsables des instituts nationaux de statistique : l'utilisation d'un modèle de superpopulation dans une procédure d'estimation est un manquement à un principe d'impartialité qui fait partie de la déontologie des statisticiens. Cet argument est issu directement de la définition actuelle de la statistique officielle. Le principe d'impartialité fait partie de cette définition comme le principe d'exactitude en faisait partie au 19^{e} siècle. Si la modélisation d'une population se conçoit aisément comme outil de recherche ou comme outil prédictif, il reste fondamentalement questionnable dans le domaine de la statistique officielle.

1.8 Tentative de synthèse

L'approche « superpopulation » a permis des recherches extrêmement fécondes. Le développement d'une approche hybride dite « approche assistée par un modèle » permet de fournir des inférences valides sous le modèle mais se veut aussi robuste dans le cas où le modèle est faux. Cette optique a été principalement développée par une école suédoise (voir Särndal *et al.*, 1992). Le modèle permet de prendre en compte des informations auxiliaires au moment de l'estimation tout en conservant des propriétés de robustesse pour les estimateurs en cas de non-adéquation du modèle. Il est en fait très difficile de construire un estimateur permettant de prendre en compte un ensemble d'informations auxiliaires *a posteriori* sans émettre une hypothèse, même simple, sur la relation existant entre les informations auxiliaires et la variable d'intérêt. La modélisation permet une conceptualisation de ce type de présomption. L'approche « assistée par un modèle » permet de construire des estimateurs intéressants et pratiques. Il est maintenant clairement acquis que l'introduction d'un modèle est une nécessité pour le traitement de certains problèmes de non-réponses et d'estimation dans des petits domaines. Dans ce type de problème, quelle que soit la technique utilisée, on postule toujours l'existence d'un modèle même si celui-ci est parfois implicite.

2. Traduit de l'anglais « That is not to say that the arguments for or against parametric inference in the usual statistical theory are not of interest in the context of theory of survey sampling. In our assessment of these arguments, however, we must pay attention to the relevant specifics of the applications. »

Le modèle mérite d'ailleurs toujours d'être clairement déterminé afin d'expliciter les idées sous-jacentes qui justifient l'application de la méthode.

1.9 Information auxiliaire

Les années 1990 ont été marquées par l'émergence du concept d'information auxiliaire. Cette notion relativement générale regroupe toute information extérieure à l'enquête proprement dite permettant d'augmenter la précision des résultats d'un sondage. Cette information peut être la connaissance des valeurs d'une ou plusieurs variables sur toutes les unités de la population ou simplement d'une fonction de ces valeurs. Pour la plupart des enquêtes, une information auxiliaire est disponible. Elle peut être donnée par un recensement ou tout simplement par la base de sondage. On peut citer comme exemple d'informations auxiliaires : le total d'une variable sur la population, des sous-totaux selon des sous-populations, des moyennes, des proportions, des variances, les valeurs d'une variable sur toutes les unités de la base de sondage. La notion d'information auxiliaire englobe donc toutes données issues de recensements ou de sources administratives.

L'objectif principal consiste donc à mettre à profit toutes ces informations pour obtenir des résultats plus précis. Comme le montre la Figure 1.1, l'information auxiliaire peut être utilisée à deux occasions : lors de la conception du plan de sondage et au moment de l'estimation des paramètres. Quand l'information auxiliaire est mise à profit pour concevoir le plan de sondage, on cherche un plan qui fournit des estimateurs précis pour un prix donné ou qui est peu coûteux pour des critères de précision donnés. Pour ces raisons, on utilisera des plans à probabilités inégales, stratifiés, équilibrés, par grappes ou à plusieurs degrés. Quand l'information est utilisée à l'étape de l'estimation, elle sert à « caler » les résultats du sondage sur l'information auxiliaire du recensement. La méthode générale de calage (en anglais : *calibration*) de Deville & Särndal (1992) permet d'utiliser des informations auxiliaires sans référence explicite à un modèle.

Tout problème de sondage traite de la manière d'utiliser l'information disponible. Avec l'idée d'information auxiliaire, on s'affranchit de toute modélisation de la population. Ce concept, qui sera le fil conducteur de cet ouvrage, permet de concevoir de manière intégrée le problème de la planification et de l'estimation.

1.10 Références et développement récents

Parmi les ouvrages importants, les livres des précurseurs sont Yates (1946, 1949, 1960, 1979), Deming (1948, 1950, 1960), Thionet (1953), Sukhatme (1954), Hansen *et al*. (1953a,b), Cochran (1953, 1963, 1977), Dalenius (1957), Kish (1965, 1989, 1995), Murthy (1967), Raj (1968), Johnson & Smith (1969), Sukhatme & Sukhatme (1970), Konijn (1973), Lanke (1975), Cassel *et al*. (1977, 1993), Jessen (1978), Hájek (1981) et Kalton (1983). Ces livres valent la peine d'être consultés, car une grande partie des idées modernes, notamment sur le calage et l'équilibrage, y sont déjà abordées.

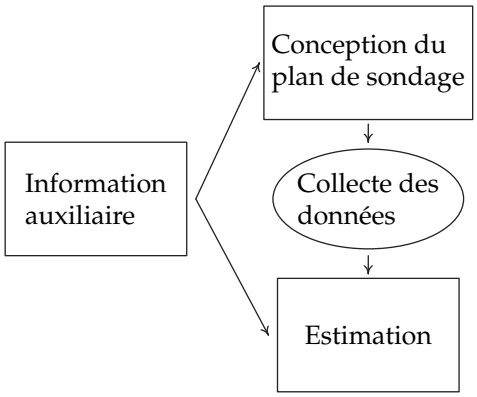

FIGURE 1.1 – L'information auxiliaire peut être mise à profit avant ou après la collecte des données pour améliorer l'estimation.

Les ouvrages de référence sont Skinner *et al*. (1989), Särndal *et al*. (1992), Lohr (1999, 2009b), Thompson (1997), Brewer (2002) et Fuller (2011). À quinze ans d'intervalle, la série *Handbook of Statistics* a d'abord consacré à l'échantillonnage un volume dirigé par Krishnaiah & Rao (1994), puis deux volumes dirigés par Pfeffermann & Rao (2009a,b). Il existe également un ouvrage collectif récent dirigé par Wolf *et al*. (2016).

Les livres de Thompson (1992, 1997, 2012) et Thompson & Seber (1996) sont consacrés à l'échantillonnage dans l'espace. Les méthodes destinées à l'échantillonnage environnemental sont développées dans Gregoire & Valentine (2007) et à la foresterie dans Mandallaz (2008). Plusieurs ouvrages sont dédiés aux tirages à probabilités inégales et aux algorithmes d'échantillonnage. On peut citer Brewer & Hanif (1983), Gabler (1990) et Tillé (2006). L'approche basée sur le modèle est clairement décrite dans Valliant *et al*. (2000), Chambers & Clark (2012) et Valliant *et al*. (2013).

De nombreux ouvrages ont été publiés et sont encore disponibles en français. On peut citer Thionet (1953), Desabie (1966), Deroo & Dussaix (1980), Gouriéroux (1981), Grosbras (1987), Dussaix & Grosbras (1992), Dussaix & Grosbras (1996), Ardilly (1994, 2006), Ardilly & Tillé (2003) et Ardilly & Lavallée (2017). En italien, on peut consulter les ouvrages de Cicchitelli *et al*. (1992, 1997), Frosini *et al*. (2011) et Conti & Marella (2012). En espagnol, il existe également les livres de Pérez López (2000), Tillé (2010) et Gutiérrez (2009) ainsi qu'une traduction du livre de Sharon Lohr (2000). En allemand, on trouve les livres de Stenger (1985) et de Kauermann & Küchenhoff (2010). Enfin, en chinois, il existe un ouvrage de Ren & Ma (1996), et en coréen de Kim (2017).

Récemment, de nouveaux champs de recherche ont été ouverts. Pour ne citer que des livres, l'estimation pour des petits domaines ou des petites régions à partir de données d'enquêtes a été un sujet de recherche majeur (Rao, 2003, Rao & Molina, 2015). Les développements récents de la méthodologie d'enquête sont décrits dans Groves (2004b) et Groves *et al*. (2009). L'échantillonnage indirect consiste à sélectionner un échantillon dans une population qui n'est pas la population d'intérêt mais qui a des

liens avec celle-ci (Lavallée, 2002, 2007). De nouveaux algorithmes d'échantillonnage ont été développés par exemple afin de sélectionner des échantillons équilibrés (Tillé, 2006). L'échantillonnage adaptatif consiste à compléter l'échantillon initial en fonction de résultats préliminaires (Thompson, 1992, Thompson & Seber, 1996). Les méthodes de capture-recapture permettent d'estimer la taille de populations animales. Des variantes de ces méthodes permettent parfois d'estimer les tailles de populations rares ou d'effectuer des enquêtes de couverture (Pollock, 2000, Seber, 2002).

Des méthodes de ré-échantillonnage ont été développées pour les populations finies (Shao & Tu, 1995, Groves, 2004b). Évidemment, les erreurs de mesure resteront toujours un sujet de recherche capital (Fuller, 1987, Groves, 2004a). Enfin, des progrès substantiels ont été réalisés pour les méthodes de traitement de la non-réponse qu'il s'agisse de méthodes de repondération ou de techniques d'imputation (Särndal & Lundström, 2005, Bethlehem *et al*., 2011, De Waal *et al*., 2011, Kim & Shao, 2013).

Un des défis qui se profile actuellement est l'intégration de données provenant de sources multiples : fichiers administratifs, registres, échantillons. Dans un article judicieux intitulé *Big data : are we making a big mistake ?*, Tim Harford (2014) rappelle que l'abondance des données n'est jamais un gage de qualité. L'accès à de nouvelles sources de données ne doit pas nous faire retomber dans les erreurs du passé, comme ce fut le cas durant l'élection présidentielle de 1936 (voir Section 1.5, page 6).

Il existe depuis des décennies des méthodes permettant d'intégrer des données issues de sources différentes. Cependant, la multiplication des sources disponibles rend ces questions d'intégration de plus en plus complexes. Il reste donc un important travail de recherche et de développement pour définir les méthodes qui permettent d'intégrer des données de multiples sources en traitant de manière appropriée les différentes erreurs de mesure.