

STATISTIQUE ET ANALYSE DES DONNÉES

PHILIPPE TASSI

De l'exhaustif au partiel : un peu d'histoire sur le développement des sondages

Statistique et analyse des données, tome 13, n° 2 (1988), p. 113-129

http://www.numdam.org/item?id=SAD_1988__13_2_113_0

© Association pour la statistique et ses utilisations, 1988, tous droits réservés.

L'accès aux archives de la revue « Statistique et analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

DE L'EXHAUSTIF AU PARTIEL : UN PEU D'HISTOIRE SUR LE DÉVELOPPEMENT DES SONDAGES

Philippe TASSI

Directeur de la division CGSA à l'ENSAE

I. Les origines

Il est généralement admis que la statistique recouvre un ensemble de méthodes contribuant à la connaissance d'une population, ce dernier vocable étant à prendre au sens large du terme, même si son usage vient de la démographie et du dénombrement d'individus.

L'apparition du mot « statistique » lui-même est relativement récent, surtout en comparaison de l'activité correspondante de recueil des données, qui remonte à la plus haute antiquité. On attribue souvent la création du terme « statistique » à l'école descriptive allemande de Gottingue, et plus particulièrement à Gottfried Achenwall (1746), comme issu du latin « statisticum » : qui a trait à l'État. Il est plus probable cependant que ce dernier a contribué seulement à imposer l'emploi du mot, puisque l'on possède une *Biblioteca Statistica* datant de 1701 et un *Microscopium Statisticum* de 1672. En remontant plus encore dans le temps, le mot « statistique » semble appartenir au langage administratif français colbertien. Jean Meyer, dans son livre *Colbert*, mentionne l'existence d'une *Déclaration des biens, charges, dettes et statistiques des communautés de la généralité de Bourgogne* établie par Claude Bouchu, intendant de Bourgogne, de 1666 à 1669.

Par contre, l'apparition du besoin « statistique » de posséder des données chiffrées précède sa dénomination de plusieurs millénaires. A son origine, il est le fait de chefs d'État — ou de ce qui en tient lieu à l'époque — désireux de connaître des éléments de leur puissance : population, potentiel militaire, richesses... Quelle meilleure connaissance que celle issue de l'observation systématique, exhaustive, de tous les éléments de la société ? L'idée de recensement, ou de liste d'inventaire, apparaît donc de façon tout à fait naturelle dans l'histoire, impliquant en outre une impression de précision de la plus haute qualité.

Les premiers recensements semblent remonter à la civilisation sumérienne, de 5 000 à 2 000 ans avant notre ère, pour laquelle on possède des listes d'hommes et de biens inscrits sur des tablettes d'argile. Le relevé des personnes et des biens a lieu régulièrement en Mésopotamie 3 000 ans avant Jésus-Christ. L'Égypte paraît avoir été la première nation à organiser des recensements systématiques de population, au moins depuis l'an 2900 avant J.-C., mais aussi à institutionnaliser des recensements à finalité fiscale (2 700 à 2 500 ans avant J.-C.). Elle semble même avoir la primeur du principe de la déclaration obligatoire et du contentieux : en effet, sous le pharaon Amasis II, au VI^e siècle avant notre ère, tout individu était tenu de déclarer ses sources de revenu et son activité. Tout manquement à cette règle était puni de mort, ce qui est une forme extrême du règlement contentieux si répandu actuellement pour les enquêtes obligatoires des services officiels de statistique.

La Chine joue également un grand rôle dans l'histoire des recensements. Par exemple, l'empereur Yao édite le relevé total de ses sujets et des productions agricoles dès l'an 2238 avant Jésus-Christ.

Il serait vain et fort long de dresser la liste des premières tentatives de dénombrement exhaustif. L'enseignement essentiel en est que les civilisations qui les engendrèrent — Japon, Rome, Inca, Inde — possédaient un système administratif fort, auto-entretenu par l'amélioration de la connaissance quantitative de l'État.

II. Les premières tentatives d'extrapolation

A partir du XIII^e siècle, les données deviennent plus nombreuses grâce à la prolifération des rôles fiscaux. Le plus célèbre est, en France, « l'état des paroisses et des feux des baillages et sénéchaussées de France » dressé en 1328. Le « feu », référence à un foyer, une famille ou une habitation, sera un élément essentiel pour estimer une population bien que devenant vite une donnée sociale plutôt que démographique (pour une référence globale, le lecteur pourra consulter l'ouvrage de J. et M. Dupâquier).

Le XIV^e siècle voit le début des enregistrements des actes d'état civil (Commune de Givry, 1334). En août 1539, François I^{er} rend obligatoire la tenue de registres de naissances, par l'article 51 de l'édit de Villers-Cotterêts; sous Henri III, l'édit de Blois (mai 1579) étend cette obligation aux mariages et aux décès.

Cependant, durant toute cette période où s'affirme la statistique administrative, les recensements sont très rares (recensement de Paris en 1590 — 200 000 habitants —, du Berry en 1565, du Bourbonnais en 1569).

Les progrès fondamentaux de la statistique vont reprendre lors de la seconde moitié du XVII^e siècle, avec le besoin que ressentent les monarques et leurs conseillers de connaître et expliquer les phénomènes économiques et sociaux. Cette époque coïncide également avec le développement de l'école d'arithmétique politique anglaise, précurseur de la statistique inférentielle. Cette école, dont les fondateurs sont John Graunt (1620-1674) et William Petty (1623-1687), est guidée par le souci de la quantification et la recherche de constantes de comportement permettant des estimations et des prévisions : nombre d'enfants par femme, temps entre deux naissances pour une même mère, nombre d'habitants par maison, nombre d'individus par feu, proportion de décès, etc. Les techniques de *multiplicateur* de l'arithmétique politique vont être utilisées au détriment des recensements, et favoriser l'apparition d'enquêtes partielles. Parallèlement, le remplacement d'une connaissance exhaustive par une extrapolation fondée sur l'examen d'une partie de la population est une attitude qui commence à trouver des éléments de justification par l'apparition des premiers résultats de probabilités (Pascal, Fermat, puis Bernoulli).

En France, deux noms dominent cette époque : Colbert et Vauban. En 1664, Jean-Baptiste Colbert ordonne aux intendants de procéder à une vaste enquête sur l'état des provinces; celle-ci est précédée d'une enquête-pilote commandée à son frère dans les provinces de l'Est. Le marquis de Vauban s'intéresse encore de plus près à la connaissance du chiffre, que ce soit par des recensements ou des enquêtes. Auteur de *Méthode générale et facile pour faire le dénombrement des peuples*, publiée en 1686, il préconise l'utilisation d'échantillons de terres arables dans chaque province pour estimer au mieux les capacités agricoles; ces procédures sont reprises dans les célèbres *mémoires dressées par les Intendants du Royaume*. Certains esprits commencent même à demander un organisme centralisateur de la collecte chiffrée. L'un des premiers, sinon le premier, est Charles Castel, dit l'Abbé de Saint-Pierre (1658-1743); loin d'être l'utopiste trop souvent dépeint, il a au contraire une vision pertinente de l'administration future. Dès le début du XVIII^e siècle, il réclame la création d'un *Bureau pour recueillir les divers dénombrements*.

En Angleterre, les méthodes de l'*arithmétique politique* poursuivent leurs progrès. A titre d'exemple, détaillons l'estimation de population de W. Petty, extrait de l'article de Jacqueline Hecht dans *Pour une histoire de la Statistique* (INSEE) :

« Il évalue tout d'abord le nombre des maisons à Londres : après plusieurs calculs, il parvient au chiffre de 88 000 pour 1686. Établissant une moyenne entre la proportion des feux par maison à Dublin et à Bristol, il évalue le nombre de feux londoniens à 105 000, à peu près équivalent au chiffre officiel du Bureau des Feux (105 315). En supposant que chaque famille compte en moyenne six membres, et que 10 % des maisons abritent deux familles chacune, les autres n'en ayant qu'une, il conclut que les 105 000 maisons londoniennes abritent 695 000 habitants. Avec d'autres procédés (par la proportion des décès qu'il estime à 1 sur 30, et la moyenne des décès de deux années), il parvient à peu près au même chiffre (696 300)... »

Ces techniques d'extrapolation où du *multiplicateur* connaissent vite un énorme succès, contribuant à mettre un frein aux recensements jugés par certains onéreux, peu précis et « monstrueux » (Saint-Simon). Elles servent de base aux calculs faits par d'Expilly, Messance (*Recherches sur la population des généralités d'Auvergne, de Lyon et de Rouen*, publié en 1766) et le Baron Moheau (*Recherches et considérations sur la population de la France*, 1778). L'abbé Jean-Joseph d'Expilly (1719-1793) utilise une extrapolation fondée sur les feux. Possédant leur nombre total, il le multiplie par une autre estimation du nombre d'habitants par feu. Il introduit en outre une « stratification campagnes-villes », et adopte un multiplicateur 4,5 pour les premières et 5 pour les secondes. Messance et Moheau préconisent des enquêtes partielles pour évaluer la population : sélectionnant un grand nombre de paroisses ou de petites villes, ils dénombrent les habitants, établissent sur la base des cinq ou dix dernières années de l'État Civil une année moyenne des naissances, et calculent le nombre de naissances par habitant, taux qui sera appelé multiplicateur des naissances. Sous l'hypothèse que celui-ci est proche de celui qui serait observé sur la France entière, il suffit de le multiplier par le total des naissances d'une année pour avoir une idée de l'ensemble de la population.

Un débat, dans lequel intervient Jacques Necker, naît alors sur la valeur de ce multiplicateur : 23, 25, 25.5, 25.75, 26, 28, 30. Il vient de ce que l'on appellerait maintenant les aléas d'échantillonnage, les valeurs dépendant des lieux utilisés pour faire le calcul.

Prenons quatre exemples de ces techniques d'extrapolation. En 1783, Pierre-Simon Laplace expose à l'Académie des Sciences une procédure de détermination du multiplicateur consistant à réaliser une enquête portant sur un million d'habitants répartis sur plusieurs régions. Il propose de calculer l'incertitude de l'estimateur — *l'erreur à craindre* — en supposant que les régions sont tirées au hasard, comme suit (*Théorie analytique des Probabilités*, 1812) : faisant l'hypothèse d'une loi a priori uniforme, il détermine la loi a posteriori de la différence entre population « vraie » et population estimée. Il fait enfin référence à son mémoire de 1783 sur ce que l'on appellera, plus tard et de façon erronée, le théorème central limite pour approximer cette loi a posteriori par une distribution normale. Par ailleurs, en l'an XI (1802), Jean-Antoine Chaptal, alors ministre de l'Intérieur, fit « sonder » 50 000 habitants dans 30 départements et obtient une valeur de 28,35. Antoine Lavoisier estime par une méthode d'échantillonnage la superficie cultivée et le cheptel en chevaux, bœufs, moutons et porcs, en valeur moyenne de 1784 à 1791. En Angleterre, Sir Frederick Morton Eden estime, en 1800, la population sur la base du nombre de naissances et de celui des habitants par maison. Ses résultats (9 millions) seront confirmés en tous points par le recensement de 1801.

Au début du XIX^e siècle, tout semble bien en place pour permettre un développement de la méthodologie des relevés partiels : la pratique, essentiellement démographique, apparaît ; le calcul des probabilités poursuit sa progression ; la volonté des États de posséder des éléments chiffrés se confirme (en mars 1800, sous le Directoire, le ministre de l'Intérieur Lucien Bonaparte crée un bureau de

Statistique, dont le premier directeur fut Adrien-Cyprien Duquesnoy, 1759-1808). Cependant, il faudra plus de 100 ans pour que l'intérêt des sondages soit reconnu, et bien plus encore pour en établir une théorie mathématique fondée sur l'aléatoire.

III. *Un temps d'arrêt : le XIX^e siècle*

Contrairement à toute attente, le XIX^e siècle est marqué par le retour en force des recensements, peu nombreux au XVIII^e siècle — sauf en Suède, très en avance pour l'époque —. L'année 1801 voit la réalisation simultanée de recensements de la population en Angleterre, au Danemark, en France et en Norvège. Dans beaucoup de pays, ils deviennent réguliers et périodiques.

Alors que la théorie statistique connaît un développement sans précédent, l'échantillonnage recule au profit de l'exhaustivité. L'un des fondateurs de la science statistique telle que nous la connaissons maintenant est le Belge Adolphe Quetelet (1796-1874). Il est certainement le premier à concevoir avec une si grande clarté que la statistique peut être fondée sur le calcul des Probabilités; au début du XX^e siècle, Lucien March, directeur de la Statistique Générale de la France, professeur à l'Institut de Statistique de l'Université de Paris, sera cependant encore d'une extrême réserve quant à l'efficacité du rôle des probabilités en statistique. Mais Quetelet élargit le champ des domaines d'application de cette dernière, passant de l'économie et des sciences sociales à l'anthropométrie; il organise à Bruxelles, en 1853, le premier congrès international de statistique.

Et pourtant, Quetelet, loin de préconiser les sondages, est mû par l'obsession du relevé total : son souci constant est d'assurer l'exhaustivité des données recueillies.

Fanatique de la rigueur, il arrive cependant à Quetelet d'accepter la généralisation de résultats partiels sans que l'on en sache trop les raisons : il accepte d'étendre à l'ensemble de la population le taux d'analphabètes calculé chez les délinquants, ou le taux de mortalité observé chez les personnes de haut revenu assurées sur la vie, mais refuse la généralisation du taux de garçons dans les naissances déclarées par les sages-femmes.

Malgré la prépondérance de l'énumération totale, les visions partielles ne disparaissent pas complètement du bagage du statisticien. Ainsi, Jean-Baptiste Fourier prend une part active au développement de la statistique. Il effectue de nombreux travaux de recherche statistique sur la ville de Paris; en 1817, il fait exécuter un contrôle par sondage dans la tentative de dénombrement individuel à Paris par liste nominative. En fin de siècle, on peut citer l'étude sur les salaires des travailleurs français de 1891-1893 réalisée sur la base d'un échantillon au 1/5^e.

Cependant, le pôle principal des utilisations du sondage se situe aux États-Unis, par l'émergence des enquêtes d'opinion publique, dont nous parlerons plus loin.

IV. *1895-1925 : le débat sur la représentativité*

Le concept de représentativité d'un résultat, ou d'un échantillon, est très récent. Bien que Condorcet y fasse parfois référence de façon allusive, il semble avoir été explicitement utilisé pour la première fois par Jean-Baptiste Say qui affirme la non-représentativité pour l'ensemble d'une population du taux de mortalité observé chez les riches assurés, extrapolation que Quetelet défend, nous l'avons vu.

Le débat sur la représentativité va avoir pour cadre l'Institut International de Statistique (IIS), qui a été créé à Londres le 24 juin 1885.

Son point de départ est la communication historique de Anders Nicolai Kiaer, directeur du Bureau Central de Statistique du Royaume de Norvège, au cours du Congrès de l'IIS de 1895, tenu

à Berne. Il présente un mémoire écrit intitulé *Observations et expériences concernant des dénombrements représentatifs* dans lequel il dresse un bilan satisfaisant de deux expériences norvégiennes d'enquêtes partielles, dont l'une sur la distribution des revenus :

« ... Après avoir fait un choix de 127 communes rurales et de 23 villes représentatives, nous avons d'abord pris les bulletins individuels recueillis par le recensement général des hommes âgés de 17, 22, 27, 32, 37, etc., jusqu'à 97 ans, et demeurant dans les communes en question. De cette façon nous aurions eu 1/5^e de la population mâle adulte des communes représentatives; mais comme nous désirions réduire encore davantage les chiffres, nous avons employé seulement les bulletins des individus commençant par les lettres A, B, C, L, M, ou N. Aux principaux renseignements donnés par le recensement (âge, état civil, profession, etc.) nous avons ajouté à l'aide d'une investigation particulière l'indication des fortunes et des revenus des individus respectifs. De cette manière nous avons obtenu un dénombrement représentatif spécial comprenant en moyenne 3,3 pour cent de la population mâle des villes et 1,6 pour cent de celle des campagnes... ».

A. Kiaer ne définit pas a priori la notion de représentativité, mais le contrôle a posteriori, par comparaison de la structure par activité de l'échantillon obtenu avec celle issue du recensement. Il pose la question :

« ... De quelle manière un dénombrement représentatif doit-il être exécuté pour qu'il puisse présenter une miniature aussi correcte que possible de la société entière?... »

Bien que Kiaer ait pris toutes les précautions oratoires nécessaires (« ... il va sans dire que les dénombrements partiels représentatifs n'ont pas pour but de remplacer les recensements généraux... »), les réactions en Assemblée Générale de l'IIS du 30 août sont extrêmement violentes. En voici quelques extraits :

Georg van Mayr (Prusse) : « ... je regarde comme très dangereux le point de vue [exposé]... On ne peut pas remplacer par le calcul l'observation réelle des faits... On n'a pas de vraie statistique... Pas de calcul là où l'observation peut être faite... ».

Luigi Bodio (Italie) : « ... il serait dangereux de recommander une méthode indirecte... On doit procéder par sondages [...] pour mesurer le fond de la mer... Cette évaluation [n'est pas] une véritable statistique... ».

Heinrich Rauchberg (Autriche) : « ... je ne crois pas que cette question ait besoin d'un examen ultérieur... Une discussion [...] est superflue... ».

Guillaume Milliet (Suisse) : « ... il n'est pas juste de donner à la méthode représentative une importance que la statistique sérieuse ne lui reconnaîtra jamais... ».

Émile Cheysson (France) : « ... il ne faut pas que [la monographie] puisse recevoir quelque atteinte du présent débat... ».

L'un des rares (ils seront deux!) à ne pas désavouer A. Kiaer est le Français Émile Levasseur : « entre les statistiques générales [...] et les monographies [...], n'y a-t-il pas la place pour un troisième procédé, l'exploration statistique, [...] à l'aide des moyens de la statistique appliqués non à la totalité [...] mais à un nombre déterminé et restreint...? ».

Malgré toutes ces critiques, l'idée de « une partie pour le tout » est lancée à un public de statisticiens. Il faut remarquer qu'a priori, aucun pont n'est jeté entre les probabilités et le dénombrement représentatif.

Après Berne, Kiaer va intervenir dans les congrès de l'IIS jusqu'en 1903, nourrissant régulièrement sa réflexion d'éléments nouveaux.

En 1897, à Saint-Petersbourg, il confirme que la représentativité est décidée après contrôle avec les résultats d'un recensement, et conteste les monographies « qui ne donnent pas une vraie miniature

de la totalité ». Il reçoit, cette même année, l'appui de la Conférence des Statisticiens Scandinaves, qui adopte une motion proclamant l'intérêt des méthodes de sondages représentatifs.

Lors du congrès de Budapest (1901), Kiaer précise :

« ... la méthode représentative demande un grand nombre d'unités d'observation réparties de telle manière que [celles] de différents caractères soient représentées autant que possible dans les mêmes rapports qu'elles existent dans l'ensemble... ».

Outre cette anticipation de la méthode des quotas, Kiaer entrevoit les procédures de rééchantillonnage :

« ... On a demandé par quel moyen on pourra contrôler que les résultats des investigations représentatives correspondent véritablement aux rapports réels de l'ensemble quand il s'agit d'une investigation absolument nouvelle? [...]. On pourra diviser l'investigation en deux ou trois parties distinctes... Si ces deux ou trois méthodes donnent des résultats conformes, on ne peut pas nier qu'il y ait là une garantie... ».

Dans la discussion qui suit, beaucoup moins négative qu'à Berne, L. von Bortkiewicz, professeur à l'Université de Berlin, suggère de recourir au calcul des probabilités pour tester statistiquement l'écart existant entre les répartitions de l'échantillon et de la population totale sur les variables-clés :

« ... il s'agit de déterminer si la différence observée peut ou ne peut pas être considérée comme étant due au jeu de hasard, aux causes fortuites ou accidentelles... »

L'intervention de von Bortkiewicz marque la première introduction sérieuse des probabilités dans les études de l'IIS consacrées à l'échantillonnage représentatif, et la première recommandation de l'utilisation d'un test de significativité.

Kiaer reçoit en outre l'appui de Carroll Wright, du Ministère du Travail des États-Unis, qui, dans une lettre adressée au rapporteur du Comité de l'IIS chargé d'étudier la méthode représentative, confirme la bonne opinion de cette technique, fréquemment utilisée avec succès.

En septembre 1903, à Berlin, Kiaer intervient de nouveau, et est appuyé par Karl Mayet (Allemagne) :

« ... la connaissance approximativement juste obtenue par une estimation systématique vaut mieux que le manque complet de connaissances... »

Le Français Lucien March appuie la méthode de Kiaer, qu'il juge proche dans son esprit des techniques de Moheau et Laplace, et insiste pour que :

« ... Quand on présente les résultats d'une enquête semblable, [on explique bien] en détail comment a été fait le choix des unités observées afin que le lecteur puisse faire la critique de ce choix... ».

La section de l'IIS clôt ses travaux « ... en recommandant [l'application correcte de la méthode représentative] pourvu que dans le compte rendu on spécifie nettement dans quelles conditions s'est fait le choix des unités observées... ».

Après 1903, l'ordre du jour de l'IIS ne mentionne plus la question du sondage représentatif jusqu'en 1925, principalement en raison de l'absence de la personnalité de A. Kiaer, qui décèdera dans l'intervalle. Par ailleurs interrogé en 1979 par W. Kruskal et F. Mosteller, P. Bjerve, directeur du Bureau central de Statistique de Norvège, répond que le recul de Kiaer par rapport à l'IIS fut causé par ses responsabilités croissantes, et que « Kiaer's work in sampling was a one-man show ».

Durant ces 22 ans, l'introduction du raisonnement probabiliste dans les méthodes partielles se poursuit.

Trois noms sont à mentionner. En 1911, von Bortkiewicz prévoit la césure entre statistique privée et statistique officielle, la première tendant à être partielle et sélective, la seconde étant générale (au sens du recensement) et complète; la statistique privée courra alors le risque d'être non

représentative. Il étend également la notion de représentativité aux tests sur les lois jointes et non seulement sur les marges.

Durant la même période, des contributions essentielles sont apportées par l'Anglais Arthur Bowley (1869-1957), tout d'abord par sa communication de 1906 à la Royal Statistical Society, puis par un livre publié en 1920, et par sa participation à la commission chargée par l'IIS, en mai 1924, d'étudier l'applicabilité de la méthode représentative. Bowley développe l'échantillonnage aléatoire, la stratification, établit l'équation d'analyse de la variance en univers stratifié, les formules de la variance d'un total dans les cas simple et stratifié. Ses travaux sont d'autant plus innovateurs que les tenants de la méthode représentative utilisent plutôt, pour sélectionner les unités observées, un « choix judicieux » — le « purposive selection » —. Bowley a une claire vision des défauts de cette méthode, tout simplement parce que dans le sondage par choix judicieux, l'origine de l'aléatoire nécessaire pour appliquer la théorie des probabilités n'est pas spécifié.

Troisième grand nom de la période inter-congrès, A. I. Tchuprow qui définit l'allocation optimale dans un échantillonnage stratifié, ce qui est en contradiction avec l'idée de représentativité courante où l'échantillon est un « modèle réduit » de la population entière.

Le point final (?) de la polémique sur la méthode représentative peut être situé lors du congrès de l'IIS de Rome en 1925, lors de la présentation faite par le rapporteur Adolphe Jensen, Chef du Département de la Statistique du Danemark, des conclusions de la Commission ad-hoc nommée en 1924 et déjà mentionnée (outre A. Jensen, cette commission est composée de Arthur Bowley, Corrado Gini, Lucien March, Verrijn Stuart et Frantz Zizek). Jensen développe deux communications intitulées *Rapport sur la méthode représentative en statistique* et *La méthode représentative en pratique* et Bowley y expose *Measurement of the precision attained in sampling*. La polémique initiale ayant disparu, la résolution adoptée par l'IIS distingue l'échantillonnage aléatoire (choix au hasard, random sampling) et l'échantillonnage raisonné (choix judicieux, purposive selection).

Extrait du Compte rendu du Congrès de Rome :

« ... l'IIS... »

1. ... [...] appelle de nouveau l'attention sur l'avantage considérable que l'on peut obtenir par l'application de la méthode représentative [...].

Pour que les résultats d'une enquête partielle puissent être légitimement généralisés, la fraction retenue comme un spécimen de l'ensemble [...] doit être suffisamment représentative de cet ensemble. A cet effet, le spécimen peut être pris dans la masse de différentes manières; cependant il convient de distinguer deux cas principaux :

A) Prise au hasard. On prend un certain nombre d'unités dans des conditions telles que les unités de l'ensemble soient également susceptibles d'être comprises dans ce nombre. Dans ce cas le degré de précision dépend en grande partie du nombre adopté, celui-ci devant être suffisant pour que les déviations accidentelles puissent être insignifiantes.

B) Choix judicieux. Les unités de l'ensemble étant réparties par groupes, le choix porte sur des groupes tels qu'ils présentent, au total, à peu près les mêmes caractéristiques que l'ensemble. Afin d'obtenir quelque idée de la précision des estimations qui résultent de ce choix, le nombre des groupes doit être suffisant pour que l'on puisse mesurer la variation des éléments caractéristiques d'un groupe à l'autre. Comme la précision des estimations dépend beaucoup de la prudence appliquée au choix du spécimen, les précautions suivantes sont recommandées :

1 — Comparer deux ou plusieurs spécimens obtenus par l'application des mêmes procédés, et, le cas échéant, joindre les spécimens choisis (cette manière de procéder peut être également recommandable en cas de prise au hasard).

2 — Réexaminer minutieusement, en faisant des enquêtes répétées de temps à autre, la relation du spécimen retenu avec l'ensemble;

II. recommande de diriger l'étude de façon à permettre autant que possible le calcul de la précision des résultats. Ces résultats devraient être fournis avec l'indication des limites de l'erreur dont ils sont susceptibles;

III. rappelle, en insistant sur son importance, le vœu formulé dans la résolution de 1903 tendant à ce que la publication des résultats d'une enquête représentative comporte obligatoirement dans tous les cas un compte rendu détaillé des procédés employés pour le choix du spécimen ».

Après 1925, le point de discussion n'est plus sur « faut-il échantillonner ou non » mais sur « comment tirer l'échantillon? ». Le débat va vite concerner les tenants du « random sampling » (la « prise au hasard » de l'IIS) et ceux de la « purposive selection » (le choix judicieux). A. Jensen en est d'ailleurs conscient :

« ... When the IIS discussed the matter twenty years ago, it was the question of the recognition of the method in principle that claimed most interest. Now it is otherwise. I think I may venture to say that nowadays there is hardly one statistician who in principle will contest the legitimacy of the representative method... ».

L'introduction du modèle probabiliste est en cours. L'étape ultérieure débutera en 1934. Cependant, depuis de nombreux lustres, aux États-Unis, un domaine d'utilisation du sondage se développe : la connaissance de l'opinion publique.

V. Un cheminement parallèle : le début des sondages d'opinion aux États-Unis

Les sondages d'opinion sont nés du désir d'obtenir des informations, si possible quantitatives, sur l'état d'esprit de la population. A cet effet, Daniel Defoe en Angleterre et le compte de Lafayette en France mettent en place, dès le début du XIX^e siècle, des réseaux de correspondants répartis géographiquement.

Le vrai berceau des sondages d'opinion se trouve cependant aux États-Unis, à l'occasion des couvertures de presse des élections présidentielles. Dès 1824, le Harrisbourg Pennsylvanian et le Raleigh Star font des enquêtes préélectorales par consultation individuelle d'électeurs (les « votes de paille »). Par la suite, d'autres journaux vont reprendre cette habitude très appréciée du public : New York Herald, Chicago American, Columbus Dispatch, et, en particulier, à partir de 1916, The Literary Digest qui va, contre son gré, jouer un rôle important.

Ces échantillons ne reposent sur aucun critère de représentativité (le mot n'apparaîtra d'ailleurs que 70 ans plus tard). Leurs propriétés sont dues essentiellement à des tailles fort élevées : 30 000 pour le New York Herald en 1905, plus de 2 000 000 pour le Literary Digest en 1936.

Une date cruciale pour l'histoire de l'échantillonnage est le mardi 3 novembre 1936, jour de la publication des résultats de l'élection présidentielle aux États-Unis. Alors que le Literary Digest a prédit la victoire de Landon, en utilisant la technique des votes de paille qui lui a fort bien réussi depuis 1916, c'est F.D. Roosevelt qui est élu. Celui-ci avait été donné favori par trois « sondages » réalisés indépendamment par Archibald Crossley, Elmo Roper et George Gallup (ce dernier a créé son propre institut en 1935). Ces trois études ont été réalisées selon la méthode représentative par choix judicieux : tirage des personnes interrogées avec contrôles par quotas, une « Amérique en microcosme » d'après le mot de E. Roper, paraphrasant Kiaer.

En dépit de sa grande taille, l'échantillon du Literary Digest s'était révélé biaisé; la technique des votes de paille disparaît alors. Dès 1937, de nombreuses revues américaines créent une rubrique d'opinion, nourrie par des consultations réalisées par des organismes privés utilisant les sondages représentatifs.

L'année 1938 voit l'apparition des premiers instituts d'études de l'opinion en Grande-Bretagne et en France; Jean Stoetzel, professeur de sociologie à la Sorbonne, fonde l'IFOP (Institut Français d'Opinion Publique). De nombreux autres pays vont suivre ce mouvement et en 1947 se tient, à Paris, le premier congrès international en ce domaine.

VI. *La situation en Russie pré-révolutionnaire*

Au XIX^e siècle, en même temps qu'apparaissent et se généralisent, aux États-Unis, les sondages d'opinion, l'Empire russe connaît un développement des méthodes fondées sur une vision partielle de la population.

Le cadre est institutionnel : en 1864, le pouvoir central crée les zemstvos, gouvernements locaux au sein desquels existent des départements statistiques chargés de recueillir, en particulier, les informations sur l'état de l'agriculture pour aider à définir les orientations de la politique économique rurale. Le père spirituel de la « statistique zemstvo » est Alexander Ivanovitch Tchuprov (1842-1908), professeur d'économie politique et de statistique à l'Université de Moscou, membre de l'IIS dès sa création en 1885, fondateur de la Société Juridique de Moscou.

L'existence d'instituts statistiques dans les zemstvos favorise l'apparition de nombreux statisticiens mathématiciens qui vont contribuer fortement au développement des méthodes : outre A.I. Tchuprov, il faut mentionner son fils Alexander Alexandrovitch Tchuprov (1874-1926), professeur au département Économie de l'Institut Polytechnique de Saint-Petersbourg, A.V. Peshekhonov (1867-1933), V.G. Groman, qui deviendra après la révolution d'octobre l'un des grands responsables de l'Office Central Statistique et du Gosplan, et A.G. Kovalevsky.

Il semble que, très tôt, les techniques d'échantillonnage aient été utilisées par les statisticiens russes, dans l'impossibilité d'observer exhaustivement les territoires de leur compétence. Dès 1894, à Moscou, A.I. Tchuprov présente une communication sur la « méthode monographique » consistant en la sélection d'éléments typiques déterminés à partir d'une connaissance exhaustive d'une population. En 1897, au congrès de l'IIS à Saint-Petersbourg, il intervient sur les méthodes représentatives que le Norvégien A. Kiaer a exposées à Berne en 1895. Si A.I. Tchuprov jouera ensuite un grand rôle pour relayer et populariser les idées de Kiaer auprès des statisticiens russes, au sens où il peut faire état de l'utilisation de ces techniques dans les pays de l'Europe de l'Ouest, il est admis que les sondages — au sens large du terme — sont néanmoins communément utilisés dans la statistique russe indépendamment de Kiaer : en 1896, Peshekhonov procède à un échantillonnage au 1/10 de zones géographiques, une sélection au 1/5 est tirée pour le district de Viatka en 1906, etc.

Une avancée méthodologique essentielle apparaît en 1910, parallèlement aux travaux de Bowley en Angleterre : A.A. Tchuprov étudie l'échantillonnage aléatoire, développe son utilisation pour mesurer la précision des estimateurs, et fait référence au tirage par grappes et à l'échantillonnage stratifié, avec et sans remise. Influencés par A. Markov, conscient de l'importance de l'utilisation des probabilités en statistique, Tchuprov et ses élèves vont obtenir de nombreux résultats qui ne seront découverts, ou popularisés, que bien plus tardivement en Europe de l'Ouest, notamment par J. Neyman.

Le meilleur exemple du niveau atteint en Russie par les sondages, en pratique et en théorie, est fourni par le contenu d'un ouvrage publié en 1924 par A.G. Kovalevsky. Au-delà d'une présentation mathématique du tirage systématique, Kovalevsky donne un traitement rigoureux du sondage par stratification, et établit l'allocation optimale par strate — ce qui est contraire à l'interprétation de la notion de représentativité admise à l'époque —. Ce résultat ne sera découvert par Neyman que dix ans plus tard.

Les difficultés dues à l'éloignement, à la langue et, bien sûr, aux événements liés à la Révolution d'Octobre, ont longtemps empêché de bien saisir l'importance de l'école russe dans la « statistique partielle ». Il semble dorénavant bien établi que, dès l'origine, le développement de la méthodologie et de l'utilisation des sondages en Russie tsariste ait été considérable, et que nombre de statisticiens russes du début du siècle peuvent être considérés comme les précurseurs et les initiateurs de la méthode, au même titre que Kiaer, Bowley et Neyman.

VII. *Les débuts du développement contemporain*

Après 1925, la discussion sur la méthode représentative ne porte plus sur son principe, mais sur ses modalités d'application. Le Congrès de l'ISS ayant distingué méthode aléatoire et méthode raisonnée, les travaux approfondissent cette double approche. Dans la lignée des résultats de Bowley et Tchuprov, Jerzy Neyman (1894-1981) peut être considéré comme l'un des fondateurs de la théorie statistique des sondages, entre autres par son exposé du 19 juin 1934 devant la Royal Statistical Society (publié dans le *Journal of Royal Statistical Society* de cette même année). L'intérêt de cette conférence déborde très largement du domaine des sondages, Neyman y détaillant en annexe le principe des intervalles de confiance, et vulgarisant le théorème de Markov sur les moindres carrés.

L'apport de Neyman va être considérable car, le premier, il va prendre parti entre échantillonnage aléatoire et échantillonnage raisonné qui sont placés sur le même plan par Bowley et la recommandation de l'ISS. Pour cela, il s'appuie sur les travaux de C. Gini et L. Galvani qui, en novembre 1926, utilisent la méthode par choix raisonné pour tirer un échantillon à 15 % du recensement général effectué en Italie en 1921, échantillon que les deux statisticiens italiens souhaitent « représentatif du pays dans ses composantes démographique, sociale, économique et géographique ». Partant des résultats sur la théorie de l'estimation de Ronald Fisher, Neyman exhibe les hypothèses à partir desquelles les estimateurs issus de la méthode par choix raisonné ont de bonnes propriétés statistiques de convergence, hypothèses qui, en pratique, ne sont jamais toutes satisfaites simultanément. Il montre donc, a contrario, que les deux techniques d'échantillonnage ne sont pas équivalentes, et qu'il convient de recommander la méthode aléatoire puisque « ... la confiance générale placée dans la sélection judicieuse devra être quelque peu réduite... ».

A Gini et Galvani qui demandent (déjà) s'il est possible de définir « un échantillon représentatif », Neyman répond qu'une méthode de sondage sera représentative si elle permet d'obtenir des estimateurs convergents et de calculer la précision des résultats.

Au-delà de cette question — fondamentale — de choix entre méthodes d'échantillonnage, Neyman contribue à de nombreux développements : à l'échantillonnage aléatoire simple, avec ou sans remise, et à l'échantillonnage stratifié, déjà envisagés par Bowley, il ajoute l'échantillon par grappes, technique qui a sa préférence; il fait grandement avancer la théorie de l'estimation en sondages; il étudie l'allocation optimale d'un échantillon par strates, ainsi que le choix de l'allocation minimisant la variance de l'estimateur, à plan d'échantillonnage fixé. Ce dernier point est important, puisqu'il remet en cause le principe de représentativité communément admis consistant en une homothétie de la population. Dans l'histoire du concept de représentativité, l'apport de Neyman sur ce point est essentiel.

Établissant la supériorité des techniques de sondage aléatoire en univers stratifié, Neyman oriente une grande partie de la recherche dans cette voie. Ironie des temps, les critiques — justifiées — que Neyman adresse à la méthode représentative à choix judicieux coïncident avec l'émergence de la vogue des sondages d'opinion effectués selon cette même technique, ou plutôt l'une de ses diverses variantes, la méthode des quotas.

En 1938, Neyman mène à son terme l'échantillonnage à deux degrés, et l'utilisation conjointe des fonctions de coût et de la variance pour définir les allocations optimales.

Toutes les bases de la statistique des sondages sont alors posées. Deux grandes voies de développement apparaissent : l'une est liée aux problèmes de la pratique (modes de tirages, estimation, variabilité, non-réponses, etc.), l'autre fait référence au modèle probabiliste sous-jacent à la théorie des sondages. Les avancées de la première catégorie ont permis la construction de la méthodologie fondamentale des sondages. Citons-en quelques grandes étapes sans aucune prétention d'exhaustivité : Kiaer (1897), Tchuprov (1923) et Neyman (1934) ont évoqué à des titres divers les tirages à probabilités inégales. Morris Hansen et William Hurwitz, en 1943, établissent les résultats d'estimation et de précision des estimateurs lorsque les unités élémentaires sont tirées avec des probabilités inégales, pour des plans de sondage avec remise. La détermination des probabilités d'inclusion optimales est due aux mêmes auteurs, en 1949. D. Horvitz et T. Thompson généraliseront, en 1952, les résultats précédents aux tirages sans remise, avec probabilités inégales. Les conditions de positivité des estimateurs de la variance dues à F. Yates et P. Grundy sont publiées en 1953.

Parallèlement, se développent les procédures d'estimation des paramètres d'intérêt lorsqu'une information supplémentaire existe. Les estimateurs par le quotient, par la différence et par la régression apparaissent en filigrane chez Neyman (1934), mais sont approfondis par W. Cochran (1942), M. Hansen, W. Hurwitz et W. Madow (1953).

La pratique laisse une place importante aux problèmes de non-réponses, totales ou partielles, auxquels des éléments de solutions commencent à être apportés par Hansen et Hurwitz (1946) et Tore Dalenius (1957).

L'étude de la variabilité des estimateurs est également un vaste champ d'investigation, essentiellement pour des statistiques non linéaires. Aux méthodes de linéarisation usuelles en statistique classique vont se superposer les procédures de réplification d'échantillons : P. Mahalanobis (1946), W. Deming (1960), les techniques de demi-échantillons équilibrés (P. Mac Carthy, 1966), qui anticipent les applications du jackknife et du bootstrap en pleine actualité.

Un dernier point technique concerne les enquêtes à passages répétés. On peut considérer que Raymond Jessen (1942) apporte, le premier, une solution à ces questions spécifiques.

VIII. *Les développements actuels*

Depuis le milieu des années cinquante, les investigations méthodologiques sont de plus en plus fondées sur les relations existant entre les sondages et le modèle de la statistique inférentielle, les « foundations » ainsi qu'aiment à les appeler les Anglo-saxons.

Un premier axe de recherche va concerner l'application aux sondages des concepts statistiques, introduits depuis Fisher, que sont l'admissibilité, l'exhaustivité, le principe de vraisemblance. Ces travaux reposent sur la considération du modèle de base des sondages, à partir de la notion de plan de sondage représenté par une loi de probabilité définie sur l'ensemble S des échantillons possibles; cette loi est déterminée par l'ensemble des probabilités élémentaires $p(s)$, $s \in S$. Si l'on s'intéresse à une variable X , le paramètre non aléatoire mais inconnu du modèle est (X_1, \dots, X_N) , vecteur des valeurs prises par X sur les N individus de la population. Pour un échantillon de taille n , les observations portent sur des individus $s = (i_1, \dots, i_n)$ et sur des valeurs $x = \{x_j = X(i_j), j = 1 \text{ à } n\}$. La loi de ces observations (s, x) est parfaitement caractérisée par la donnée de la loi de l'échantillon s et celle des observations x conditionnellement à s .

La structure est donc paramétrique, paramétrée par (X_1, \dots, X_N) . Malheureusement, le modèle statistique de la théorie des sondages se révèle complexe : d'une part, le support de la loi des observations dépend du paramètre, ce qui crée toujours des difficultés techniques, phénomène bien connu en statistique mathématique; d'autre part, ce modèle est dominé par la mesure de comptage sur l'ensemble des observations possibles, mesure non σ - finie. Le recours à la théorie statistique sera, pour cette raison également, difficile. Néanmoins un certain nombre d'auteurs ont tenté d'étendre les critères de choix d'un estimateur au cas de la théorie des sondages.

Les concepts de statistique exhaustive ou exhaustive minimale et l'amélioration d'un estimateur par le théorème de Rao-Blackwell sont utilisés pour la première fois par D. Basu en 1958.

Les restrictions apportées au modèle statistique des sondages n'entrent pas en ligne de compte, et il est possible d'utiliser un critère de factorisation de type Fisher-Neyman pour déterminer des statistiques exhaustives. Cependant, leur utilisation est d'un intérêt faible, car les seules réductions par exhaustivité consistent à ne pas tenir compte de l'ordre dans lequel les individus ont été tirés, réduction déjà implicitement retenue dans l'écriture du modèle.

L'approche fondée sur le maximum de vraisemblance (V. Godambe, 1966) se révèle peu appropriée. En effet, en général, on possède peu d'observations par rapport à la dimension du paramètre (X_1, \dots, X_N) . En outre, certains paramètres sont parfaitement connus, ceux des n individus appartenant à l'échantillon, mais on ne possède aucune information sur les $N - n$ paramètres restants. Le maximum de vraisemblance se heurte donc à un problème de non-estimabilité.

Il est par contre possible d'estimer sans biais certaines fonctions d'intérêts $g(X)$ du paramètre; le fait que le modèle ne soit pas dominé par une mesure σ - finie ni exponentiel implique cependant que n'existe aucun résultat général concernant l'estimation sans biais optimale. Dès 1955, V. Godambe démontre qu'il n'existe pas d'estimateur linéaire sans biais de la moyenne \bar{X} sur la population qui soit uniformément de variance minimale. Le résultat est étendu à la classe de tous les estimateurs sans biais en 1965 par V. Godambe et V. Joshi. Le premier théorème d'existence d'un estimateur sans biais uniformément de variance minimale est dû à C. Cassel, C. Särndal et J. Wretman, en 1977, sous un certain nombre de contraintes sur le plan de sondage.

L'échec relatif de l'application aux sondages des principes statistiques qui viennent d'être énoncés va conduire certains auteurs à introduire, vers la fin des années soixante-dix, une approche plus globale des sondages par les modèles dits de « superpopulation » qui figurent déjà en filigrane dans les travaux de Cochran en 1939, ceux de W. Deming et F. Stephan en 1941, ou ceux de W. et L. Madow en 1944. Les premières publications fondamentales sur les modèles de superpopulation datent de 1976 et 1977 (C. Cassel, C. Särndal, J. Wretman); citons encore C. Särndal en 1984.

Cette optique consiste à considérer les valeurs (X_1, \dots, X_N) de X sur la population comme des réalisations de N variables aléatoires dont la loi jointe Q appartient à une famille de lois de probabilités spécifiée. Alors que dans la démarche classique des sondages, l'objectif est souvent de rechercher, pour un plan de sondage p donné, l'estimateur h du paramètre $g(X)$ qui minimise

$$EQM(h) = \sum_{s \in S} p(s)(h - g(X))^2,$$

d'autres critères sont privilégiés; par exemple, si $g(X) = \bar{X}$, on pourra rechercher l'estimateur h qui minimise

$$EQE(h - \bar{X})^2 \text{ sous la contrainte } E(h) = \bar{X}$$

ou bien

$$EQE(h - \bar{X})^2 \text{ sous la contrainte } EQ(h - \bar{X}) = 0.$$

L'apport essentiel des modèles de superpopulation est, au niveau de la théorie, de proposer une réflexion globale sur le choix des plans de sondage, des estimateurs, et non des critères d'optimalité. En pratique, la spécification de modèles adaptés, par exemple le modèle de régression, fournit des éléments intéressants pour prendre en compte les non-réponses, étudier des enquêtes répétées dans le temps, ou encore estimer $g(X)$ dans un domaine spécifique de la population.

Un autre exemple du lien entre statistique et sondages est le développement récent de l'application du bootstrap aux données issues d'un sondage. Apparue en 1979, la technique du bootstrap, due à B. Efron, renoue avec la tradition des replications d'échantillons telles que les utilisait l'Institut Indien de Statistique de P. Mahalanobis dans l'immédiat après-guerre. Le principe du bootstrap est fort simple : possédant un échantillon $E = (Y_1, \dots, Y_n)$ extrait d'une même loi P , on s'intéresse à un paramètre $\theta = g(P)$, estimé par $\hat{\theta}(Y_1, \dots, Y_n)$.

Désirant avoir une idée du comportement de $\hat{\theta}$ pour estimer θ (biais, variance, ...), Efron propose de ré-échantillonner avec remise dans E et, ayant obtenu un nouvel échantillon $E^* = (Y_1^*, \dots, Y_n^*)$, d'approximer la loi de $\hat{\theta} - \theta$ par celle de $\theta^* - \hat{\theta}$, où $\theta^* = \theta(Y_1^*, \dots, Y_n^*)$. Dès 1981, les spécialistes des sondages entrevoient l'intérêt du bootstrap pour obtenir des estimations de la précision d'estimateurs du paramètre d'intérêt $g(X)$, en particulier lorsque celui-ci n'est pas simple (statistique hautement non linéaire, ratios, corrélation, fractiles), et lorsque le sondage est stratifié ou « complexe ». Les articles de D. Krewski et J.N.K. Rao (1981), J.N.K. Rao et Wu (1983) et P. Bickel et D. Freedman (1984) permettent d'entrevoir de futurs développements du plus haut intérêt pour les praticiens des sondages. Mentionnons également J.-C. Deville qui, en 1986, étudie l'application du bootstrap à un sondage avec probabilités inégales et à un algorithme d'échantillonnage.

Ce retour vers des considérations pratiques apparaît également dans les développements récents des algorithmes de tirage d'un échantillon ou dans les modélisations des non-réponses utilisant les modèles économétriques sur variables qualitatives. Cette dernière approche, cohérente avec les modèles de superpopulation, postule l'existence d'une variable latente inobservable l_i régissant, pour chaque individu i , le fait de répondre ou ne pas répondre; l_i dépend de variables exogènes $(Z_{i1}, \dots, Z_{ik}) = Z_i'$:

$$l_i = Z_i' c + v_i \quad (i = 1 \text{ à } n).$$

Si par exemple, on postule un modèle général de la forme régression sur des variables auxiliaires $(Y_{i1}, \dots, Y_{ip}) = Y_i'$:

$$x_i = Y_i' b + u_i \quad (i = 1 \text{ à } n),$$

on peut envisager une solution globale du système (l_i, x_i) sous certaines conditions techniques portant sur (u_i, v_i) . Ce modèle est dit tobit généralisé lorsque le couple (u_i, v_i) est gaussien; il a, par exemple, été utilisé à l'INSEE par F. Chicoineau, J.-F. Payen et C. Thélot en 1985 pour redresser les données de salaires relevées dans l'enquête annuelle sur l'emploi. Une alternative logistiquе bidimensionnelle à la loi normale a été proposée par C. Thélot en 1985.

Pour conclure, nous reviendrons à la méthode statistique. La lecture des « grands » ouvrages ou articles qui ont marqué l'histoire de la statistique à la fin du XIX^e siècle et au début du XX^e siècle montre que sont passées généralement sous silence les conditions d'obtention des données, ce qui est parfaitement compréhensible au vu des principales étapes du développement des sondages. Néanmoins, les procédures statistiques sont habituellement utilisées comme si les propriétés de l'échantillon étaient semblables à celles de la population, c'est-à-dire, en simplifiant, comme si l'on procédait toujours à un tirage équiprobable. Ne faudrait-il pas modifier les méthodes usuelles pour tenir compte de la procédure aléatoire de recueil des données et de la complexité croissante des plans de sondage?

Des réflexions récentes ont donc commencé à apparaître sur la validité — la « robustesse » — des procédures statistiques lorsque les données proviennent de sondages complexes. Ainsi, en 1981, J.N.K. Rao et A. Scott ont montré que, dans le problème classique d'adéquation à une loi, la statistique

du khi-deux fonctionne bien dans le cas d'un tirage aléatoire simple, mais que, pour un échantillonnage stratifié, cette même statistique ne suit plus asymptotiquement une loi du χ^2 , mais un mélange de lois du khi-deux à un degré de liberté. La statistique à considérer pour fonder le test d'adéquation doit être modifiée en conséquence. De même, quelques auteurs ont établi les conditions sous lesquelles l'assimilation échantillon-population est fondée pour l'usage du modèle de régression linéaire.

Cette tendance à ré-examiner la robustesse des outils de la statistique mathématique au mode d'obtention des données devrait s'accroître rapidement, et promettre de futurs développements pour la science statistique dans son ensemble.

L'ÉTYMOLOGIE DU MOT « SONDAGE »

L'emploi du substantif « sonde » dans la langue française semble remonter au treizième siècle. Son origine est probablement nordique, du mot *sund* qui signifie « mer, détroit », et il représente un instrument servant à mesurer la profondeur de l'eau. Le verbe « sonder » au sens maritime est rencontré postérieurement à son substantif (1382). Au XVI^e siècle apparaissent la sonde chirurgicale, et l'usage du verbe « sonder » pour qualifier l'action de celui qui cherche à pénétrer un secret (1556, Ronsard). Le « sondeur » naît en 1690, le « sondage » en 1769.

Le terme « sondage », au sens figuré d'enquête ou d'investigation rapide, est utilisé par Balzac en 1841, mais on trouve dans un article du Baron de Bielfeld publié en 1760 (mais écrit, selon toute vraisemblance, bien avant) et intitulé « Arithmétique Politique » un plaidoyer pour la création d'un bureau central de statistique : « l'homme d'État qui prendra toujours en main cette sonde ne marchera point au hasard et évitera mille écueils... ».

Popularisé depuis les années 1930 par les enquêtes d'opinion publique, l'utilisation, en France, de l'unique terme « sondage » est récente, et diffère de la pratique des pays anglo-saxons. En effet, ces derniers distinguent le sondage maritime (*sounding* en anglais, *lotung* en allemand), le sondage géologique (*boring, probebohrung*), le sondage médical (*probing, stichprobe*), le sondage d'opinion (*poll, umfrage*), l'échantillonnage statistique (*sampling, statistische untersuchung*).

Fallait-il imposer des vocables distincts selon la méthodologie ou le domaine d'application? La Commission Française de Normalisation du vocabulaire statistique décida, en 1947, de l'emploi du seul terme sondage pour désigner toutes les opérations d'échantillonnage quel que soit le domaine.

BIBLIOGRAPHIE

- D. BASU (1958) : On sampling with and without replacement, *Sankhya*, 20, pp. 287-294.
- D. BASU (1969) : Rôle of the sufficiency and likelihood principles in sampling survey theory, *Sankhya*, 31, pp. 441-454.
- P. BICKEL, D. FRIEDMAN (1984) : Asymptotic normality and the bootstrap in stratified sampling, *Ann. Statist.*, Vol. 12, pp. 470-482.
- C. BLUM-GIRARDEAU (1976) : La Statistique, miroir de l'Histoire, *Économie et Statistique*, n° 83, novembre 1976.
- A. BOWLEY (1960) : *Elements of statistics* (4^e édition), King, Londres.
- A. BOWLEY (1920) : Measurement of the precision attained in sampling, *Bulletin de l'IS*, 22, pp. 1-62 (Congrès de Rome, 1925).
- C. CASSEL, C. SARNDAL, J. WRETMAN (1976) : Some results on generalized difference estimation and generalized regression estimation for finite populations, *Biometrika*, Vol. 63, pp. 615-620.
- C. CASSEL, S. SARNDAL, J. WRETMAN (1979) : *Foundations of inference in survey sampling*, Wiley, New York.
- Wei-Ching CHANG (1974) : Statistical theories and sampling practice, in *On the history of statistics and Probability*, éd. D. Owen, Marcel Dekker, New York.
- F. CHICOINEAU, J.-F. PAYEN, C. THÉLOT (1985) : Modélisation et redressement des non-réponses, le cas du salaire, *Notes de l'INSEE*, n° 0071/430.
- W. COCHRAN (1939) : The use of analysis of variance in enumeration by sampling, *J. Amer. Statist. Assoc.*, 34, pp. 492-510.
- W. COCHRAN (1942) : Sampling theory when the sampling units are of unequal sizes, *J. Amer. Statist. Assoc.*, 37, pp. 199-212.
- T. DALENIUS (1957) : *Sampling in Sweden. Contributions to the methods and theories of sample survey practice*, Almqvist et Wiksell, Stockholm.
- W. DEMING, F. STEPHAN (1941) : On the interpretation of censuses as samples, *J. Amer. Statist. Assoc.*, 36, pp. 45-49.
- W. DEMING (1960) : *Sample design in business research*, Wiley, New York.
- M. DEROO, A.-M. DUSSAIX (1980) : *Pratique et analyse des enquêtes par sondages*, PUF, Paris.
- J. DESABIE (1966) : *Théorie et pratique des sondages*, Dunod, Paris.
- J.-C. DEVILLE (1986) : Replications d'échantillons : demi-échantillons, jackknife et bootstrap, in *Les Sondages*, publication de l'ASU, *Economica* 1987, Paris.
- J. et M. DUPAQUIER (1984) : *Histoire de la démographie*, Librairie Académique Perrin, Paris.
- B. EFRON (1979) : Bootstrap methods : another look to the jackknife, *Ann. Statist.*, 7, pp. 1-26.
- C. GINI, L. GALVANI (1929) : *Annali di Statistica*, Ser. VI, Vol. IV, p. 1.
- V. GODAMBE (1955) : A unified theory of sampling from finite populations, *J. Roy. Statist. Assoc.*, 17, B, pp. 269-278.
- V. GODAMBE, V. JOSHI (1965) : Admissibility and Bayes estimation in sampling finite populations, *Ann. Math. Statist.*, 36, pp. 1707-1722.
- V. GODAMBE (1966) : A new approach to sampling from finite populations, *J. R. Statist. Soc.*, B, 28, pp. 310-328.
- V. GODAMBE et D. SPROTT (1971) : *Foundations of statistical inference*, Wiley, New York.
- C. GOURIEROUX (1981) : *Théorie des sondages*, *Economica*, Paris.
- J.-M. GROSBKAS (1987) : *Méthodes statistiques des sondages*, *Economica*, Paris.
- M. HANIF, K. BREWER (1980) : Sampling with unequal probabilities without replacement : a review, *Int. Statist. Rev.*, 48, pp. 317-335.
- M. HANSEN (1987) : Some history and reminiscences on survey sampling, *Statistical Science*, 2, pp. 180-190.
- M. HANSEN, W. HURWITZ (1943) : On the theory of sampling from finite populations, *Ann. Math. Statist.*, 14, pp. 332-362.
- M. HANSEN, W. HURWITZ (1946) : The problem of non-reponse in sample surveys, *J. Amer. Statist. Assoc.*, 41, pp. 517-529.
- M. HANSEN, W. HURWITZ, W. MADOW (1953) : *Sample survey methods and theory* (Vol. I, II), Wiley, New York.
- M. HANSEN, W. MADOW (1974) : Some important events in the historical development of sample survey, in *On the history of Statistics and Probability*, éd. D. Owen, Marcel Dekker, New York.
- M. HANSEN, T. DALENIUS, B. TEPPING (1985) : The development of sample surveys of finite populations, A celebration of statistics, *The ISI Centenary Volume*, A. Atkinson and S. Fienberg éditeurs, Springer-Verlag.
- H. HARTLEY, J.N.K. RAO (1962) : Sampling with unequal probabilities and without replacement, *Ann. Math. Statist.*, 33, pp. 350-374.

DE L'EXHAUSTIF AU PARTIEL : UN PEU D'HISTOIRE SUR LE DÉVELOPPEMENT DES SONDAGES

- D. HORVITZ et D. THOMPSON (1952) : A generalization of sampling without replacement from a finite universe, *J. Amer. Statist. Assoc.*, 47, pp. 663-685.
- INSEE : Pour une histoire de la Statistique, Actes des Journées d'Études sur l'histoire de la Statistique, 23, 24, 25 juin 1976, INSEE, Paris.
- A. JENSEN (1926, a) : Report on the representative method in statistics, *Bulletin de l'IS*, 22, livraison 1, pp. 359-380 (Discussion pp. 58-69, 185-186, 212-213).
- A. JENSEN (1926, b) : The representative method in practice, *Bulletin de l'IS*, 22, livraison 1, pp. 381-439.
- R. JESSEN (1942) : Statistical investigation of a sample survey for obtaining farm facts, Iowa State College of Agriculture and Mechanics Arts, 304.
- N. JOHNSON, H. SMITH, éditeurs (1969) : *New Development in survey sampling*, Wiley, New York.
- N. JOHNSON, S. KOTZ (1985) : *Encyclopedia of Statistics*, « article Neyman, Jersy », Vol. 6.
- A. KIAER (1896) : Observations et expériences concernant des dénombrements représentatifs, *Bulletin de l'IS*, 9, liv. 2, pp. 176-183 (Congrès de Berne, 1895).
- A. KIAER (1899) : Sur les méthodes représentatives ou typologiques appliquées à la statistique, *Bulletin de l'IS*, 11, liv. 1, pp. 180-185 (Congrès de Saint-Petersbourg, 1897).
- A. KIAER (1903) : Sur les méthodes représentatives ou typologiques appliquées à la statistique, *Bulletin de l'IS*, 13, liv. 1, pp. 66-78 (Congrès de Budapest, 1901).
- A. KIAER (1905) : Discours sans intitulé sur la méthode représentative, *Bulletin de l'IS*, 14, liv. 1, pp. 119-134 (Congrès de Berlin, 1903).
- D. KREWSKI, J.N.K. RAO (1981) : Inference from stratified sample, *Ann. Statist.*, 9, pp. 1010-1019.
- W. KRUSKAL, F. MOSTELLER (1979, a) : Representative Sampling, I : Non-scientific literature, *Int. Statist. Rev.*, 47, pp. 13-24.
- W. KRUSKAL, F. MOSTELLER (1979, b) : Representative Sampling, II : Scientific literature, excluding statistics, *Int. Statist. Rev.*, 47, pp. 111-127.
- W. KRUSKAL, F. MOSTELLER (1979, c) : Representative Sampling, III : The current statistical literature, *Int. Statist. Rev.*, 47, pp. 5-265.
- W. KRUSKAL, F. MOSTELLER (1980) : Representative Sampling, IV : The history of the concept in statistics, 1895-1939, 48, pp. 169-195.
- P.S. LAPLACE (1847) : *Théorie analytique des probabilités*, Paris, Imprimerie Royale.
- P. MAC CARTHY (1966) : Replication : An approach to the analysis of data from complex surveys, *Vital and Health Statistics*, Ser. 2, 14, Washington, US Government Printing Office.
- W. MAIDOW, L. MAIDOW (1944) : On the theory of systematic sampling, *Ann. Math. Statist.*, 15, pp. 1-24.
- P. MAHALANOBIS (1946) : On large-scale sample surveys, *Philos. Trans. Roy. Soc.*, London, 231, B, pp. 329-451.
- L. MARCH (1930) : *Les principes de la méthode statistique*, Paris, Librairie Felix Alcan.
- J. MEYER (1981) : *Colbert*, Hachette, Paris.
- M. MOUCHART, J.-M. ROLIN (1981) (éditeurs) : Enquêtes et Sondages, Comité de Statistique, Série « Recyclage en statistique », Université Catholique de Louvain.
- J. NEYMAN (1934) : On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection, *J. Roy. Statist. Assoc.*, 97, pp. 558-606, discussion pp. 607-625.
- J. NEYMAN (1938) : Contribution to the theory of sampling human populations, *J. Amer. Statist. Assoc.*, 33, pp. 101-116.
- J. NEYMAN et L. LE CAM (1965) : *Bernouilli, Bayes, Laplace*, Anniversary Volume, Proceedings of an International Research Seminar, Statistical Laboratory, University of California, Berkeley 1963, Springer Verlag.
- I. OLKIN (1987) : A conversation with Morris Hansen, *Statistical Science*, 2, pp. 162-179.
- J.N.K. RAO, A. SCOTT (1981) : The analysis of categorical data from complex surveys : Chi square tests for goodness of fit and independence in two way tables, *J. Amer. Statist. Assoc.*, 76, pp. 221-230.
- J.N.K. RAO, WU (1983) : Bootstrap inference with stratified samples, Technical Report, Carletons University, Ottawa.
- R. ROYALL (1968) : An old approach to finite population sampling theory, *J. Amer. Statist. Assoc.*, 63, pp. 1269-1279.
- R. ROYALL (1970) : On finite population sampling theory under certain linear regression modes, *Biometrika*, 57, pp. 337-387.
- C. SARNDAL (1984) : *Inférence statistique et analyse des données sous des plans d'échantillonnage complexes*, Presses de l'Université de Montréal.
- F. STEPHAN (1948) : History of the uses of modern sampling procedures, *J. of Amer. Statist. Assoc.*, 43, pp. 12-39.
- S. STIGLER (1977) : Eight centuries of sampling inspection : the trial of the Pyx, *J. Amer. Statist. Assoc.*, 72, pp. 493-500.

DE L'EXHAUSTIF AU PARTIEL UN PEU D'HISTOIRE SUR LE DEVELOPPEMENT DES SONDAGES

- M. THIONET et R. LEVY-BRUHL (1947-1948): *Théorie et pratique des sondages*, INSEE, Ecole d'Application, Ministère de l'Économie Nationale, Paris.
- C. THELOT (1985): Lois logistiques à deux dimensions, *Annales de l'INSEE*, 58.
- A. TCHUPROV (1923): On the mathematical expectation of the moments of frequency distributions in the case of correlated observations, *Metron*, 2, pp. 461-493 et pp. 646-683.
- F. YATES, P. GRUNDY (1953): Selection without replacement from within strata with probability proportional to size, *J. Roy. Statist. Soc.*, B, 15, pp. 253-261.
- H. WESTERGAARD (1932): *Contributions to the history of statistics*, King, Londres.
- S. ZARCOVICH (1966): *Sondages et recensements*, FAO, Rome.
- S. ZARCOVICH (1967): *La qualité des données statistiques*, FAO, Rome.