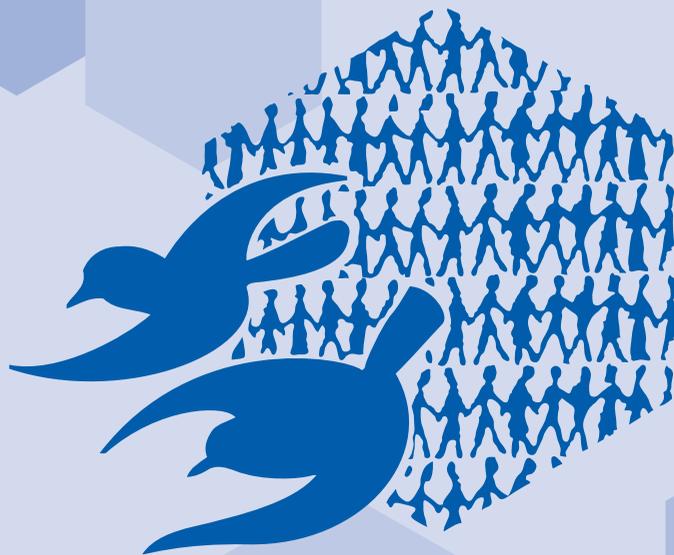


démographie et destin des sous-populations

Colloque de Liège (21-23 septembre 1981)



ASSOCIATION INTERNATIONALE DES DÉMOGRAPHES DE LANGUE FRANÇAISE

AIDELF

AIDELF. 1983. Démographie et destin des sous-populations - Actes du colloque de Liège, septembre 1981,
Association internationale des démographes de langue française, ISBN : 2-7332-7004-4, 452 pages.

QUELQUES DIFFICULTES SOULEVEES PAR L'ETUDE

DES SOUS-POPULATIONS

GILDAS ROY

(Institut national de la statistique et des études économiques, France)

I. INTRODUCTION

La théorie statistique et la pratique acquise depuis de nombreuses années en matière d'enquête permettent aujourd'hui aux statisticiens d'étudier des populations humaines suffisamment nombreuses, définies par des critères aussi objectifs que simples et aisément observables : il n'est pour s'en convaincre que d'observer le grand nombre d'enquêtes par sondage que réalisent annuellement les instituts nationaux de statistiques, pour ne citer qu'eux, auprès des populations nationales.

En revanche, l'étude de sous-populations n'a fait, jusqu'à présent, l'objet d'aucune recherche systématique et demeure, en quelque sorte, le parent pauvre de la statistique d'enquête. En conséquence, ce type d'étude oblige le statisticien à surmonter, au coup par coup, un certain nombre de difficultés ; ainsi, les critères à partir desquels est définie une sous-population, son approche et son observation, la méthode d'échantillonnage retenue, les estimations qui y font suite, sont autant d'obstacles d'une nature particulière par rapport à ceux que l'on rencontre dans les enquêtes à "grande échelle".

On se propose d'examiner ici ces différents problèmes ; on essaiera aussi d'en déduire les implications, si logiques soient-elles, sur les phases de préparation de l'étude d'une sous-population.

II. UNE SOUS-POPULATION EST-ELLE TOUJOURS DEFINISSABLE ?

Après avoir indiqué la position du problème, nous examinerons les difficultés de définition d'une sous-population : sur le plan formel dans un premier temps, sur le plan opérationnel dans un second temps.

1) Position du problème

La définition d'une sous-population est, dans l'ordre naturel des choses, le premier problème auquel se trouve confronté le chargé d'étude.

Une sous-population peut se définir comme l'ensemble des individus d'une population de référence, ayant en commun une ou plusieurs caractéristiques particulières. Cette définition appelle quelques commentaires qu'il n'est sans doute pas inutile de mentionner, même s'ils appartiennent à l'environnement quotidien du praticien. En effet, s'il est aisé de définir objectivement certaines sous-populations, il en est d'autres dont la définition achoppe sur le critère d'objectivité. Ainsi, la sous-population des femmes d'un pays ayant à une date donnée entre 30 et 35 ans révolus se trouve définie à la fois sur le plan formel et opérationnel par la seule expression qui la dénomme ; ce n'est d'ailleurs pas un hasard si le sexe et l'âge ont, de tous temps, retenu l'attention des démographes et, aujourd'hui, des statisticiens.

En revanche, si "la (ou les) caractéristiques particulière(s)" permettant de définir une sous-population font référence à une pratique, une situation ou un comportement social, culturel, économique... alors se pose le problème de la définition objective de cette pratique, de cette situation, de ce comportement : qu'est-ce qu'un pauvre, qu'est-ce qu'un riche (pas nécessairement un non-pauvre), qu'est-ce qu'un catholique, qu'est-ce qu'un bas salaire, qu'est-ce qu'un chômeur... ?

Les différents exemples mettent en évidence la large dimension subjective pesant sur la définition de certaines sous-populations.

2) Difficultés d'ordre formel

Nous distinguerons deux cas selon que la subjectivité précédemment évoquée peut, ou non, être en partie contournée.

. Un premier type de situation autorise le chargé d'étude à recourir, pour pallier cette subjectivité, à un ensemble de critères apparemment plus objectifs, se référant à des faits matériels, concrets ou quantifiables ; la conjonction de ces critères permet alors une approche de la sous-population recherchée. Nous ne parlons que d'approche" car cette conjonction ne coïncide pas nécessairement avec la définition avancée a priori ; en effet, l'objectivité des critères ainsi retenus ne tient le plus souvent qu'à leur contenu numérique : niveau de revenu, niveau de salaire, temps de pratique... et l'on retrouve le problème de la subjectivité quant à la fixation du (ou des) seuil(s) partageant la population de référence en sous-populations.

Une personne ainsi classée dans une sous-population donnée se serait-elle spontanément considérée comme telle ? Pour esquisser cette question, observons simplement qu'aucun seuil ne peut susciter un accord complet : le chargé d'étude est donc amené à trancher, la réflexion laissant à l'arbitraire une part aussi faible que possible.

. Dans un deuxième type de situation, les références du type précédent sont moins fructueuses, sinon impossibles, et ne permettent d'approcher une sous-population que d'assez loin : comment en effet définir formellement des sous-populations telles que celles des catholiques, des cinéphiles, des militants syndicaux... ? On voit que le biais des pratiques religieuses, culturelles ou sociales peut constituer un palliatif, lequel n'est souvent que peu satisfaisant sur le plan formel - nous en verrons les raisons- et, surtout, pose un problème d'ordre opérationnel : l'observation.

3) Difficultés d'ordre opérationnel

Ces difficultés surgissent dès lors qu'il s'agit de passer de la définition formelle d'une sous-population à sa mise en évidence : les critères retenus pour la définir sont-ils observables ? Il y a lieu ici de distinguer trois sources possibles d'achoppement : considérations d'ordre technique, d'ordre confidentiel, d'ordre juridique.

a) Considérations d'ordre technique

Nous classons comme difficultés d'ordre technique les différents obstacles que rencontre la constitution d'une base de sondage, même partielle, même imparfaite.

Citons cependant un contre-exemple : les Déclarations annuelles de salaires (DAS) constituent un recensement (annuel) des salariés des secteurs privés et semi-public, permettant le repérage des "bas-salaires", une fois arrêtée la définition de ceux-ci par référence à un niveau de salaire.

Ne disposant généralement pas d'une base de sondage se rapportant à la sous-population que l'on souhaite étudier, on en vient à utiliser les méthodes d'échantillonnage, aléatoire ou non. Nous les examinerons ci-dessous ; disons pour l'instant qu'elles ne sont pas dépourvues d'inconvénients.

b) Considérations d'ordre confidentiel

Les difficultés d'ordre confidentiel se rencontrent dès lors que la caractéristique classant une personne dans une sous-population est plus ou moins infamante, socialement mal connotée, ou, simplement mal acceptée par la personne elle-même : ainsi en est-il des sous-populations d'anciens détenus, de toxicomanes, de femmes ayant subi un avortement, de personnes atteintes de certains handicaps.

Qu'il nous soit permis de relater ici une expérience au cours de laquelle il était demandé à un groupe d'étudiants de l'ENSAE de préparer un protocole d'enquête sur les loisirs des jeunes sourds-muets ; à l'évidence, la principale difficulté du travail demandé résidait dans la détermination de l'ensemble des jeunes sourds-muets, ensemble dans lequel on puiserait un échantillon. Les étudiants se sont adressés à des institutions spécialisées dans la scolarisation, la formation et, plus généralement, l'insertion des jeunes sourds-muets dans la société ; ils ont rencontré un écho tout à fait favorable auprès de ces institutions, qui ont tenté de les mettre en contact, à fins d'enquête, avec des jeunes sourds-muets ; les étudiants ont très souvent, à ce niveau de leur travail, dû faire face à une réticence quasiment rédhibitoire : il convient de préciser que cette réticence était le fait des parents plus que des enfants. Ceci montre l'exemple d'une sous-population à peu près inobservable par nature, par souci de confidentialité.

c) Considérations d'ordre juridique

Il convient enfin de mettre l'accent, sur un plan plus général que précédemment, sur le respect de l'anonymat des personnes physiques.

Si une personne peut de façon générale être sélectionnée pour appartenir à un échantillon (aléatoire ou non), peut-elle l'être au vu de la possession d'un attribut particulier ? C'est en toute humilité que nous posons ici ce problème de nature juridique qui nous semble avoir sa place dans cette discussion.

Sans tenter d'apporter une réponse à la question posée, nous observons simplement qu'une réponse négative militerait pour les méthodes d'échantillonnage dont nous allons maintenant parler, méthodes qui supposent l'accord préalable des personnes sollicitées.

III. PROBLEMES DE SONDAGE

Lorsqu'on a à extraire un échantillon d'une sous-population préalablement définie, on a vu qu'on disposait rarement d'une base de sondage ; nous passerons d'ailleurs sous silence le cas où on en dispose car, alors, le problème d'échantillonnage est de nature classique, sauf peut-être s'il s'agit

d'une sous-population rare.

Cela dit, nous mentionnerons principalement deux méthodes d'échantillonnage propres à réaliser une enquête par questionnaire auprès d'une sous-population ; une troisième méthode semble s'imposer lorsqu'il s'agit seulement de dénombrer une sous-population, mais pose quelques problèmes de mise en oeuvre.

1) Sondage en plusieurs phases (screening method)

La méthode est bien connue dans son principe. Ayant tiré un échantillon dans la population de référence, on observe, dans une première phase, les caractéristiques que présentent les individus-échantillons ; dans une deuxième phase, on réalise l'enquête auprès du sous-échantillon constitué par les individus présentant les caractéristiques voulues.

La mise en oeuvre de cette méthode est différente selon que les critères définissant la sous-population sont définis avant la première phase, ou à partir de (c'est-à-dire) après cette première phase.

a) Critères définis avant la première phase

Dans ce premier cas (exemple : sous-population des femmes de 30 à 35 ans révolus à la date de l'enquête), le questionnaire comporte un jeu de filtres à la suite des questions concernant le sexe et l'âge, de sorte que l'enquêteur puisse, sur place, poursuivre ou ne pas poursuivre l'enquête selon que la personne présente ou non les caractéristiques voulues.

Cette méthode est d'un coût d'autant plus élevé que la sous-population recherchée est plus rare : l'obtention d'un sous-échantillon de 1 000 personnes nécessite un échantillon initial (première phase) de, en moyenne, 2 000 personnes si la sous-population est en proportion de 50 % de la population de référence ; si cette proportion est de 1 % l'échantillon initial devra comporter 100 000 personnes à contacter.

On observe donc que, pour des populations rares, la méthode est d'un coût rédhibitoire.

b) Critères définis à partir de la première phase

Dans ce deuxième cas (exemple : sous-population des bas-salaires définis par référence à un niveau relatif de salaire), le déroulement de la deuxième phase suppose réunies plusieurs conditions :

- tout d'abord que les personnes contactées dans la première phase aient accepté d'être, le cas échéant, à nouveau sollicitées ;

- que soient dépouillés les résultats de la première phase pour formaliser le critère retenu : par exemple, au vu des résultats d'ensemble, un bas salaire sera un salaire inférieur à So.

L'enquêteur retourne ensuite sur le terrain afin de réaliser l'enquête auprès de la sous-population ainsi déterminée.

Par rapport au cas précédent, le coût de l'opération se trouve évidem-

ment accru : nécessité de dépouiller toute la première phase, nécessité de retourner sur le terrain. La méthode présente par ailleurs un inconvénient supplémentaire : le flou créé par les entrées et sorties de la sous-population entre les deux phases d'enquête.

Entre également dans cette seconde catégorie le cas où les résultats de la première phase doivent, non seulement être dépouillés, mais analysés de façon fine pour pouvoir exhiber une sous-population qu'on n'avait pas su définir explicitement a priori, autrement que par référence à une pratique. La difficulté se situe alors en amont : la référence à une pratique est souvent nécessaire, mais non suffisante, pour cerner la réalité que représente une situation ou un comportement : problème de la définition d'une confession à partir de pratiques religieuses. Qui plus est, la nature de l'information à recueillir pour exhiber une pratique nécessite un lourd investissement préalable : les études et enquêtes-pilotes semblent aujourd'hui reconnues comme tout à fait indispensables au succès d'études sur des domaines (sous-population) inexplorés.

Nous retenons donc que le sondage en plusieurs phases, pour séduisant qu'il soit a priori, est d'une mise en oeuvre parfois délicate, toujours onéreuse.

2 Méthode "boule de neige" (snowball sampling)

La méthode précédente avait l'avantage d'ouvrir la voie théorique de l'estimation statistique : c'est ce que permettent les sondages aléatoires, où toutes les unités de la population de référence ont une probabilité connue d'appartenir à l'échantillon.

Les obstacles qu'elle oblige cependant à surmonter conduisent à imaginer une autre méthode, dite boule de neige, qui consiste à constituer un échantillon d'une sous-population en utilisant comme informateurs un petit nombre de ses membres : ainsi, on pourrait constituer un échantillon d'Anglais vivant en France à partir d'un petit noyau de telles personnes indiquant les noms et adresses de leurs compatriotes connus et en même situation.

La méthode n'est certes pas probabiliste (toute personne de la sous-population n'a pas une probabilité connue d'appartenir à l'échantillon) mais l'inconvénient n'est pas rédhibitoire. Il est plus préoccupant d'observer que certains individus connus de tous les autres ont considérablement plus de chance d'appartenir à l'échantillon, que d'autres, connus de personne. L'échantillon obtenu ne sera donc que peu représentatif de la sous-population dont il est issu, de sorte que les résultats d'enquête sont à interpréter avec la plus grande prudence.

3 Méthode des réponses anonymisées (randomized response method)

Il nous semble opportun d'évoquer ici cette méthode, qui se prête tout à fait, en théorie, au dénombrement de sous-populations définies par une caractéristique plus ou moins infamante. Cette méthode ne date que de 1965 et de nombreux statisticiens lui ont donné, depuis, des développements intéressants sur le plan théorique, mais qui ne permettent malheureusement pas d'esquiver ses difficultés de mise en oeuvre.

Le principe de la méthode des réponses anonymisées est parfaitement probabiliste et permet d'estimer, dans une population de référence, la proportion des individus présentant une caractéristique particulière (par exemple : être un ancien détenu, être toxicomane, avoir subi un avortement...).

Ce principe est le suivant. Ayant tiré un échantillon dans la population de référence, l'enquêteur demande à chaque individu échantillon de tirer (au hasard), un carton dans une urne : cette urne contient une proportion p de cartons portant la mention, disons "je suis toxicomane", et une proportion $(1-p)$ de cartons portant la mention "je ne suis pas toxicomane". Ayant tiré un carton, l'enquêté n'en dévoile pas l'inscription à l'enquêteur, et répond "vrai" ou "faux" selon que l'inscription correspond ou non à son cas ; le carton est ensuite remis dans l'urne (dont la composition reste ainsi inchangée) et l'enquêteur enregistre la réponse, sans avoir donc eu connaissance de la question posée. Le décompte des "vrai" et des "faux" permet (méthode du maximum de vraisemblance) d'estimer la proportion de toxicomanes dans la population.

Pour séduisante qu'elle soit sur le plan théorique cette méthode pose à l'évidence le problème de la confiance qu'y met l'enquêté : il doit d'abord accepter le principe du jeu probabiliste qui lui est proposé, puis, ayant accepté ce principe, répondre en toute honnêteté ; remarquons à ce sujet qu'aucune des deux réponses "vrai" ou "faux" n'est plus révélatrice que l'autre.

A notre connaissance, cette méthode n'a pas été expérimentée en France, en particulier pas par l'INSEE. Des expériences menées aux Etats-Unis ont cependant montré qu'elle conduisait à des résultats intéressants, mettant en particulier en évidence les sous-estimations consécutives à des méthodes plus traditionnelles.

*
* *

Ce qui précède met l'accent sur certaines difficultés propres à l'étude des sous-populations. On connaît généralement des solutions aux problèmes rencontrés, mais ces solutions achoppent souvent sur des considérations de coût et sur des difficultés de mise en oeuvre.

On a cependant exhibé un cas de figure pour lequel il ne semble pas y avoir de solution systématique : il s'agit des sous-populations définies à partir d'une pratique ou d'un comportement, que, au mieux, on ne saura qu'approcher.

Nous avons par ailleurs délibérément omis d'évoquer l'estimation pour des sous-populations, car les modalités de cette estimation supposent résolues les difficultés d'observation ce qui, nous l'avons vu, ne doit généralement pas être tenu pour acquis.

BIBLIOGRAPHIE

BAUDELLOT (Ch.) : "Bas salaires : état transitoire ou permanent" ? Economie et statistiques - n° 131 - mars 1981

BLANC (R) et PACCOU (Y) : "Le recensement des nomades Mauritanien". Population 2 - mars-avril 1979

Groupe de Travail ENSAE "Les loisirs des Sourd-Muets". Disponible auprès de G. ROY (INSEE Paris)

PURCELL (N.J.) and KISH (L) : "Post censal Estimates for Local Areas (or Domains)" International Statistical Review - 48 (1980) 3-18

"Randomized Response Model" International Statistical Review - 44 (1976) : le numéro est entièrement consacré au "Randomized Response Model".

TRACY (P.E.) and FOX (J.A.) : "The validity of randomized response for sensitive measurements" American Sociological Review 1981. Vol. 46 (pp. 187-200)

SCPR (Social and Community Planning Research). London Spring 81 Newsletter : "Sampling Minority Populations".