UN ESSAI DE MODÉLISATION SIMPLE DU RECOURS À DES ÉCHANTILLONS DE RÉSERVE

Marc CHRISTINE

Insee, Direction de la méthodologie et de la coordination statistique et internationale

marc.christine@insee.fr

Mots-clés : échantillonnage, non-réponse, biais, taux de réponse, deux phases

Résumé

Dans la plupart des enquêtes de la statistique publique, la non-réponse constitue un fléau dont la prévalence s'amplifie au cours du temps. Ceci est problématique pour deux raisons principales : d'une part, cela diminue la taille de l'échantillon utile, donc cela détériore la précision, et, d'autre part et surtout, cela crée des biais car la non-réponse est potentiellement corrélée aux variables d'intérêt de l'enquête.

Pour atténuer les difficultés liées à ce type de situation, les instituts statistiques - dont l'Insee - développent de plus en plus des stratégies de recours à des ***échantillons de réserve*** : schématiquement, il s'agit de prévoir d'augmenter l'échantillon courant (« principal ») au moyen d'un échantillon additionnel (« de réserve »). Cet échantillon additionnel est mobilisé si le responsable d'enquête estime que les taux de réponse à l'enquête sur l'échantillon principal sont « insuffisants ». Bien entendu, cela consiste à fixer un seuil de taux de réponse admissible, souvent arbitraire ; dans les meilleurs cas, le but est d'augmenter la taille de l'échantillon pour compenser la perte due à la non-réponse, dans le but de diminuer la variance et remédier ainsi à la première difficulté mentionnée dans l'introduction (mais en omettant le traitement du biais).

Sur le plan pratique, le responsable d'enquête doit aussi définir les conditions de déclenchement de ces échantillons : par fractions ou en bloc, à partir de quels seuils, strate par strate séparément … La collecte de l'enquête sur l'échantillon de réserve se fait en général postérieurement à celle relative à l'échantillon principal ou, en tout cas, elle commence en général de manière décalée dans le temps. Mais, bien entendu, les unités collectées sont ensuite indiscernables du point de vue de l'appartenance à l'échantillon initial ou à celui de réserve, et traitées identiquement dans le processus d'estimation des paramètres d'intérêt (avec les probabilités d'inclusion *ad hoc*). Il faut mentionner aussi un écueil à éviter, notamment lorsque le producteur d'enquête a recours à un prestataire privé pour réaliser l'enquête sur le terrain : c'est celui de s'arrêter dès lors que les taux de réponse ont atteint une valeur jugée suffisante après mobilisation de l'échantillon de réserve. Une telle pratique n'est évidemment pas pertinente compte tenu des biais de sélection qui en résulteraient.

Le papier s'attachera à un examen de la question sous deux angles.

Dans un premier temps, seront rappelées les modalités usuelles de constitution d'échantillons de réserve, en soulignant les difficultés sous-jacentes par rapport aux contraintes que se donne le statisticien : taille fixe, conditions d'équilibrage, probabilités d'inclusion données... D'une manière générale, il apparaît que, si la constitution d'échantillons de réserve n'a pas été anticipée, la sélection *ex-post* d'un nouvel échantillon venant s'adjoindre à l'échantillon principal peut s'avérer problématique vis-à-vis de certaines de ces contraintes. Dans ce cadre, c'est la théorie des échantillons successifs conditionnels qui entre en jeu.

Si, au contraire, on anticipe bien l'éventualité de recourir à un tel échantillon, sur la base d'observations des taux de réponse à des éditions antérieures de l'enquête ou sur des enquêtes similaires, la plupart des problèmes de constitution disparaissent. Dans ce cas, la solution la plus simple consiste à réaliser un échantillon en deux phases, la seconde phase constituant l'échantillon principal et la réserve étant le complément par rapport à la 1ère phase : ainsi, en cas de mobilisation de l'échantillon de réserve, c'est l'intégralité de l'échantillon de 1ère phase qui est utilisée.

Dans un second temps, on essaie de modéliser, d'un point de vue statistique, l'effet du recours à un échantillon de réserve. La plupart du temps, en effet, on « oublie » le processus de recours à la réserve et les estimateurs utilisés ne tiennent pas compte de la stratégie de mobilisation de l'échantillon final en deux temps. Ici, on essaiera donc de mettre en œuvre des formules générales traduisant la constitution de chacun des deux échantillons tirages (principal + réserve) et la décision de recourir ou non à l'échantillon de réserve.

Cette modélisation sera limitée à des cas simples, qui ne prétendent pas épuiser toute la complexité du problème, ni toutes les subtilités de leur mise en œuvre.

L'objectif final est de pouvoir répondre à des questions telles que : vaut-il mieux mettre en œuvre une stratégie en deux temps avec une étape consistant à décider du recours à la réserve ou bien, vaut-il mieux décider d'emblée d'utiliser un plus gros échantillon, quitte à ce que le taux de réponse soit en définitive plus élevé qu'anticipé ?

Outre les questions de précision, peut intervenir un paramètre de coût puisque le budget d'une enquête dépend fortement - lorsque l'enquête se fait par voie postale et, *a fortiori*, en face à face - de la taille de l'échantillon, avec, la plupart des temps, des différentiels de rémunération selon que l'enquête est effectivement réalisée ou non.