Méthodes de sondage – SOL3017 et SOL 6448

Notes de cours

Stratégies d'échantillonnage : deux exemples

Département de sociologie
Université de Montréal

Professeur : Claire Durand

Exemples de stratégie d'échantillonnage

Exemple 1.

Supposons que je veux faire un échantillon de directeurs et directrices d'écoles du Québec en vue de faire une enquête auprès d'eux/elles mais je n'ai pas de liste des directions d'école, seulement une liste des écoles.

La première étape consiste à identifier l'information dont j'ai besoin. Combien y a t-il d'écoles au Québec. Comment l'enseignement est-il organisé : privé vs public, Centre de services scolaires, primaire vs secondaire vs éducation des adultes, nombre d'élèves par école, etc. Normalement, je devrais pouvoir trouver l'ensemble de ces informations sur le site du ministère de l'éducation.

Une fois l'information recueillie, je dois tenter de voir quel type d'échantillon convient le mieux à ma problématique et quels sont les avantages et inconvénients de chaque type d'échantillon. A titre d'exemple.

Echantillon aléatoire à probabilités égales:

Dans ce cas, je tire un échantillon d'écoles, sans me soucier de la taille des écoles, de leur type (privé, public, primaire, secondaire, etc). **Je m'intéresse aux directeurs/directrices d'école** et tous et toutes, quel que soit le type d'école dont ils ou elles sont responsables sont équivalents.

Pour procéder, je dois me rappeler que je peux tirer un ***échantillon aléatoire simple...***

- je peux prendre une table de nombres aléatoires (que l'on trouve dans tout bon livre de statistique),
- je peux imprimer la liste de noms, découper la liste en un papier par école, mettre le tout dans un chapeau et piger jusqu'à atteindre le nombre pré-défini (voir estimation de la taille requise)
- je peux, plus simplement, récupérer la liste de noms dans une base de données (SPSS, Excel,...) et utiliser une procédure d'échantillonnage (SAMPLE dans SPSS).

ou aléatoire systématique:

- J'imprime la liste des écoles et je m'assure qu'il n'y a pas un ordre dans la liste susceptible de biaiser le hasard. Je tire au hasard un premier chiffre compris entre 1 et la fraction de sélection calculée, ce qui me permet de choisir la première école sélectionnée dans la liste. Je continue ensuite en suivant la liste et en choisissant les écoles en fonction de l'intervalle de sélection. Concrètement, si je dois choisir une école sur dix, je prends au hasard aléatoire un nombre compris entre un et dix inclusivement. S'il s'agit de 6 par exemple, je garderai dans l'échantillon la sixième école, puis la seizième, la vingt-sixième, etc. Cette procédure est très pratique et appropriée quand il faut procéder manuellement.

- Je peux encore là utiliser une base de données. Dans ce cas, il est habituellement possible de faire "mêler" la liste de noms pour s'assurer qu'elle n'ait pas d'ordre susceptible de biaiser l'échantillon. Ensuite, on peut sélectionner via des fonctions de sélection du type (MODULO (ID,10) égale 6) où ID est le numéro séquentiel du cas.

J'obtiendrai alors un échantillon d'écoles que je pourrai contacter pour obtenir le nom et les coordonnées de la direction. Je peux vérifier, à partir des données que j'ai déjà recueillies sur la "population" des écoles, que l'échantillon reflète assez bien cette population, sans trop de distorsions : Les répartitions public-privé, primaire-secondaire, selon les régions administratives, sont-elles similaires dans la population d'origine et dans l'échantillon? Ceci me permet de m'assurer que le hasard n'a pas fait mal les choses, ce qui arrive parfois (au moins une chance sur 20).

Ceci ne m'assure pas d'une bonne représentation des caractéristiques des directions -- qui pourraient être influencées par d'autres paramètres que je ne peux pas contrôler -- mais je peux affirmer que j'ai tenté dans la mesure du possible de m'assurer de contrôler tout ce qu'il est possible de contrôler. Les caractéristiques des directions dans l'échantillon risquent donc de s'approcher fortement de celles de la population. Dans cet exercice, je ne connais pas les paramètres de la population des directions d'écoles et je tente de les estimer au mieux, c'est-à-dire avec une marge d'erreur acceptable, à partir de l'échantillon.

L'inconvénient de cet échantillon est que les directions de petites écoles auront autant de chances d'être dans dans l'échantillon que les directions d'écoles de grande taille. Si je suis intéressée par les directions d'écoles de grandes tailles, je n'en aurai peut-être pas assez, d'où la stratégie suivante.

Echantillon aléatoire à probabilités inégales

Dans le cas de ce type d'échantillon, je voudrais que l'échantillon des directions reflète l'importance de leurs responsabilités. En d'autres termes, plutôt que de donner la même importance à une direction d'école primaire de village et à une direction d'école polyvalente du niveau secondaire, je veux que l'échantillon de directions que je vais constituer reflète l'importance de la population scolaire dont elles ont la responsabilité.

Les données sur la population des écoles étant disponibles, l'idée est de constituer l'échantillon en accordant aux directions une probabilité d'être choisies proportionnelle à l'importance de leurs responsabilités. Il s'agit donc de constituer la liste de telle manière que, par exemple, la direction responsable d'une école de 100 élèves voit son nom apparaître une seule fois alors que la direction d'une école de 600 élèves voit son nom apparaître six fois et ait donc six fois plus de chances d'être sélectionnée dans l'échantillon. Avec un tel échantillon, les directions d'écoles plus importantes seront représentés dans l'échantillon en fonction de l'importance de leurs clientèles. Comment procéder?

- Il faut d'abord que je constitue une liste "pondérée" en fonction du nombre d'élèves
  - Pour cela, si je veux procéder par tirage dans un chapeau (de type loto), je dois doubler, tripler, etc. les noms en fonction du nombre d'élèves par école
  - Si j'ai récupéré la liste dans une base de données, je peux utiliser une procédure de pondération. Dans SPSS, par exemple, la commande Weight (pondérer les observations dans le menu Données) permet de définir que la variable donnant le nombre d'élèves est utilisée pour pondérer.
- Ensuite, il suffit d'utiliser la même procédure (tirage au sort, SAMPLE, systématique) que pour l'échantillon aléatoire à probabilités égales.

L'avantage et l'inconvénient de cet échantillon réside dans ce qui est recherché, soit une surreprésentation des directions des écoles de grande taille.

Echantillon aréolaire, échantillon en grappes

Rappelons que les deux termes sont parfois utilisés l'un pour l'autre mais recouvrent des procédures légèrement différentes.

- Dans le cas de l'échantillon aréolaire, exemple que nous allons donner, il s'agit de sélectionner des "aires" et de procéder ensuite si nécessaire à une sélection à l'intérieur de ces aires.
- Dans le cas de l'échantillon en grappes, il s'agit de sélectionner des cas et de constituer les grappes en prenant un certain nombre de cas voisins des cas sélectionnés.

Supposons que l'enquête auprès des directions d'école me demande d'aller les rencontrer en face à face, soit pour conduire l'entrevue proprement dite, soit pour effectuer des observations sur le milieu de l'école, ou pour d'autres raisons. Ce type de situation amène fréquemment à utiliser des échantillons aréolaires de façon à éviter les coûts associés à des déplacements importants. On pourrait ainsi, dans notre cas, présélectionner des Centres de services scolaires (CSS), des régions ou sous-régions administratives, etc.

Supposons que je décide d'y aller par Centre de service scolaire, en incluant les écoles privées qui sont sur le territoire de ces Centres. En recueillant les informations, je m'aperçois qu'il y a 72 CSS, de taille toutefois très variable. La CSS de Montréal (CSSDM), par exemple, regroupait environ sept pour cent de la clientèle scolaire des écoles publiques du Québec en 2021. De plus, ses écoles sont proportionnellement plus grosses que celles des autres CSS. Il s'agit d'un problème important lorsque l'on veut constituer un échantillon aréolaire et que les unités de la population ne sont pas de même taille. Imaginons ce que se passerait si, en sélectionnant au hasard, la CSSDM ne se trouve pas incluse dans mon échantillon : sept pour cent de la population scolaire, des situations d'écoles, spécifiques à une partie importante de l'île de Montréal, se trouvent exclues de mon échantillon. Celui-ci se retrouve irrémédiablement biaisé. Par ailleurs, si la CSSDM est sélectionnée et que j'ai décidé de prendre toutes les écoles des CSS sélectionnées, je biaise également l'échantillon en surreprésentant la région de Montréal et en constituant un échantillon comprenant trop d'unités vivant des situations similaires non

indépendantes (puisqu'elles sont dans la même CSS et sur un territoire ayant nombre de caractéristiques communes). Ceci illustre les dangers de l'utilisation irréfléchie d'un échantillon aréolaire.

Ayant réfléchi à l'ensemble des problèmes auxquels je risque de me heurter, et comme la première raison pour laquelle j'ai opté pour un échantillon aréolaire dans ce cas est la nécessité de réduire les coûts de transport, j'opte pour un échantillon mixte: Sur l'île de Montréal, je constituerai un échantillon au hasard à probabilités égales parmi la liste de toutes les écoles des cinq CSS (3 francophones, 2 anglophones). Hors de l'île de Montréal, je constituerai un échantillon de CSS au hasard à probabilités inégales, en fonction de la taille des CSS. Ensuite, dans chacune des CSS sélectionnées, je prends une fraction des écoles au hasard aléatoire.

Ce type d'échantillon me permettra normalement de m'assurer de représenter l'ensemble du Québec et la variété des situations de la façon la plus efficiente possible, en réduisant les coûts associés à la nécessité de visiter des écoles très éloignées les unes des autres.

Un des avantages de cet échantillon demeure la possibilité de recueillir des informations de deuxième niveau sur les unités -- les caractéristiques de la CSS d'appartenance par exemple ou celles des populations desservies par celle-ci -- ce qui pourrait permettre d'estimer si certaines relations ou certains comportements ou directives des directions d'école seraient dues à des politiques présentes au niveau des CSS, plutôt qu'au niveau des écoles elles-mêmes.

Ce faisant, j'ai stratifié mon échantillon en deux strates (Ile de Montréal, reste du Québec) en utilisant une stratégie d'échantillonnage et une fraction de sélection différentes dans chaque strate. Je pourrais stratifier davantage et différemment (voir suite).

Echantillon stratifié

Revenons au problème mentionné dans le premier type d'échantillon, à savoir la représentation en fonction de la taille des écoles. Il pourrait arriver que, étant donné le sujet de mon enquête, j'accorde une forte importance à ce qu'il y ait une bonne représentation des écoles selon leur taille ou selon, par exemple, qu'elles sont situées en milieu rural ou urbain. Je pourrais ainsi distinguer les écoles se trouvant dans des grands centres urbains -- Montréal, Québec, Gatineau -- ou hors de ces centres. J'ai donc deux grandes strates pour lesquelles je pourrai décider d'appliquer des fractions de sélection et même des modes de sélection différents de façon par exemple, à obtenir la même marge d'erreur pour les deux strates.

Exemple:

Données 2021 du ministère de l'Éducation du Québec:

Strate	**Nb écoles publiques**		
	Total	**primaires**	**secondaires**
Strate 1 Régions urbaines - Total	**502**	**398**	**103**
Montréal	409	324	85
Québec	59	47	12
Gatineau	33	27	6
Strate 2 Reste du Québec -Total	**1672**	**1391**	**281**
Total Québec	**2174**	**1789**	**384**

Ainsi dans le cas précédent, pour tirer un échantillon d'écoles publiques égal dans les trois centres urbains et dans le reste du Québec, on pourrait décider de prendre une école sur 10 dans la strate 2 (reste du Québec), ce qui donnerait 167 écoles, et de prendre une école sur 5 dans Gatineau (6 écoles), une sur 8 à Québec (7 écoles) et une sur 6 à Montréal (68 écoles). On aurait donc un échantillon final comprenant 81 écoles dans les trois centres urbains et 167 écoles dans le reste du Québec. *Une fois les données en main, on devrait les pondérer par l'inverse de la fraction de sélection lorsque l'on présenterait les données pour l'ensemble du Québec.*

Si on avait pris un échantillon proportionnel – une école sur 10 partout –, on aurait 167 écoles dans le reste du Québec (strate 2) mais 50 écoles seulement dans les trois régions urbaines (502 divisé par 10).

Exemple 2.

J'ai d'abord donné la définition suivante de la population qui m'intéresse:

Population : *Ensemble des élèves âgés de 13 ans à 16 ans de Montréal*

- Où les trouver de façon à ce que toute personne répondant à ce critère ait une chance connue et égale aux autres de faire partie de l'échantillon?

Réponse "évidente": Écoles secondaires publiques et privées de l'Île de Montréal

Problèmes:

- Est-ce que je veux inclure les élèves vivant à Montréal (même s'ils étudient ailleurs). Dans ce cas, ces personnes ne seraient pas dans l'échantillon puisque je prends celui-ci dans les écoles situées sur l'Ile de Montréal; il y aura donc un biais →réviser la définition de la population?
- Est-ce que je veux inclure les élèves qui étudient sur l'île de Montréal même s'ils vivent hors de l'île de Montréal? Si je considère qu'ils ne font pas partie de la population, je devrai filtrer pour les éliminer de mon échantillon.

Donc: "*Les élèves de Montréal*": ce n'est pas assez précis. → révision:
Population : *L'ensemble des personnes âgées de 13 à 16 ans étudiant dans une école publique ou privée située sur l'île de Montréal et résidant à Montréal.*

Problème:

Est-ce que j'inclus les personnes en situation de handicap qui sont dans des écoles spéciales ou dans les écoles régulières mais dans des classes spécifiques? Si non, ...

> *Les personnes de 13 à 16 ans fréquentant le cursus régulier d'une école secondaire publique ou privée...*

Problème:

De 13 à 16 ans: S'ils ont 12 ans et qu'ils sont en Secondaire 1, pourquoi est-ce que je les exclurait? Et s'ils ont 16 ans et un mois et sont en Secondaire 5? En fait pourquoi est-ce que j'ai fixé ce critère d'âge? Et si je maintiens le critère d'âge, est-ce inclusif – 13, 14, 15 et 16 ans – ou exclusif --13, 14, 15? Nouvelle révision de la définition de la population:

> Supposons que la population est définie comme *l'ensemble des personnes résidant à Montréal et étudiant au secteur régulier dans une école secondaire privée ou publique située sur l'île de Montréal.*

Je n'exclurai que les élèves qui étudient sur l'île de Montréal mais ne résident pas à Montréal.

Maintenant, regardons la base de sondage, c'est-à-dire la liste des unités à partir de laquelle on peut tirer l'échantillon:

- Liste de tous les élèves inscrits dans les écoles....

Problème 1: Les décrocheurs encore inscrits doivent-ils être conservés dans l'échantillon?

- Normalement, si ma population est celle des personnes **étudiant**, je vais les considérer comme non-éligibles.

Problème 2: Où trouver la liste de tous les élèves inscrits ? Cette liste est disponible uniquement dans les CSS et les écoles privées *MAIS la loi de protection des renseignements personnels interdit aux CSS et aux écoles de donner accès à cette liste*. Elles ne peuvent l'utiliser que pour leur propres fins.

→Stratégie d'échantillonnage:

Stratégie 1: Échantillonnage des élèves, de type aléatoire simple ou systématique

Pour utiliser cette stratégie, il faudrait que j'obtienne la liste de tous les élèves inscrits dans les écoles secondaires privées et publiques de l'île de Montréal.

- Je devrais d'abord **trouver la liste des CSS** de l'île (il y en a cinq) et la liste des écoles privées dispensant un enseignement secondaire. C'est une information publique que je peux obtenir de diverses sources, comme le Conseil scolaire de l'île de Montréal, le site Web du Ministère de l'éducation (MELS), l'Association des écoles privées.

- Il faudrait ensuite **obtenir la collaboration de chaque CSS et de chaque école privée**. Je pourrais tenter de convaincre le Conseil scolaire de l'île de Montréal et l'Association des écoles privées d'appuyer ma recherche et d'inviter leurs membres à y collaborer. (ATTENTION: tous ces organismes sont très sollicités. A la CSSDM, il faut passer par le service de recherche qui filtre les demandes).

- Le problème demeure toutefois. **Ils ne peuvent pas légalement fournir la liste** de leurs élèves (Loi de protection des renseignements personnels). Je ne pourrais donc utiliser un échantillon complètement aléatoire à moins que toutes les unités décident de collaborer avec moi et reprennent l'étude à leur compte.

Si seulement certaines unités acceptaient, je devrais me demander si elles sont représentatives de l'ensemble des élèves. La réponse scientifique à cette question est **non**. Les élèves de ces unités ont comme caractéristique spécifique de faire partie d'unités dont la direction est intéressée au projet de recherche. Si l'étude porte sur la perception que les élèves ont de la direction de

l'école, l'échantillon peut apparaître biaisé puisque les directions intéressées par le projet ont plus de chances d'être un type de direction qui est mieux perçu par les étudiants.

Stratégie 2 : Échantillon à plusieurs degrés.

Au premier degré, j'échantillonne les écoles à partir de la liste des écoles. →Chaque école a une probabilité connue d'être choisie. L'*unité d'échantillonnage*, c'est l'école.

Je contacte ces écoles sélectionnées et j'essaie d'avoir la collaboration de toutes les directions d'écoles. Je leur demande de me fournir la liste des classes de français obligatoire de leur école et le nombre d'élèves par classe.

Au deuxième degré, j'échantillonne des classes; chaque classe -- *unité d'échantillonnage* de deuxième niveau -- a une probabilité connue d'être choisie.

Au troisième degré, je peux décider d'utiliser un échantillon de type aréolaire et prendre tous les élèves des classes sélectionnées. Je pourrais aussi décider de sélectionner des élèves à l'intérieur de chaque classe de façon aléatoire simple ou systématique.

Note : Ce type d'échantillon a des désavantages dont celui d'entraîner possiblement une plus grande homogénéité de l'échantillon. Il peut toutefois avoir des avantages énormes, particulièrement en sociologie, celui de permettre de recueillir des données sur les caractéristiques collectives des unités de premier degré – niveau socio-économique du quartier où est situé l'école; proportion d'élèves d'origine ethnique "autre", etc -- et du deuxième degré – caractéristiques de l'enseignant, niveau de la classe, etc.

Stratégie 3: Échantillon à plusieurs phases.

J'échantillonne un nombre important d'écoles. J'adresse un questionnaire à la direction d'école demandant des informations quant à certaines caractéristiques. J'échantillonne par la suite de nouveau les écoles mais sur la base des informations que j'ai recueillies sur les caractéristiques.

Exemple: Je veux échantillonner les écoles selon le niveau d'activité du comité d'école. J'échantillonne d'abord 1,000 écoles; je leur expédie un questionnaire demandant un certain nombre de questions me permettant de déterminer le niveau d'activités du comité d'école (fréquence des réunions, présence aux réunions, nombre d'activités organisées, etc). A partir de ces indicateurs, je détermine un indice d'activité. Les résultats montrent qu'il y a 300 écoles dont le comité est considéré comme très actif, 150 dont le comité est moyennement actif et 150 dont le comité est peu actif. Je décide d'échantillonner 50 écoles dont le comité d'école est très actif (soit une sur 6), 50 écoles dont le comité est moyennement actif (soit une sur 3) et 50 écoles donc le comité est peu actif (soit une sur 3). ***Je fais alors un échantillon stratifié, c'est à dire un échantillon où la probabilité d'être choisi est déterminée par strate (niveau d'activité dans***

l'exemple présenté) et peut différer selon les strates. A l'intérieur des écoles sélectionnées, je déciderai par la suite d'une sélection qui peut comprendre diverses populations: enseignants, membres du comité d'école, parents d'élèves, élèves...).

La méthodologie, c'est l'art de poser toutes les questions.

Le choix de la stratégie d'échantillonnage dépend de divers critères:

- Efficacité de la stratégie : recueillir le maximum d'informations au moindre coût en temps et en argent.

- Disponibilité des informations : A-t-on accès à une liste d'adresses ou de numéros de téléphones, le cas échéant?.

- Mode de cueillette :
    - Par la poste → liste d'adresses, hasard aléatoire ou systématique
    - Par téléphone → numéros de téléphone ou génération aléatoire de numéros de téléphone (RDD), hasard aléatoire ou systématique.
    - Entrevues sur place → sélection d'ilôts, d'écoles, de Collèges, échantillons aréolaires ou en grappes.

L'IMAGINATION doit être au rendez-vous : Une stratégie originale permettant de combiner divers modes d'enquêtes ou d'échantillonnage lorsque nécessaire permettra d'arriver à des résultats plus intéressants, permettant une meilleure analyse de la situation.