

ESSAIS



Outil méthodologique 22
Méthodes d'échantillonnage probabiliste

TÉLUQ

OUIL MÉTHODOLOGIQUE 22 : *Méthodes d'échantillonnage probabiliste*¹

Objectif : Distinguer les différentes méthodes d'échantillonnage probabiliste.

L'échantillonnage probabiliste entraîne la sélection d'un échantillon à partir d'une population, sélection qui repose sur le principe de la randomisation (la sélection au hasard ou aléatoire) ou la chance. Il est plus complexe, prend plus de temps et est habituellement plus coûteux que l'échantillonnage non probabiliste. Toutefois, comme les unités de la population sont sélectionnées au hasard et qu'il est possible de calculer la probabilité d'inclusion de chaque unité dans l'échantillon, on peut, grâce à l'échantillonnage probabiliste, produire des estimations fiables, de même que des estimations de l'erreur d'échantillonnage et faire des inférences au sujet de la population.

Il existe plusieurs méthodes différentes permettant de sélectionner un échantillon probabiliste. La méthode qu'on choisira dépendra d'un certain nombre de facteurs, comme la base de sondage dont on disposera, la façon dont la population sera distribuée, ce que sonder les membres de la population coûtera et la façon également dont les utilisateurs analyseront les données. Lorsque vous choisirez un plan d'échantillonnage probabiliste, votre but devrait consister à réduire le plus possible l'erreur d'échantillonnage des estimations pour les variables d'enquête les plus importantes, tout en réduisant le plus possible également le délai et le coût de réalisation de l'enquête.

Voici les méthodes d'échantillonnage probabiliste les plus courantes :

- l'échantillonnage aléatoire simple;
- l'échantillonnage systématique;
- l'échantillonnage avec probabilité proportionnelle à la taille;
- l'échantillonnage stratifié;
- l'échantillonnage en grappes;
- l'échantillonnage à plusieurs degrés;
- l'échantillonnage à plusieurs phases.

Échantillonnage aléatoire simple

Dans un *échantillonnage aléatoire simple* (EAS), chaque membre d'une population a une chance égale d'être inclus à l'intérieur de l'échantillon. Chaque combinaison de membres de la population a aussi une chance égale de composer l'échantillon. Ces deux propriétés sont ce qui définit un échantillonnage aléatoire simple. Vous devez dresser une liste de toutes les unités incluses dans la population observée pour sélectionner un échantillon aléatoire simple.

Exemple n° 1 : Il faudrait numéroter dans un ordre séquentiel chaque entrée ou inscription pour prélever un échantillon aléatoire simple d'un annuaire téléphonique. S'il y avait 10 000 entrées dans l'annuaire téléphonique et si la taille de l'échantillon était 2 000 numéros, un ordinateur devrait alors générer au hasard 2 000 numéros entre 1 et 10 000. Chaque numéro aurait la même chance qu'un autre d'être généré par l'ordinateur (ce qui respecterait l'exigence de l'échantillonnage aléatoire simple : une chance égale pour chaque unité). Les 2 000 entrées dans l'annuaire téléphonique correspondant aux 2 000 numéros aléatoires générés par l'ordinateur composeraient l'échantillon.

Un échantillonnage aléatoire simple peut s'effectuer avec ou sans remplacement. Un échantillon avec remplacement signifie qu'il est possible que l'entrée dans l'annuaire téléphonique échantillonnée soit sélectionnée

1. Source : Statistique Canada.

deux fois ou plus. Habituellement, l'échantillonnage aléatoire simple est effectué sans remplacement, parce qu'il est plus pratique et donne des résultats plus précis. Nous ferons référence à l'échantillonnage sans remplacement lorsque nous traiterons de l'échantillonnage aléatoire simple aux fins des présentes descriptions.

L'échantillonnage aléatoire simple est la méthode d'échantillonnage la plus facile à appliquer et la plus couramment utilisée. L'avantage de cette technique tient au fait qu'elle n'exige pas de données additionnelles dans la base de sondage (comme des régions géographiques) autres que la liste complète des membres de la population observée et l'information pour les contacter. Également, puisque l'échantillonnage aléatoire simple est une méthode simple et que la théorie qui la sous-tend est bien établie, il existe des formules-types pour déterminer la taille de l'échantillon, les estimations, etc., et ces formules sont faciles à utiliser.

D'un autre côté, cette technique ne fait aucunement appel aux données auxiliaires contenues dans la base de sondage (p. ex., le nombre d'employés de chaque entreprise) qui pourraient rendre le plan d'échantillonnage plus efficace. En outre, même s'il est facile d'appliquer l'échantillonnage aléatoire simple à de petites populations, le faire peut être coûteux et irréalisable pour de grandes populations, parce qu'il faut en identifier et en étiqueter toutes les unités avant l'échantillonnage. Son application peut également être coûteuse s'il faut effectuer des interviews sur place, puisqu'il est possible que l'échantillon soit géographiquement distribué dans toute la population.

Un tirage à la loterie est un bon exemple d'échantillonnage aléatoire simple. Par exemple, lorsqu'un échantillon de six numéros est généré au hasard à partir d'une population de 49 numéros, chacun de ces derniers a une chance égale d'être sélectionné et chaque combinaison de six numéros a la même chance d'être la combinaison gagnante. Même si les gens tendent à éviter une combinaison comme 1-2-3-4-5-6, cette combinaison a la même chance d'être la série gagnante de numéros que la combinaison 8-15-21-28-32-40.

Exemple n° 2 : Supposez que votre école compte 500 élèves et que vous devez mener une courte enquête sur la qualité des aliments servis à sa cafétéria. Vous déterminez qu'un échantillon de 10 élèves devrait suffire à vos fins. Pour obtenir votre échantillon, vous attribuez à chaque élève de votre école un numéro compris entre 1 et 500. Pour sélectionner cet échantillon, vous utilisez une table de numéros générés au hasard. Tout ce que vous avez à faire consiste à prendre un point de départ à l'intérieur de la table (un numéro de rangée et un numéro de colonne) et à examiner les numéros aléatoires qui y figurent. Dans ce cas, puisque les données ne dépassent pas trois chiffres, les numéros aléatoires devraient renfermer trois chiffres également. Ne tenez pas compte des numéros aléatoires supérieurs à 500, parce qu'ils ne correspondent à aucun des élèves de votre école. Rappelez-vous que votre échantillon est un échantillon sans remplacement et que, si un numéro se répète, vous devez le sauter et utiliser le numéro aléatoire suivant. Les 10 premiers numéros différents entre 001 et 500 composent votre échantillon.

Exemple n° 3 : Imaginez que vous êtes propriétaire d'un cinéma et que vous y organiserez un festival de films d'horreur le mois prochain. Pour déterminer quels films d'horreur vous y présenterez, vous voulez demander à des cinéphiles lesquels parmi les films que vous leur énumérerez ils préfèrent. Pour dresser la liste des films nécessaire à votre sondage, vous décidez d'échantillonner 100 des 1 000 meilleurs films d'horreur de tous les temps. La population des films d'horreur se divise en proportions égales entre les films classiques (tournés en ou avant 1969) et les films modernes (produits en ou après 1970). L'une des façons d'obtenir un échantillon consisterait à écrire tous les titres des films sur des bouts de papier et à les placer dans une boîte et à tirer ensuite 100 titres et vous auriez alors votre échantillon. En utilisant cette méthode, vous auriez l'assurance que chaque film courrait une chance égale d'être sélectionné.

Vous pourriez aussi calculer la probabilité de sélection de tel ou tel film. Puisque nous connaissons la taille de l'échantillon (**n**) et la population totale (**N**), calculer la probabilité pour un film d'horreur d'être inclus dans votre échantillon deviendrait une simple question de division :

Probabilité de sélection (la même pour chaque film)

$$\begin{aligned}
 &= (n \div N) \times 100 \% \\
 &= (100 \div 1\,000) \times 100 \% \\
 &= 10 \%
 \end{aligned}$$

Cela signifie que chaque titre de film inscrit sur votre liste aurait 10 % de chances ou 1 chance sur 10 d'être sélectionné.

Vous pouvez constater que l'un des inconvénients de l'échantillonnage aléatoire simple (parce que ce n'est pas le seul, mais c'en est un important) est le fait que vous savez que la population se compose de 500 films classiques et de 500 films modernes et que vous connaissez la date de parution sur les écrans de chaque film à partir de la base de sondage, mais qu'on n'utilise aucunement cette information. L'échantillon en question pourrait renfermer 77 films classiques et 23 films modernes, qui ne seraient pas représentatifs de la population entière des films d'horreur.

Il existe des moyens de surmonter ce problème de non-représentativité (moyens dont nous traiterons brièvement dans la section Estimation). Cependant, il y a aussi des moyens de tenir compte de cette information (moyens dont nous traiterons aussi plus loin, dans la section Échantillonnage stratifié).

Échantillonnage systématique

Parfois appelé *échantillonnage par intervalles*, l'*échantillonnage systématique* (SYS) signifie qu'il existe un écart, ou un intervalle, entre chaque unité sélectionnée qui est incluse dans l'échantillon. Vous devez suivre les étapes énumérées ci-dessous pour sélectionner un échantillon systématique :

1. Numéroté de 1 à **N** les unités incluses dans votre base de sondage (où **N** est la taille de la population totale).
2. Déterminer l'intervalle d'échantillonnage (**K**) en divisant le nombre d'unités incluses dans la population par la taille de l'échantillon que vous désirez obtenir. Par exemple, pour sélectionner un échantillon de 100 unités à partir d'une population de 400, vous auriez besoin d'un intervalle d'échantillonnage de $400 \div 100 = 4$. $K = 4$, par conséquent. Vous devrez sélectionner une unité sur 4 pour avoir finalement au total 100 unités à l'intérieur de votre échantillon.
3. Sélectionner au hasard un nombre entre 1 et **K**. Ce nombre s'appelle l'origine choisie au hasard et serait le premier nombre inclus dans votre échantillon. À l'aide de l'échantillon fourni ci-dessus, vous sélectionneriez un chiffre entre 1 et 4 à partir d'une table de nombres aléatoires (pris au hasard). Si vous choisissiez 3, la troisième unité incluse dans votre base de sondage serait la première unité comprise dans votre échantillon; si vous choisissiez 2, le début de votre échantillon serait la deuxième unité incluse dans votre base de sondage.
4. Sélectionner chaque **K^e** (dans ce cas, chaque 4^e) unité après ce premier nombre. L'échantillon pourrait, par exemple, se composer des unités suivantes de façon à constituer un échantillon de 100 : 3 (l'origine choisie au hasard), 7, 11, 15, 19... 395, 399 (jusqu'à **N**, qui est 400 dans ce cas).

Vous pouvez constater, à l'aide de l'exemple fourni ci-dessus, que dans le cas d'un échantillonnage systématique, seuls quatre échantillons possibles, qui correspondent aux quatre origines choisies au hasard également possibles, peuvent être sélectionnés :

- 1, 5, 9, 13... 393, 397
- 2, 6, 10, 14... 394, 398
- 3, 7, 11, 15... 395, 399
- 4, 8, 12, 16... 396, 400

Chaque membre de la population ne fait partie que de l'un des quatre échantillons et chaque échantillon a une chance égale d'être sélectionné. Cela nous permet de constater que chaque unité a une chance sur quatre d'être sélectionnée à l'intérieur de l'échantillon. Sa probabilité d'être sélectionnée est la même que si l'on sélectionnait un échantillon aléatoire simple de 100 unités. La principale différence tient au fait que dans le cas d'un échantillonnage aléatoire simple, toute combinaison de 100 unités aurait une chance de constituer l'échantillon, tandis que dans celui d'un échantillonnage systématique, il n'y a que quatre échantillons possibles. Cela nous permet aussi de constater à quel point l'échantillonnage systématique est précis comparativement à l'échantillonnage aléatoire simple. L'ordre de la population incluse dans la base de sondage déterminera les échantillons possibles pour l'échantillonnage systématique. Si la population est distribuée au hasard dans la base de sondage, un échantillonnage systématique devrait alors produire des résultats similaires à ceux d'un échantillonnage aléatoire simple.

On utilise souvent cette méthode dans l'industrie, où l'on sélectionne une unité pour des essais à partir d'une chaîne de production afin de s'assurer que la machinerie et l'équipement sont d'une qualité uniforme. Un essayeur à l'intérieur d'une usine pourrait, par exemple, soumettre à un contrôle de la qualité chaque 20^e produit sur une ligne de montage. L'essayeur pourrait choisir une origine au hasard entre les nombres 1 et 20. Cela déterminerait le premier produit à essayer; chaque 20^e produit serait ensuite soumis à des essais.

Les intervieweurs peuvent utiliser cette technique d'échantillonnage lorsqu'ils interrogent des gens pour une enquête-échantillon. Le responsable d'une étude de marché pourrait sélectionner, par exemple, chaque 10^e personne qui entrerait dans un magasin, après avoir sélectionné au hasard la première personne. Un enquêteur peut interviewer les occupants de chaque 5^e maison d'une rue, après avoir sélectionné au hasard l'une des cinq premières maisons.

Exemple n° 4 : Imaginez que vous devez mener une enquête pour votre collège ou votre université sur les logements pour les étudiants. Dix mille (10 000) étudiants sont inscrits dans votre établissement d'enseignement et vous voulez en prélever un échantillon systématique de 500. Pour ce faire, vous devez premièrement déterminer ce que serait votre intervalle d'échantillonnage (**K**) :

Population totale ÷ taille de l'échantillon = intervalle d'échantillonnage

$$N \div n = K$$

$$= 10\,000 \div 500$$

$$= 20$$

Il faudrait attribuer un numéro séquentiel à chaque étudiant pour entreprendre cet échantillonnage systématique. On choisirait le point de départ en sélectionnant un numéro au hasard entre 1 et 20. Si ce numéro était 9, on sélectionnerait alors le 9^e étudiant inscrit sur la liste et chaque 20^e étudiant par la suite. L'échantillon d'étudiants serait constitué de ceux qui correspondraient aux numéros d'étudiant 9, 29, 49, 69... 9 929, 9 949, 9 969 et 9 989.

Dans les exemples utilisés jusqu'ici, l'intervalle d'échantillonnage **K** était toujours un nombre entier, mais ce n'est pas toujours le cas. Par exemple, si vous prélevez un échantillon de 30 unités d'une population qui en compte 740, votre intervalle d'échantillonnage (ou **K**) sera 24,7. Dans de tels cas, il existe quelques possibilités de faire en sorte que le nombre soit plus facile à utiliser. Vous pouvez arrondir le nombre, c'est-à-dire au nombre entier supérieur le plus rapproché (le plus près) ou au nombre entier inférieur le plus rapproché. Le fait de l'arrondir au nombre entier inférieur le plus rapproché vous assurera de sélectionner au moins le nombre d'unités que vous vouliez à l'origine (et vous pourrez ensuite supprimer certaines unités pour obtenir la taille exacte de l'échantillon de votre choix). Il existe des techniques pour adapter l'échantillonnage systématique dans les cas où **N** (la population totale) n'est pas un multiple de **n** (la taille de l'échantillon), mais

donne encore un échantillon qui est exactement le même que les unités n . Nous ne traiterons pas ici de ces techniques.

Les avantages de l'échantillonnage systématique tiennent au fait que la sélection de l'échantillon ne peut être plus facile (vous n'obtenez qu'un seul nombre aléatoire – l'origine choisie au hasard – et le reste de l'échantillon suit automatiquement) et que l'échantillon est distribué dans des proportions égales à l'intérieur de la population répertoriée. Le plus gros inconvénient de la méthode d'échantillonnage systématique tient au fait que les échantillons possibles risquent de ne pas être représentatifs de la population s'il existe un certain cycle sur le plan du mode d'ordonnement de la population inscrite sur une liste et si ce cycle coïncide d'une quelconque façon avec l'intervalle d'échantillonnage. C'est ce que l'on peut constater dans l'exemple qui suit :

Exemple n° 5 : Supposez que vous dirigez une épicerie de grande surface et que vous possédez une liste des employés de chacune de ses sections. L'épicerie est divisée entre les 10 sections suivantes : le comptoir de charcuterie, la boulangerie, les caisses, les stocks, le comptoir des viandes, les fruits et légumes, la pharmacie, le magasin de photographie, le magasin de fleurs et le nettoyage à sec. Chaque section compte 10 employés, y compris un gérant (ce qui fait 100 employés au total). Votre liste est ordonnée par section, le gérant y étant énuméré le premier et les autres employés y étant ensuite inscrits dans l'ordre décroissant d'ancienneté.

Si vous voulez sonder vos employés au sujet de leurs réflexions sur leur milieu de travail, vous pourriez choisir un petit échantillon pour répondre à vos questions. Si vous utilisiez un échantillonnage systématique et si votre intervalle d'échantillonnage était 10, vous pourriez alors ne sélectionner finalement que les gérants ou que les employés de chaque section ayant le moins d'ancienneté. Ce type d'échantillon ne vous donnerait pas un portrait complet ni approprié des réflexions de vos employés.

Échantillonnage avec probabilité proportionnelle à la taille

Pour l'échantillonnage probabiliste, il faut que chaque membre de la population observée ait une chance d'être inclus dans l'échantillon, mais il n'est pas nécessaire que cette chance soit la même pour tous. Si la base de sondage renferme de l'information sur la taille de chaque unité (comme le nombre d'employés de chacune des entreprises qui y sont inscrites) et si la taille de ces unités varie, on peut utiliser cette information dans le cadre de la sélection de l'échantillonnage afin d'en accroître l'efficacité. Cela s'appelle l'échantillonnage avec *probabilité proportionnelle à la taille* (PPT). Dans le cas de cette méthode, plus la taille de l'unité est grande, plus sa chance d'être incluse dans l'échantillon est élevée. Il faut que la mesure de la taille soit exacte pour que cette méthode accroisse l'efficacité. C'est une méthode d'échantillonnage plus complexe dont nous ne traiterons pas ici davantage.

Échantillonnage stratifié

Lorsqu'on utilise l'*échantillonnage stratifié*, on divise la population en groupes homogènes (appelés strates), qui sont mutuellement exclusifs, puis on sélectionne à partir de chaque strate des échantillons indépendants. On peut utiliser n'importe laquelle des méthodes d'échantillonnage mentionnées dans la présente section (et il en existe d'autres) pour sélectionner l'échantillon à l'intérieur de chaque strate. La méthode d'échantillonnage peut varier d'une strate à une autre. Lorsqu'on utilise l'échantillonnage aléatoire simple pour sélectionner l'échantillon à l'intérieur de chaque strate, on appelle le plan d'échantillonnage un plan d'échantillonnage aléatoire simple stratifié. On peut stratifier avant l'échantillonnage une population au moyen de toute variable dont on dispose pour la totalité des unités incluses dans la base de sondage (comme l'âge, le sexe, la province de résidence, le revenu, etc.).

Pourquoi doit-on créer des strates? Pour bien des raisons, la principale étant que leur création peut rendre la stratégie d'échantillonnage plus efficace. Nous avons mentionné précédemment que vous aviez besoin d'un échantillon plus grand pour obtenir une estimation plus exacte d'une caractéristique qui varie beaucoup d'une unité à l'autre. Si chaque personne incluse dans une population, par exemple, avait le même salaire, il suffirait alors d'un échantillon d'une seule unité pour obtenir une estimation précise du salaire moyen des membres de cette population.

C'est l'idée qui sous-tend le gain d'efficacité qu'on réalise grâce à la stratification. Si vous créez des strates à l'intérieur desquelles des unités auraient des caractéristiques qui seraient similaires (comme le revenu) et qui différeraient considérablement de celles d'unités incluses dans d'autres strates (comme la profession et le type de logement), vous n'auriez alors besoin que d'un petit échantillon tiré de chaque strate afin d'obtenir une estimation précise du revenu total pour la strate en question. Vous pourriez ensuite combiner ces estimations afin d'obtenir une estimation précise du revenu total de l'ensemble de la population. Si vous deviez utiliser un échantillonnage aléatoire simple de la population entière sans effectuer de stratification, il vous faudrait un échantillon plus grand que la totalité de tous les échantillons de strate afin d'obtenir pour le revenu total une estimation du même degré de précision.

L'échantillonnage stratifié nous assure d'obtenir une taille d'échantillon suffisante pour des sous-groupes de la population à laquelle nous nous intéressons. Étant donné que chaque strate devient une population indépendante lorsque vous stratifiez une population, vous devrez déterminer pour chaque strate la taille de l'échantillon.

Exemple n° 6 : Supposez que vous voulez estimer combien d'élèves des écoles secondaires ont un emploi à temps partiel, et ce, tant au niveau national qu'à celui de chaque province. Si vous deviez sélectionner un échantillon aléatoire simple de 25 000 personnes à partir d'une liste de tous les élèves des écoles secondaires du Canada (en supposant que vous disposiez d'une telle liste pour effectuer cette sélection), vous auriez finalement en moyenne un peu plus de 100 personnes de l'Île-du-Prince-Édouard, puisque cette province représente moins de la moitié de 1 % de toute la population canadienne. Cet échantillon ne serait probablement pas assez important pour le genre d'analyse détaillée auquel vous songeriez. Le fait de stratifier votre liste par province, en supposant encore une fois que vous disposeriez de cette information, puis de sélectionner une taille d'échantillon pour chacune des provinces vous permettrait de déterminer la taille d'échantillon exacte qu'il vous faudrait pour l'Île-du-Prince-Édouard. Ainsi, afin d'obtenir une bonne représentation de l'Île-du-Prince-Édouard, vous utiliseriez un échantillon plus important que celui que la méthode d'échantillonnage aléatoire simple lui attribuerait.

Exemple n° 7 : Un conseil scolaire ontarien voulait évaluer l'opinion des élèves sur la suppression de la 13^e année du programme d'études secondaires. Il a décidé de sonder les élèves de l'école secondaire de l'Érablière et a utilisé une technique d'échantillonnage stratifié pour s'assurer de sélectionner un échantillon représentatif d'élèves de toutes les années d'études offertes dans cet établissement d'enseignement.

Dans ce cas, les strates étaient les cinq années d'études (les 9, 10, 11, 12 et 13^e années). Le conseil scolaire a ensuite sélectionné un échantillon à l'intérieur de chaque strate. Les noms des élèves sélectionnés à l'intérieur de cet échantillon en ont été extraits à l'aide de la méthode d'échantillonnage aléatoire simple ou d'échantillonnage systématique, ce qui a donné un échantillon total de 100 élèves.

La stratification est des plus utiles lorsque les variables de stratification sont :

- simples à utiliser;
- faciles à observer;
- étroitement reliées au thème de l'enquête.

Échantillonnage en grappes

Il est parfois trop dispendieux de disséminer un échantillon dans l'ensemble de la population. Les coûts de déplacement risquent de devenir élevés lorsque les intervieweurs doivent sonder des gens d'un bout à l'autre du pays. Les statisticiens peuvent choisir la technique de l'échantillonnage en grappes pour réduire les coûts.

La technique de l'échantillonnage en grappes entraîne la division de la population en groupes ou en grappes comme son nom l'indique. Suivant cette technique, on sélectionne au hasard un certain nombre de grappes pour représenter la population totale, puis on englobe dans l'échantillon toutes les unités incluses à l'intérieur des grappes sélectionnées. On n'inclut dans l'échantillon aucune unité de grappes non sélectionnées; ces unités sont représentées par celles tirées de grappes sélectionnées. La technique en question diffère de la technique d'échantillonnage stratifié, qui entraîne la sélection d'unités de chaque groupe.

Mentionnons, entre autres exemples de grappes, les usines, les établissements d'enseignement et les régions géographiques telles que les subdivisions électorales. On utilise les grappes sélectionnées pour représenter la population.

Exemple n° 8 : Supposez que vous représentez une organisation d'athlétisme désirant déterminer quels sports pratiquent les élèves de 11^e année au Canada. Il serait trop coûteux et trop long d'interroger chaque élève canadien de 11^e année ou même deux ou trois élèves de chaque classe de 11^e année au Canada. On sélectionne plutôt au hasard 100 écoles de tout le pays.

Ces écoles fournissent des grappes d'échantillons. On sonde ensuite chaque élève de 11^e année de chacune des 100 grappes. Les élèves inclus dans ces grappes représentent, en effet, tous les élèves de 11^e année au Canada.

Exemple n° 9 : Imaginez que le conseil municipal d'une petite localité veut faire enquête sur l'utilisation par ses résidents des services de soins de santé.

Le conseil demande premièrement à Statistique Canada des cartes des subdivisions électorales qui identifient et étiquettent chaque îlot. Il dresse à partir de ces cartes une liste de tous les îlots. Cette liste servira de base de sondage.

Chaque ménage de la localité en question fait partie d'un îlot et chaque îlot représente une grappe de ménages. Le conseil sélectionne au hasard un certain nombre d'îlots. Il dresse ensuite, à l'aide de la méthode d'échantillonnage aléatoire simple, une liste de tous les ménages inclus dans les îlots sélectionnés; ces ménages composent l'échantillon de l'enquête.

Comme nous l'avons mentionné, réduire les coûts est l'une des raisons d'utiliser l'échantillonnage en grappes. Ce dernier crée des « poches » d'unités échantillonnées, au lieu de disséminer l'échantillon sur tout le territoire. Le fait que l'on ne dispose pas parfois d'une liste de toutes les unités incluses dans la population (une nécessité lorsqu'on effectue un échantillonnage aléatoire simple, un échantillonnage systématique ou un échantillonnage avec probabilité proportionnelle à la taille), tandis qu'une liste de toutes les grappes est disponible ou facile à dresser, constitue une autre raison d'utiliser l'échantillonnage en grappes.

Dans la plupart des cas, une perte d'efficacité comparativement à ce qu'il en est lorsqu'on utilise un échantillonnage aléatoire simple représente le principal inconvénient de l'emploi de l'échantillonnage en grappes. Il est habituellement préférable de sonder un grand nombre de petites grappes, plutôt qu'un petit nombre de grandes grappes. Pourquoi? Parce que les unités avoisinantes tendent à se ressembler davantage, ce qui donne un échantillon ne représentant pas l'éventail complet d'opinions ou de situations de l'ensemble de la population. Dans les deux exemples précédents, les élèves de la même école ont tendance à pratiquer les mêmes types de sports (suivant les installations dont dispose leur établissement d'enseignement); de la même façon, les gens âgés tendent à vivre dans des quartiers bien précis et à être de grands utilisateurs des services de santé.

L'échantillonnage en grappes ne permet pas de contrôler totalement la taille finale de l'échantillon, ce qui constitue un autre inconvénient de son utilisation. Puisque toutes les écoles ne comptent pas le même nombre d'élèves de 11^e année et que les îlots ne renferment pas non plus le même nombre de ménages, et que vous devez sonder chaque élève ou ménage inclus dans votre échantillon, ce dernier peut être finalement d'une taille supérieure ou inférieure à ce que vous espériez.

Échantillonnage à plusieurs degrés

La méthode d'échantillonnage à plusieurs degrés ressemble à la méthode d'échantillonnage en grappes, sauf qu'il faut dans son cas prélever un échantillon à l'intérieur de chaque grappe sélectionnée, plutôt que d'inclure toutes les unités dans la grappe. Ce type d'échantillonnage exige au moins deux degrés. On identifie et sélectionne au premier degré de grands groupes ou de grandes grappes. Ces grappes renferment plus d'unités de la population qu'il n'en faut pour l'échantillon final.

Pour obtenir un échantillon final, on prélève au second degré des unités de la population à partir des grappes sélectionnées (à l'aide de l'une des méthodes d'échantillonnage probabiliste possibles). Si l'on utilise plus de deux degrés, le processus de sélection d'unités de la population à l'intérieur des grappes se poursuit jusqu'à l'obtention d'un échantillon final.

Exemple n° 10 : Dans l'exemple n° 8 d'échantillonnage en grappes, on choisissait 100 écoles, puis on interviewait chaque élève de 11^e année de ces écoles. Dans le cas de l'échantillonnage à plusieurs degrés, vous pourriez plutôt sélectionner davantage d'écoles, vous procurer une liste de tous les élèves de 11^e année de ces écoles sélectionnées et choisir un échantillon au hasard (un échantillon aléatoire simple, par exemple) d'élèves de chaque école. Ce serait un plan d'échantillonnage à deux degrés.

Vous pourriez aussi obtenir une liste de toutes les classes de 11^e année des écoles sélectionnées, prélever un échantillon aléatoire de classes de chacune de ces écoles, vous procurer une liste de tous les élèves des classes sélectionnées et finalement choisir un échantillon au hasard d'élèves de chaque classe. Ce serait un plan d'échantillonnage à trois degrés. Le processus se complique chaque fois que nous ajoutons un degré.

Imaginez maintenant que chaque école compte en moyenne 80 élèves de 11^e année. L'échantillonnage en grappes donnerait alors à votre organisation un échantillon d'environ 8 000 élèves (100 écoles x 80 élèves de 11^e année). Si vous vouliez un échantillon de plus grande taille, vous pourriez sélectionner des écoles comptant davantage d'élèves et, pour obtenir un échantillon plus petit, sélectionner des écoles comptant moins d'élèves.

L'un des moyens de contrôler la taille de l'échantillon consisterait à stratifier les écoles en grande taille, en taille moyenne et en petite taille (tailles ici signifiant le nombre d'élèves de 11^e année) et à sélectionner un échantillon d'écoles de chaque strate. On appelle cette méthode la méthode d'échantillonnage en grappes stratifiées.

Si vous utilisiez un plan d'échantillonnage à trois degrés, vous pourriez sélectionner un échantillon de 400 écoles, puis sélectionner deux classes de 11^e année par école (en supposant qu'il y ait deux classes de 11^e année ou plus dans chaque établissement d'enseignement). Finalement, vous pourriez sélectionner 10 élèves par classe. De cette façon, vous finiriez quand même par avoir un échantillon d'environ 8 000 élèves (400 écoles x 2 classes x 10 élèves), mais l'échantillon serait davantage disséminé.

Vous pouvez constater à partir de cet exemple que l'échantillonnage à plusieurs degrés vous offre quand même l'avantage d'un échantillon plus concentré, ce qui réduit les coûts. Cependant, l'échantillon n'est pas aussi concentré que d'autres grappes et sa taille est quand même plus grande que celle d'un échantillon aléatoire simple. Vous n'avez pas non plus besoin de disposer d'une liste de tous les élèves membres de la population. Tout ce qu'il vous faut, c'est une liste des classes des 400 écoles et une liste également des élèves des 800 classes. Il est vrai que ce type d'échantillon exige plus d'information qu'en nécessite l'échantillonnage en grappes. L'échantillonnage à plusieurs degrés épargne cependant quand même beaucoup de temps et d'efforts, parce qu'il ne nécessite pas la création d'une liste de toutes les unités incluses dans une population.

Échantillonnage à plusieurs phases

Un *échantillonnage à plusieurs phases* entraîne la collecte de données de base auprès d'un échantillon d'unités de grande taille et ensuite, pour un sous-échantillon de ces unités, la collecte de données plus détaillées. La forme la plus courante d'échantillonnage à plusieurs phases est l'échantillonnage à deux phases (ou l'échantillonnage double), mais il est également possible d'effectuer un échantillonnage à trois phases ou plus.

L'échantillonnage à plusieurs phases est assez différent de l'échantillonnage à plusieurs degrés, malgré les similarités entre eux sur le plan de leur appellation. Même si l'échantillonnage à plusieurs phases suppose aussi le prélèvement de deux échantillons ou plus, dans son cas, tous les échantillons sont tirés de la même base de sondage et les unités sont structurellement les mêmes à chaque phase. Comme dans le cas de l'échantillonnage à plusieurs degrés, plus l'on utilisera de phases, plus le plan d'échantillonnage et l'estimation deviendront complexes.

L'échantillonnage à plusieurs phases est utile lorsqu'il manque à l'intérieur de la base de sondage des données auxiliaires qui pourraient servir à stratifier la population ou à rejeter à la sélection une partie de la population.

Exemple n° 11 : Supposez qu'une organisation a besoin d'information sur des éleveurs de bétail de l'Alberta, mais que dans la base de sondage sont énumérés tous les types d'exploitations agricoles : d'élevage de bétail et de production laitière, de grains, de porcs, de volailles et de fruits et de légumes. Pour compliquer les choses, la base de sondage ne fournit aucune donnée auxiliaire sur les exploitations agricoles qui y sont énumérées.

On pourrait mener une enquête toute simple dont la seule question serait : « Votre exploitation agricole est-elle en partie ou en totalité consacrée à l'élevage du bétail? » Si elle ne comptait qu'une seule question, cette enquête devrait entraîner un faible coût par interview (surtout si elle était faite par téléphone), ce qui, par conséquent, permettrait à l'organisation de prélever un grand échantillon. Une fois le premier échantillon prélevé, on pourrait en extraire un second, plus petit, à partir des éleveurs de bétail et poser à ces derniers des questions plus détaillées. En utilisant cette méthode, l'organisation éviterait de dépenser de l'argent à sonder des unités ne faisant pas partie du champ d'observation (c'est-à-dire les producteurs agricoles autres que les éleveurs de bétail).

On pourrait utiliser l'échantillonnage à plusieurs phases lorsqu'on disposerait d'un budget insuffisant pour recueillir des données auprès de l'échantillon entier ou lorsque recueillir des données auprès de l'échantillon

entier imposerait un fardeau excessif aux répondants ou même quand les coûts de collecte pour différentes questions d'une enquête différeraient énormément.

Exemple n° 12 : On pose aux participants à une enquête sur la santé des questions de fond au sujet de leur régime alimentaire et de leurs habitudes au niveau de la consommation de tabac et d'alcool et de la pratique de l'exercice physique. Cette enquête oblige, en outre, les répondants à se soumettre à certains examens médicaux, comme courir sur un tapis roulant ou faire mesurer leur tension artérielle et leur taux de cholestérol.

Remplir des questionnaires ou interviewer des participants sont des procédures relativement peu dispendieuses, mais les examens médicaux exigent la supervision et l'aide d'un professionnel ou d'une professionnelle de la santé qualifié(e), de même que l'utilisation d'un laboratoire équipé, ce qui peut dans un cas comme dans l'autre être assez coûteux. La meilleure façon de mener l'enquête susmentionnée consisterait à utiliser une méthode d'échantillonnage à deux phases. À la première phase, on soumettrait aux interviews un échantillon d'une taille appropriée. On prélèverait à partir de cet échantillon un échantillon plus petit. Ce sont les membres de ce second échantillon qui passeraient des examens médicaux.

Source : <http://www.statcan.gc.ca/edu/power-pouvoir/ch13/prob/5214899-fra.htm>

Date de modification : 2009-04-20