Cours Méthodes de sondage

La stratégie: Les types d'échantillon

© Claire Durand, 2024
Département de sociologie, Université de Montréal

Qu'est-ce que la stratégie?

- Développer une stratégie, c'est décider comment exactement, concrètement, se fera l'enquête.
  - La stratégie d'échantillonnage et **le type d'échantillon**
  - Le mode d'administration
- La stratégie doit être pertinente, réaliste et réalisable.
- Elle doit être appropriée à la question de recherche.
- **Quel type d'échantillon?**

Qu'est-ce qu'un échantillon?

- C'est une partie d'un tout...
  - On veut goûter la soupe, on en prend une cuillerée.
  - On prend un échantillon de bulletins de nouvelles pour voir comment une information est traitée.
  - On prend une personne sur dix qui sortent de la cafétéria pour connaître leur opinion sur la nourriture.
  - On prend un sac de bonbons pour connaître la proportion de bonbons rouge...
- ***De préférence prise au hasard...***

A quoi ça sert de prendre un échantillon plutôt que l'ensemble de la "population"?

- A réduire le fardeau des répondants: ils ne pourraient pas répondre à tous les questionnaires.
- A réduire les coûts.
- À préserver les énergies du chercheur. Il ne peut pas collecter des informations sur l'ensemble de la population. Elle est trop importante. Son temps est mieux utilisé à élaborer de bons outils de collecte et à bien gérer celle-ci.
- Ca ne sert à rien de manger toute la soupe pour savoir si elle est bonne. Une cuillerée -- ou deux -- suffisent!

4

A quoi ça sert de prendre un échantillon plutôt que l'ensemble de la "population"?

- À partir de l'échantillon,
  - Je peux généraliser à l'ensemble de la population,
  - ET je peux connaitre la *marge d'erreur* de mon estimation,
  - ET cette marge d'erreur devrait être suffisamment négligeable pour le niveau de précision dont j'ai besoin.
- La théorie de l'échantillonnage s'applique seulement aux échantillons aléatoires, probabilistes.
  - Elle ne s'applique pas si on prend toute la population.
  - Ni, strictement parlant, si on prend un échantillon par quotas, de volontaires, etc.

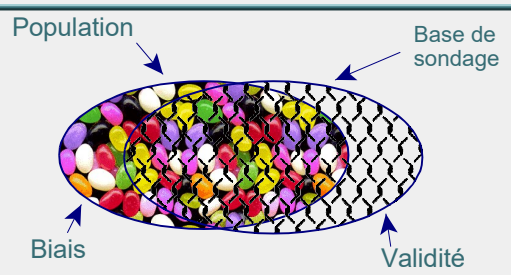
5

Quelques définitions

- *Population*: Ensemble des unités auxquelles on veut généraliser les résultats.
- *Base de sondage*: **Liste** des unités à partir de laquelle se fera la sélection.
  - Chaque unité de la population devrait **idéalement** y apparaître **une fois** et **une seule fois**.
- *Unité d'échantillonnage*: Chaque unité de la base de sondage.
  - A différencier de :
    - Unité d'analyse, sur laquelle portent les analyses.
    - Unité déclarante: relative à l'informateur. EX: un parent à qui on demande des informations sur son enfant.
    - Unité de référence: sur laquelle les informations sont demandées. Ex: l'enfant dont on analysera les réponses données par le parent.

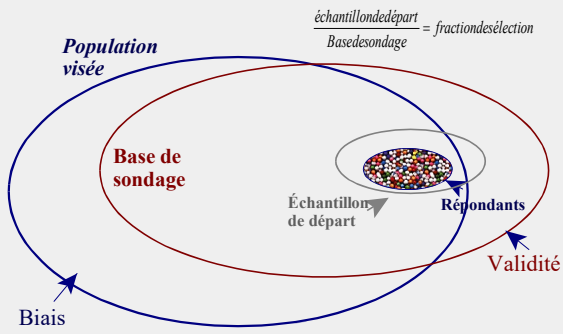
6

Relation population - base de sondage



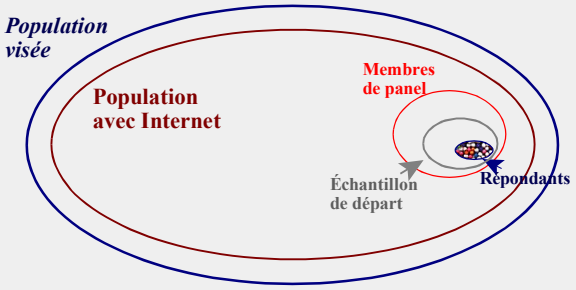
7

Relation population - base de sondage - échantillon - répondants



8

Relation population - base de sondage pour les sondages WEB avec panels de volontaires



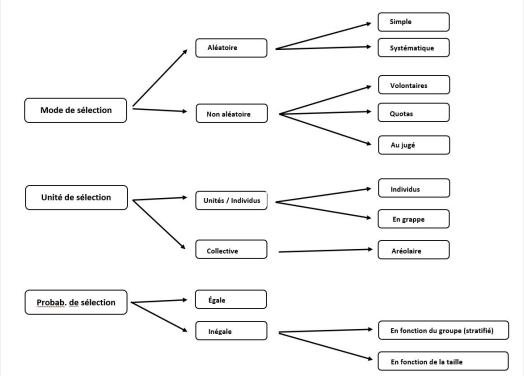
Adapté de Tourangeau et coll. (2013). The Science of Web surveys, p. 98

9

Les types d'échantillon

10

Synthèse



11

Les échantillons aléatoires

12

Les échantillons probabilistes (aléatoires)

Le mode de sélection peut être

- Aléatoire **simple** :
  - C'est comme si on mettait dans un sac les identifiants de tous les membres de la population et qu'on pigeait.
  - En pratique, les logiciels font ça...
- Aléatoire **systématique** :
  - On fait une liste, on "aligne" tous les membres de la population sur une ligne.
  - Ensuite, on tire un premier nombre au hasard et on prend par la suite une unité à toutes les "n" unités en fonction de la *fraction de sélection*.
  - Par exemple, si je veux prendre une unité sur 10, je prends au hasard un nombre entre 1 et 10. Si ce nombre est 4, par exemple, je prendrai les unités numéro 4, 14, 24, 34, 44, etc...

13

Problèmes de l'aléatoire

- Il **faut** avoir une "liste" de la population, une base de sondage la plus complète possible.
- Pour ce qui est de l'aléatoire systématique, il ne faut pas que cette liste ait un *patron* qui entraîne un biais.
  - Exemple: Si je prends un appartement sur 3 dans un quartier de triplex (maisons à 3 appartements superposés), je ne choisirai que des ménages habitant soit tous au rez-de-chaussée (souvent propriétaires), soit tous au deuxième ou au troisième étage (souvent locataires).
  - **La liste doit avoir un ordre relativement aléatoire -- on peut la mélanger -- qui n'entraîne pas de biais.**

14

Les échantillons aléatoires - aréolaires

Types d'unités de sélection

- L'unité de sélection n'est pas nécessairement l'individu...
  - S'il s'agit d'un sondage téléphonique sur ligne fixe ou en face à face auprès de la population, *l'unité de sélection* est le *ménage*...
  - Si je fais un sondage dans une institution d'enseignement, *l'unité de sélection* pourra être la classe ou le cours.
  - *L'unité de sélection* de départ pourrait également être un village, un quartier, un pâté de maison, une école,...
- L'unité de sélection peut donc être **aréolaire**, c'est-à-dire que l'on sélectionne des groupes.

15

Les échantillons aréolaires

Avantages

- Avantage des échantillons aréolaires en sciences sociales, particulièrement en sociologie:
  - Il est souvent possible d'obtenir des informations sur le groupe
    - niveau socio-économique moyen du quartier de résidence,
    - homogénéité/hérérogénéité de la population du groupe,
    - caractéristiques du professeur dans une classe -- expérience, pédagogie, etc. --,
  - Ce qui permet d'utiliser cette information pour les analyses subséquentes, d'analyser les effets de contexte.

16

Les échantillons aléatoires - en grappe

On peut sélectionner...

- Un nombre plus restreint d'unités et prendre les unités "adjacentes": il s'agit d'un ***échantillon en grappe.***
  - *L'unité de sélection* pourrait être un individu et on prend, par exemple, la classe à laquelle l'individu appartient, le village/le quartier/le pâté de maison où il demeure, etc.
  - *L'unité de sélection* peut être un pâté de maison et on prend également les quatre pâtés de maison adjacents situés au nord, au sud, à l'est et à l'ouest du pâté de maison choisi.
- On fait ce choix quand l'information disponible entraîne que c'est la solution la plus efficiente.

17

Problèmes des échantillons aréolaires ou en grappe

- La plus grande homogénéité possible de l'échantillon:
  - Les personnes d'une même classe, d'un même village, d'un même pâté de maison, sont plus similaires entre elles que des gens pris au hasard dans différents villages, quartiers, classes.
  - De même, les unités adjacentes se ressemblent plus, en moyenne, que les unités éloignées.
  - Toutefois, on peut en tenir compte dans les analyses statistiques (modèles multi-niveaux, correction dans le calcul de la variance).

18

Les échantillons aléatoires - probabilité de sélection

La probabilité de sélection peut être...

- *Égale,* soit la même pour toutes les unités ou
- *Inégale*, **différente** selon les unités, en fonction
  - De l'appartenance à un groupe: **échantillon stratifié**
    - Exemple 1: Si 3/4 des employés d'une organisation de 10,000 employés sont des hommes et que je sélectionne avec une probabilité égale un employé sur 10, j'aurai au final un échantillon de 1000 répondants dont environ 750 hommes et 250 femmes.
    - Si je veux avoir autant d'hommes que de femmes parmi les répondants (pour être certain de bien représenter les femmes), je devrai prendre un homme sur 15 (≈ 500 H) et une femme sur 5 (≈500 F). Il s'agit d'un échantillon *stratifié* selon le sexe.
    - Exemple 2: Universités du Québec→ échantillon stratifié selon les univers.
  - De la taille des unités: par exemple, un échantillon d'écoles devrait-il prendre en compte le nombre d'élèves par école dans la probabilité de choisir l'école? Si on ne le fait pas, on peut sous-représenter les "grosses" écoles.

L'échantillon peut être à plusieurs degrés ou à plusieurs phases...

Toute combinaison...

- Plusieurs degrés (exemples):
  - On sélectionne des ménages, puis des personnes à l'intérieur des ménages (fréquent dans les sondages auprès de la population).
  - On sélectionne des écoles, puis des classes à l'intérieur des écoles (fréquent dans les sondages auprès des élèves).
- Plusieurs phases (exemples):
  - On interroge tous les présidents de comité pour catégoriser le comité et avoir le nom des membres.
  - Ensuite, en fonction des informations obtenues, on sélectionne des comités au hasard et on envoie un questionnaire à tous les membres des comités sélectionnés.

Les échantillons non probabilistes

Les échantillons non probabilistes

- Échantillons de volontaires:
  - Courant en psychologie, en sciences de la santé. “Cherche volontaire pour participer à une expérience...”
  - D’une certaine manière, on postule que les êtres humains constituent tous et chacun un échantillon valide des “êtres humains”.
- Échantillons au jugé:
  - Courant en méthodologie qualitative. On cherche à couvrir toutes les situations possibles.
  - On postule qu’il y a un certain nombre de situations différentes et que pour chacune, il faut un “représentant”.

22

Les échantillons non probabilistes

- Échantillons par quotas ou “échantillons représentatifs”:
  - L’idée est de recréer une “mini-société” basée sur ce que l’on sait des caractéristiques de la société.
    - Exemple:
      - On a choisi un certain nombre de caractéristiques que l’on croit liées aux informations qui nous intéressent -- comme le vote par exemple -- et on constitue l’échantillon de telle sorte que les proportions de personnes ayant ces caractéristiques soient les mêmes que dans la population.
      - On arrête d’interviewer des personnes ayant certaines caractéristiques dès que les quotas sont atteints.
      - Les quotas peuvent être très spécifiques: “Nombre attendu d’hommes et de femmes dans les ménages dont le chef de ménage a un certain type d’occupation, pour chacun des groupes d’âge dans chaque type d’agglomération dans chaque région de France.”
    - Les sondages internet de type “opt-in panel” (panels de volontaires) courants en ce moment fonctionnent avec des quotas age-sexe-région-langue maternelle au Québec et parfois, scolarité ou présence d’enfants dans le ménage.

23

Quelques problèmes des quotas

- Chaque membre de la population n’a pas une probabilité connue d’être choisi (une fois une cellule de quota remplie, un membre de ce groupe ne peut plus être choisi).
- Le biais d’interviewer: celui-ci aura tendance à chercher les répondants les plus sympatiques.
- Le biais d’auto-sélection: les répondants peuvent tricher dans les informations qu’ils donnent s’ils sont intéressés à répondre au sondage ou à faire mousser leurs opinions.
- En théorie, si j’ai besoin de 5 hommes de 18 à 24 ans, je pourrais tous les prendre dans la même classe de sociologie. Ils ne seraient pas nécessairement représentatifs des Québécois de 18-24 ans.
- S’il n’y a pas de relations entre les quotas et les variables d’intérêt, mais qu’il y a des relations avec d’autres variables qui n’entrent pas dans les quotas et qui sont mal représentées (scolarité, pratique religieuse), l’échantillon pourrait être biaisé et ne donnerait pas une bonne estimation.

24

Quelques moyens pris pour pallier aux problèmes des quotas

- En pratique, on introduit de l'aléatoire... comme
  - Choisir au hasard les numéros de téléphone dans un sondage téléphonique qui se fait ensuite par quota.
  - Choisir des quartiers ou donner des routes aléatoires à suivre pour les interviewers qui font des sondages en face à face.
  - Choisir de façon aléatoire l'échantillon de départ dans les panels de volontaires des sondages Web.
  - Sélectionner les répondants qui sont actifs sur internet à des moments aléatoires sur des sites choisis aléatoirement.
    - Développement de nouvelles méthodes *random device engagement, river sampling, RIWI (présentation dans le cours du 16 octobre).*

25

Les échantillons mixtes

Boule-de neige et respondent-driven sampling

- Boule-de-neige: Il s'agit de partir d'une ou de quelques personnes qui ont les caractéristiques de la population désirée et de leur demander de nous référer d'autres personnes ayant les mêmes caractéristiques.
  - Exemples:
    - professionnels utilisant tel type de technologie,
    - population autochtone de Chicago,
    - Communauté LGBT,
    - Population d'origine roumaine du Québec,etc.
- Problème: Très dépendant de la première personne contactée, non aléatoire.

26

Les échantillons mixtes

Boule-de neige et respondent-driven sampling

- Respondent-driven sampling: Il s'agit de partir de plusieurs personnes qui ont les caractéristiques de la population désirée
  - On leur demande de nommer jusqu'à sept personnes, par exemple, qui ont les mêmes caractéristiques.
  - On choisit *au hasard* deux personnes parmi les sept.
  - On donne au répondant des codes pour les deux répondants choisis. Il les contacte, leur donne l'information et le code. Ces répondants reçoivent un montant d'argent pour leur collaboration à l'enquête.
  - Voir Webinar de Angelo Cozzubo sur le site de WAPOR: https://wapor.org/resources/wapor-webinars/webinar-october-2023/ et présentation de Eric Lacroix (SOM) sur le site du cours.

27

Errésumé,...

- Toujours privilégier les échantillons probabilistes MAIS...
- Ne pas penser qu'il n'y a que des échantillons aléatoires simples pris à même des listes d'individus. Plusieurs possibilités existent, il faut trouver la plus efficace.

28

En résumé, les questions à se poser

- Quel accès aux informations sur la population?
- Accès aux coordonnées de la personne elle-même ou seulement d'un groupe auquel elle appartient?
- Est-ce une sous-population (les jeunes, les vieux, les femmes seulement, les immigrants, etc.)
- Échantillon probabiliste est-il possible?
- Probabilité égale ou inégale de sélection? Selon quels critères?
- Accès direct ou plusieurs degrés/ phases?

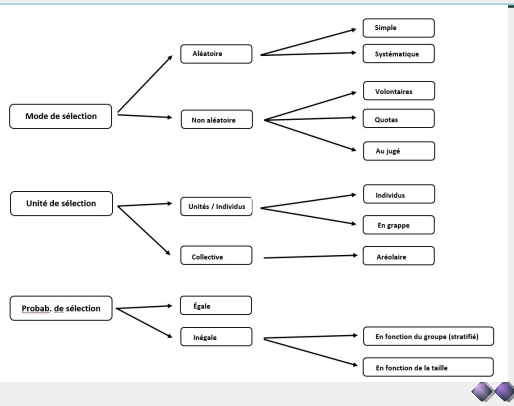
29

A lire!

- Notes de cours sur les types d'échantillon et les stratégies.
- **Chapitres du livre Recherche sociale**
  - **Durand, C. (2021) Le sondage.**
  - **Beaud, J. P. (2021) L'échantillonnage.**
- Dussaix, A.M. et Grosbras, J. M. (1994). Sondages et qualité des données.

30

Synthèse



Vous êtes maintenant prêts...

Pour votre travail de cours

- A préciser la population visée, soit...
  - Ensemble des personnes qui...
- A recueillir les informations sur votre population **et** sur votre base de sondage.
  - Combien de personnes, combien de ménages, combien de classes, combien d'institutions, etc. selon la définition de la population.
  - Quelle base de sondage est -- ou pourrait être -- accessible?
- A venir: élaborer une stratégie efficace, efficiente et le plus probabiliste possible.