Sondages empiriques

Pierre Duchesne

August 1, 2017

Méthodes d'échantillonnage non-probabilistes

- Avec ce genre de méthodes, habituellement,
  1. Moins grande exactitude qu'avec les méthodes probabilistes;
  2. Impossible de mesurer la précision.
- Avantage important des méthodes non-probabilistes
  1. Ces méthodes sont souvent moins coûteuses.
  2. Les méthodes ne sont pas farfelues, mais elles reposent fortement sur l'expertise des ressources en cause.

Inconvénients statistiques des méthodes non-probabilistes

- Avec les méthodes probabilistes, le hasard dans la sélection est pris en compte: notre connaissance des mécanismes aléatoires permet de faire des affirmations reposant sur notre connaissance de la théorie des probabilités.
- Avec les méthodes non-probabilistes: des concepts comme le biais des estimateurs, ou encore les marges d'erreurs, ne peuvent pas être considérés ou calculés.
- En particulier, les écarts-types des estimateurs ne peuvent pas être calculés.

Autres conséquences d'un échantillonnage non-probabiliste

- Les éléments sont choisis de façon arbitraire.
  1. Impossible de trouver les $\pi_k = P(s \ni k)$;
  2. Il n'est pas possible de savoir si $\pi_k > 0$;
  3. Rappelons que des probabilités d'inclusion nulles impliquent la possibilité que des unités ne puissent être échantillonnées.
- Sans π_k et encore moins les probabilités π_{kl}, on ne peut pas savoir si les estimateurs utilisés sont sans biais et il est impossible de formuler un estimateur de variance.

Méthode des quotas (échantillonnage dirigé; échantillonnage par choix raisonné)

- Utilisé dans les enquêtes d'opinion, les études de marché.
- On demande aux interviewers de faire un nombre déterminé d'interviews dans divers groupes définis dans la population. Ces groupes sont définis en fonction de variables socio-démographiques (âge, sexe, géographie, démographie, etc.).
- Les quotas sont souvent établis de façon à être sensiblement poportionnels à la fraction de la population représentée par chaque groupe.

Méthode des quotas (suite)

Pour mettre en oeuvre la méthode:

- Pas besoin d'une base de sondage!
- Pas besoin de plan d'échantillonnage!

Inconvénients de la méthode des quotas

- La méthode des quotas est toujours sujette à des problèmes de biais de sélection.
- Cette méthode masque le problème de la non-réponse.
  1. Si un individu refuse de participer ou s'il est absent de son domicile, l'interviewer ira voir un autre individu car il a à combler son quota. Ainsi, il y a un sérieux risque de ne jamais rejoindre certaines catégories de la population ayant de la réticence à répondre ou difficile à rejoindre.
- De manière générale, il faut surtout retenir qu'avec les méthodes non-probabilistes, on ne peut pas vraiment mesurer la précision des estimations.

Exemple

Dans une certaine population de personnes, on désire une représentativité de l'ensemble de la population sur les variables âge et sexe.

	Âge			
	< 30	$30 - 50$	$50 >$	
H	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
F	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Exemple (suite)

On doit déterminer la taille de l'échantillon n.

- On doit déterminer les n_{ij} de sorte que la contrainte suivante soit respectée:

$$\sum\sum n_{ij} = n.$$

- On peut utiliser la règle proportionnelle suivante:

$$n_{ij} = n\frac{N_{ij}}{N},$$

  présumant que l'on dispose des effectifs N_{ij} connus de la population.
- Même dans ce cas-ci on doit disposer d'une information auxiliaire, obtenue par exemple de sources administratives ou d'un recensement.

Exemple (suite): Formation des quotas et estimation

- On forme les quotas, où des gens sont interrogés au hasard où selon une autre méthode.
- On pourrait considérer l'estimateur:

$$\hat{\bar{y}}_{U,quotas} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{N_{ij}}{N} \bar{y}_{s_{ij}}$$

où s_{ij} représente l'ensemble des uités sélectionnées dans la cellule (i, j).

Propriétés de l'estimateur $\hat{\bar{y}}_{U,quotas}$

- En fait on ne peut savoir grand chose: est-il sans biais? On ne peut le savoir.
- L'estimateur n'est pas farfelu.
- Le plan ressemble à unm sondage stratifié, sauf que l'on n'est pas assuré que ce soit obtenu selon le plan SI dans chaque cellule U_{ij}.
- Compte tenu de ce que l'on sait dans les plans stratifier, certains pourraient être tentés de reporter certaines formules connues.

Mesure de précision

On retrouve comme estimateur de variance de $\hat{\bar{y}}_{U,quotas}$ la formule familière suivante:

$$\hat{V}(\hat{\bar{y}}_{U,quotas}) = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{N_{ij}}{N}\right)^2\left(\frac{1}{n_{ij}}-\frac{1}{N_{ij}}\right)S^2_{ys_{ij}},$$

avec

$$S^2_{ys_{ij}} = \frac{1}{n_{ij}-1}\sum\nolimits_{s_{ij}}(y_k-\bar{y}_{s_{ij}})^2.$$

Remarques

Avec la théorie que nous connaissons, nous ne pouvons pas dire si les estimateurs proposés sont biaisés ou non, tant concernant l'estimateur de la moyenne ou l'estimateur de la variance.

Remarques (suite)

- La théorie vue dans le cours STT-2000 repose sur la *théorie de la randomisation* (approche *design-based*). Le plan de sondage induit des probabilités d'inclusion et les propriétés des estimateurs reposent sur le hasard engendré par la mise en oeuvre d'un plan d'échantillonnage.
- Il faut noter qu'il existe d'autres théories, qui ne reposent pas autant sur le plan de sondage mais sur l'utilisation de modèles. On parle de l'apprcche fondée sur les modèles ou *model-based*.