# Section 4: Types d'erreur de sondage

Pierre Duchesne

August 1, 2017

# Non-réponse

- Problème survenant dans plusieurs sondages.
- Peut varier considérablement dans le temps, même pour un même sondage effectué à plusieurs reprises dans le temps.
- Pour la contrôler: entrainement des interviewers, déroulement de l'entrevue (longueur de l'entrevue, choix et formulation des questions, etc.)

# Exemples de non-réponse

- Incapacité de contacter un individu, un ménage, ou de manière générale une unité.
- La personne peut décider de ne pas répondre; refus catégorique; incompréhension (langue, analphabétisme).
- Idéalement, l'interviewer tente d'obtenir de l'information démographique sur le non-répondant (âge, sexe, ethnie, lieu de résidence: rural ou urbain, etc.)

# Effets de la non-réponse

- Biais potentiel dans les estimateurs des paramètres de la population.
- La non-réponse est particulièrement grave car la volonté de réponse est souvent liée à la variable d'intérêt $y$.

# Facteurs qui contribuent à de petits biais

- Dans l'estimation du paramètre moyenne, si la moyenne chez les non-répondants est similaire à la moyenne chez les répondants.
- Taux de non-réponse est relativement petit (pour EPA = enquête sur la population active, le taux de non-réponse est de l'ordre d'environ 7%; moins de 5% est souvent jugé 'acceptable').

# Formulation de la non-réponse

- Contexte: On sélectionne $s$ dans $U$ avec probabilité $p(s)$, avec $\pi_k > 0$ et $\pi_{kl} > 0$.
- Après la collecte des données, $y$ est mesurée mais s'avère disponible seulement pour les unités $k \in r$, avec $r \subseteq s$.
- L'ensemble $r$ est l'ensemble des répondants.
- Dans une telle situation, nous sommes en présence de données manquantes ou encore en présence de non-réponse.

# Exemple

***Échantillon s***

***Réponse aux deux questions***

***Réponse à une question mais pas aux deux questions***
***Non-réponse par item***

**Non-réponse aux deux questions**
***Non-réponse par unité***

***Partie non-échantillonnée partie U - s***

| *y* | *x* |
|---|---|
| OK! | OK! |
| OK! | OK! |
| OK! | OK! |
| OK! | OK! |
| OK! | OK! |
| OK! | OK! |
| … | … |
| OK! | OK! |
| OK! | N.D. |
| N.D. | OK! |
| OK! | N.D. |
| … | … |
| N.D. | OK! |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| … | … |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| N.D. | N.D. |
| … | … |
| N.D. | N.D. |

# Illustration du problème de biais

- Supposons une enquête sur le revenu, où tous les hauts revenus ont une probabilité plus faible de répondre.
- Dans l'exemple, il y a un lien entre la volonté de répondre et la variable d'intérêt.
- Dans l'échantillon des répondants $r$, il y aura donc sous-représentativité des hauts revenus, $r \subseteq s$.
- Supposons que le plan SI est mis en oeuvre. On voudrait calculer l'estimateur d'Horvitz-Thompson $\bar{y}_s = n^{-1} \sum_s y_k$.
- Comme il n'est pas calculable, on pourrait considérer la moyenne des répondants:

$$\bar{y}_r = m^{-1} \sum_r y_k,$$

où $m$ est le nombre de répondants. Il est attendu que $E(\bar{y}_r) < \bar{y}_U$.

# Techniques pour contrôler le problème de la non-réponse

- Sous-échantillonnage des non-répondants.
- Techniques de réponses randomisées.
- Techniques basées sur la modélisation de la non-réponse.
- Imputation.

# Imputation

- La technique générale consiste à poser des valeurs plausibles pour les valeurs manquantes.
- Supposons que l'on dispose d'information auxiliaire $x_k$ pour $k \in s$.
- Une technique est l'imputation par le ratio. Si une bonne relation linéaire existe en $y$ et $x$ on considère:

$$\hat{y}_k = x_k \frac{\sum_r y_k}{\sum_r x_k} = x_k \hat{B}, \ k \in s - r.$$

- On obtient ainsi un ensemble de données complétées:

$$\begin{cases} y_k, & k \in r, \\ \hat{y}_k, & k \in s - r \end{cases}$$

# Autres exemples d'imputation

- *Hot Deck*: On choisit au hasard une valeur déjà obtenue pour remplir les trous.
- *Cold Deck*: Utilisation de données provenant de sources externes.
- *Imputation par la moyenne:* Chacun des trous est remplacé par la valeur moyenne des répondants.
- Imputation utilisant des modèles: imputation par le ratio, par la régression, etc.

# Repondération comme méthode d'ajustement pour la non-réponse

- Plutôt que de chercher à remplir les trous créer par la non-réponse, on décide de ne travailler qu'avec l'information disponible.
- Ainsi on accepte la non-réponse, on n'imputera pas, mais on va charger la méthode d'estimation.
- On cherche à redresser l'échantillon des répondants.
- Pour considérer cette technique, on a besoin d'information auxiliaire permettant un découpage en catégories.
- Par exemple: effectifs selon des variables comme l'âge ou le sexe.

# Repondération

- On identifie $G$ groupes, $g = 1, \ldots, G$.
- L'échantillon $s$ se trouve donc divisé en $g$: $s_1, \ldots, s_g, \ldots, s_G$.
- Ainsi, ceci implique que l'échantillon des répondants $r$ est décougé en $r_1, \ldots, r_g, \ldots, r_G$.
- Pour $k \in s$, on doit donc pouvoir observer les caractéristiques qui définissent les regroupements.

# Repondération (suite)

- Pour le groupe $g$, soit $m_g/n_g$ le taux de réponse, où $m_g$ est la taille de $r_g$, et $n_g$ est la taille de $s_g$.
- Par repondération, on veut dire que le poids d'échantillonnage $1/\pi_k$ (poids sans non-réponse) est remplacé (ou redressé) par la valeur $(n_g/m_g)(1/\pi_k)$ (poids ajusté pour la non-réponse).

# Intuition derrière la méthode de redressement (plan SI)

- Le poids de sondage initial est $1/\pi_k = N/n$.
- On remarque que $\sum_s \pi_k^{-1} = \sum_s \frac{N}{n} = N$.
- Cependant basé uniquement sur les répondants $\sum_r \pi_k^{-1} = \sum_r \frac{N}{n} = m\frac{N}{n}$.
- On cherche $\theta$ tel que $\sum_r \theta\frac{N}{n} = m\theta\frac{N}{n}$.

# Estimateur par repondération

- L'estimateur est:

$$\hat{t} = \sum_{g=1}^{G} \left( \frac{n_g}{m_g} \sum\nolimits_{r_g} \frac{y_k}{\pi_k} \right)$$

- Pour le plan SI, avec $\bar{y}_{r_g} = m_g^{-1} \sum_{r_g} y_k$ et $w_g = n_g/n$, on a

$$\hat{t} = \sum_{g=1}^{G} \left( \frac{n_g}{m_g} \frac{N}{n} \sum\nolimits_{r_g} y_k = N \sum_{g=1}^{G} w_g \bar{y}_{r_g}. \right)$$

# Justification de l'estimateur par repondération

- On se rappelle que le plan de sondage induit $\pi_k = P(s \ni k)$. En quelque sorte, c'est ce qui se passe au niveau de $s$.
- Pour la création de $r$, on pose $\theta_{k|s} = P(r \ni k|s)$.
- Dans le cas de l'estimateur par repondération par groupes, on suppose:

$$\theta_{k|s} = \theta_g, \ \forall k \text{ dans le groupe } g.$$

- L'estimateur est:

$$\hat{t} = \sum_{g=1}^{G} \sum_{r_g} \frac{y_k}{\pi_k \theta_g}$$

# Justification (suite)

- Sous les hypothèses suivantes: pour chaque unité $k$ dans le groupe $g$, les unités répondent avec probabilité $\theta_g$ indépendamment (plan BE conditionnellement à $s$).
- Sous ces hypothèses, l'estimateur précédent est alors sans biais *sous ce mécanisme de réponse*.
- Puisque $\theta_g$ est inconnu, on l'estime par:

$$\hat{\theta}_g = \frac{m_g}{n_g}.$$

# Exemple, T.P.9, no. 5

| | Hommes | Femmes | Total |
|---|---|---|---|
| $n_g$ | 500 | 500 | 1000 |
| $m_g$ | 300 | 450 | 750 |
| Répondants en faveur | 165 | 360 | 525 |

# Solution

- L'estimateur est $\hat{t}/N = \sum_{g=1}^{G} w_g \bar{y}_{r_g}$, avec $w_g = n_g/n$.
- On pose

$$y_k = \begin{cases} 1, & \text{en faveur,} \\ 0, & \text{sinon.} \end{cases}$$

- On obtient $\bar{y}_{r_1} = 165/300$ et $\bar{y}_{r_2} = 360/450$, avec $n_1 = n_2 = \frac{1}{2}$.
- L'estimateur repondéré est alors:

$$\frac{\hat{t}}{N} = \frac{1}{2}\left(\frac{165}{300}\right) + \frac{1}{2}\left(\frac{360}{450}\right) = 0.675;$$

- On remarque que la moyenne des répondants est $\bar{y}_r = \frac{525}{750} = 0.7$.

# Solution (suite et fin)

- L'estimateur repondéré est mieux que la moyenne des répondants dans la mesure où l'estimateur repondéré tient compte des différences qu'il y a entre les groupes en rapport avec la volonté de répondre.
- Dans l'exemple, les hommes répondaient moins que les femmes.