Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles?

Jean-François Beaumont

Colloque francophone sur les sondages
Lyon, 24-26 octobre 2018

Un peu d'histoire…

- Jusqu'au début du 20è siècle, on privilégie les recensements
  - **Coûteux (en argent et en temps)**
- **Une alternative:** tirer un échantillon de la population
  - **Comment?** Aléatoire ou non?
  - Plusieurs débats … jusqu'à Neyman (1934)
  - Rao (2005); Bethlehem (2009)
- Les enquêtes probabilistes se sont **ensuite** graduellement établies dans les agences nationales de statistique **(1ère EPA au Canada en 1945)**

Pourquoi les enquêtes probabilistes …?

- Théorie de Neyman (1934) est attrayante:
  - Méthode objective pour tirer les échantillons
  - Inférence “design-based”: validité ne dépend pas d’hypothèses de modèle (approche non paramétrique)
  - Propriété indispensable pour les agences nationales de statistique (Deville, 1991) **historiquement** réticentes à la prise inutile de risques
- Quelques exemples marquants d’échantillons non probabilistes qui ont mené à des conclusions erronées (ex.: sondage pré-électoral de 1936 aux É-U)

Sont-elles une panacée?

- Estimations imprécises si n est petite
- Fondées sur l'hypothèse que les erreurs non dues à l'échantillonnage sont négligeables
  - Beaucoup de ressources pour minimiser les erreurs de non-réponse, de mesure et de couverture
- Pas parfaites mais **généralement** reconnues comme une source fiable sauf peut-être si les erreurs non dues à l'échantillonnage deviennent prépondérantes (Brick, 2011)

Vent de changement …

- On considère de plus en plus d'autres sources de données
- **Quatre raisons principales:**
  - Déclin des taux de réponse dans les enquêtes ➡ biais
  - Coûts de collecte élevés + fardeau sur les répondants
  - Désir d'avoir des statistiques en "temps réel" (Rao, 2018)
  - Prolifération des sources non probabilistes (ex.: enquêtes Web, administratives, médias sociaux, …)
    - Moins coûteuses, *n* plus grande

… sont-elles une panacée?

- Biais (couverture, sélection)
  - Devient dominant à mesure que n augmente (Meng, 2018)
  - Échantillon de grande taille n'est pas un gage de qualité
  - **Exemple:** Sondage pré-électoral de 1936 aux É.-U. mené par la revue *Literary Digest* avec $n > 2000000$ et échantillon fortement non représentatif de la population d'électeurs
- Erreurs de mesure (ex.: enquêtes Web menées auprès de volontaires)

Une question pertinente dans le contexte actuel

- Comment peut-on utiliser les données d'une source non probabiliste afin de
  - **minimiser les coûts de collecte et le fardeau sur les répondants d'une enquête probabiliste**
  - **tout en conservant un cadre d'inférence statistique valide et une qualité acceptable?**
    - Validité statistique: Zhang (2012)

Cadre d'inférence statistique?

- Cadre d'inférence statistique est caractérisé par une **distribution de référence** et une **liste d'hypothèses**
- Que nous permet-il?
  - Définir les propriétés comme le biais et la variance
  - Choisir un estimateur selon un critère objectif (ex.: sans biais avec la plus petite variance possible)
  - Déterminer la qualité des estimations et faire des inférences valides
- Sans un cadre d'inférence statistique valide, on peut calculer des estimations mais pas leur qualité et pas d'inférences possibles

Dans ce qui suit …

- Méthodes d'intégration de données
- Contexte et notation
- Approches "design-based"
- Approches "model-based"
  - Calage
  - Appariement statistique
  - Pondération par l'inverse du score de propension
  - Modèle de Fay-Herriot
- Quelques réflexions supplémentaires

Étapes de production

- Déterminer les besoin d'informations avec les utilisateurs
  - Définir la population cible: U
  - Définir les paramètres d'intérêt: $\theta = \sum_{k \in U} y_k$
- Déterminer les procédures d'estimation en tenant compte du budget, fardeau, qualité, …
  - Identifier la ou les sources de données, probabilistes ou non
  - Déterminer le cadre d'inférence statistique
- Éviter de choisir une source et ensuite déterminer les besoins en fonction de la source

Notation

- Échantillon non probabiliste: s_{NP}
  - Sous-ensemble de U
  - Contient une variable y^* et possiblement d'autres variables
  - Indicateur d'inclusion dans s_{NP} : δ_k ➡ $\boldsymbol{\delta}$
- Deux scénarios:
  - $y_k^* = y_k$
  - $y_k^* \neq y_k$: différences conceptuelles ou erreurs de mesure
- Variable d'intérêt: y_k ➡ $\mathbf{Y}$

Notation

- Échantillon probabiliste: s_P
  - Sous-ensemble de U sélectionné aléatoirement avec probabilité $p(s_P|\mathbf{Z})$
  - Indicateur d'inclusion dans s_P : I_k ➡ $\mathbf{I}$
  - Probabilité d'inclusion: $\pi_k = \Pr\left(I_k = 1|\mathbf{Z}\right) > 0$
  - Contient ou non la variable y
- $\boldsymbol{\Omega}$: Ensemble de toutes les informations utilisées pour faire les inférences sauf $\mathbf{I}$, $\boldsymbol{\delta}$ et $\mathbf{Y}$
- Inférence "design-based": Tout est considéré comme fixe sauf $\mathbf{I}$

Inférence design-based

- Distribution de référence: $F(\mathbf{I} \mid \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega})$
- Pour l'estimation du total $\theta = \sum_{k \in U} y_k$, on considère des estimateurs de la forme

$$\hat{\theta} = \sum_{k \in s_P} w_k y_k$$

- Si $w_k = \pi_k^{-1}$ alors $E\left(\hat{\theta} - \theta \middle| \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega}\right) = 0$
- **Alternative**: Calage (Deville & Särndal, 1992)
- Aucune hypothèse de modèle requise sauf pour tenir compte des **erreurs non dues à l'échantillonnage**
  - Suppose que le biais n'est pas trop grand (Brick, 2011)

Approches design-based

- Caractéristiques:
  - **Rôle de l'échantillon non probabiliste**: Réduire la variance
  - Réduction de variance peut être utilisée pour justifier une réduction de la taille d'échantillon
  - **Variable d'intérêt doit être recueillie dans l'échantillon probabiliste et mesurée sans erreur**
  - Estimation sur petits domaines possède les trois mêmes caractéristiques
    - On s'attend à des gains d'efficacité plus modestes que ceux obtenus avec les méthodes d'estimation sur petits domaines

Scénario 1: $y_k^* = y_k$

- Contexte:
  - s_{NP} est un sous-ensemble de U : **sous-couverture**
  - Le rôle de l'échantillon probabiliste est d'éliminer le biais de couverture
  - Plus la taille de s_{NP} sera grande, plus la taille de s_P pourra être petite sans compromettre la précision
- **Idée:**
  - On veut utiliser les données de $s = s_P \cup s_{NP}$
  - On pondère chaque unité $k \in s$ par

$$\left[\Pr(k \in s \mid \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega})\right]^{-1}$$

Scénario 1: $y_k^* = y_k$

- Estimateur:

$$\hat{\theta} = \sum_{k \in s_{NP}} y_k + \sum_{k \in s_P} \frac{1}{\pi_k}(1-\delta_k)y_k$$

  - δ_k doit être disponible pour $k \in s_P$
  - $E\left(\hat{\theta} - \theta \middle| \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega}\right) = 0$
  - Équivalent à la méthode de Bankier (1986) pour traiter le problème des bases de sondage multiples
  - On peut améliorer l'estimateur en remplaçant les poids π_k^{-1} par des poids calés

Scénario 2: $y_k^* \neq y_k$

- y_k^* ne peut être utilisé en remplacement de y_k ; seulement comme variable auxiliaire
- Vecteur de variables auxiliaires: $\mathbf{x}_k^*,\ k \in s_{NP}$
- Total: $\mathbf{T}_{\mathbf{x}^*} = \sum_{k \in s_{NP}} \mathbf{x}_k^* = \sum_{k \in U} \delta_k \mathbf{x}_k^*$
- **Calage**: On trouve des poids $w_k,\ k \in s_P$ tels que

$$\sum_{k \in s_P} w_k \begin{pmatrix} \mathbf{x}_k \\ \delta_k \mathbf{x}_k^* \end{pmatrix} = \begin{pmatrix} \mathbf{T}_{\mathbf{x}} \\ \mathbf{T}_{\mathbf{x}^*} \end{pmatrix}$$

- $\delta_k \mathbf{x}_k^*$ doit être disponible pour $k \in s_P$ ➡ ajout de questions à l'enquête probabiliste

Approches model-based

- Objectif des trois prochaines méthodes:
  - Réduire le fardeau et les coûts en éliminant la collecte de variables d'intérêt dans s_P
- Inférences valides si:
  - Hypothèses tiennent la route
  - $y_k^* = y_k$
- Estimateur naïf: $\hat{\theta}^{NP} = N \sum_{k \in s_{NP}} y_k / n^{NP}$
  - Peut être très biaisé (Bethlehem, 2016)
  - Les méthodes réduisent le biais au moyen d'un vecteur de variables auxiliaires $\mathbf{x}_k$

Calage de s_{NP}

- **Idée** (Royall, 1970):
  - Modéliser la relation entre y_k et $\mathbf{x}_k$ en utilisant l'échantillon non probabiliste
  - Prédire y_k pour les unités $k \in U - s_{NP}$
- **Inférences**: conditionnelles à $\boldsymbol{\delta}$ et $\mathbf{X}$
- **Hypothèse de sélection non informative**:
  - $F(\mathbf{Y} \mid \boldsymbol{\delta}, \mathbf{X}) = F(\mathbf{Y} \mid \mathbf{X})$
  - Essentielle pour éliminer le biais
  - Plus $\mathbf{X}$ est riche, plus l'hypothèse devient réaliste

Calage de s_{NP}

- Modèle linéaire: $E\left(y_k \middle| \mathbf{X}\right) = \mathbf{x}_k'\boldsymbol{\beta}$
- BLUP du total θ : $\hat{\theta}^{BLUP} = \sum_{k\in s_{NP}} y_k + \sum_{k\in U-s_{NP}} \mathbf{x}_k'\hat{\boldsymbol{\beta}}$
- Peut être ré-écrit: $\hat{\theta}^{BLUP} = \sum_{k\in s_{NP}} w_k^C y_k$
- Le poids de calage satisfait: $\sum_{k\in s_{NP}} w_k^C \mathbf{x}_k = \mathbf{T}_\mathbf{x}$
- Le calage implique un modèle linéaire
- Si $\mathbf{T}_\mathbf{x}$ n'est pas connu, on le remplace par un estimateur sans biais (enquête probabiliste):

$$\hat{\mathbf{T}}_\mathbf{x} = \sum_{k\in s_P} w_k \mathbf{x}_k$$

Calage de S_{NP}

- BLUP est sans biais: $E\left(\hat{\theta}^{BLUP} - \theta \middle| \boldsymbol{\delta}, \mathbf{X}\right) = 0$
- **Réduction du biais de sélection:**
  - Considérer un grand nombre de variables auxiliaires
  - Une grande enquête probabiliste peut être utile pour obtenir des estimations des totaux auxiliaires
  - Méthodes de sélection de variables (LASSO, …)

Calage de s_{NP}

- **Modèle de poststratification**:
  - $E\left(y_k \middle| \mathbf{X}\right) = \mu_h$, k dans la poststrate h
  - Poststrates peuvent être obtenues par le croisement de plusieurs variables catégorielles
- **Réduction du biais de sélection:**
  - Considérer un grand nombre de poststrates
  - Arbres de régression peuvent être utiles
- Modèle linéaire n'est pas toujours approprié
  - **Solution**: Calage assisté par un modèle de Wu et Sitter (2001)

Appariement statistique

- **Idée**:
  - Modéliser la relation entre y_k et $\mathbf{x}_k$ en utilisant l'échantillon non probabiliste
  - Prédire (imputer) y_k dans un échantillon probabiliste qui contient les variables auxiliaires
- **Inférences**: conditionnelles à $\boldsymbol{\delta}$ et $\mathbf{X}$
- Hypothèse de sélection non informative:
- Prédicteur du total θ : $\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp}$
- Sans biais si: $E\left(y_k^{imp} - y_k \middle| \boldsymbol{\delta}, \mathbf{X}\right) = 0$

Appariement statistique

- Pour un modèle linéaire, l'appariement statistique est dans la plupart des cas équivalent à un calage de s_{NP} sur les totaux estimés $\hat{\mathbf{T}}_{\mathbf{x}}$
  - Ex.: modèle de poststratification
- Imputation par donneur est souvent considérée (ex.: Rivers, 2007)
  - Méthode non paramétrique
- Imputation fractionnelle par donneur (Kim et Fuller, 2004)
  - Équivalent à Lavallée et Brisbane (2016)

Appariement statistique

- Imputation linéaire: $y_k^{imp} = \sum_{l \in s_{NP}} \omega_{kl} y_l$
  - **Cas particuliers**: Régression linéaire, donneur, …
  - Beaumont et Bissonnette (2011)
  - On peut ré-écrire $\hat{\theta}^{SM}$ sous une forme pondérée:

$$\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp} = \sum_{k \in s_{NP}} W_k \, y_k$$

- **Pondérer ou imputer? Appariement statistique ou calage?**
  - Est-ce que c'est le contenu de la source non probabiliste qui est d'intérêt ou c'est plutôt le contenu de l'enquête probabiliste?

Pondération par l'inverse …

- **Idée:**
  - Modéliser la relation entre δ_k et $\mathbf{x}_k$
  - Estimer la probabilité de participation $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right)$ par $\hat{p}_k$
  - Estimateur: $\hat{\theta}^{PS} = \sum_{k \in s_{NP}} w_k^{PS} y_k$, où $w_k^{PS} = 1/\hat{p}_k$
- **Avantage principal:**
  - Simplifie l'effort de modélisation quand il y a plusieurs variables d'intérêt (un seul indicateur de participation à modéliser)

Pondération par l'inverse …

- **Hypothèses:**
  - Sélection non informative:

$$\Pr\left(\delta_k = 1 \middle| \mathbf{Y}, \mathbf{X}\right) = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right)$$

  - $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right) > 0$
- **Inférences**: conditionnelles à $\mathbf{Y}$ et $\mathbf{X}$
- **Modèle paramétrique** (ex.: logistique):

$$p_k(\boldsymbol{\alpha}) = g(\mathbf{x}_k; \boldsymbol{\alpha})$$

- $\hat{p}_k = g(\mathbf{x}_k; \hat{\boldsymbol{\alpha}})$: Comment estimer $\boldsymbol{\alpha}$ de telle sorte que $E\left(\hat{\theta}^{PS} - \theta \middle| \mathbf{Y}, \mathbf{X}\right) \approx 0$?

Pondération par l'inverse …

- **Maximum de vraisemblance (logistique):**
  - $\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$
  - Requiert de connaître $\mathbf{x}_k$ pour toute la population
- **Iannacchione, Milne et Folsom (1991):**
  - $\sum_{k \in s_{NP}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in U} \mathbf{x}_k = \mathbf{0}$
  - Requiert de connaître $\sum_{k \in U} \mathbf{x}_k$
  - Propriété de calage: $\sum_{k \in s_{NP}} w_k^{PS} \mathbf{x}_k = \sum_{k \in U} \mathbf{x}_k$

Pondération par l'inverse …

- **Chen, Li et Wu (2018):**
  - $\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$
  - Requiert de connaître $\mathbf{x}_k$ pour un échantillon probabiliste
- **Alternative:**
  - $\sum_{k \in s_{NP}} \frac{\mathbf{x}_k}{p_k(\boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k = \mathbf{0}$
  - Requiert de connaître $\sum_{k \in s_P} w_k \mathbf{x}_k$
  - Propriété de calage: $\sum_{k \in s_{NP}} w_k^{PS} \mathbf{x}_k = \sum_{k \in s_P} w_k \mathbf{x}_k$
  - Lesage (2017)

Pondération par l'inverse …

- Formation de classes homogènes par rapport à $\hat{p}_k$
  - Taux de participation dans la classe c: $\dfrac{n_c^{NP}}{\hat{N}_c}$
  - w_k^{PS} : inverse du taux de participation
  - Estimateur a la même forme que l'estimateur poststratifié
  - Robuste par rapport à une mauvaise spécification du modèle logistique (Haziza et Lesage, 2016)

Pondération par l'inverse …

- **Quelques remarques:**
  - Choix des variables auxiliaires (ou des classes homogènes) est clé pour réduire le biais de sélection
  - Arbres de régression?
  - Si $\Pr\left(\delta_k = 1 \middle| \mathbf{Y}, \mathbf{X}\right) = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right)$, on peut considérer le calage généralisé (Deville,1998):

$$\sum_{k \in s_{NP}} \frac{\mathbf{x}_k^I}{g(y_k, \mathbf{x}_k; \boldsymbol{\alpha})} - \sum_{k \in s_P} w_k \mathbf{x}_k^I = \mathbf{0}$$

  - **Hypothèse**: $F(\boldsymbol{\delta} \mid \mathbf{Y}, \mathbf{X}, \mathbf{X}^I) = F(\boldsymbol{\delta} \mid \mathbf{Y}, \mathbf{X})$
  - Plus possible de former des classes homogènes

Estimation sur petits domaines

- Quand considérer l'EPD?
  - On veut des estimations pour des domaines qui contiennent peu d'unités échantillonnées dans l'enquête probabiliste
    - Problème de variance mais pas de biais
- Méthodes EPD
  - Compensent le manque de données observées dans un domaine par des **hypothèses de modèle** qui relient des **données auxiliaires** aux données de l'enquête

Estimation sur petits domaines

- Modèle de Fay-Herriot
  - On a m domaines disjoints (m grand)
  - Variables auxiliaires disponibles au niveau des domaines: $\mathbf{x}_d$
    - Ex.: Estimations d'une source non probabiliste
  - Inférences conditionnelles à $\mathbf{X}$
  - On veut prédire le total dans le domaine d: θ_d
  - Estimateur direct (enquête probabiliste): $\hat{\theta}_d$
  - **Modèle**: $E\left(\hat{\theta}_d \middle| \mathbf{X}\right) = \mathbf{x}_d'\boldsymbol{\beta}$

Estimation sur petits domaines

- Modèle de Fay-Herriot
  - BLUP de θ_d :

$$\hat{\theta}_d^{BLUP} = \gamma_d \hat{\theta}_d + (1-\gamma_d)\mathbf{x}_d'\hat{\boldsymbol{\beta}} \quad , \quad 0 \le \gamma_d \le 1$$

  - Si $\hat{\theta}_d$ est précis, γ_d est près de 1
  - Gains d'efficacité sont plus importants quand γ_d est près de 0 mais le risque de biais due à une mauvaise spécification du modèle est plus grand…
  - **Risque doit être contrôlé par une modélisation minutieuse du modèle**

Taille d'échantillon	Moyenne des Diff. Rel. Abs. entre estimations directes (EPA) et Recensement	Moyenne des Diff. Rel. Abs. entre estimations EPD et Recensement
28 plus petits domaines	70.4%	17.7%
28 suivants	38.7%	18.9%
28 suivants	26.2%	13.8%
28 suivants	20.9%	12.7%
28 plus grands domaines	13.2%	10.2%
Total	**33.9%**	**14.7%**

Conclusion

- Discuté de quelques méthodes qui:
  - Utilisent des données de sources non probabilistes
  - **Conservent un cadre d'inférence statistique valide**
  - Estimation de variance: pas discuté mais ne pose pas de difficultés particulières en général
- Pour les approches "model-based":
  - Essentiel de planifier du temps et des ressources à la modélisation (ex.: analyse des résidus du modèle, …)
  - Baker et al. (2013)

Conclusion

- Est-ce que les enquêtes probabilistes sont vouées à disparaître pour la production de statistiques officielles?
  - L'avenir à court et moyen terme réside dans l'intégration de données d'échantillons non probabilistes à des données d'enquêtes probabilistes
  - Certaines enquêtes sont de qualité douteuse (et pourraient être éliminées) mais ce n'est pas le cas de la plupart des enquêtes menées par Statistique Canada
  - Peut plutôt s'attendre à une réduction de leur utilisation pour contrôler les coûts et le fardeau