Inférence à partir d'échantillons non probabilistes

Jean-François Beaumont, Statistique Canada
Webinaire, Groupe Enquête, SFdS
16 novembre 2023

Delivering insight through data, for a better Canada

Un peu d'histoire sur les enquêtes probabilistes

- Jusqu'au début du 20è siècle les recensements sont privilégiés
  - **Coûteux (en argent et en temps)**
- **Une alternative:** tirer un échantillon de la population
  - **Comment?** Aléatoirement ou pas?
  - Nombreux débats … jusqu'à Neyman (1934)
  - Rao (2005); Bethlehem (2009)
- **Ensuite**, les enquêtes probabilistes sont devenues graduellement la norme dans les ONS

  Au Canada: 1ère Enquête (sur la pop. active) en 1945

Pourquoi les enquêtes probabilistes dans les ONS?

- La théorie de Neyman est attrayante:
  - Méthode objective pour tirer des échantillons
  - Inférence “design-based”: validité ne dépend pas d’hypothèses de modèle (approche non paramétrique)
- Quelques exemples percutants d’échantillons non probabilistes qui ont mené à des conclusions totalement erronées (ex.: sondage pré-électoral de 1936 aux U.S.A.)

Est-ce que les enquêtes probabilistes sont une panacée?

- Estimations instables quand n est petite
- Basées sur l'hypothèse que les erreurs non dues à l'échantillonnage sont négligeables
  - Des efforts considérables sont souvent employés pour minimiser les erreurs de non-réponse, de couverture et de mesure
- Imparfaites mais **généralement** reconnues comme une source fiable, sauf peut-être pour les cas où les erreurs non dues à l'échantillonnage deviennent dominantes
- Brick (2011)

Vent de changement

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

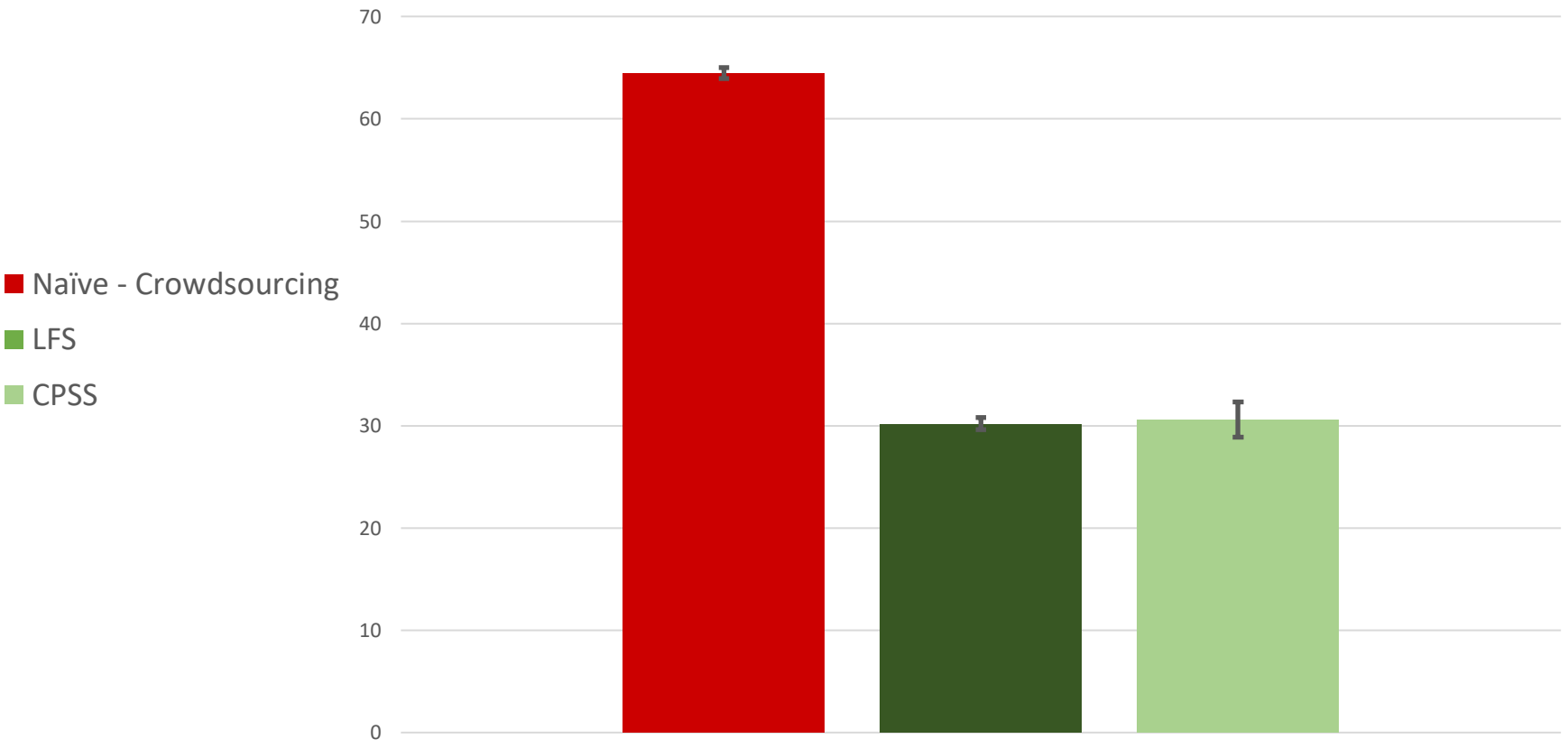
Est-ce que les enquêtes non probabilistes sont une panacée?

- Biais (sélection, couverture)
  - Devient dominant à mesure que n augmente (Meng, 2018)
  - Une grande taille d'échantillon n'est pas une garantie d'estimations de grande qualité
  - **Exemple:** sondage pré-électoral de 1936 aux É.-U. mené par la revue *Literary Digest* avec $n > 2,000,000$ et un échantillon fortement non représentatif de la population des voteurs
- Erreurs de mesure (ex.: enquêtes volontaires par panel Web)

Illustration du biais de sélection/couverture

- **Calculé des estimations de la proportion de personnes qui ont un diplôme universitaire au Canada selon 3 sources (Juin 2020):**
  - “Crowdsourcing” (échantillon non probabiliste avec 31,415 participants)
  - LFS (échantillon probabiliste avec 87,779 répondants et un taux de réponse autour de 70%)
  - CPSS (échantillon probabiliste avec 4,209 répondants et un taux de réponse autour de 15%)

Proportion de personnes qui ont un diplôme universitaire



Contexte

- **Objectif:** utiliser les données crowdsourcing pour réduire le fardeau, les coûts et le temps de production (en réduisant les efforts de collecte de données d'enquête)
- **Une question pertinente:**
  - Comment des données d'un échantillon non probabiliste peuvent-elles être utilisées pour produire des estimations fiables en les combinant avec des données existantes d'un échantillon probabiliste (qui ne contient pas les variables d'intérêt)?

Notation

- Paramètre de population: $\theta = \sum_{k \in U} y_k$
- Variable d'intérêt: $y_k \Rightarrow \mathbf{Y}$
- Échantillon non probabiliste: s_{NP}
  - Sous-ensemble de U
  - y_k est observé (sans erreur)
  - On observe également un vecteur de variables auxiliaires:

$$\mathbf{x}_k \Rightarrow \mathbf{X}$$

  - Indicateur d'inclusion dans s_{NP} : $\delta_k \Rightarrow \boldsymbol{\delta}$

Notation

- Échantillon probabiliste: s_P
  - Sous-ensemble de U tiré aléatoirement
  - Indicateur d'inclusion dans s_P : I_k ⇒ $\mathbf{I}$
  - Poids de sondage: w_k
  - Ne contient pas y_k mais $\mathbf{x}_k$ est observé
  - Hypothèse: Estimations pondérées sont approximativement sans biais (biais non dus à l'échantillonnage sont petits)
- Approches d'intégration de données diffèrent selon ce qu'elles traitent comme étant fixe et aléatoire ($\mathbf{Y}$, $\mathbf{I}$ | $\boldsymbol{\delta}$, $\mathbf{X}$) ou ($\boldsymbol{\delta}$, $\mathbf{I}$ | $\mathbf{Y}$, $\mathbf{X}$)

Méthodes d'intégration de données

- Estimateur naïf: $\hat{\theta}^{NP} = N\sum_{k\in s_{NP}} y_k / n^{NP}$
  - Peut être très biaisé (Bethlehem, 2016)
- Objectif des méthodes d'intégration de données:
  - Réduire le biais au moyen d'un vecteur de variables auxiliaires $\mathbf{x}_k$ observé dans les deux échantillons
  - Trois méthodes seront présentées: Prédiction/Calage, Appariement statistique et Pondération par l'inverse de la probabilité
  - Requiert la validité d'hypothèses de modèle

Prédiction / Calage de s_{NP}

- **Idée** (Royall, 1970):
  - Modéliser la relation entre y_k et $\mathbf{x}_k$ à partir de 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 participation non informative**:
  - $F(\mathbf{Y} \mid \boldsymbol{\delta}, \mathbf{X}) = F(\mathbf{Y} \mid \mathbf{X})$
  - Clé pour enlever le biais
  - Plus l'information auxiliaire est riche, plus l'hypothèse est réaliste

Prédiction / 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: $\sum_{k\in s_{NP}} w_k^C \mathbf{x}_k = \mathbf{T}_\mathbf{x}$
- Propriété de calage: seulement pour un modèle linéaire
- Si $\mathbf{T}_\mathbf{x}$ n'est pas connu, il peut être remplacé par un estimateur sans biais (enquête prob.): $\hat{\mathbf{T}}_\mathbf{x} = \sum_{k\in s_P} w_k \mathbf{x}_k$

Prédiction / Calage de s_{NP}

- Un modèle linéaire n'est pas toujours approprié
  - Ex. 1: Variables d'intérêt catégoriques
  - Ex. 2: Domaines ($y_k = 0$ en dehors du domaine)
- **Calage sur le modèle** (Wu and Sitter, 2001):
  - Considère un modèle non linéaire: $E\left(y_k \middle| \mathbf{X}\right) = \mu_k = h(\mathbf{x}_k)$
  - Obtient des valeurs prédites $\hat{\mu}_k$
  - Cale: $\sum_{k \in s_{NP}} w_k^{MC} \begin{pmatrix} 1 \\ \hat{\mu}_k \end{pmatrix} = \sum_{k \in s_P} w_k \begin{pmatrix} 1 \\ \hat{\mu}_k \end{pmatrix}$
  - Peut être généralisé à plusieurs variables d'intérêt

Appariement statistique

- **Idée**:
  - Modéliser la relation entre y_k et $\mathbf{x}_k$ à partir de 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 participation non informative
- Prédicteur du total θ : $\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp}$

Appariement statistique

- Pour un modèle linéaire, l'appariement statistique est équivalent dans la plupart des cas au calage de s_{NP} sur les totaux estimés $\hat{\mathbf{T}}_{\mathbf{x}}$
- On considère souvent l'imputation par donneur
  - Rivers (2007): appariement d'échantillon (sample matching)
  - Méthode non paramétrique
- Yang, Kim and Hwang (2021)

Appariement statistique

- **Imputation linéaire:** $y_k^{imp} = \sum_{l \in s_{NP}} \omega_{kl} y_l \ , \ k \in s_P$
  - Cas particuliers: Régression linéaire, donneur, …
  - Beaumont et Bissonnette (2011)
  - $\hat{\theta}^{SM}$ peut être ré-écrit 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?**
  - Quel contenu est d'intérêt? Le contenu de la source non probabiliste ou de l'enquête probabiliste?

Illustration avec des données réelles

- S_P : Enquête sur la santé dans les communautés canadiennes (ESCC)
- S_{NP} : Grand panel web de volontaires
- Variables d'intérêt sont observées dans les deux échantillons
  - On compare les estimations par calage et par appariement statistique avec les estimations de ESCC
- Variables auxiliaires: région, âge, sexe, état civil et niveau d'instruction
- Calage: effets principaux et quelques interactions
- Appariement: Imputation par donneur "le plus proche"
- Chatrchi, Beaumont, Gambino and Haziza (2018)

Variable	Estimations de proportions			
	ESCC (±1.96*s.e.)	Naïf	Calage	Appariement
Haute pression	19.3% (±0.8%)	14.3%	22.1%	28.6%
Très fort sentiment d'appartenance à la communauté	19.5% (± 0.8%)	8.4%	10.9%	14.8%
Plutôt faible sentiment d'appurtenance à la communauté	22.1% (± 1.0%)	36.4%	33.6%	30.2%
Excellente santé	23.3% (±0.9%)	7.8%	8.9%	11.7%
Très bonne santé	35.9% (±1.0%)	29.4%	33.8%	33.0%
Excellente santé mentale	33.5% (±1.1%)	13.7%	17.0%	21.4%
Senté mentale acceptable	6.0% (±0.5%)	17.1%	13.1%	11.4%

Pondération par l'inverse de probabilité

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

$$\sum_{k \in s_{NP}} w_k^{IPW,CAL} \tilde{\mathbf{x}}_k = \sum_{k \in s_P} w_k \tilde{\mathbf{x}}_k$$

Pondération par l'inverse de probabilité

- **Hypothèses:**
  - Participation 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 param. (ex.: logistique): $p_k(\boldsymbol{\alpha}) = \left[1 + \exp(-\mathbf{x}_k'\boldsymbol{\alpha})\right]^{-1}$
- Probabilité estimée: $\hat{p}_k = p_k(\hat{\boldsymbol{\alpha}})$
- Comment estimer $\boldsymbol{\alpha}$ tel que: $E\left(\hat{\theta}^{IPW} - \theta \middle| \mathbf{Y}, \mathbf{X}\right) \approx 0$

Pondération par l'inverse de probabilité

- Si $s_P = U$, on connaît $\mathbf{x}_k$ pour toute la population et on peut appliquer la méthode du **maximum de vraisemblance**:
  - Modèle logistique:

$$\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$$

  - Problème semblable à la pondération pour la non-réponse dans une enquête probabiliste

- Quoi faire si $\mathbf{x}_k$ est connu seulement dans s_P et s_{NP} ?

Pondération par l'inverse de probabilité

- **Chen, Li and Wu (2020):** Pseudo maximum de vraisemblance

$$\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$$

  - Il faut connaître $\mathbf{x}_k$ dans les deux échantillons
  - Pas besoin de connaître δ_k , $k \in s_P$
  - Une solution peut ne pas exister
  - Idée ingénieuse et valide mais inefficace. **Pourquoi?**

Pondération par l'inverse de probabilité

- Une équation d'estimation plus efficace peut être obtenue en utilisant les $\mathbf{x}_k$ observés dans s_P et s_{NP}
- Beaumont et al. (2024b) considèrent le meilleur estimateur linéaire sans biais (BLUE) du total $\sum_{k\in U} p_k(\boldsymbol{\alpha})\mathbf{x}_k$
- L'équation d'estimation sans biais qui en résulte prend la forme générale suivante (pour un modèle logistique):

$$\sum_{k\in s_{NP}} \lambda\left[\pi_k, p_k(\boldsymbol{\alpha})\right]\mathbf{x}_k - \sum_{k\in s_P} w_k\, p_k(\boldsymbol{\alpha})\lambda\left[\pi_k, p_k(\boldsymbol{\alpha})\right]\mathbf{x}_k = \mathbf{0}$$

Pondération par l'inverse de probabilité

- Plusieurs méthodes proposées dans la littérature sont un cas particulier de l'équation d'estimation générale

Méthode	$\lambda\left[\pi_k, p_k(\boldsymbol{\alpha})\right]$
Pseudo-ML (Chen, Li and Wu, 2020)	1
Pseudo-ILR (Gershunskaya and Beresovsky, 2024; Wang, Valliant and Li, 2021)	$\dfrac{1-p_k(\boldsymbol{\alpha})}{1+p_k(\boldsymbol{\alpha})}$
ILR (Implicit Logistic Reg.) (Beresovsky, 2019; Elliott, 2009)	$\dfrac{\pi_k\left[1-p_k(\boldsymbol{\alpha})\right]}{\pi_k+p_k(\boldsymbol{\alpha})}$
BLUE (optimale) (Beaumont et al., 2024b)	$\dfrac{\pi_k\left[1-p_k(\boldsymbol{\alpha})\right]}{\pi_k\left[1-p_k(\boldsymbol{\alpha})\right]+p_k(\boldsymbol{\alpha})\left[1-\pi_k\right]}$

Pondération par l'inverse de probabilité

- Remarques:
  - Toutes les méthodes sont équivalentes si $p_k(\boldsymbol{\alpha})/\pi_k$ sont en majorité petits
  - Si l'échantillon probabiliste est petit par rapport à l'échantillon non probabiliste, les méthodes Pseudo-ML et Pseudo-ILR sont inefficaces (ex.: Savitsky et al., 2022)
  - Si $s_P = U$, seulement les méthodes Pseudo-ML et optimales (BLUE) sont équivalentes à la méthode du maximum de vraisemblance

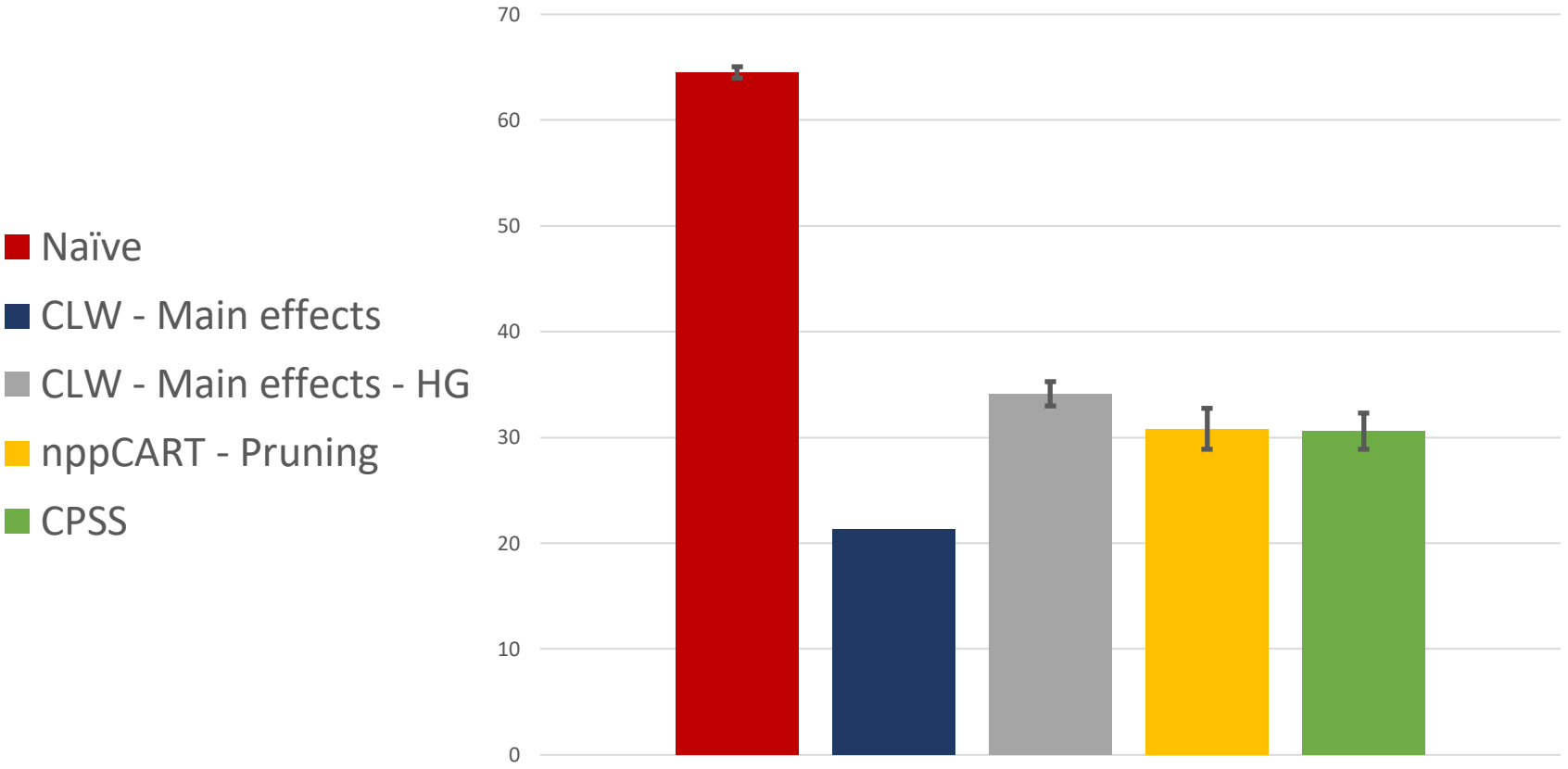
Pondération par l'inverse de probabilité

- En pratique, on va préférer créer des groupes homogènes par rapport à $\hat{p}_k^{\text{logistic}}$:
  - Robustesse par rapport à une mauvaise specification du modèle logistique (Haziza and Lesage, 2016)
  - Évite l'instabilité causée par des probabilités estimées très petites
- nppCART (Beaumont et al., 2024a)
  - version de CART qui tient compte de la structure des données et du plan de sondage probabiliste (fonction R disponible)

Illustration

- Échantillon non probabiliste: Crowdsourcing (31,415 participants)
- Échantillon probabiliste: CPSS (4,209 répondants et un taux de réponse autour de 15%)
- Variables auxiliaires: Niveau d'instruction (8), région (56), âge (13), sexe (2), statut d'immigration (3), statut d'emploi (3), état civil (6) et taille du ménage (6)
- Les deux échantillons contiennent les variables auxiliaires et les variables d'intérêt. Donc, CPSS peut être utilisée comme référence pour les comparaisons

Proportion de personnes ayant un diplôme universitaire



Conclusions principales

- Importance de créer des groupes homogènes pour réduire les poids extrêmes
- nppCART a bien performé avec ces données
- Méthodes de pondération réduisent le biais mais parfois un biais significatif persiste (tout comme le calage et l'appariement statistique)
- Un échantillon probabiliste grand (LFS) est plutôt bénéfique et atténue les différences entre les méthodes
- Effets principaux (en particulier, le niveau d'instruction) sont plus importants que les interactions avec ces données

Conclusion

- Présenté trois méthodes d'intégration de données qui:
  - Ne requièrent pas d'observer les variables d'intérêt dans un échantillon probabiliste
    - Réduction des coûts et du temps de production
  - Requièrent la validité d'hypothèses de modèle
    - Essentiel de planifier suffisamment de temps et de ressources à la modélisation (ex.: analyses de résidus, …): Baker et al. (2013)
  - Peuvent réduire le biais mais pas toujours l'éliminer
- Estimation de la variance: Pas discuté mais des méthodes existent (en supposant que le biais est éliminé)

Conclusion

- Ces méthodes devraient-elles être utilisées pour la production de statistiques officielles?
  - Avantages principaux:
    - Réduction des coûts et du fardeau sur les répondants
    - Réduction de temps de production
  - Inconvénient principal:
    - Plus grand risque de biais (sauf si les hypothèses sont raisonnables … mais difficile à vérifier)
  - Ça dépend des objectifs et à quel point un faible biais est important en comparaison des coûts et du temps de production

Références citées

- **Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K., and Tourangeau, R. (2013).** Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, 1, 90-143.
- **Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J., and Chu, K. (2024a).** Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 49 (to appear in the June 2024 issue).
  - **Gershunskaya, J., and Beresovsky, V. (2024).** Comments on Beaumont et al. (2024a). *Survey Methodology*, 49 (to appear in the June 2024 issue).
  - **Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J., and Chu, K. (2024b).** Response to comments on Beaumont et al. (2024a). *Survey Methodology*, 49 (to appear in the June 2024 issue).

Références citées

- **Beaumont, J. F., and Bissonnette, J. (2011).** Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology*, 37, 171-179.
- **Beresovsky, V. (2019).** On application of a response propensity model to estimation from web samples. In ResearchGate.
- **Bethlehem, J. (2009).** The rise of survey sampling. Discussion paper (09015), Statistics Netherlands, The Hague.
- **Bethlehem, J. (2016).** Solving the nonresponse problem with sample matching. *Social Science Computer Review*, 34, 59-77.
- **Brick, J. M. (2011).** The future of survey sampling. *Public Opinion Quarterly*, 75, 872-888.

Références citées

- **Chatrchi, G., Beaumont, J.-F., Gambino, J. and Haziza, D. (2018).** An investigation into the use of sample matching for combining data from probability and non-probability samples. *Proceedings of the Survey Methods Section*, Statistical Society of Canada.
- **Chen, Y., Li, P., and Wu, C. (2020).** Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association,* 115, 2011-2021.
- **Elliott, M. R. (2009).** Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813–845.
- **Haziza, D., and Lesage, É. (2016).** A discussion of weighting procedures for unit nonresponse. *Journal of Official Statistics*, 32, 129-145.
- **Lumley, T., and Scott, A. (2015).** AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology*, 3, 1-18.

Références citées

- **Meng, X.-L. (2018).** Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12, 685-726.
- **Neyman, J. (1934).** On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-625.
- **Rao, J.N.K. (2005).** Interplay between sample survey theory and practice: an appraisal. *Survey Methodology*, 31, 117-138.
- **Rivers, D. (2007).** Sampling from web surveys. In *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- **Royall, R. M. (1970).** On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

Références citées

- **Savitsky, T. D., Williams, M.R., Gershunskaya, J., Beresovsky, V., and Johnson, N.G. (2022).** Methods for combining probability and nonprobability samples under unknown overlaps. https://doi.org/10.48550/arXiv.2208.14541.
- **Wang, L., Valliant, R., and Li, Y. (2021).** Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.
- **Wu, C., and Sitter, R.R. (2001).** A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96, 185-193.
- **Yang, S., Kim, J.K. and Hwang, Y. (2021).** Integration of data from probability surveys and big found data for finite population inference using mass imputation. *Survey Methodology*, 47, 29-58.

Références: Articles de synthèse

- **Beaumont, J.-F. (2020).** Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1-28.
- **Elliott, M., and Valliant, R. (2017).** Inference for non-probability samples. *Statistical Science*, 32, 249-264.
- **Lohr, S. (2021).** Multiple-frame surveys for a multiple-data-source world. *Survey Methodology*.
- **Lohr, S., and Raghunathan, T.E. (2017).** Combining survey data with other data sources. *Statistical Science*, 32, 293-312.
- **Rao, J. N. K. (2021).** On making valid inferences by integrating data from surveys and other sources. *Sankhya B*, *83*, 242-272.
- **Valliant (2020).** Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, 8, 231-263.

Références: Articles de synthèse

- **Wu, C. (2022).** Statistical inference with non-probability survey samples. *Survey Methodology*, *48*, 283-311 (with discussion).
- **Yang, S., and Kim, J. K. (2020).** Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 1-26.