Making inferences from non-probability samples through data integration

or

Are probability surveys bound to disappear for the production of official statistics?

Jean-François Beaumont, Statistics Canada

IASS Webinar

January 26, 2022

Delivering insight through data, for a better Canada

A quick history of probability surveys

- Up to the beginning of the 20th century, censuses are the preferred tool
  - **Costly (in terms of money and time)**
- **An alternative:** draw a sample from the population
  - **How?** Random or not?
  - Many debates … until Neyman (1934)
  - Rao (2005); Bethlehem (2009)
- **Then**, probability surveys gradually became the standard in National Statistical Offices

  In Canada: First Labour Force Survey in 1945

Why probability surveys for official statistics?

- Neyman's theory is attractive:
  - Objective method for drawing samples
  - Design-based inference: validity does not depend on model assumptions (nonparametric approach)
- Some striking examples of nonprobability samples that led to dramatically wrong conclusions (ex.: 1936 U.S. pre-electoral poll)

Are probability surveys a panacea?

- Unreliable estimates when n is small
- Based on the assumption that nonsampling errors are negligible
  - Many resources are used to minimize nonresponse, measurement and coverage errors
- Imperfect but **generally** known to be a reliable source except perhaps for cases where nonsampling errors become dominant
- Brick (2011)

Wind of change

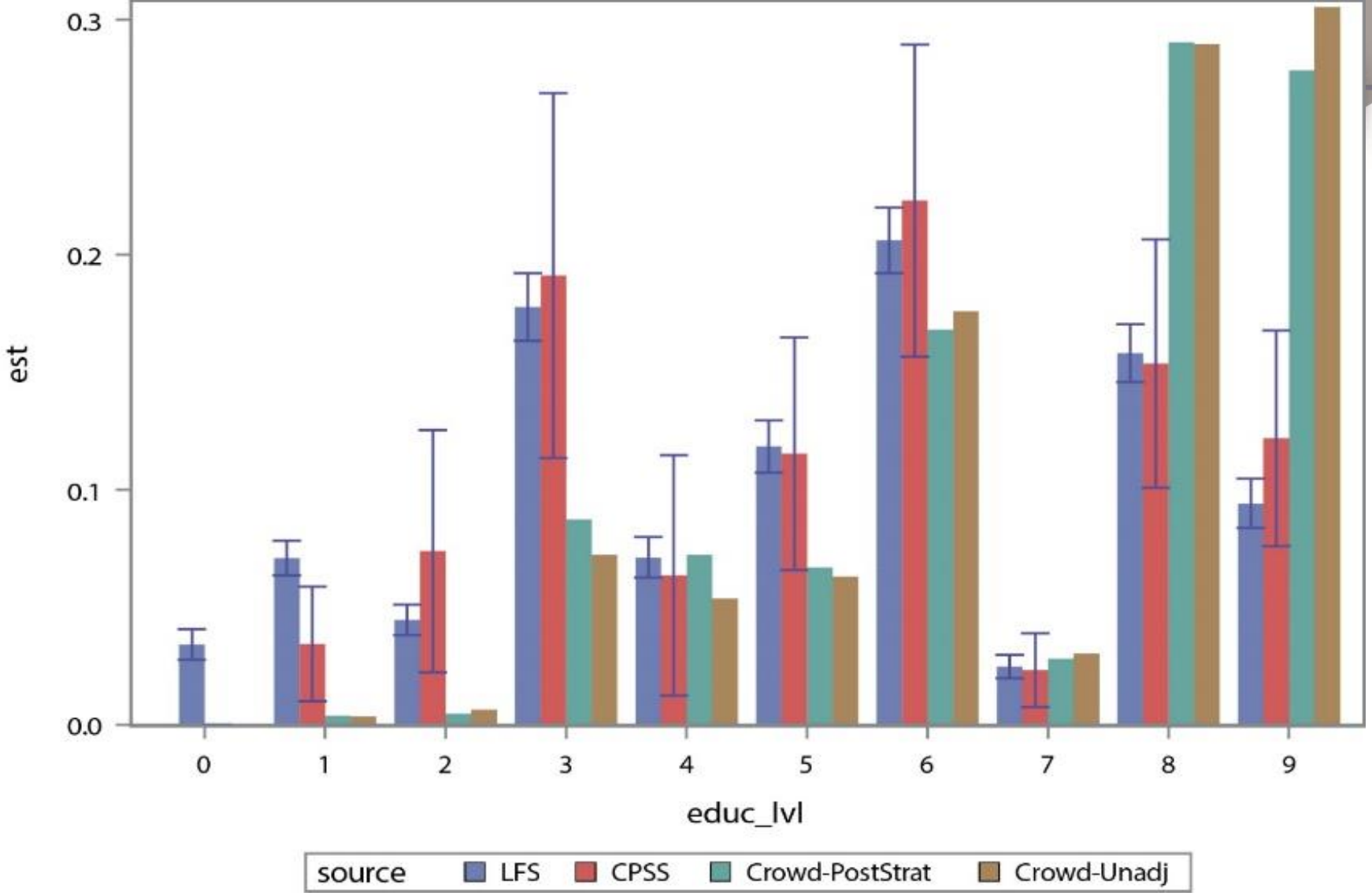
- Other types of data sources are increasingly considered
- **Four main reasons:**
  - Decline of survey response rates ➡ bias
  - High data collection costs + burden on respondents
  - Desire to have “real time” statistics (Rao, 2021)
  - Proliferation of nonprobability sources (ex.: Web panel surveys, administrative data, social medias, …)
    - Less costly, larger sample size

Are nonprobability surveys a panacea?

- Bias (selection, coverage)
  - Becomes dominant as the sample size n increases (Meng, 2018)
  - Large sample size is not a guarantee of high quality estimates...
  - **Example:** 1936 U.S. pre-electoral poll conducted by the magazine *Literary Digest* with $n > 2,000,000$ and highly nonrepresentative sample of the population of voters
- Measurement errors (ex.: Web panel surveys administered to volunteers)

An illustration of selection/coverage bias

- **Crowdsourcing experiments to obtain quickly information about the Canadian population**
  - Non-probability sample of volunteers who provide information through an online application
  - Bias accounted for through post-stratification weighting by province, age group and sex
- **Computed estimates of proportions in different education categories** (Beaumont and Rao, 2021)**:**
  - LFS estimates (probability survey with 88,000 respondents and response rate around 80%)
  - CPSS estimates (probability survey with 4,209 respondents and response rate around 15%)
  - Unadjusted crowdsourcing estimates (31,505 participants)
  - Post-stratified crowdsourcing estimates



Estimates of proportions in different education categories for a Canadian province

A relevant question in the current context

- How can data from a nonprobability source be used to
  - **minimize data collection costs and burden on respondents of a probability survey**
  - **while preserving a valid statistical inference framework and an acceptable quality?**
- **Statistical inference framework**: characterized by a **reference distribution** and a **list of assumptions**
  - Provides criteria for measuring the quality of estimates and make statistical inferences

In what follows …

- Review data integration methods
  - Background and notation
  - Design-based approaches
  - Model-based approaches
    - Calibration
    - Statistical matching
    - Inverse probability weighting
    - Small Area Estimation through the Fay-Herriot model
  - Some additional thoughts

Notation

- Population parameter: $\theta = \sum_{k \in U} y_k$
- Variable of interest: y_k ⟹ $\mathbf{Y}$
- Nonprobability sample: s_{NP}
  - Subset of U
  - Contains a variable y^* and possibly other variables
  - Indicator of inclusion in s_{NP} : δ_k ⟹ $\boldsymbol{\delta}$
- Two scenarios:
  - $y_k^* = y_k$
  - $y_k^* \neq y_k$: conceptual differences or measurement errors

Notation

- Probability sample: s_P
  - Subset of U randomly drawn with probability $p(s_P|\mathbf{Z})$
  - Indicator of inclusion in s_P : I_k ➡ $\mathbf{I}$
  - Inclusion probability: $\pi_k = \Pr\left(I_k = 1|\mathbf{Z}\right) > 0$
  - Contains or not the y variable
- $\boldsymbol{\Omega}$: Set of all the auxiliary data used to make inferences (including $\boldsymbol{Z}$)
- Approaches differ in what they treat as fixed and random ($\mathbf{I}$, $\boldsymbol{\delta}$, $\mathbf{Y}$, $\boldsymbol{\Omega}$)

Design-based inference

- Reference distribution: $F(\mathbf{I} \mid \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega})$
- For the estimation of the total $\theta = \sum_{k \in U} y_k$, estimators with a weighted form are often used:

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

- If $w_k = \pi_k^{-1}$ then $E\left(\hat{\theta} - \theta \middle| \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega}\right) = 0$
- **Alternative**: Calibration (Deville & Särndal, 1992): $\sum_{k \in s_P} w_k \mathbf{x}_k = \mathbf{T}_\mathbf{x}$
- No model assumption is required except for dealing with **nonsampling errors**

  - Assume that nonsampling biases are not too large (Brick, 2011)

Characteristics of design-based approaches

- The variable of interest must be **collected in the probability sample** and **measured without error**
- Role of nonprobability sample:
  - **Variance reduction**
  - **Sample size reduction** may be preferred to variance reduction if costs and burden must be reduced
- Small Area Estimation (SAE) has the same characteristics and is expected to yield **larger efficiency gains** but requires **model assumptions**

Scenario 1: $y_k^* = y_k$

- Context:
  - s_{NP} is a subset of U: **undercoverage**
  - The use of a probability sample allows us to get rid of the coverage bias
  - Generally, the larger the size of s_{NP}, the larger the variance reduction
- **Idea:**
  - Use data of the combined sample $s = s_P \cup s_{NP}$
  - Each unit $k \in s$ is weighted by $\left[\Pr(k \in s \mid \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega})\right]^{-1}$

Scenario 1: $y_k^* = y_k$

- Estimator:

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

  - δ_k must be available for $k \in s_P$
  - $E\left(\hat{\theta} - \theta \middle| \boldsymbol{\delta}, \mathbf{Y}, \boldsymbol{\Omega}\right) = 0$
  - Equivalent to the Bankier (1986) method for **multiple frame surveys**: Here the two frames are U and s_{NP} (see also Kim and Tam, 2020 ; Lohr, 2021)
  - The estimator can be improved by replacing weights π_k^{-1} with calibrated weights
  - **Efficiency gains are modest unless** s_{NP} is so large that **the overlap between both samples is not small**

Scenario 2: $y_k^* \neq y_k$

- y_k^* cannot be used as a replacement of y_k ; only as auxiliary variable
- Vector of auxiliary variables: $\mathbf{x}_k^*, \ k \in s_{NP}$
- Total: $\mathbf{T}_{\mathbf{x}^*} = \sum_{k \in s_{NP}} \mathbf{x}_k^* = \sum_{k \in U} \delta_k \mathbf{x}_k^*$
- **Calibration**: Find weights $w_k, \ k \in s_P$ such that

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

- $\delta_k \mathbf{x}_k^*$ must be available for $k \in s_P$ ⟹ May need **linkage** or **a few more questions** in the probability survey

Scenario 2: $y_k^* \neq y_k$

- Studied in Kim and Tam (2020)
- Efficiency gains are again modest unless the overlap between both samples is not small
- **Possible application**: Unemployment estimation
  - Probability survey collects the employment status: y_k
  - Administrative files contain employment insurance beneficiaries
  - The probability survey must contain the employment insurance status

Model-based approaches: Cal. , SM and IPW

- **Objective:**
  - Reduce burden and costs by eliminating collection of some variables of interest in s_P : y_k is not observed in s_P
- Assumption: $y_k^* = y_k$
- **Naïve estimator**: $\hat{\theta}^{NP} = N \sum_{k \in s_{NP}} y_k / n^{NP}$
  - Can be very biased (Bethlehem, 2016)
- **Objective of Calibration, SM and IPW:**
  - Bias reduction through a vector of auxiliary variables $\mathbf{x}_k$ observed in both samples
  - Require the validity of model assumptions

Calibration of s_{NP}

- **Idea** (Royall, 1970):
  - Model the relationship between y_k and $\mathbf{x}_k$ by using a nonprobability sample
  - Predict y_k for units $k \in U - s_{NP}$
- **Inferences**: conditional on $\boldsymbol{\delta}$ and $\mathbf{X}$
- **Noninformative selection/participation assumption**:
  - $F(\mathbf{Y} \mid \boldsymbol{\delta}, \mathbf{X}) = F(\mathbf{Y} \mid \mathbf{X})$
  - Key to removing bias
  - The richer the auxiliary information, the more realistic the assumption

Calibration of S_{NP}

- Linear model: $E\left(y_k \middle| \mathbf{X}\right) = \mathbf{x}_k'\boldsymbol{\beta}$
- BLUP of the total θ : $\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} y_k + \sum_{k \in U - s_{NP}} \mathbf{x}_k'\hat{\boldsymbol{\beta}}$
- Can be rewritten as: $\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} w_k^C y_k$
- The calibration weight satisfies: $\sum_{k \in s_{NP}} w_k^C \mathbf{x}_k = \mathbf{T}_\mathbf{x}$
- Calibration property only for a linear model
- If $\mathbf{T}_\mathbf{x}$ is unknown, it can be replaced with an unbiased estimator (probability survey): $\hat{\mathbf{T}}_\mathbf{x} = \sum_{k \in s_P} w_k \mathbf{x}_k$

Calibration of S_{NP}

- BLUP is unbiased if noninformative selection/participation assumption holds:

$$E\left(\hat{\theta}^{BLUP} - \theta \middle| \boldsymbol{\delta}, \mathbf{X}\right) = 0$$

- **Reduction of selection bias:**
  - Consider a large number of auxiliary variables
  - A large probability survey can be useful to obtain estimates of auxiliary totals
  - Variable selection methods (LASSO, …)
    Chen, Valliant and Elliott (2018)

Calibration of s_{NP}

- **Post-stratification model**:
  - $E\left(y_k \middle| \mathbf{X}\right) = \mu_h \quad , \quad k \in U_h$
  - Post-strata can be obtained by crossing many categorical variables
  - BLUP of the total θ : $\hat{\theta}^{BLUP} = \sum_{h=1}^{H} N_h \hat{\mu}_h$
- **Reduction of selection bias:**
  - Consider a large number of post-strata
  - Regression trees can be useful
  - Alternative: Multilevel Regression and Post-stratification (MRP)

Calibration of S_{NP}

- Idea behind MRP (Gelman and Little, 1997):
  - Form a very large number of post-strata by crossing many categorical variables
  - $\hat{\mu}_h$ may be unstable (small sample size)
  - Idea is to use a multilevel model to obtain more stable estimators of μ_h (or small area estimation model)
  - MRP estimator: $\hat{\theta}^{MRP} = \sum_{h=1}^{H} N_h \tilde{\mu}_h$
  - **Issue:** Population size in each post-stratum must be available
  - Is it more efficient than a simple post-stratification where post-strata are determined using regression trees?

Calibration of s_{NP}

- Linear model is not always appropriate
  - Ex. 1: Categorical variables of interest
  - Ex. 2: Domain estimation ($y_k = 0$ outside the domain)
- **Model Calibration** (Wu and Sitter, 2001):
  - Use a nonlinear model: $E\left(y_k \middle| \mathbf{X}\right) = \mu_k = h(\mathbf{x}_k)$
  - Obtain predicted values $\hat{\mu}_k$
  - Calibrate: $\sum_{k \in s_{NP}} w_k^{MC} \begin{pmatrix} 1 \\ \hat{\mu}_k \end{pmatrix} = \begin{pmatrix} \hat{N} \\ \hat{T}_{\hat{\mu}} \end{pmatrix}$
  - Can be generalized to multiple variables of interest

Statistical matching

- **Idea**:
  - Model the relationship between y_k and $\mathbf{x}_k$ using the nonprobability sample
  - Predict (impute) y_k in a probability sample that contains the auxiliary variables
- **Inferences**: conditional on $\boldsymbol{\delta}$ and $\mathbf{X}$
- Noninformative selection/participation assumption
- Predictor of the total θ : $\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp}$
- Unbiased if: $E\left(y_k^{imp} - y_k \middle| \boldsymbol{\delta}, \mathbf{X} \right) = 0$

Statistical matching

- For a linear model, statistical matching is equivalent in most cases to calibration of s_{NP} on estimated totals $\hat{\mathbf{T}}_{\mathbf{x}}$
  - Ex.: post-stratification model
- Donor imputation is often considered
  - Rivers (2007): Sample matching
  - Nonparametric method
- Yang, Kim and Hwang (2021)

Statistical matching

- **Linear imputation:** $y_k^{imp} = \sum_{l \in s_{NP}} \omega_{kl} y_l$
  - Beaumont and Bissonnette (2011)
  - Special cases: Linear regression, donor, …
  - $\hat{\theta}^{SM}$ can be rewritten in a weighted form:

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

- **To weight or to impute? Statistical matching or calibration?**
  - Which content is of interest? The content of the nonprobability source or the probability survey?

Empirical illustration

- S_P : Canadian Community Health Survey (CCHS)
- S_{NP} : Large web panel of volunteers
- Variables of interest are observed in both samples
  - Calibration and sample matching can be compared with CCHS estimates
- Auxiliary variables: health region, age, sex, marital status, and education
- Calibration: main effects and some interactions
- Sample matching: "Nearest" donor imputation
- Chatrchi, Beaumont, Gambino and Haziza (2018)

Variable	Estimates of proportions			
	CCHS (±1.96*s.e.)	Panel	Calibration	Sample Matching
High blood pressure	19.3% (±0.8%)	14.3%	22.1%	28.6%
Very strong sense of belonging to the community	19.5% (± 0.8%)	8.4%	10.9%	14.8%
Somewhat weak sense of belonging to the community	22.1% (± 1.0%)	36.4%	33.6%	30.2%
Excellent health	23.3% (±0.9%)	7.8%	8.9%	11.7%
Very good health	35.9% (±1.0%)	29.4%	33.8%	33.0%
Excellent mental health	33.5% (±1.1%)	13.7%	17.0%	21.4%
Fair mental health	6.0% (±0.5%)	17.1%	13.1%	11.4%

Inverse probability weighting

- **Idea:**
  - Model the relationship between δ_k and $\mathbf{x}_k$
  - Estimate the participation probability $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right)$ by $\hat{p}_k$
  - Estimator: $\hat{\theta}^{IPW} = \sum_{k \in s_{NP}} w_k^{IPW} y_k$, where $w_k^{IPW} = 1/\hat{p}_k$
- **Main advantage:**
  - Simplify the modelling effort when there are many variables of interest (only one participation indicator to model)
  - w_k^{IPW} can be further calibrated to improve precision:

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

Inverse probability weighting

- **Assumptions:**
  - Noninformative participation: $\Pr\left(\delta_k = 1 \middle| \mathbf{Y}, \mathbf{X}\right) = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right)$
  - $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right) > 0$
- Inferences: conditional on $\mathbf{Y}$ and $\mathbf{X}$
- Parametric model (ex.: logistic): $p_k(\boldsymbol{\alpha}) = \left[1 + \exp(-\mathbf{x}_k'\boldsymbol{\alpha})\right]^{-1}$
- Estimated probability: $\hat{p}_k = p_k(\hat{\boldsymbol{\alpha}})$
- How to estimate $\boldsymbol{\alpha}$ such that: $E\left(\hat{\theta}^{IPW} - \theta \middle| \mathbf{Y}, \mathbf{X}\right) \approx 0$

Inverse probability weighting

- **Maximum likelihood (logistic):**
  - $\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in U} p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$
  - Require knowing $\mathbf{x}_k$ for the entire population
  - Similar to weighting for survey nonresponse

Inverse probability weighting

- **Chen, Li and Wu (2020):** Pseudo Maximum Likelihood
  - $\sum_{k \in s_{NP}} \mathbf{x}_k - \sum_{k \in s_P} w_k p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$
  - A solution may not exist
  - Requires knowing $\mathbf{x}_k$ for both $k \in s_{NP}$ and $k \in s_P$
  - Does not require knowing δ_k , $k \in s_P$

Inverse probability weighting

- A simple alternative: Stack both samples and use weighted logistic regression
  - $\sum_{k \in s_{NP}} \phi_k^{NP} \left[1 - p_k(\boldsymbol{\alpha})\right] \mathbf{x}_k - \sum_{k \in s_P} w_k \, p_k(\boldsymbol{\alpha}) \mathbf{x}_k = \mathbf{0}$
  - Lee (2006); Valliant and Dever (2011)
  - Implicit assumption: n_{NP} / N is small
  - If assumption is reasonable and $\phi_k^{NP} = 1$ then the method is approximately equivalent to Chen, Li and Wu (2020)
- Another option (small n_{NP} / N): Elliott and Valliant (2017)

Inverse probability weighting

- Wang, Valliant and Li (2021)
  - Extension of Valliant and Dever (2011) to account for a large sampling fraction n_{NP}/N
  - Proposed a different estimating equation than Chen-Li-Wu
  - Participation probability: $p_k(\boldsymbol{\alpha}) = \exp(\mathbf{x}_k'\boldsymbol{\alpha})$ (not bounded)
  - Show significant efficiency gains compared with Chen-Li-Wu
  - **Why?**
  - If $\mathbf{x}_k = 1$ or only one categorical auxiliary variable: Both estimators are identical

Inverse probability weighting

- Creation of homogeneous groups with respect to $\hat{p}_k^{\text{logistic}}$ is common:
  - Robust with respect to a misspecification of the logistic model (Haziza and Lesage, 2016)
  - Avoids very small estimated probabilities
  - w_k^{IPW} for k in group g : $w_k^{IPW} = \dfrac{\hat{N}_g}{n_g^{NP}}$
  - Estimator has the same form as the post-stratified estimator
  - If homogeneous groups are used, both Chen-Li-Wu and Wang-Valliant-Li are expected to be roughly equivalent

Inverse probability weighting

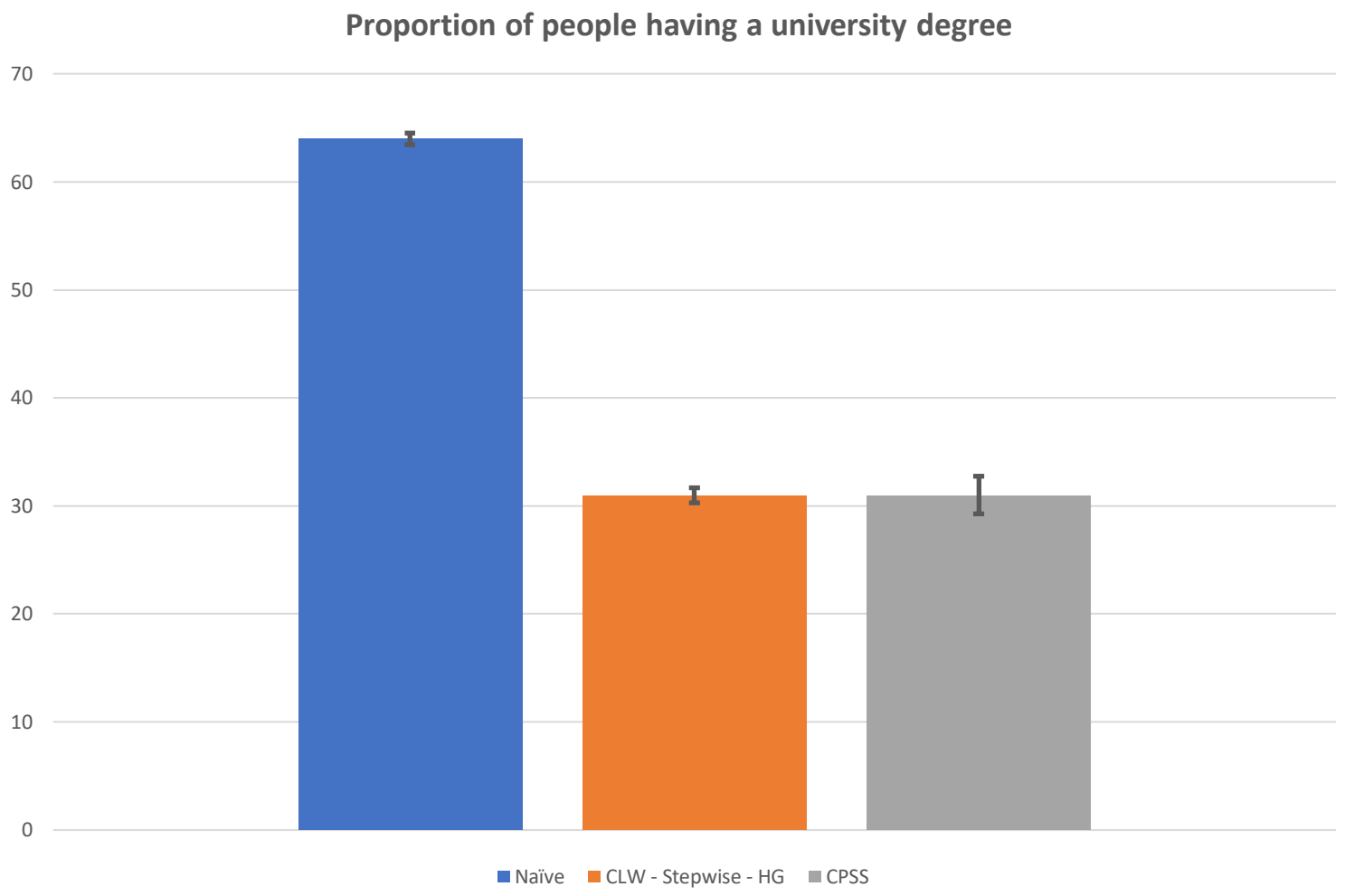
- Choice of auxiliary variables and interactions (or homogeneous groups) is key to reduce bias
- We are currently doing research and experimentations:
  - Variable selection: stepwise procedure that minimizes an AIC
  - CART (trees)
- Standard procedures cannot be used:
  - The pooled sample is not an i.i.d. sample
  - The probability sampling design must be taken into account

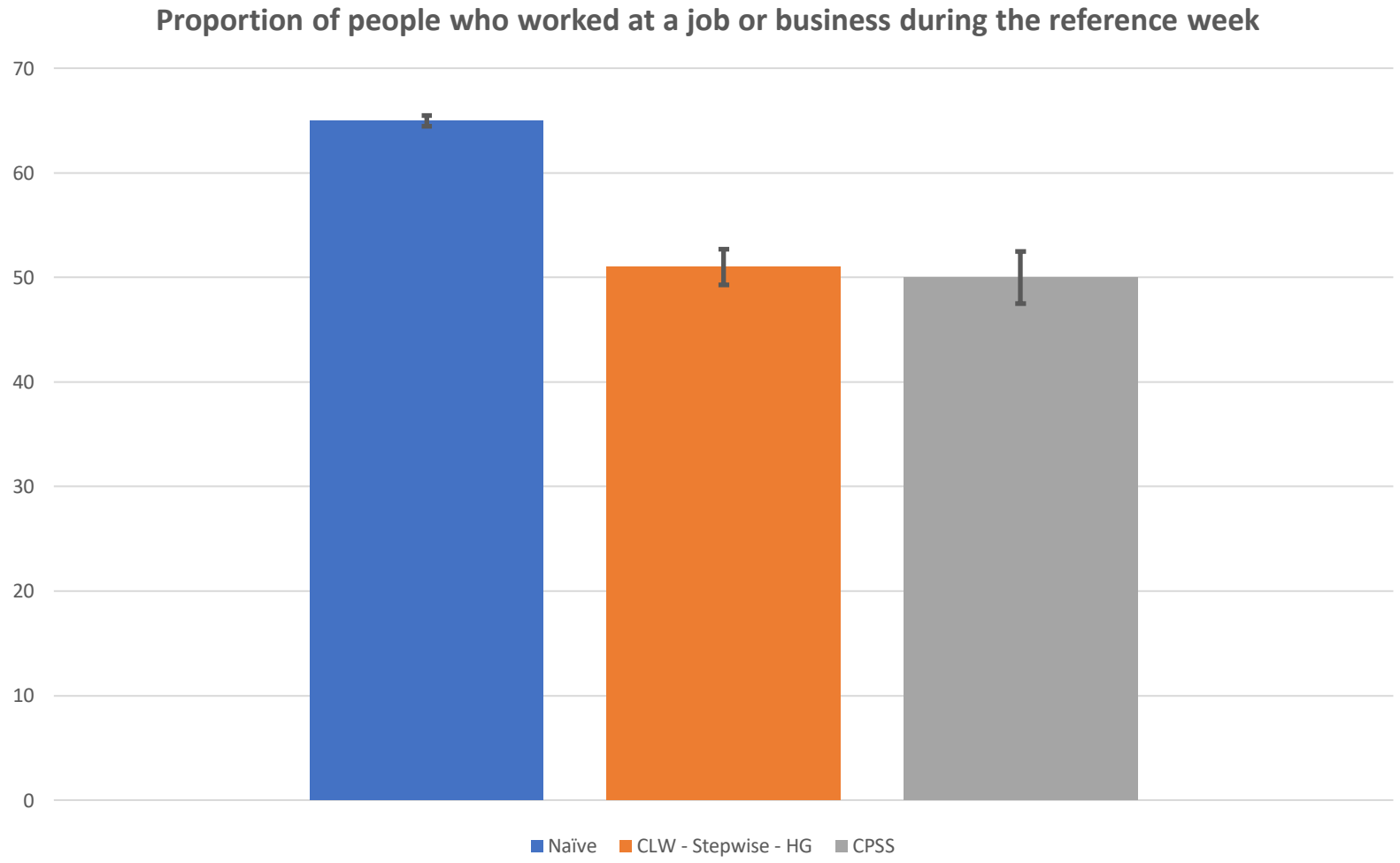
Inverse probability weighting

- Develop an AIC, similar to Lumley and Scott (2015), that penalizes the pseudo log likelihood for
  - The number of model parameters
  - The selection of a probability sample
- K-fold cross-validation could be an alternative:
  - Not straightforward: Requires to partition the probability sample carefully (like random groups method for variance estimation) and to repeat weighting adjustments (Wieczorek, 2019)

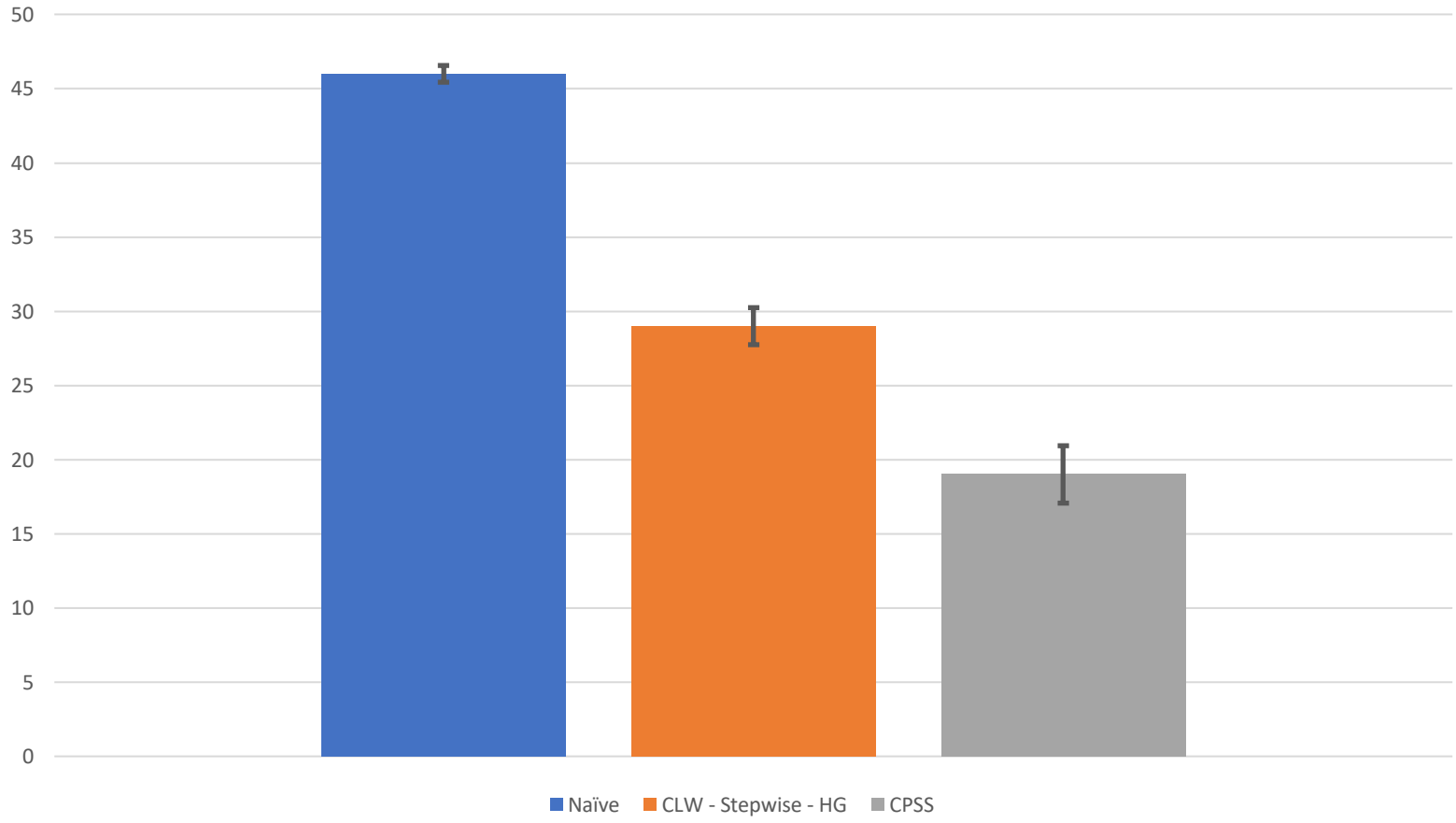
Inverse probability weighting

- Main conclusions of our experimentations using social data:
  - Main effects (educ., region, age, sex, immig., employ., marital, household size) are more important than first-order interactions to reduce the AIC
  - The variable Education is by far the most important to explain participation in a volunteer online survey (crowdsourcing)
  - AIC: the penalty for the selection of a probability sample is not negligible compared with the penalty for the number of model parameters
  - IPW methods reduce bias but sometimes a significant bias remains (like calibration and sample matching)

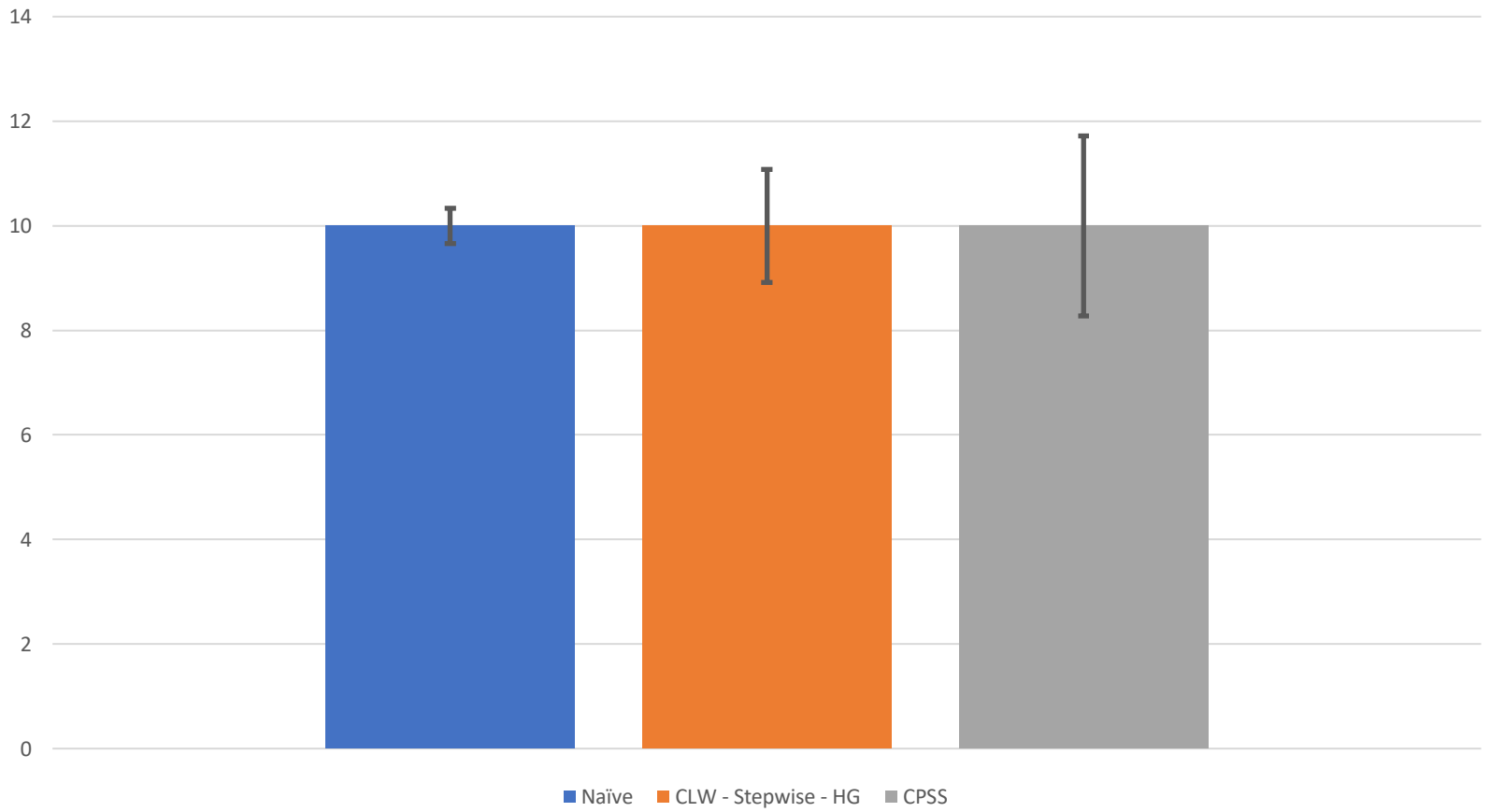

Proportion of people having a university degree
70
60
50
40
30
20
10
0
Naïve
CLW - Stepwise - HG
CPSS


Proportion of people who worked at a job or business during the reference week
70
60
50
40
30
20
10
0
Naïve
CLW - Stepwise - HG
CPSS

Proportion of people who worked most of their hours at home during the reference week



Proportion of people who “fear being a target for putting others at risk” because they do not always wear a mask in public



Small area estimation

- When to consider Small Area Estimation (SAE)?
  - The variable of interest is collected in a probability sample
  - The non-probability sample only provides auxiliary data
  - Domain estimates are desired but some domains contain a small probability sample size

→ Variance may be large for some domain estimates

- SAE methods
  - Compensate for the lack of observed data in a domain through **model assumptions** that link auxiliary data to survey data

Small area estimation

- Fay-Herriot model
  - m disjoint domains of interest (m not small)
  - Auxiliary variables $\mathbf{x}_d$ available at the domain level
    - Ex.: Estimates from a nonprobability source
  - We want to predict the total in domain d : θ_d
  - From s_P : Direct estimator: $\hat{\theta}_d$ (assumed unbiased)
  - **Model**: $\hat{\theta}_d = \mathbf{x}_d'\boldsymbol{\beta} + v_d + e_d$
  - Inferences conditional on $\mathbf{X}$

Small area estimation

- **Empirical Bayes (or EBLUP) of θ_d :**

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

  - If $\hat{\theta}_d$ is precise, $\hat{\gamma}_d$ should be close to 1
  - Efficiency gains tend to be larger when $\hat{\gamma}_d$ is close to 0 but risk of bias due to model misspecification is larger
  - Risk of bias can be controlled by careful modelling
- **Example:** Estimation of the unemployment rate by area
  - Direct estimate: Labour Force Survey
  - Auxiliary information: Administrative data

Sample size	Average of Abs. Rel. Dif. between direct estimates (LFS) and Census 2016 estimates	Average of Abs. Rel. Dif. between EB estimates and Census 2016 estimates
28 smallest areas	70.4%	17.7%
28 next smallest areas	38.7%	18.9%
28 next smallest areas	26.2%	13.8%
28 next smallest areas	20.9%	12.7%
28 largest areas	13.2%	10.2%
Total	**33.9%**	**14.7%**

Conclusion

- Presented a few methods that:
  - Use data from nonprobability sources
  - Preserve a “valid” statistical inference framework
  - Variance estimation: Not discussed but methods exist for most estimators presented
- For the model-based approaches:
  - Essential to plan sufficient time and resources for modelling (ex.: analyses of model residuals, …)
  - Baker et al. (2013)

Conclusion

- Are probability surveys bound to disappear for the production of official statistics?
  - The short and mid-term future is in the integration of data from probability and nonprobability samples
  - The quality of some surveys may be doubtful (and could be eliminated) but it is not the case of most surveys conducted by Statistics Canada
  - Can rather expect a reduction of their use to control burden and costs

Selected References

- Design-based approaches:
  - **Kim, J. K., and Tam, S. M. (2020).** Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*.
- Calibration of the non-probability sample:
  - **Chen, J.K.T., Valliant, R.L., and Elliott, M.R. (2018).** Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, 44, 117-144.
  - **Elliot, M., and Valliant, R. (2017).** Inference for non-probability samples. *Statistical Science*, 32, 249-264.
  - **Gelman, A., and Little, T.C. (1997).** Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, 23, 127-135.

Selected References

- Statistical Matching:
  - **Rivers, D. (2007).** Sampling from web surveys. In *Proceedings of the Survey Research Methods Section*, American Statistical Association.
  - **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.

Selected References

- Inverse probability weighting:
  - **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.
  - **Lee, S. (2006).** Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.
  - **Valliant, R., and Dever, J. A. (2011).** Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40, 105-137.
  - **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.

Selected References

- Small Area Estimation:
  - **Rao, J.N.K., and Molina, I. (2015).** *Small area estimation*. Second Edition, Wiley, Hoboken, NJ.
- Review papers:
  - **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*.

Selected References

- Review papers:
  - **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.
  - **Yang, S., and Kim, J. K. (2020).** Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science*, 1-26.

Other Cited References

- **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.
- **Bankier, M. D. (1986).** Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- **Beaumont, J. F., and Bissonnette, J. (2011).** Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology*, 37, 171-179.
- **Beaumont, J.-F., and Rao, J.N.K. (2021).** Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, 83, 11-22.

Other Cited References

- **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.
- **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.
- **Deville, J.-C., and Särndal, C.E. (1992).** Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

Other Cited References

- **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.
- **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.

Other Cited References

- **Rao, J.N.K. (2005).** Interplay between sample survey theory and practice: an appraisal. *Survey Methodology*, 31, 117-138.
- **Royall, R. M. (1970).** On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.
- **Wieczorek, J. (2019).** K-fold cross-validation for complex survey data. Poster presented at the 21st Meeting of New Researchers in Statistics and Probability, July 2019, Fort Collins, https://web.colby.edu/jawieczo/files/2019/08/NRC_Poster.pdf.
- **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.

Disclaimer

The content of this presentation represents the authors' opinions and not necessarily those of Statistics Canada.