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: \( y_k \rightarrow Y \)

• Nonprobability sample: \( s_{NP} \)
  • Subset of \( U \)
  • Contains a variable \( y^* \) and possibly other variables
  • Indicator of inclusion in \( s_{NP} \): \( \delta_k \rightarrow \delta \)

• 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 $p(s_P|Z)$
  • Indicator of inclusion in $s_P$: $I_k \rightarrow I$
  • Inclusion probability: $\pi_k = \Pr(I_k = 1|Z) > 0$
  • Contains or not the $\gamma$ variable

• $\Omega$: Set of all the auxiliary data used to make inferences (including $Z$)

• Approaches differ in what they treat as fixed and random ($I, \delta, Y, \Omega$)
Design-based inference

- Reference distribution: $F(\mathbf{I} | \delta, \mathbf{Y}, \Omega)$
- 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 $E\left(\hat{\theta} - \theta \mid \delta, \mathbf{Y}, \Omega\right) = 0$
- **Alternative**: Calibration (Deville & Särndal, 1992): $\sum_{k \in s_p} w_k x_k = \mathbf{T}_x$
- 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 $\left[ \Pr(k \in s \mid \delta, Y, \Omega) \right]^{-1}$
Scenario 1: \( y_k^* = y_k \)

- Estimator:
  \[
  \hat{\theta} = \sum_{k \in s_{NP}} y_k + \sum_{k \in s_p} \frac{1}{\pi_k} (1 - \delta_k) y_k
  \]

- \( \delta_k \) must be available for \( k \in s_p \)

- \( E\left(\hat{\theta} - \theta \mid \delta, Y, \Omega\right) = 0 \)

- 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 \( \pi_k^{-1} \) 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: \( x_k^* \), \( k \in s_{NP} \)

- Total: \( T_x = \sum_{k \in s_{NP}} x_k^* = \sum_{k \in U} \delta_k x_k^* \)

- **Calibration**: Find weights \( w_k \), \( k \in s_P \) such that

\[
\sum_{k \in s_P} w_k \begin{pmatrix} x_k \\ \delta_k x_k^* \end{pmatrix} = \begin{pmatrix} T_x \\ T_x^* \end{pmatrix}
\]

- \( \delta_k x_k^* \) must be available for \( k \in s_P \) \( \rightarrow \) 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 $x_k$ observed in both samples
  - Require the validity of model assumptions
Calibration of $S_{NP}$

- **Idea** (Royall, 1970):
  - Model the relationship between $y_k$ and $x_k$ by using a nonprobability sample
  - Predict $y_k$ for units $k \in U - s_{NP}$

- **Inferences**: conditional on $\delta$ and $X$

- **Noninformative selection/participation assumption**:
  - $F(Y | \delta, X) = F(Y | X)$
  - Key to removing bias
  - The richer the auxiliary information, the more realistic the assumption
Calibration of $S_{NP}$

- Linear model: $E\left(y_k \mid X\right) = x'_k \beta$

- BLUP of the total $\theta$: $\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} y_k + \sum_{k \in U - s_{NP}} x'_k \hat{\beta}$

- Can be rewritten as: $\hat{\theta}^{BLUP} = \sum_{k \in s_{NP}} w^C_k y_k$

- The calibration weight satisfies: $\sum_{k \in s_{NP}} w^C_k x_k = T_x$

- Calibration property only for a linear model

- If $T_x$ is unknown, it can be replaced with an unbiased estimator (probability survey): $\hat{T}_x = \sum_{k \in s_p} w_k x_k$
Calibration of $S_{NP}$

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

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

- 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:**

- $E\left( y_k \bigg| X \right) = \mu_h$, $k \in U_h$
- Post-strata can be obtained by crossing many categorical variables
- BLUP of the total $\theta$: $\hat{\theta}^{BLUP} = \sum_{h=1}^{H} N_h \hat{\mu}_h$

**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 $\mu_h$ (or small area estimation model)
  - MRP estimator: $\hat{\theta}^{MRP} = \sum_{h=1}^{H} N_h \hat{\mu}_h$
  - **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: $E\left(y_k \mid X\right) = \mu_k = h(x_k)$
  - 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 $x_k$ using the nonprobability sample
  - Predict (impute) $y_k$ in a probability sample that contains the auxiliary variables

• **Inferences:** conditional on $\delta$ and $X$

• **Noninformative selection/participation assumption**

• **Predictor of the total $\theta$:**
  \[
  \hat{\theta}^{SM} = \sum_{k \in s_p} w_k y_k^{imp}
  \]

• **Unbiased if:**
  \[
  E\left( y_k^{imp} - y_k \mid \delta, X \right) = 0
  \]
Statistical matching

• For a linear model, statistical matching is equivalent in most cases to calibration of $S_{NP}$ on estimated totals $\hat{T}_x$
  • 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)
<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimates of proportions</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CCHS (±1.96*s.e.)</td>
<td>Panel</td>
<td>Calibration</td>
<td>Sample</td>
</tr>
<tr>
<td>High blood pressure</td>
<td>19.3% (±0.8%)</td>
<td>14.3%</td>
<td>22.1%</td>
<td>28.6%</td>
</tr>
<tr>
<td>Very strong sense of belonging to the community</td>
<td>19.5% (±0.8%)</td>
<td>8.4%</td>
<td>10.9%</td>
<td>14.8%</td>
</tr>
<tr>
<td>Somewhat weak sense of belonging to the community</td>
<td>22.1% (±1.0%)</td>
<td>36.4%</td>
<td>33.6%</td>
<td>30.2%</td>
</tr>
<tr>
<td>Excellent health</td>
<td>23.3% (±0.9%)</td>
<td>7.8%</td>
<td>8.9%</td>
<td>11.7%</td>
</tr>
<tr>
<td>Very good health</td>
<td>35.9% (±1.0%)</td>
<td>29.4%</td>
<td>33.8%</td>
<td>33.0%</td>
</tr>
<tr>
<td>Excellent mental health</td>
<td>33.5% (±1.1%)</td>
<td>13.7%</td>
<td>17.0%</td>
<td>21.4%</td>
</tr>
<tr>
<td>Fair mental health</td>
<td>6.0% (±0.5%)</td>
<td>17.1%</td>
<td>13.1%</td>
<td>11.4%</td>
</tr>
</tbody>
</table>
Inverse probability weighting

• **Idea:**
  - Model the relationship between $\delta_k$ and $x_k$
  - Estimate the participation probability $p_k = \Pr(\delta_k = 1 | X)$ 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:
    $$\sum_{k \in s_{NP}} w_k^{IPW, CAL} \tilde{x}_k = \hat{T}_x$$
Inverse probability weighting

• **Assumptions:**
  
  • Noninformative participation: \( \Pr(\delta_k = 1|Y, X) = \Pr(\delta_k = 1|X) \)
  
  • \( p_k = \Pr(\delta_k = 1|X) > 0 \)

• **Inferences:** conditional on \( Y \) and \( X \)

• **Parametric model** (ex.: logistic): \( p_k(\alpha) = [1 + \exp(-x_k'\alpha)]^{-1} \)

• Estimated probability: \( \hat{p}_k = p_k(\hat{\alpha}) \)

• How to estimate \( \alpha \) such that: \( E\left( \hat{\theta}^{IPW} - \theta \middle| Y, X \right) \approx 0 \)
Inverse probability weighting

• **Maximum likelihood (logistic):**

\[ \sum_{k \in S_{NP}} x_k - \sum_{k \in U} p_k(\alpha)x_k = 0 \]

• Require knowing \( x_k \) for the entire population

• Similar to weighting for survey nonresponse
Inverse probability weighting

- **Chen, Li and Wu (2020):** Pseudo Maximum Likelihood
  \[
  \sum_{k \in s_{NP}} x_k - \sum_{k \in s_P} w_k p_k(\alpha) x_k = 0
  \]
  - A solution may not exist
  - Requires knowing \( x_k \) for both \( k \in s_{NP} \) and \( k \in s_P \)
  - Does not require knowing \( \delta_k, \ k \in s_P \)
Inverse probability weighting

• A simple alternative: Stack both samples and use weighted logistic regression

\[ \sum_{k \in s_{NP}} \phi_k^{NP} [1 - p_k(\alpha)] x_k - \sum_{k \in s_p} w_k p_k(\alpha) x_k = 0 \]

• 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: $p_k(\alpha) = \exp(x'_k \alpha)$ (not bounded)
  - Show significant efficiency gains compared with Chen-Li-Wu
  - Why?
    - If $x_k = 1$ 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
  
  • $w_k^{\text{IPW}}$ for $k$ in group $g$: $w_k^{\text{IPW}} = \frac{\hat{N}_g}{n_g^{\text{NP}}}$
  
  • 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

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 \textit{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 $x_d$ available at the domain level
    • Ex.: Estimates from a nonprobability source
  • We want to predict the total in domain $d$ : $\theta_d$
  • From $s_P$ : Direct estimator: $\hat{\theta}_d$ (assumed unbiased)
  • Model: $\hat{\theta}_d = x_d' \beta + v_d + e_d$
  • Inferences conditional on $X$
Small area estimation

- **Empirical Bayes (or EBLUP) of** $\theta_d$:

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

- 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
<table>
<thead>
<tr>
<th>Sample size</th>
<th>Average of Abs. Rel. Dif. between direct estimates (LFS) and Census 2016 estimates</th>
<th>Average of Abs. Rel. Dif. between EB estimates and Census 2016 estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>28 smallest areas</td>
<td>70.4%</td>
<td>17.7%</td>
</tr>
<tr>
<td>28 next smallest areas</td>
<td>38.7%</td>
<td>18.9%</td>
</tr>
<tr>
<td>28 next smallest areas</td>
<td>26.2%</td>
<td>13.8%</td>
</tr>
<tr>
<td>28 next smallest areas</td>
<td>20.9%</td>
<td>12.7%</td>
</tr>
<tr>
<td>28 largest areas</td>
<td>13.2%</td>
<td>10.2%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>33.9%</strong></td>
<td><strong>14.7%</strong></td>
</tr>
</tbody>
</table>
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:

• Calibration of the non-probability sample:
Selected References

• **Statistical Matching:**
  
  
Selected References

• **Inverse probability weighting:**
  


Selected References

- **Small Area Estimation:**

- **Review papers:**
Selected References

• Review papers:
  
  
  
  
Other Cited References


Other Cited References


Other Cited References


Other Cited References


Disclaimer

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