Handling non-probability sample data: An overview of issues and methods

Jean-François Beaumont, Statistics Canada

Webinar of the Inter-Secretariat Working Group on Household Surveys

December 9, 2025

Delivering insight through data, for a better Canada

Probability surveys

- After Neyman (1934), probability surveys gradually became the standard in National Statistical Offices
  - Example: First probability survey in Canada in 1945 (Labour Force Survey)
- Why?
  - Objective method for drawing samples
  - Nonparametric approach to inference (Design-based): validity does not depend on model assumptions
  - Some striking examples of nonprobability samples that led to dramatically wrong conclusions (ex.: 1936 U.S. pre-electoral poll)

Wind of change

- Other types of data sources are increasingly considered
- **Three main reasons:**
  - Decline of survey response rates ➡ bias
  - High cost of conducting probability surveys
  - Proliferation of nonprobability sources (ex.: Web panel surveys, administrative data, social media data, …)
    - Less costly, larger sample size, speed up the production of estimates

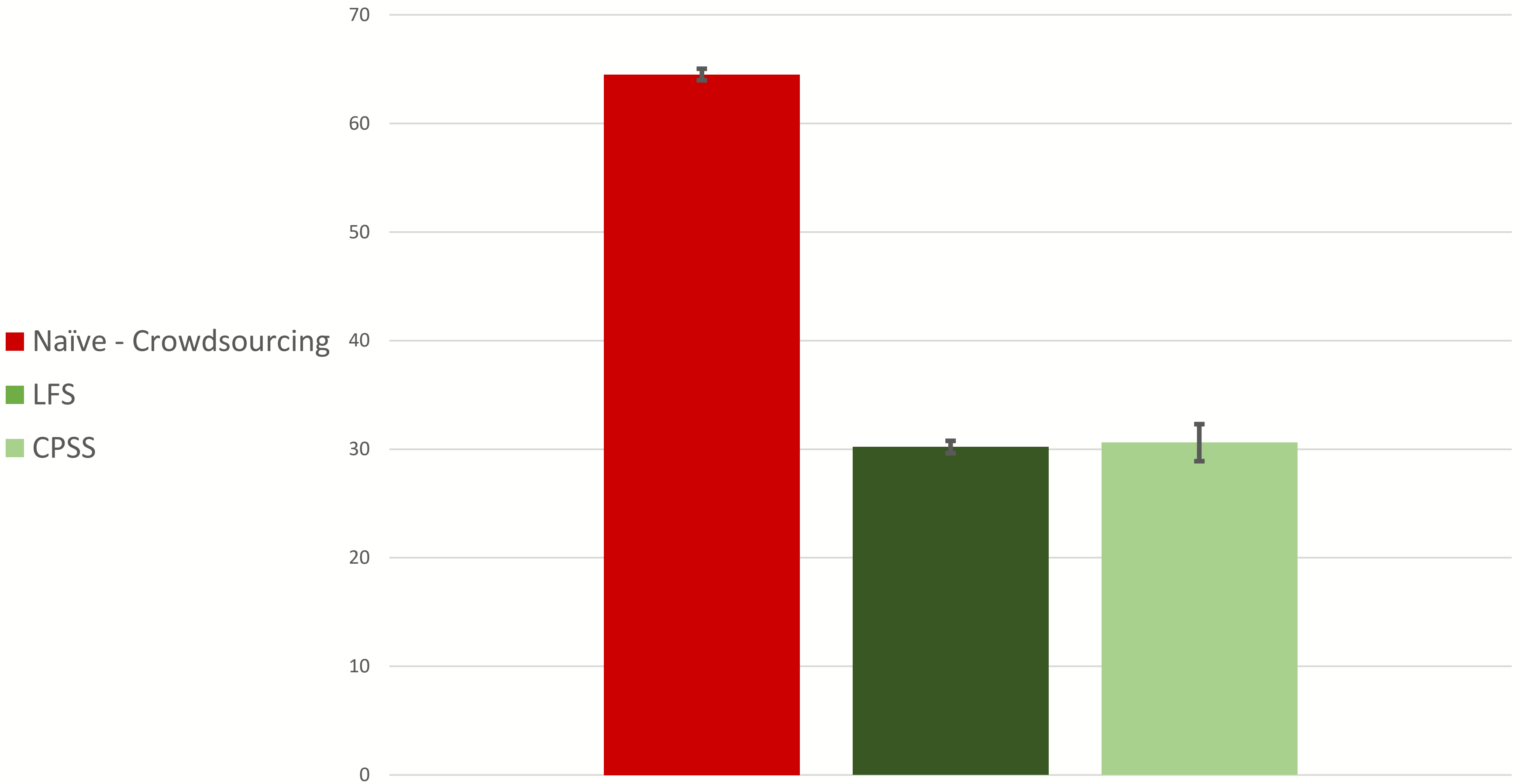
Are nonprobability surveys a panacea?

- “Representativity” Bias
  - Selection/Coverage bias
  - Large sample size is not a guarantee of high-quality estimates (Meng, 2018): does not address bias
- Measurement errors
  - Ex.: Online nonprobability surveys (Kennedy, Mercer and Lau, 2024)

Illustration of representativity bias

- **Computed estimates of the proportion of people having a university degree in Canada from three data sources (June 2020):**
  - Crowdsourcing sample (nonprobability sample with 31,415 participants)
  - LFS (probability sample with 87,779 respondents and response rate around 70%)
  - CPSS (probability sample with 4,209 respondents and response rate around 15%)

Proportion of people having a university degree



Why and how to use non-probability sample data?

- **Why?** To reduce costs, time and burden on survey respondents (by reducing survey data collection efforts)
- **A relevant question:**
  - How can data of a non-probability sample be used to produce accurate estimates ?
- A possible answer:
  - Through data integration methods: integration of nonprobability sample data with existing data from a probability sample (that does not contain the variables of interest)

Available data

- Population parameter: $\theta = \sum_{k \in U} y_k$
- Variable of interest: y_k
- Nonprobability sample: s_{NP}
  - Subset of U
  - y_k is observed (assuming without error)
  - A vector of auxiliary variables is also observed: $\mathbf{x}_k$
  - Indicator of inclusion in s_{NP} : δ_k

Available data

- Probability sample: s_P
  - Subset of U randomly drawn
  - Survey weight: w_k
  - Assumption: survey weighted estimates are approximately unbiased (nonsampling biases are small)
  - Does not contain y_k but $\mathbf{x}_k$ is observed

Model-based methods

- Naïve estimator: $\hat{\theta}^{NP} = N \sum_{k \in s_{NP}} y_k / n^{NP}$
  - Can be very biased (Bethlehem, 2016)
- Objective of data integration methods:
  - Bias reduction through a vector of auxiliary variables observed in both samples $\mathbf{x}_k$
  - Review three methods: Prediction/Calibration, Statistical Matching and Inverse Probability Weighting
  - Require the validity of model assumptions

A key assumption for all the methods

- **Noninformative selection/participation**:
  - $F(y_k \mid \delta_k, \mathbf{x}_k) = F(y_k \mid \mathbf{x}_k)$ or $\Pr(\delta_k = 1 \mid y_k, \mathbf{x}_k) = \Pr(\delta_k = 1 \mid \mathbf{x}_k)$
  - Key to removing bias
  - Bias reduction is achieved by considering auxiliary variables that are associated with both δ_k and y_k
  - The richer the auxiliary information, the more realistic the assumption

A key assumption for all the methods

- What can be done at the design stage (before data are collected in the NP sample) to tend to non-informativeness?
- What auxiliary information would be useful to have in the NP sample that is already available in an existing probability sample?
  - Add (a few) questions to the NP sample
  - Add variables to the NP sample through record linkage?

Prediction / Calibration

- **Idea** (Royall, 1970; Elliott and Valliant, 2017):
  - Model the relationship between y_k and $\mathbf{x}_k$ by using a nonprobability sample
  - Predict y_k for units $k \in U - s_{NP}$ (provided $\mathbf{x}_k$ is available for the entire population)
  - Predictor:

$$\hat{\theta}^{PRED} = \sum_{k \in s_{NP}} y_k + \sum_{k \in U - s_{NP}} y_k^{PRED}$$

Prediction / Calibration

- If a linear model is used, the resulting predictor is equivalent to a calibration predictor of θ :

$$\hat{\theta}^{PRED} = \sum_{k \in s_{NP}} w_k^C y_k$$

- These calibration weights minimize a (weighted) sum of squares subject to

$$\sum_{k \in s_{NP}} w_k^C \mathbf{x}_k = \mathbf{T}_\mathbf{x}$$

- If $\mathbf{T}_\mathbf{x}$ is unknown, it can be replaced with an unbiased estimator (probability survey): $\hat{\mathbf{T}}_\mathbf{x} = \sum_{k \in s_P} w_k \mathbf{x}_k$

Prediction / Calibration

- The calibration predictor is unbiased provided that
  - Noninformative selection/participation assumption holds
  - Linear model is correctly specified
- If the linear model does not hold, model calibration can be considered (Wu and Sitter, 2001)

Statistical matching / Mass imputation

- **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
- Predictor of the total θ : $\hat{\theta}^{SM} = \sum_{k \in s_P} w_k y_k^{imp}$

Statistical matching / Mass imputation

- For a linear model, statistical matching is equivalent in most cases to calibration of the NP sample using estimated totals $\hat{\mathbf{T}}_{\mathbf{x}}$
- Donor imputation is often considered
  - Sample matching (Rivers, 2007)
  - Nonparametric method
- Other nonparametric methods: Yang, Kim and Hwang (2021), Chen, Xu and Cutler (2025)

Inverse probability weighting

- **Idea:**
  - Model the relationship between δ_k and $\mathbf{x}_k$
  - Estimate the participation probability $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{x}_k\right)$
  - Estimator: $\hat{\theta}^{IPW} = \sum_{k \in s_{NP}} w_k^{IPW} y_k$, where $w_k^{IPW} = 1/\hat{p}_k$
  - w_k^{IPW} can be further calibrated to improve precision and obtain a double robustness property:

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

Inverse probability weighting

- **Main advantage of IPW:**
  - Simplifies the modelling effort when there are many variables of interest (only one participation indicator to model)
- **Main assumptions:**
  - Noninformative selection/participation
  - $p_k = \Pr\left(\delta_k = 1 \middle| \mathbf{X}\right) > 0$
- Parametric model (ex.: logistic): $p_k(\boldsymbol{\alpha}) = \left[1 + \exp(-\mathbf{x}_k'\boldsymbol{\alpha})\right]^{-1}$
- Estimated probability: $\hat{p}_k = p_k(\hat{\boldsymbol{\alpha}})$

Inverse probability weighting

- **How to estimate $\boldsymbol{\alpha}$?**
- Maximum likelihood
  - Requires knowing $\mathbf{x}_k$ for the entire population
- Pseudo maximum likelihood (Chen, Li and Wu, 2020)
  - Requires knowing $\mathbf{x}_k$ in a probability sample
  - Inefficient when the probability sample is small
- More efficient alternatives:
  - Beaumont et al. (2024); Kim and Kwon (2024)
  - Better use of available auxiliary information

Inverse probability weighting

- Robustness to model misspecifications may be achieved by
  - creating homogeneous groups
  - using machine learning methods
- Machine learning methods:
  - Easier to justify if the overlap between both samples is negligible (Beaumont et al., 2024; Elliott and Valliant, 2017)
  - Stack both samples and ignore overlap

Conclusions from empirical experiments

- Conducted several experiments with StatCan data
- General conclusion:
  - Data integration methods reduce bias but do not eliminate it: sometimes a significant bias remains

Conclusion

- Data integration methods require the validity of a model/assumptions
  - Essential to plan sufficient time and resources for modelling: Baker et al. (2013)
- Should they be used?
  - Main advantages:
    - Reduce burden and costs, Improve timeliness
  - Main disadvantage:
    - Lower accuracy (unless assumptions are satisfied)
  - It depends on the objectives and how important accuracy is compared with costs and timeliness

Review papers

- **Beaumont, J.-F. (2020).** Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, 46, 1-28.
- **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.
- **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.

Review papers

- **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.
- **Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J., and Chu, K. (2024).** Authors' response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data": Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples. *Survey Methodology*, 50, 123-141.
- **Bethlehem, J. (2016).** Solving the nonresponse problem with sample matching. *Social Science Computer Review*, 34, 59-77.

Other Cited References

- **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.
- **Chen, S., Xu, C., and Cutler, J. (2025).** Integrating probability and non-probability samples through deep learning-based mass imputation. *Survey Methodology*, 51 (to appear).
- **Kennedy, C., Mercer, A., and Lau, A. (2024).** Exploring the assumption that commercial online nonprobability survey respondents are answering in good faith. *Survey Methodology*, 50, 3-21.
- **Kim, J.K., and Kwon, Y. (2024).** Comments on “Exchangeability assumption in propensity-score based adjustment methods for population mean estimation using non-probability samples”. *Survey Methodology*, 50, 57-63.

Other Cited References

- **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.
- **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.

Other Cited References

- **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.

Selected Topics in Estimation for Nonprobability Sampling

Jay Breidt

NORC at the University of Chicago

Designing and Integrating Nonprobability Samples for Official Statistics

December 9, 2025

Probability sampling

- Finite population, $U = \{1, 2, \ldots, k, \ldots, N\}$
- Inferential target

$$T_y = \sum_{k \in U} y_k$$

for a study variable of interest, y_k

- For $A \subset U$, define sample membership indicators

$$A_k = \begin{cases} 1, \text{ with probability } \pi_k^A, & \text{if } k \in A \\ 0, \text{ with probability } 1 - \pi_k^A, & \text{if } k \notin A \end{cases}$$

- A is a **probability sample** if $\pi_k^A > 0$ for all $k \in U$

Unbiased estimation under probability sampling

- Minimal conditions for unbiased estimation:
  1. All elements in the universe have **positive** probabilities of selection, $\pi_k^A > 0$ for $k \in U$
  2. Sampled elements have **known** probabilities of selection, $\{\pi_k^A\}_{k \in A}$
- Under repeated sampling, an **unbiased** estimator of the population total $T_y = \sum_{k \in U} y_k$ is

$$\sum_{k \in A} \frac{y_k}{\pi_k^A} = \sum_{k \in U} y_k \frac{A_k}{\pi_k^A},$$

because

$$\mathrm{E}\left[\sum_{k \in U} y_k \frac{A_k}{\pi_k^A}\right] = \sum_{k \in U} y_k \frac{\mathrm{E}\left[A_k\right]}{\pi_k^A} = \sum_{k \in U} y_k \frac{\pi_k^A}{\pi_k^A} = T_y$$

Nonprobability sampling

- All samples that have either ...
  1. **Zero** probabilities of inclusion for some population elements, or
  2. **Unknown** probabilities of inclusion for some sampled elements

  ... can be considered **nonprobability samples**
- Failing to account for nonprobability sampling yields **biased estimators**
- For $B \subset U$, **model** the membership indicators as independent random variables:

$$B_k = \begin{cases} 1, \text{ with probability } \pi_k^B, & \text{if } k \in B \\ 0, \text{ with probability } 1 - \pi_k^B, & \text{if } k \notin B \end{cases}$$

  - π_k^B is **unknown** and **might be zero**
  - sometimes called **quasi-randomization model**
  - $\{A_k\}$ uses **randomization** and does not require a model

Nonprobability examples

- **Convenience samples:** easier to access, more likely to respond, etc.
- **Judgment samples:** field crews may use their judgment to “improve” a sample or substitute for missing units
- **Snowball/respondent-driven samples:** participants recruit additional participants from among their acquaintances
- **Quota samples, administrative/commercial databases, broken probability samples, opt-in online samples, . . .**
- In each example, what does the nonprobability sample represent?

Concerns about representation of nonprobability samples

- Good probability samples are **representative**
  - sampling error is precisely controlled and described
  - other errors are carefully studied and mitigated
  - **sampling weights** reflect the part of the population represented by the sample
  - safe, defensible inferences
  - often **time-consuming and expensive**
- Nonprobability samples are usually **not representative**
  - typically have minimal control of non-observation errors (coverage errors, sampling/selection, and nonresponse)
  - not clear what part of the population is represented by the sample
  - dangerous for inference due to selection bias
  - **often fast and cheap**

Combining probability and nonprobability samples

- Assume that we have **both *A* and *B*** and look for a trade-off:
  - low bias/high cost/small sample size of prob sample A
  - high bias/low cost/large sample size of nonprob sample B
- For both $k \in A$ and $k \in B$, we have an auxiliary vector $\boldsymbol{x}_k$
  - assumed sufficiently rich to explain participation in B
- Consider two versions of this problem:
  - if y_k is observed **only for *B***, we are doing **data integration**
  - if y_k is observed **for both *A* and *B***, we are doing **data fusion**
- Methods for both problems are related
- **General idea:** "borrow representation" from the probability sample and apply it to the nonprobability sample

Data integration via mass imputation

- For data integration, y_k is missing on A:

Sample	Probability?	$\boldsymbol{x}_k$	y_k	Weight
A	Yes	✓	•	$\left(\pi_k^A\right)^{-1}$
B	No	✓	✓	•

- **Mass imputation:** impute **all** the missing $\{y_k\}_{k \in A}$

Sample	Probability?	$\boldsymbol{x}_k$	y_k	Weight
A	Yes	✓	$\boldsymbol{y_k^*}$	$\left(\pi_k^A\right)^{-1}$
B	No	✓	✓	•

- ...then apply A-weights to **this specific $\{\boldsymbol{y_k^*}\}_{k \in A}$**:

$$\widehat{T}_{y,MI} = \sum_{k \in A} \frac{y_k^*}{\pi_k^A}$$

Data integration via inverse probability weighting

- For data integration, B has no weights:

Sample	Probability?	$\boldsymbol{x}_k$	y_k	Weight
A	Yes	✓	•	$\left(\pi_k^A\right)^{-1}$
B	No	✓	✓	•

- **Inverse probability weighting:** estimate missing $\{\pi_k^B\}_{k\in B}$

Sample	Probability?	$\boldsymbol{x}_k$	y_k	Weight
A	Yes	✓	•	$\left(\pi_k^A\right)^{-1}$
B	No	✓	✓	$\boldsymbol{\left(\widehat{\pi}_k^B\right)^{-1}}$

- ...then apply B-weights to **any** $\boldsymbol{\{y_k\}_{k\in B}}$:

$$\widehat{T}_{y,IPW} = \sum_{k\in B} \frac{y_k}{\widehat{\pi}_k^B}$$

Data fusion example: Large Pelagics Intercept Survey

- US National Marine Fisheries Service is interested in fishing trips that target pelagic species (tuna, sharks, billfish, etc.)
- How many Wahoo were caught by recreational anglers along the US Atlantic coast in 2025?

Sampling the large pelagics fishery

- Sample from population of site-days: $U = \{\text{access sites}\} \times \{\text{days in season}\}$
- Send field staff to selected site-days, A
- Count the number of pelagics trips, $\{z_k\}_{k \in A}$
- Collect catch by species for pelagics trips, generically denoted $\{y_k\}_{k \in A}$

Judgment sampling in LPIS

- Large Pelagics Intercept Survey (LPIS) data are used to estimate **catch rate**: average recreational catch per large pelagic trip, by species: T_y / T_z
- **Problem:** Many site-days have no pelagics trips: $z_k = 0$
  - field crews want to choose their own site-days!
- **Designed compromise**: select an initial probability sample of site-days $S \subset U$ and randomly divide it into A and B
  - A is maintained as a strict probability sample, with **known** inclusion probabilities $\pi_k^A > 0$
  - field crew can leave B as-is or move anywhere in $U \setminus A$
  - B is a nonprobability sample because it relies on field crew **judgment** and has **unknown** inclusion probabilities π_k^B

LPIS is an ideal data fusion problem

- **Data fusion:** Obtain number of pelagics trips z_k and catch by species y_k for **both** probability sample A and nonprobability judgment sample B
- From a **total survey error** perspective, LPIS example is an ideal data fusion problem!
- On the **measurement** side,
    - same mode: in-person interviewing
    - same data collection instrument and protocols
    - same interviewers
    - unified process within one agency for editing data
- On the **representation** side,
    - same population, frame, and coverage issues
    - **different** selection of A versus B
    - same nonresponse of anglers within site-days
    - unified process within one agency for estimation

Dual-frame approach for LPIS data fusion

- Site-days can enter the combined sample, $A \cup B$, via two paths:

$$\begin{aligned} \mathrm{P}[k \in A \cup B] &= \mathrm{P}[k \in A] + \mathrm{P}[k \in B] - \mathrm{P}[k \in A \cap B] \\ &= \pi_k^A + (1 - \pi_k^A)\rho_k \end{aligned}$$

- If we knew the combined probability above for all $k \in A \cup B$, we could construct the unbiased **dual-frame estimator**

$$\widetilde{T}_y = \sum_{k \in U} \frac{A_k + (1 - A_k)B_k}{\pi_k^A + (1 - \pi_k^A)\rho_k} y_k$$

Dual-frame IPW estimator for LPIS

- Model ρ_k as logistic function of auxiliary vector $\boldsymbol{x}_k$ and fit using combined $A \cup B$ data to obtain $\widehat{\rho}_k$
- **Dual-frame IPW estimator** from combined sample is

$$\widehat{T}_y = \sum_{k \in A \cup B} \frac{y_k}{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k}$$

  - **Advantage:** even if $\widehat{\rho}_k$ are small or zero, dual-frame weights are stable:
$$1 \leq \frac{1}{\pi_k^A + (1 - \pi_k^A)\widehat{\rho}_k} \leq \frac{1}{\pi_k^A}$$
  - **Challenge 1:** need estimated ρ_k (and hence estimated π_k^B) for $k \in A \cup B$, not just $k \in B$
  - **Challenge 2:** need to know π_k^A for $k \in B$, not just $k \in A$

Monte Carlo evaluation of dual-frame IPW approach

- Developed methodology in joint work with Chien-Min Huang and tested via extensive Monte Carlo
- Used historical LPIS data to create population with 30 strata and 57,388 site-days, each with known **"pressure"** (expected fishing activity)
- Given pressure, simulate **trips z_k** using zero-inflated Poisson
- Simulate **catch $y_k \mid z_k$** for 11 different "fish species" with various relationships to trips
- Given the simulated population, **draw 1000 samples** following traditional LPIS design (stratified probability proportional to pressure)

Monte Carlo evaluation, continued

- Given simulated sample S, split into 75% pure probability (A) and 25% judgment (B)
- **No Move:** keep B sample as originally selected
- **Unskilled:** move the sample completely at random
- **Skilled:** seven judgment variants
  - **Finding some trips instead of zero trips:** field crew reduces zero-trip site-days, without affecting non-zero-trip site days
  - **Finding more trips when there are some trips:** field crew increases trips on non-zero-trip site-days, without affecting zero-trip site-days
  - Field crew improves at both finding some trips and more trips when there are some trips
- Across (11 catch characteristics)$\times$(9 judgment types), data fusion with dual-frame IPW has **lower mean squared error** than 100% probability sample

Pilot study evaluation of dual-frame IPW approach

- National Marine Fisheries Service field-tested the judgment sampling and data fusion approach
  - 10 northern Atlantic US states
  - fishing seasons 2020–2023
  - across all states and seasons, $|A| = 2410$ and $|B| = 957$
- Judgment sample leads to **more pelagic boat trips**

Productivity Measure	in A	in B	Increase
at least one eligible trip	29.7%	50.1%	up 69%
private boats per hour	0.17	0.22	up 29%
charter boats per hour	0.11	0.19	up 73%

- Data fusion with dual-frame IPW improves productivity while yielding defensible inferences

Recommendations for nonprobability sampling

- Whenever possible, combine nonprobability sample B with a probability sample A
  - fall-back position if B is a disaster!
  - allows assessment of selection bias in B
  - allows adjustment to mitigate selection bias in B
- Compare A and B via total survey error framework
  - carefully assess trade-offs in timing/cost/bias/variance
  - wherever possible, minimize measurement and representation differences at the design stage
- Whenever possible, opt for data fusion over data integration
  - always safer inferences if we have y_k from both probability and nonprobability
  - at a minimum, collect a rich auxiliary vector $\boldsymbol{x}_k$ as similarly as possible across A and B
- **Proceed with caution!** Inherently more dangerous inferences