COLUMBIA | MAILMAN SCHOOL OF PUBLIC HEALTH

BIOSTATISTICS

Enhancing Inference for Nonprobability Samples with Administrative Data and Probability Surveys

Qixuan Chen, PhD

Department of Biostatistics
Columbia University

September 26, 2024

Background

- Inference about a target population based on sample data relies on the assumption
  - the sample is representative
  - the sample can be adjusted to account for nonrepresentativeness
- Probability samples are expensive to collect and often not available in real data problems
  - probability surveys with low response rates are often nonrepresentative
- Nonprobability samples are more widely available
  - unknown inclusion mechanisms and not representative of the population

Nonprobability samples

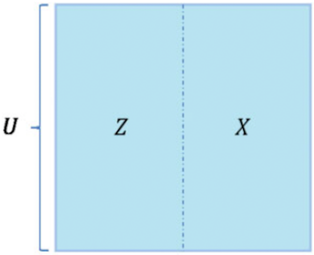
- Types of nonprobability sampling (Baker et al. 2013; Elliott and Valliant 2017)
  - convenience sampling (e.g. volunteer panels, mall intercepts, river samples, observational studies)
  - sampling matching (e.g. quota sampling)
  - network sampling (e.g. snowball sampling)

Data integration

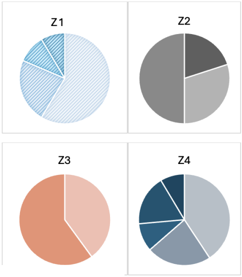
- Using nonprobability samples for population inference requires additional data information
- Such data can include
  - population data, e.g. administrative records, electronic health records
  - well designed and executed probability surveys

Incorporating different types of population data

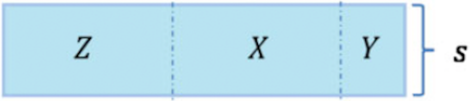
Unit-level population data



Aggregated population data



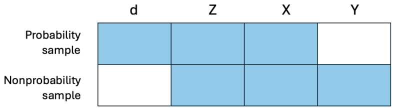
Nonprobability samples



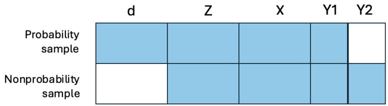
Y: survey outcomes of interest; X: continuous auxiliary variables;
Z: discrete auxiliary variables; U: finite population; s: nonprobability sample.

Integrating with probability surveys

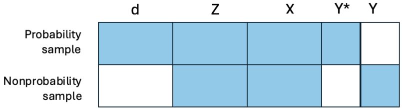
Scenario 1



Scenario 2



Scenario 3



d denotes design variables in the probability sample.

Weighting methods

- Inverse propensity weighting
  - predict the probability being in the nonprobability sample
  - use unit-level population data or a probability survey that is not subject to coverage or other types of bias (Elliott and Davis, 2005; Elliott 2009; Chen et al. 2020)
- Calibration weighting
  - calibrated estimator (Deville & Särndal 1992; Kott 2006)
  - raking and poststratification
  - use aggregated population data or probability surveys

Prediction approaches

- Consider the simple case of estimating a population total (Valliant, Dorfman & Royall, 2000)
  - fit a model of Y on X and Z using the sample
  - predict the values of Y in the population that are not included in the sample
  - estimate the population total: $\hat{t}_1 = \sum_{i \in s} y_i + \sum_{j \notin s} \hat{y}_j$ or $\hat{t}_2 = \sum_{i \in U} \hat{y}_i$.
- Regularized regression approach
  - penalized spline regression (Zheng and Little 2005; Chen, Elliott, and Little 2010)
  - multilevel regression and poststratification (MRP; Wang et al. 2015)

Leveraging high-dimensional auxiliary variables

- In the era of "big data", more and more auxiliary information became available
- Novel methods are needed to incorporate the high-dimensional auxiliary variables
  - pseudo-likelihood approach for combining multiple non-survey data with high dimensionality (Gao and Carroll, 2017)
  - model-based calibration approach using LASSO (Chen et al. 2018)
  - a doubly robust variable selection and estimation strategy (Yang et al. 2019)

Machine learning in high-dimensional contexts

- Machine learning algorithms
  - effectively process large amounts of continuous and discrete high-dimensional data
  - automatically select features associated with sample inclusion and survey outcomes
  - excel in making predictions, incorporating nonlinear relationships and interactions
- Bayesian machine learning
  - leverage Bayesian statistics to model uncertainty and make probabilistic predictions

Prediction inference using BART

Journal of Survey Statistics and Methodology (2023) *11*, 433–455

INFERENCE FROM NONRANDOM SAMPLES USING BAYESIAN MACHINE LEARNING

YUTAO LIU
ANDREW GELMAN
QIXUAN CHEN*

- Estimate population mean by integrating with individual-level population data
- Consider Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2010) and soft BART (Linero and Yang 2018). With continuous y,

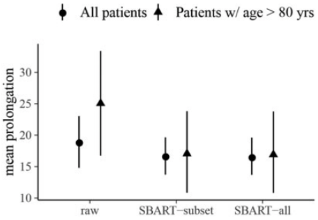
$$y = G(\mathbf{z}, \mathbf{x}) + \epsilon = \sum_{m=1}^{M} g(\mathbf{z}, \mathbf{x}; T_m, \mu_m) + \epsilon, \quad \epsilon \sim N(0, \sigma^2) \qquad (1)$$

Prediction inference using BART (Cont.)

- Inspired by Little and An (2004) in missing data literature, we extended the BART prediction to a doubly robust approach
  - estimate $\pi = \Pr(I = 1|\mathbf{z}, \mathbf{x})$ using probit BART
  - model y using $y = G(\mathbf{z}, \mathbf{x}, \hat{\pi}) + \epsilon$
- Key findings
  - the regularized prediction methods using (soft) BART
    - effectively reduce selection bias in the nonrandom sample
    - yield efficient estimates of population quantities
    - with close to the nominal level coverage rate
  - adding estimated propensity score as a covariate can offer protection from model misspecification, when important predictors are omitted from the model.

Application example

- Application to a COVID-19 study
  - estimand of interest: mean QTc prolongation of the 470 COVID-19 patients who received hydroxychloroquine treatments during 03/01/20 - 05/01/20 at CUIMC (Rubin et al. 2021)



- nonprobability sample: 244 patients had ECG QTc prolongation measurements
- admin data: EHR data of all 470 patients on demographic characteristics and relevant biomarker characteristics

Some extensions

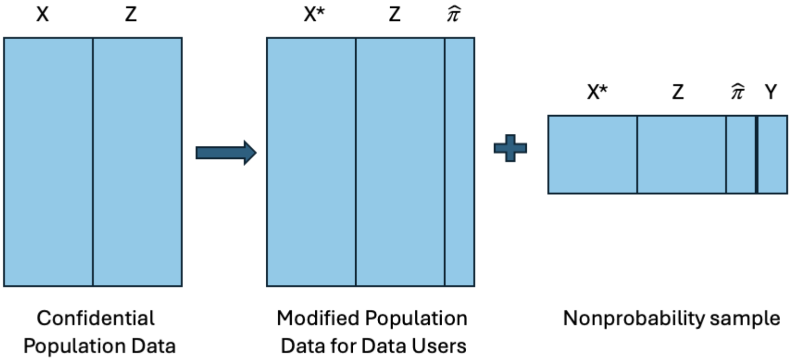
- Two-phase design: phase I with probability sample and phase II with nonprobability sample (Wang et al. 2024)
- Multilevel regression and poststratification using margins of high-dimensional post-stratifiers (Pitts et al. 2024)

Data privacy concerns

- Inference for nonprobability samples relies on access to rich auxiliary information
- Data privacy is often a concern when releasing auxiliary information
- An application example
  - a nonprobability sample of national guard service members was used to study psychological wellbeing
  - demographic details and years of service for all service members were available through an administrative file
  - the confidential population data with individual-level continuous data cannot be released due to disclosure risks

Improving survey inference using administrative records without releasing individual-level continuous data

Sharifa Z. Williams, DrPH[1,2,3] | Jungang Zou, MS[1] | Yutao Liu, PhD[1] | Yajuan Si, PhD[4] | Sandro Galea, MD, DrPH[5] | Qixuan Chen, PhD[1,3]



Summary

- Nonprobability samples are widely used for research purposes.
- Data integration offers an effective solution to improve inference for nonprobability samples.
- Machine learning algorithms are powerful tools for robust and efficient data analysis.

Key challenges in data integration

- Confidentiality risks increase with the release of more granular auxiliary information
  - synthetic data can help mitigate disclosure risks, but adding noise may reduce data utility
  - balancing data utility and privacy remains a critical area for future research
- Heterogeneity among data sources poses significant challenges to data integration
  - covariate shift problem
  - varied data structures
  - differences in data quality
  - efficient integration of diverse data sources is a crucial research area

Refining study design and data collection

- How can we improve the utility of probability surveys for inference of nonprobability samples?
  - for example, with the growing popularity of internet and social media-based sample recruitment, adding questions about internet access and social media usage to probability surveys can increase their relevance
- Can we improve the design and data collection process for nonprobability samples?
  - for example, implementing control during sample recruitment can help reduce the covariate shift problem

Statistical methods and software advances

- In addition to selection bias, nonprobability samples are also prone to measurement error and missing data.
  - there is a need for methods that can address all these issues simultaneously
- Potential of large language models
  - enhanced data imputation and synthetic data generation
  - improved robustness and efficiency in data analysis
- Workflow and software tools needed to facilitate
  - the design of nonprobability surveys with generalizability considerations for post-survey analysis
  - inference from nonprobability surveys through data integration

Other areas for future research

- Extend from the estimation of descriptive statistics to analytic inference, e.g., regression, small area estimation
- Combine regression modeling and inverse propensity weighting (Gelman, Si, and West 2024)
- Other aspects of generalization
  - causal inference

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.

Chen, J.K.T., Valliant, R., Elliott, M.R. (2018). “Model-assisted calibration of non-probability sample survey data using adaptive LASSO”. *Survey Methodology*, 44, 117-144.

Chen, Q., Elliott, M., Little, R.J.A. (2010). “Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling”. *Survey Methodology*, 36(1), 23-34.

Chen, Y., Li, P., Wu, C. (2020). “Doubly robust inference with nonprobability survey samples”. *Journal of the American Statistical Association*, 115, 2011-21.

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

Elliott, M.R. and Davis, W.W. (2005). “Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys”. *Journal of the Royal Statistical Society: Series C*, 54, 595-609.

Elliott, M. (2009). “Combining data from probability and nonprobability samples using pseudo-weights”. *Survey Practice*, 2(6), https://doi.org/10.29115/SP-2009-0025.

Elliott, M., Valliant, R. (2017). “Inference for nonprobability samples”, *Statistical Science*, 32(2), 249-264.

Gelman, A., Si, Y., West, B. (2024). “Regression, poststratification, and small-area estimation with sampling weights”,
`http://stat.columbia.edu/~gelman/research/unpublished/weight_regression.pdf`

Gao, X., Carroll, R. J. (2017). “Data integration with high dimensionality”. *Biometrika*, 104, 251-272.

Kott, P.S. (2006). “Using calibration weighting to adjust for nonresponse and coverage errors”. *Survey Methodology*, 32, 133-142.

Liu, Y. Gelman, A., Chen, Q. (2023). “Inference from nonrandom samples using Bayesian machine learning”, Journal of Survey Statistics and Methodology, 11, 433-435.

Pitts, A.J., Yomogida, M., Aidala, A., Gelman, A., Chen, Q. (2024+). “Multilevel regression and poststratification using margins of post-stratifiers: improving inference for HIV outcomes during the COVID-19 pandemic”, *submitted*.

Rubin, G. A., A. D. Desai, Z. Chai, A. Wang, Q. Chen, A. S. Wang, C. Kemal, et al. (2021).“COVID-19 Infection Is Associated with QTc Prolongation”, *JAMA Network Open*, 4, e216842

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Wang, X., Kennedy, L., Chen, Q. (2024+). “Improving Survey Inference in Two-phase Designs Using Bayesian Machine Learning”, *Journal of the Royal Statistical Society: Series A*, revision submitted.

Wang, W., Rothschild, D., Goel, S. and Gelman, A. (2015). “Forecasting elections with non-representative polls”. *Int. J. Forecast*, 31, 980-991.

Williams, S.Z., Zou, J., Liu, Y., Si, Y., Galea, S., and Chen, Q. (2024). “Improving survey inference using administrative records without releasing individual-level continuous data”, *Statistics in Medicine*, provisionally accepted.

Yang, S., Kim, J. K., Song, R. (2019). “Doubly robust inference when combining probability and nonprobability samples with high-dimensional data”. *Journal of the Royal Statistical Society, Series B*, 82, 445-465.

Zheng, H., and Little, R.J.A. (2005). “Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model”. *Journal of Official Statistics*, 21, 1-20.