A Review of Methods to Combine Probability Surveys and/or Calibrate One Probability Survey with Official Federal Statistical Survey Data

2023-10-24

NORC Team + NCHS Team

NORC at the University of Chicago

Views expressed in this presentation are those of the authors, and do not represent CDC, NCHS or NORC.

NCHS Team

Paul Scanlon

James Dahlhamer

Katherine Irimata

NORC Literature Review Team

David Dutwin

Ipek Bilgen

Stas Kolenikov

Michael Yang

Chien-Min Huang

Margrethe Montgomery

NORC R code team

Stas Kolenikov

Soubhik Barari

Nuria Adell Raventos

Emerson Berry

Chien-Min Huang

Jiazhi Yang

Matt Gunther

Agenda

01 Why combined data

02 Methodological considerations

03 Statistical methods: Combine probability sources

04 Statistical methods: Combine probability and non-probability sources

05 Current work

Why combined data

NORC at the University of Chicago

Fitness for use and various aspects of quality

FCSM 2020 Data Quality Framework

- Three domains, 11 dimensions

Gold standard in-person surveys

- NHIS: National Health Interview Survey; NHANES: National Health and Nutrition Examination Survey
- Better accuracy due to decades of experience, underlying methods research, higher response rates in active interview modes

Online panels

- RANDS: Research and Development Survey; RSS: Rapid Survey System
- Better timeliness (and hence, potentially, relevance) due to faster turnaround times

Variations in granularity, accuracy (specimen instrumentation in NHANES vs. self-report in NHIS, RSS).

The question of combining the data is that of coherence between sources.

Methodological considerations

NORC at the University of Chicago

Methodological considerations in online panels

- Probability vs. non-probability nature of recruitment
- Sampling frames
  - Coverage of the U.S. population
- Panel refresh frequency
- Panelist interview load / burden
- Retention, incentivization strategies

- Documentation of the above in public-facing documents, on demand, not at all

How can one compare the different panels?

- Response rates
- Panel composition, representation of key subpopulations
  - Racial and ethnic minorities
  - Young adults
  - Non-English speakers
- Panel recruitment
  - Coverage
  - Protocols
- Panel maintenance and retainment
- Weighting methodology
  - Complex sampling designs
  - Replenishment
  - Adjustments for eligibility
  - Adjustments for nonresponse in recruitment
  - Adjustments for nonresponse in individual surveys
- Transparency

See also: ESOMAR 37 questions

- The European Society for Opinion and Marketing Research
- https://esomar.org/code-and-guidelines/37-questions-to-help-buyers-of-online-samples

Total survey error and sources of biases

- Coverage of the non-Internet population
  - Additional response mode(s) vs. access provision
- Recruitment mode(s)
  - Mixed modes more likely to recruit diverse panels
  - Recruitment mode ≠ data collection mode
- Recruitment materials design features
- Within household selection
  - Recruitment of every HH member
- Main data collection mode effects
- Panel conditioning
- Low quality respondents
  - Ineligible (e.g. does not reside in the U.S.)
  - Cheaters
  - Speeders
  - Bots

Combining survey data

NORC at the University of Chicago

Review of reviews

- Citro (2014)
- Elliott and Valliant (2017)
- Lohr and Raghunathan (2017)
- Rao (2020)
- Yang and Kim (2020)
- Beaumont (2020)
- Wu (2022)

Combining multiple probability surveys

NORC at the University of Chicago

Dual/multiple frame estimation

- Each source is considered a separate frame
- For each combined data observation, determine (potential) frame membership
- Compositing: $\hat{Y}_\lambda = \hat{Y}_a + \lambda\hat{Y}^A_{ab} + (1-\lambda)\hat{Y}^B_{ab} + \hat{Y}_b$
  - Optimize λ to minimize sampling variance or coverage bias
- Single frame: $\hat{Y} = \sum_i \frac{y_i}{\pi_{i;\text{frame 1}} + \pi_{i;\text{frame 2}} - \pi_{i;\text{both}}}$
- More than two frames: frame count estimator

Hartley (1962), Kalton and Anderson (1986), Bankier (1986), Lohr (2009)

Regression- and calibration-type methods

	Outcome y	Covariate X	Covariate Z
Survey 1	✓	✓	✓
Survey 2		✓	
Population			✓

- First survey regression estimator:

$$\hat{y}_R = \hat{y}_{HT} + \hat{B}'\left[T_Z - \hat{Z}_{HT}\right];\ \hat{B} = (Z'WZ)^{-1}Z'WY$$

- Second survey adjustment:

$$\hat{y}_{AR} = \hat{y}_R + \hat{D}'\left[\hat{X} - \hat{X}_R\right];\ \hat{D} = (Z'RZ)^{-1}Z'RY; R = W - WZ(Z'WZ)^{-1}Z'W$$

- +1/-1 raking trick
- Empirical likelihood family of approaches

Renssen and Nieuwenbroek (1997), Hidiroglou (2001), Wu (2004), Fu et al. (2008)

Multiple and mass imputation

	Outcome y	Covariate X
Survey 1	✓	✓
Survey 2		✓

- Fit imputation models on survey 1
- Impute inside survey 2
- Rubin's combine formula:

$$\hat{\theta}_M^{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m \; ; \; \hat{V}_M^{MI} = \frac{1}{M}\sum_{m=1}^{M}\hat{v}_m + \left(1+\frac{1}{M}\right)\sum_{m=1}^{M}(\hat{\theta}_m - \hat{\theta}_M^{MI})(\hat{\theta}_m - \hat{\theta}_M^{MI})'$$

- Complex surveys:
  - Subsample survey 1 according to complex survey design per each imputation
  - Internal $\hat{v}_m$ must account for the complex survey design (of survey 2)

Rubin (1976), Shao and Sitter (1996), Raghunathan (2006), Schenker et al. (2010), Kim and Rao (2012), van Buuren (2018)

Bayesian methods

Macro-level

- Bayesian updating / combining:

$$y_1 \sim N(\mu, \sigma_1^2)\ \&\ y_2 \sim N(\mu, \sigma_2^2) \Rightarrow y_c | y_1, y_2 \sim N(\mu_c, \sigma_c^2), \mu_c = \frac{y_1/\sigma_1^2 + y_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2}, \sigma_c^2 = \frac{1}{1/\sigma_1^2 + 1/\sigma_2^2}$$

Micro-level

- Model estimation for survey 2 with priors obtained from survey 1

Erciulescu, Opsomer and Breidt (2021)

Small area estimation methods

$$y_i = \theta_i + \text{sampling error}_i; \theta_i = x_i'\beta + \text{model error}_i$$

Asymmetric roles of data

- Survey 1 provides covariates for SAEs built on Survey 2

Symmetric roles of data

- Different sources are random area-level effects (possibly with bias)

Raghunathan et al. (2007), Ybarra and Lohr (2008), Kim et al. (2015)

Combining probability and non-probability surveys

NORC at the University of Chicago

Applicable prob + prob methods

- Superpopulation modeling
- Mass imputation
- Calibration
  - Base weights?
  - Lasso selection for the outcome model (Chen et al. 2018)
  - Behavior variables

Small area estimation methods

Joint detailed domain modeling:

$$y_d^P = x_d'\beta + \nu_d + \varepsilon_d^P$$
$$y_d^{NP} = x_d'\beta + \nu_d + \varepsilon_d^{NP} + \alpha_d^{NP}$$

- ν_d: domain model error
- $\varepsilon_d^P, \varepsilon_d^{NP}$: domain sampling error (estimable)
- α_d^{NP}: systematic error in the low quality source (may have nonzero mean)

Ganesh et al. (2017)

Propensity score adjustments

- Model membership in the non-probability sample (over combined data set):

$$\Pr[\delta_i = 1 | x_i] = \text{parametric or machine learning model}$$

- Estimating equations:

$$\sum_{i \in S_{NP}} [1 - p_i(\alpha)]\, x_i - \sum_{i \in S_P} w_i p_i(\alpha) x_i = 0 \text{ or } \sum_{i \in S_{NP}} \frac{x_i}{p_i(\alpha)} - \sum_{i \in U} x_i = 0$$

- Non-prob sample weights:
  - Inverse pseudo-probabilities of inclusion
  - Propensity classes / cells
  - Weight imputation from PS-matched donors
  - Kernel estimates

Kim and Wang (2019), Chen et al. (2020), Wang et al. (2022), Shin et al. (2022)

Doubly robust methods

$$\hat{y}_{DR} = \sum_{i \in S_{NP}} w_i^{NP} \{y_i - m(x_i, \hat{\beta})\} + \sum_{i \in S_P} w_i^P \, m(x_i, \hat{\beta})$$

where $m(x_i, \hat{\beta})$ is the model for outcome (parametric or machine learning) and w_i^{NP} is the pseudo-weight for the non-probability sample (usually obtained via PS-type methods)

- If the propensity model is right, $\sum_{i \in S_{NP}} w_i^{NP} \, m(x_i, \hat{\beta})$ and $\sum_{i \in S_P} w_i^P \, m(x_i, \hat{\beta})$ cancel one another, and the estimate is essentially $\sum_{i \in S_{NP}} w_i^{NP} \, y_i$
- If the outcome model is right, $\sum_{i \in S_{NP}} w_i^{NP} \{y_i - m(x_i, \hat{\beta})\}$ is zero, and the estimate is essentially $\sum_{i \in S_P} w_i^P \, m(x_i, \hat{\beta})$

Chen et al. (2020); Kim and Wang (2019); Valliant (2020); Yang et al. (2020)

Simulation study

NORC at the University of Chicago

Simulation study

Estimators

- Calibration to standard demographics
- Calibration to demographics + health
- Calibration with lasso selection
- Propensity score
- Double robust
- NORC SAE-type estimator

Scenarios for online panels

- SRS
- Mild to strong correctable nonresponse
- Non-correctable nonresponse
- Mild to strong coverage error

Thank you.

Stas Kolenikov
Principal Statistician
kolenikov-stas@norc.org

Research You Can Trust™

NORC at the University of Chicago

References - I

Bankier, M. D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association* 81(396), 1074–1079.

Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology* 46(1), 1–29.

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2nd ed.). Interdisciplinary Statistics. Boca Raton, FL: Chapman and Hall/CRC.

Chen, J. K. T., R. L. Valliant, and M. R. Elliott (2018). Model-assisted calibration of non-probability sample survey data using adaptive lasso. *Survey Methodology* 44(1), 117–144.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association* 115(532), 2011–2021.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology* 40(2), 137–162.

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

Erciulescu, A. L., J. D. Opsomer, and F. J. Breidt (2021). A bridging model to reconcile statistics based on data from multiple surveys. *The Annals of Applied Statistics* 15(2), 1068–1079.

Fu, Y., Wang, X. and Wu, C. (2008). Weighted Empirical Likelihood Inference for Multiple Samples. *Journal of Statistical Planning and Inference*, 139, 1462–1473.

References - II

Ganesh, N., V. Pineau, A. Chakraborty, and J. M. Dennis (2017). Combining probability and non-probability samples using small area estimation. In *JSM Proceedings, Survey Research Methods Section*, 1657–1667. Alexandria, VA: American Statistical Association.

Hartley, H. O. (1962). Multiple frame surveys. *In JSM Proceedings, Social Statistics Section*, 203–206. Alexandria, VA: American Statistical Association.

Hidiroglou, M. (2001). Double sampling. *Survey methodology* 27(2), 143–154.

Kalton, G. and D. W. Anderson (1986). Sampling rare populations. *Journal of the Royal Statistical Society: Series A (General)* 149(1), 65–82.

Kim, J. K., S. Park, and S.-Y. Kim (2015). Small area estimation combining information from several sources. *Survey Methodology* 41(1), 21–36.

Kim, J. K. and J. Rao (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika* 99(1), 85–100.

Kim, J. K. and Z. Wang (2019). Sampling techniques for big data analysis. *International Statistical Review* 87(S1), S177–S191.

Lohr, S. L. (2009). Multiple-frame surveys. In *Handbook of Statistics; Sample Surveys: Design, Methods and Applications*, Volume 29, 71–88. Elsevier.

References - III

Lohr, S. L. and T. E. Raghunathan (2017). Combining survey data with other data sources. *Statistical Science* 32(2), 293–312.

Raghunathan, T. E. (2006). Combining information from multiple surveys for assessing health disparities. *Allgemeines Statistisches Archiv* 90(4), 515–526.

Raghunathan, T. E., D. Xie, N. Schenker, V. L. Parsons, W. W. Davis, K. W. Dodd, and E. J. Feuer (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association* 102(478), 474–486.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B* 83(1), 242–272.

Renssen, R. H. and N. J. Nieuwenbroek (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association* 92(437), 368–374.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63(3), 581–592.

Schenker, N., T. E. Raghunathan, and I. Bondarenko (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine* 29(5), 533–545.

Shao, J. and R. R. Sitter (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association* 91(435), 1278–1288.

Shin, H.-C., J. Parker, V. Parsons, Y. He, K. Irimata, B. Cai, and V. Beresovsky (2022). Propensity-score adjusted estimates for selected health outcomes from

References - IV

Shin, H.-C., J. Parker, V. Parsons, Y. He, K. Irimata, B. Cai, and V. Beresovsky (2022). Propensity-score adjusted estimates for selected health outcomes from the research and development survey. *Vital and Health statistics. Ser. 1, Programs and Collection Procedures* (196), 1–20.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology* 8(2), 231–263.

Wang, L., B. I. Graubard, H. A. Katki, and Y. Li (2022). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *International Statistical Review* 90(1), 146–164.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics* 32(1), 15–26.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology* 48(2), 283–311.

Yang, M., N. Ganesh, E. Mulrow, and V. Pineau (2019). Evaluating estimation methods for combining probability and nonprobability samples through a simulation study. In *JSM Proceedings, Survey Research Methods Section*, 1714–1727. Alexandria, VA: American Statistical Association.

Yang, S. and J. K. Kim (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science* 3(2), 625–650.

Ybarra, L. M. and S. L. Lohr (2008). Small area estimation when auxiliary information is measured with error. *Biometrika* 95(4), 919–931.