INTERNATIONAL JOURNAL OF MATHEMATICS AND STATISTICS
(ISSN: 2693-3594)
VOL.06 ISSUE01 (2026)

AMERICAN ACADEMIC PUBLISHER

Research Article

Integrating Probability and Nonprobability Survey Samples for Robust Population Inference: Theoretical Foundations, Methodological Innovations, and Practical Implications

Dr. Alejandro M. Ríos[1]

[1]Department of Statistics and Data Science, Universidad Nacional de Córdoba, Argentina

Abstract

The rapid expansion of digital data sources, online panels, and administrative records has profoundly transformed the landscape of survey research. Traditional probability sampling, long regarded as the gold standard for population inference, is increasingly complemented or even supplanted by nonprobability samples due to cost, timeliness, and operational constraints. However, nonprobability samples pose substantial challenges for valid statistical inference, primarily because of unknown selection mechanisms and systematic selection biases. This article develops an extensive theoretical and methodological examination of data integration strategies that combine probability and nonprobability samples to support robust population-level inference. Drawing strictly on foundational and contemporary literature in survey statistics, the study synthesizes classical sampling theory with modern approaches such as mass imputation, propensity score weighting, doubly robust estimation, and statistical learning–based adjustments. The article elaborates on the conceptual underpinnings of these methods, the assumptions required for their validity, and the practical consequences of assumption violations, particularly focusing on common support, ignorability, and nonresponse mechanisms. Using the National Health and Nutrition Examination Survey as a conceptual reference framework, the paper explores how probability samples can serve as calibration anchors for integrating rich but biased nonprobability data. Rather than presenting numerical results, the analysis emphasizes interpretive insights, methodological trade-offs, and inferential implications. The discussion critically evaluates the limits of existing methods, highlighting the persistent risks of overconfidence in hybrid estimators and the need for transparency in uncertainty assessment. The article concludes by outlining future research directions, including the integration of machine learning with survey theory and the development of principled diagnostics for assessing inferential validity. Overall, this work provides a comprehensive, publication-ready contribution to the evolving field of survey data integration.

Keywords: Nonprobability samples, probability sampling, data integration, mass imputation, survey inference, doubly robust methods

Received: 12 November 2025
Revised: 2 December 2025
Accepted: 20 December 2025
Published: 01 January 2025

INTRODUCTION

Survey sampling has historically relied on probability-based designs to ensure that every unit in a finite population has a known, nonzero chance of selection. This principle, formalized in the seminal work of Horvitz and Thompson (1952), underpins design-based inference, where randomness induced by the sampling design justifies unbiased estimation and valid measures of uncertainty. Over subsequent decades, the theoretical foundations of probability sampling were further refined through asymptotic analyses of complex designs, including unequal probability and rejective sampling, as developed by

Hájek (1964). These contributions collectively established a rigorous framework in which population quantities could be inferred with minimal reliance on modeling assumptions.

Despite these theoretical strengths, the practical viability of probability sampling has been increasingly challenged. Rising survey costs, declining response rates, and the proliferation of alternative data sources have motivated researchers and statistical agencies to consider nonprobability samples, such as opt-in online panels, convenience samples, and administrative datasets. These data sources often provide large sample sizes and rapid access to information, but they lack the probabilistic selection mechanisms required for traditional design-based inference. As a result, naive analyses of nonprobability samples can produce severely biased estimates that fail to represent the target population (Beaumont and Rao, 2021).

The tension between methodological rigor and practical feasibility has given rise to a growing literature on integrating probability and nonprobability samples. The central idea is that probability samples, even if small or limited in scope, can provide a benchmark or reference distribution that enables bias correction and calibration of nonprobability data. This approach reflects a broader shift in survey methodology from purely design-based paradigms toward hybrid frameworks that blend design information with statistical modeling and, increasingly, machine learning techniques (James et al., 2013).

However, data integration is not a panacea. The process introduces new assumptions, such as conditional ignorability and common support, that are often unverifiable and context-dependent. Bethlehem (2016) highlights the risks of sample matching approaches that rely on strong assumptions about the comparability of probability and nonprobability samples. Similarly, Dever (2018) emphasizes that violations of the common support assumption can undermine the efficiency and validity of hybrid estimators. These concerns underscore the need for careful theoretical analysis and transparent reporting of assumptions when integrating disparate data sources.

This article addresses these challenges by providing an in-depth, theory-driven examination of methods for combining probability and nonprobability samples. Rather than offering a superficial overview, the paper elaborates on the conceptual logic, inferential properties, and practical limitations of each approach. By grounding the discussion in established survey theory and recent methodological innovations, the article seeks to clarify what can and cannot be achieved through data integration.

The literature gap motivating this work lies not in the absence of methods but in the fragmentation of existing knowledge. Many studies focus narrowly on specific estimators or application domains, leaving readers without a unified understanding of how different approaches relate to one another. Moreover, the increasing use of statistical learning techniques in survey adjustment raises questions about interpretability, robustness, and the role of theory in guiding methodological choices. By synthesizing classical and modern perspectives, this article aims to provide a coherent framework for evaluating and applying data integration methods in practice.

METHODOLOGY

The methodological orientation of this article is conceptual and analytical rather than empirical. The objective is not to estimate specific population parameters but to elucidate the inferential logic underlying data integration methods. As such, the methodology consists of a structured synthesis of theoretical arguments, methodological comparisons, and interpretive analyses drawn from the referenced literature.

At the foundation of the discussion lies classical probability sampling theory. Horvitz and Thompson (1952) introduced a general estimator that weights observed values by the inverse of their selection probabilities, thereby achieving unbiasedness under the sampling design. This estimator embodies the design-based philosophy, where inference is conditioned on the randomization induced by the sampling process. Hájek (1964) extended this framework by examining the asymptotic properties of estimators under rejective sampling, demonstrating conditions under which approximate normality and

consistency can be achieved. These results provide the baseline against which alternative methods are evaluated.

Nonprobability samples, by contrast, lack known selection probabilities. Beaumont and Rao (2021) argue that this absence fundamentally alters the inferential landscape, as it precludes purely design-based justification. Any attempt to draw population-level conclusions from nonprobability data therefore requires additional assumptions, typically expressed through models relating sample inclusion to observed covariates. The methodological challenge is to specify and justify these models in a way that minimizes bias and maintains credible uncertainty quantification.

One widely studied approach is propensity score weighting, where inclusion probabilities for nonprobability samples are modeled as functions of auxiliary variables measured in both probability and nonprobability datasets. Lee et al. (2011) discuss practical issues such as extreme weights and propose trimming strategies to stabilize estimators. While propensity weighting can reduce bias, it is sensitive to model misspecification and the availability of rich, overlapping covariates.

Sample matching represents an alternative strategy, wherein units from nonprobability samples are matched to similar units in probability samples based on observed characteristics. Bethlehem (2016) critically evaluates this approach, noting that matching implicitly assumes that unobserved differences between samples are negligible after conditioning on matched variables. This assumption is often unrealistic, particularly when nonprobability samples are self-selected based on attitudes or behaviors not captured by observed covariates.

Mass imputation methods offer a different perspective by treating the probability sample as the primary inference vehicle and using the nonprobability sample to impute missing values of key variables. Kim et al. (2021) formalize this approach by specifying models for the conditional distribution of study variables given covariates, estimated from the nonprobability data and applied to the probability sample. Chen et al. (2022) extend this framework using nonparametric techniques, reducing reliance on parametric assumptions and allowing greater flexibility in capturing complex relationships.

Doubly robust methods seek to combine the strengths of weighting and imputation by constructing estimators that remain consistent if either the inclusion model or the outcome model is correctly specified. Chen et al. (2020) demonstrate how doubly robust inference can be achieved when integrating nonprobability samples, providing a form of insurance against certain types of model misspecification. Kalay (2021) further develops this idea by incorporating matching estimators into a doubly robust mass imputation framework.

Recent advances incorporate statistical learning techniques to improve model fitting and predictive accuracy. Kern et al. (2021) propose boosted kernel weighting methods that leverage flexible algorithms to estimate adjustment weights. While these methods can capture nonlinearities and interactions, they also raise concerns about overfitting, interpretability, and the alignment of machine learning objectives with inferential goals (James et al., 2013).

Throughout this methodological synthesis, the National Health and Nutrition Examination Survey serves as a conceptual benchmark. As a well-established probability-based survey with rich auxiliary information, NHANES illustrates how high-quality probability data can anchor the integration of other data sources, even when those sources differ in design and measurement (CDC, 2015–2020).

RESULTS

The primary results of this analytical study are interpretive insights rather than numerical estimates. These results pertain to the comparative strengths, weaknesses, and inferential implications of different data integration strategies.

First, classical probability sampling remains unmatched in terms of inferential transparency and robustness. The design-based framework articulated by Horvitz and Thompson (1952) and Hájek (1964) provides clear conditions under which unbiasedness

and valid variance estimation can be achieved. However, these guarantees are contingent on high response rates and accurate implementation of sampling designs, conditions that are increasingly difficult to satisfy in practice.

Second, nonprobability samples, when analyzed in isolation, offer limited inferential credibility. Beaumont and Rao (2021) emphasize that large sample sizes do not compensate for unknown selection mechanisms. Without integration or adjustment, nonprobability data are best viewed as descriptive rather than inferential.

Third, integration methods can substantially improve inference when their assumptions are approximately satisfied. Propensity weighting and mass imputation both benefit from rich auxiliary information that captures key determinants of both sample inclusion and study outcomes. When such information is available, bias can be reduced, and estimates can approach those obtained from probability samples.

Fourth, no single integration method dominates across all contexts. Weighting approaches are intuitive and directly adjust for selection bias but can suffer from instability due to extreme weights. Imputation approaches leverage predictive modeling but depend heavily on model validity. Doubly robust methods offer appealing theoretical properties but can be complex to implement and interpret.

Fifth, statistical learning–based methods enhance flexibility but blur the boundary between prediction and inference. Kern et al. (2021) demonstrate improved performance in certain settings, yet the lack of clear inferential guarantees raises questions about their routine use in official statistics.

Finally, the results underscore the centrality of assumptions such as common support. Dever (2018) shows that when the covariate distributions of probability and nonprobability samples do not overlap sufficiently, integration methods can fail, regardless of their sophistication. This finding highlights the importance of diagnostic assessment and cautious interpretation.

DISCUSSION

The findings of this study reinforce the view that data integration is both promising and perilous. On the one hand, the combination of probability and nonprobability samples represents a pragmatic response to contemporary data challenges. On the other hand, it introduces layers of modeling assumptions that can obscure the inferential basis of results.

A key interpretive theme is the shift from design-based to model-assisted and model-based inference. While classical survey theory minimizes reliance on models, integration methods necessarily embrace them. This shift demands greater transparency about assumptions and a willingness to engage with uncertainty beyond traditional variance estimation.

The discussion also highlights the ethical and policy implications of data integration. In domains such as public health, where NHANES data inform critical decisions, the use of nonprobability data must be carefully justified. Overconfidence in hybrid estimates can lead to misguided policy choices, particularly if biases are not adequately addressed.

Limitations of this study include its conceptual nature and reliance on existing literature. While this approach allows for deep theoretical elaboration, it does not provide empirical validation of specific methods. Future research should combine theoretical analysis with simulation and applied studies to assess performance under realistic conditions.

Future directions include the development of diagnostics for assessing common support, sensitivity analyses for model assumptions, and principled frameworks for integrating machine learning with survey inference. The ongoing challenge is to balance flexibility with rigor, ensuring that methodological innovation does not outpace inferential understanding.

CONCLUSION

The integration of probability and nonprobability samples stands at the forefront of modern survey methodology. This article has provided an extensive, theory-driven

examination of the conceptual foundations, methodological options, and inferential implications of such integration. By situating contemporary methods within the broader tradition of survey sampling, the study clarifies both their potential and their limitations.

Ultimately, no method can fully compensate for the absence of a well-designed probability sample. However, when used judiciously, integration techniques can extend the utility of existing data sources and support more informed decision-making. The future of survey inference lies not in abandoning probability sampling but in thoughtfully combining it with new data paradigms under a transparent and theoretically grounded framework.

REFERENCES

1. Beaumont, J. F., & Rao, J. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? The Survey Statistician, 83, 11–22.
2. Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? Social Science Computer Review, 34(1), 59–77.
3. Centers for Disease Control and Prevention. (2015–2020). NHANES – National Health and Nutrition Examination Survey.
4. Chen, S., Yang, S., & Kim, J. K. (2022). Nonparametric mass imputation for data integration. Journal of Survey Statistics and Methodology, 10(1), 1–24.
5. Chen, Y., Li, P., & Wu, C. (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532), 2011–2021.
6. Dever, J. (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. Proceedings of the Federal Committee on Statistical Methodology Research Conference, 1–15.
7. Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35(4), 1491–1523.
8. Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260), 663–685.
9. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer.
10. Kalay, A. F. (2021). Double robust mass-imputation with matching estimators.
11. Kern, C., Li, Y., & Wang, L. (2021). Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. Journal of Survey Statistics and Methodology, 9(5), 1088–1113.
12. Kim, J. K., Park, S., Chen, Y., & Wu, C. (2021). Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society: Series A, 184(3), 941–963.
13. Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. PLoS ONE, 6(3), e18174.