American Journal of
Biomedical Science & Research

www.biomedgrid.com
ISSN: 2642-1747

Mini Review

Methodology of Non-Probability Samples Through Data Integration

Kyu-Seong Kim*

Department of Statistics, University of Seoul, South Korea

***Corresponding author:** Kyu-Seong Kim, Professor of Department of Statistics, University of Seoul, South Korea.

To Cite This Article: *Kyu-Seong Kim*, Methodology of Non-Probability Samples Through Data Integration. Am J Biomed Sci & Res. 2024 21(5) AJBSR.MS.ID.002880,* DOI: 10.34297/AJBSR.2024.21.002880

Received: : February 26, 2024 ; **Published:** March 01, 2024

Introduction

In survey research, a non-probability sample is obtained by a method of selecting units from a population using a non-random selection mechanism. It occurs when either the sample is not selected randomly, or the inclusion probability attached to the sample unit is unknown even under random sampling. Quota sample, judgment sample, and volunteer sample are considered as typical non-probability samples [1]. Moreover, new data sources have emerged because of the use of digital technologies by both individuals and business units. These sources encompass extensive amounts of digital information, including web surveys, website visits, social media activity, online purchases, and other online interactions [2]. Non-probability sampling is not free from selection bias by researcher and does not provide randomization distribution where theoretical inference takes place. Therefore, these two things should be considered in developing theories of non-probability sampling [3]. Unlikely the probability sampling framework, a single framework that encompasses the non-probability sampling has not been established yet. Making inferences for any probability and non-probability samples requires some reliance on modeling assumptions. If non-probability samples are widely accepted among survey researchers, there must be a coherent framework and accompanying set of measures for evaluating their quality [4].

Data Integration for Non-Probability Samples

Survey researchers have been responding with intensified explorations on statistical inference with non-probability samples. Statistical inference with non-probability samples is part of a more general topic on combining data from multiple sources. Combining information from independent probability samples has been extensively studied [5, 6]. Discussions are provided on combining survey data with other data sources [7, 8]. In addition, some authors discuss data integration by combining big data and survey sample data for finite population inference [9]. Data integration can provide a way to construct a useful framework based on both probability sample and non-probability samples, where the probability samples represent the population, but the non-probability sample does not. Data integration techniques depend on the information that can be combined with the sample. A common consideration for data integration is to assume that auxiliary information for the same population can be used in both non-probability and probability sample [10-12]. As for the target variable, two cases can be considered. First, the target variable is observed only in the non-probability sample and not in the probability sample. Second, conversely, target variable is observed only in the probability sample and not in the non-probability sample [2].

Methodology of Data Integration

In the first situation, we assume that the target variable is observed only in the non-probability sample and not in the probability sample. Then, since there is no target variable in the probability sample, the target variable is missing, and we can apply the statistical techniques for handing missing data. Existing methods for data integration can be classified into three types: mass imputation, propensity score weighting and calibration weighting. In mass imputation, a probability sample is considered to have missing values for all units for the target variable. We can then use the non-probability sample as training data to develop an imputation model and construct synthetic data for the target variable in the probability sample. Although the observations in the non-probability sample are not necessarily representative of the target population, the relationship between auxiliary variables in two samples can be used to develop a predictive model for mass imputation. For given auxiliary variables, we can build a model for the probability of being included

in the sample and use this probability to construct the propensity score weights for the non-probability sample [12,13]. One of the drawbacks of the propensity score method is that it relies on an explicit propensity score model, which causes the propensity score estimator to be biased when the model is mis-specified. Moreover, if the estimated propensity score is close to zero, the estimate become very unstable [10].

The next method of weighting is calibration weighting. We can use this technique to calibrate auxiliary information of a non-probability sample with the auxiliary information of the probability sample so that the non-probability sample after calibration resembles the target population [14]. In this method, calibration weights can be obtained by solving the covariate balancing constraints. Combination of weights and imputation approaches can be considered to improve robustness against model misspecification [15]. The doubly robust estimator uses both the propensity score and the outcome models. The estimator is doubly robust in that it is consistent if either the propensity score model or the outcome model is correctly specified, not necessarily both. In another situation, it is assumed that the target variable is observed only in the probability sample, whereas in the non-probability sample, it is (a) observed correctly, (b) observed with error, or (c) predicted using covariate from large non-probability sample. The pseudo-calibration estimators are proposed in this situation [2]. The pseudo-calibration estimator aims to compute the weights of units in the non-probability sample. Unlike previous estimators developed in a model-assisted framework in the first situation, the pseudo-calibration estimators are developed within a model-based framework.

Discussion

In this article, we briefly reviewed the methodology of non-probability samples through data integration. Prior to data integration, we need to consider the following concerns. The first concern is the possibility that measurements of units in probability sample and non-probability sample may differently depending on the survey mode and characteristics of the auxiliary data sources. For example, differences in measurement may arise when probability sample is from one survey and non-probability sample is from a big data source. The second concern is that the quality of the data in probability sample and non-probability sample can also be different. So, quality and error frameworks need to be developed for integrated data [16]. It is expected that inference using integrated data can produce better results than inference using probability samples or non-probability samples only.

Acknowledgements

None.

Conflict of Interest

None.

References

1. Statistics Canada (2010) Survey methods and practices. Catalogue no 12-587-X.
2. Golini N, Righi P (2024) Integrating probability and big non-probability samples data to produce official statistics. Statistical Methods & Applications.
3. Kim KS (2022) Methodology of non-probability sampling in survey research. American Journal of Biomedical Science & research 15(6): 616-618.
4. Baker R, Brick JM, Bates NA, Battaglia M, Couper MP, et al. (2013) Summary report of the AAPOR task force on non-probability sampling. J Sur Stat Methodol 1(2): 90-143.
5. Wu C (2004) Combining information from multiple surveys through the empirical likelihood method. The Canadian Journal of Statistics 32(1): 15-26.
6. Kim JK, Rao JNK (2012) Combining data from two independent surveys: a model-assisted approach. Biometrika 99(1): 85-100.
7. Lohr SL, Raghunathan TE (2017) Combining survey data with other data sources. Statistical Science 32(2): 293-312.
8. Thompson ME (2019) Combining data from new and traditional sources in population surveys. International Statistical Review 87(S1): S79-S98.
9. Yang S, Kim JK, Hwang Y (2021) Integration of data from probability surveys and big found data for finite population inference using mass imputation. Survey Methodology 47(1): 29-58.
10. Yang S, Kim JK (2020) Statistical data integration in survey sampling: a review. Japanese Journal of Statistics and data Science 3: 625-650.
11. Rivers D (2007) Sampling for web surveys. ASA proceeding of the section of survey research methods.
12. Elliott MR, Valliant R (2017) Inference for nonprobability samples. Statist Sci 32(2): 249-264.
13. Chen Y, Li P, Wu C (2020) Doubly robust inference with non-probability survey samples. Journal of the American Statistical Association 115(532): 2011-2021.
14. Lee S, Valliant R (2009) Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. Sociological Methods and Research 37(3): 319-343.
15. Kim JK, Haziza D (2014) Doubly robust inference with missing data in survey sampling. Statistica Sinica 24(1): 375-394.
16. Salvatore C (2023) Inference with non-probability samples and survey data integration: a science mapping study. Metron 81(1): 83-107.