Recovering Prevalences under Missing Data and Sample Selection using Causal Inference, External Data and Bias Analysis

Siddartha Aradhya
Max Thaning

November 2025

Abstract

Surveys are fundamental to demographic and social science research, yet missing data and sample selection bias compromise basic descriptive estimates like prevalences and population means. Despite a conventional view that descriptive research requires no causal reasoning, we argue that even the most fundamental descriptive analysis can be informed by causal inference when data are incomplete or unrepresentative. We use directed acyclic graphs for missing data (m-DAGs), methods combing external data with survey data, Probabilistic Bias Analysis to show how to recover target prevalences using Monte Carlo simulations. We evaluate ideal-type data scenarios ranging from simplistic Missing Completely At Random (MCAR) to complicated Missing Not At Random (MNAR). Conventional complete case (or list-wise deletion) analysis produce severe bias across all scenarios. Multiple imputation recovered unbiased estimates only when data were Missing At Random (MAR) with fully observed confounders. De-biased estimation with external data (e.g., census information) successfully recovered true prevalence across all MAR and sample selection scenarios. For the taxing Missing-Not-At-Random (MNAR) conditions, Probabilistic Bias Analysis (PBA) with simualted validation studies recovers the true value. These findings demonstrate that causal inference is critical for descriptive demographic research whenever missingness or selection occur — which is a nearly universal condition in survey practice. By leveraging m-DAGs, external data, and sensitivity analysis, researchers can recover valid population estimates for key demographic indicators.

1 Introduction

Surveys have been one of the most powerful tools for demographers and social science researchers as well as policymakers, because they allow users to study factors that are often unobservable through other data sources. Specifically, surveys uniquely capture self-reported information about behaviors, beliefs, and

experiences that cannot be observed directly. Understanding people's attitudes, values, health behaviors, and subjective well-being requires asking them directly, making surveys irreplaceable for research on topics ranging from voting behavior to social inequality and quality of life assessments. Some of the most important quantities that inform the development and adjustment of policies come from simple descriptives — such as the prevalence or the mean — derived from surveys, e.g. the poverty or unemployment rate for social policy, and prevalences of various diseases or health characteristics in medical decision-making.

Nevertheless, surveys come with a number of limitations. For fundamental descriptive quantities, most notably, missing data and selection bias, which are increasingly common in an age characterized by increasing survey fatigue and lower response rates. In relation to surveys, two commonly discussed forms of missing data occur from respondents not answering specific items in a survey (*item missing*), or subsets of the target population entirely not responding to surveys for which they have been sampled (*unit missing*). In the case of the former, respondents may answer some share of the questionnaire but not others, providing analysts with a potential idea of characteristics of the survey respondents who experience item missing. In the case of the latter, users have no information whatsoever on the characteristics of the population that is missing in the survey data. In both cases, if missingness occurs at random, it does not bias survey estimates. However, if, as is more likely in applied cases, missingness occurs not at random (either based on observables or unobservables), estimates will be biased.

Given the threat that missing data poses to inferences from surveys, researchers have focused significant attention on developing methods to help alleviate this problem — and thus improving the *recoverability* of the target quantity. Directed Acyclic Graphs (DAGs), a common tool used to inform causal inferences, have more recently been shown to be one approach that helps encode theoretical insights about the data-generating process in relation to missing data (Moreno-Betancur et al., 2018; Mohan and Pearl, 2021). The conventional view that if our study aims to measure a descriptive quantity, we do not have to consider causal relationships, thus has to be revisited. It turns out, even for the most fundamental descriptive quantities — a sample prevalence or a mean (e.g. poverty or unemployment rate) — we can leverage causal inference to gauge or remedy processes behind sample selection and/or missing data. Moreover, recent research has shown that we can leverage external data (e.g., from population registers) on covariate distribution(s) to recover the target quantity in a given (and biased) survey (Bareinboim et al., 2014; Schuessler and Selb, 2025).

In this paper, we reconcile sample selection (unit missing) and missing data (item missing) scenarios using causal graphs for missing data (m-DAGs) and exploit state-of-the-art methods to recover target prevalences. We conduct a simulation study to examine the performance of various methods to adjust for bias when estimating the prevalence of both binary and continuous variables. We present approaches to address bias from all of the ideal-type missing data scenarios identified by Rubin (2018). We thus cover data Missing-Completely-At-Random (MCAR), data Missing-At-Random (MAR) and, more importantly,

from the more common and critical situation of data Missing-Not-At-Random (MNAR). For MNAR scenarios, which are notoriously complicated to address, we further showcase the use of Quantitative Bias Analysis (QBA) and Probabilistic Bias Analysis (PBA) using more or less informed priors and the use of validation studies for sample selection.

2 Results

We conduct Monte Carlo simulations (n=1,000 per replication, 1,000 repetitions) across five missing data and sample selection scenarios to evaluate different methods for recovering unbiased prevalence estimates. When missingness depends on fully observed confounders, a conventional, or naive complete case (also called list-wise deletion) analysis produces substantial to highly problematic bias, while both multiple imputation (MI) and the causal inference informed de-biased estimation recover unbiased estimates.

In pure sample selection scenarios where covariates are also plagued by missing, MI (which is the most common method used for missing data) is, however, much less useful since there is unit-level missingness. Nevertheless, the de-biased approach using external population data on covariates recover unbiased estimates, while complete case analysis exhibited substantial to highly problematic bias. Under the simulated assumptions, we show that bias scales proportionally with both the causal effect of X on Y and the severity of selection.

With both unit and item missingness, again, complete case and MI approaches produce severe to highly problematic bias, while the mean of the de-biased method in the MC simulations recovers the true prevalence. Bias is particularly pronounced when confounders have strong effects on both the outcome and missingness processes.

While we also extend the aforementioned scenarios to continuous outcome settings and showcase the complications that arise under such conditions, our final example is Missing-Not-At-Random (MNAR). Under MNAR (when e.g., the outcome of interest causes its own missingness) conditions, standard methods cannot recover true values. Here, however, we leverage Probabilistic Bias Analysis (PBA), which is a method developed under the broader Quantitative Bias Analysis approach (Fox et al., 2021). We test two PBA approaches using assumptions about how Y is related to its own missingness as well as a simulation of a validation study. The first approach yields sub-optimal results with (severely to heavily) biased estimates. However, the latter, where we simulate a validation study of 5 percent of non-respondents (19 cases) recovers the true value. Interestingly, an increase to a 10 percent sub-sample provides minimal additional benefit.

References

Bareinboim, E., Tian, J., and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. pages 2410–2416.

Fox, M. P., MacLehose, R. F., and Lash, T. L. (2021). *Applying quantitative bias analysis to epidemiologic data*, volume 10. Springer.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037. Publisher: Taylor & Francis.

Moreno-Betancur, M., Lee, K. J., Leacy, F. P., White, I. R., Simpson, J. A., and Carlin, J. B. (2018). Canonical causal diagrams to guide the treatment of missing data in epidemiologic studies. *American journal of epidemiology*, 187(12):2705–2715. Publisher: Oxford University Press.

Rubin, D. B. (2018). Multiple imputation. In *Flexible imputation of missing data, second edition*, pages 29–62. Chapman and Hall/CRC.

Schuessler, J. and Selb, P. (2025). Graphical causal models for survey inference. *Sociological Methods & Research*, 54(1):74–105. Publisher: Sage Publications Sage CA: Los Angeles, CA.