Introduction to data integration for combining probability and nonprobability samples

Sixia Chen, PhD

OSCTR BERD Novel Methodology Unit Workshop

9/29/2023

HUDSON
COLLEGE OF PUBLIC HEALTH
The UNIVERSITY of OKLAHOMA HEALTH SCIENCES CENTER

Acknowledgement

This workshop was supported by the Oklahoma Shared Clinical and Translational Resources (U54GM104938) with an Institutional Development Award (IDeA) from NIGMS. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health

- Please keep presentation slides, data files, and computational codes confidential and don’t share with others
- Please use the data files only for the purpose of this short course
- We will only use R for real data applications and in-class exercise

Course Objectives

- Obtain basic knowledge of probability samples, nonprobability samples, and data integration
- Apply data integration approaches (e.g. Calibration, Inverse Propensity Score Weighting, Mass Imputation) in practice
- Learn pros and cons for different data integration methods

Outline

1. Introduction
2. Calibration weighting approach
3. Inverse propensity score weighting approaches
4. Mass imputation approaches
5. Variance estimation
6. Real data applications and in-class exercise
7. Discussion

1. Introduction

Introduction - Types of survey data

- Probability sampling: every unit in the target population has non-zero probability of being selected
- Nonprobability sampling: not every unit in the target population has non-zero probability of being selected. Sometimes it is called convenient sample in practice
- References: Cochran (1977), Lohr (2021). Wu and Thompson (2020)

Probability sample VS Nonprobability sample

Measure\Type of Survey	Prob	NonProb
Selection Bias	Small	Large
Representativeness	High	Low
Cost	High	Low
Time	Long	Short
Lack of Frame Survey	Impossible	Possible

Commonly used probability sampling designs

- Simple random sampling with/without replacement
- Stratified sampling
- Multi-stage sampling design
- Probability proportional to size sampling design
- Two-Phase sampling
- Multi-Frame sampling design

Probability Sample - Example

- National Health and Nutrition Examination Survey (NHANES): stratified multi-stage complex sampling design
- Sample design:
  - Draw stratified systematic PPS sample of 60 counties from US
  - Within each selected county, draw independent segment sample by using stratified systematic PPS
  - Within each selected segment, draw systematic sample of households
  - Within each selected household, draw people randomly
  - Oversampling of certain groups such as older people, Asians and so on
- NHANES has clustering, stratification, PPSWOR and oversampling

Probability Sample - Example

- The Behavioral Risk Factor Surveillance System (BRFSS): Stratified dual frame sampling design (Cell and Landline)
- The National Health Interview Survey (NHIS): Stratified multi-stage sampling design
- Population Assessment of Tobacco and Health (PATH) Study: Stratified multi-stage longitudinal sampling design

Nonprobability Sample - Example

- 2019 Tribal Behavioral Risk Factor Surveillance System conducted by Tribal Epidemiology Center
- Target population: American Indian Adults who lived in OK, KS, TX
- Sampling design:
  - Tribal event sampling
  - Email sampling
  - Social media sampling
- Sample size improved from about 300 in 2015 to about 800 in 2019

Nonprobability Sample - Example

- Strong Heart Study: A longitudinal study of cardiovascular disease and its risk factors among American Indians. Field Centers are located in Arizona, North and South Dakotas, and Oklahoma
- Online surveys: Griswold and Wright (2004), Martin (2009), O'Brien (2017), Sagar et al. (2016)

Motivations for Nonprobability Sample

- Costs have been increasing for all types of surveys (Willems et al., 2006)
- Response rates have been falling (Bethlehem, 2016), particularly for random digit dialing (RDD) phone surveys (Curtin et. al., 2005)
- Lack of sampling frame information (e.g. rare disease population, minority population)
- Availability of rich information from nonprobability samples (online surveys)

Examples for rich information from nonprobability samples

- 2019 Tribal Behavioral Risk Factor Surveillance System is a nonprobability sample which collected health and behavior information for Native American Population living at Oklahoma, Kansas, and Texas
- Data collection includes a combination of event sampling, email sampling, and social media sampling

Strategies for handling nonprobability sample

- Reason: statistical analysis based on nonprobability sample without further adjustment might be biased (Baker, 2013)
- Strategy: reducing the selection bias by combining the information of nonprobability sample with another probability sample
- Commonly used methods:
  - Calibration
  - Inverse propensity score weighting
  - Pseudo weight
  - Mass imputation
  - Hybrid method

Real Data Example – Calibration (DiSogra et. al., 2011)

- Probability sample: A representative study sample drawn from a probability-based Web panel, after post-stratification weighting
- Nonprobability sample: opt-in Internet panel
- Study interest: early adopter (EA) behavior
- Opt-in samples tend to proportionally have more EA characteristics compared to probability samples
- Results: a reduction in the average mean squared error from 3.8 to 1.8 can be achieved with calibration. The average estimated bias is also reduced from 2.056 to 0.064

Real Data Example – Inverse Propensity Score Weighting (PSW) (Wang et. al., 2021)

- Nonprobability sample: The adult household interview part of The Third U.S. National Health and Nutrition Examination Survey (NHANES III) III conducted in 1988 to 1994 (Ignoring design features)
- Probability sample: 1994 U.S. National Health Interview Survey (NHIS) respondents to the supplement for monitoring achievement of the Healthy People Year 2000 objectives
- Objective: They estimated prospective 15-year all-cause, all-cancer, and heart disease mortality rates for adults in the US
- Results: PSW methods reduced the relative biases from 22% to 80%

Real Data Example – Mass Imputation (Chen et. al., 2023)

- Nonprobability sample: 2019 Tribal Behavioral Risk Factor Surveillance System for collecting health and behavior information for American Indian adults in OK, TX, and KS. Data collection is a combination of event, email, and social media sampling procedures
- Probability sample: 2019 Behavioral Risk Factor Surveillance System which used stratified random digit dialing to collect data
- Results: Mass imputation methods reduced the biases of nine health related outcome variables

Questions

Q1. What nonprobability samples you have worked with?

Q2. How did you handle it in the analysis?

Q3. Did you use any data integration methods for combining probability sample and nonprobability sample?

Notations

- U: Target population with population size N
- $\boldsymbol{x}_i$: k dimensional covariate vector for unit $i \in U$
- y_i: study variable of interest for unit $i \in U$
- S_A: Probability sample with size n_A
- I_i: Sampling indicator for probability sample (1/0)
- π_i: first order inclusion probability for unit $i \in S_A$
- $w_i = \pi_i^{-1}$: design weight for probability sample

Notations (2)

- S_B: Nonprobability sample with size n_B
- r_i: Sampling indicator for nonprobability sample (1/0)
- $p(\boldsymbol{x}_i) = \Pr(r_i = 1|\boldsymbol{x}_i, y_i) = \Pr(r_i = 1|\boldsymbol{x}_i)$: unknown selection probability for nonprobability sample
- θ_N: Unknown population parameter of interest. In this presentation, we consider estimating population mean $\theta_N = \bar{Y}_N$ for simplicity, where $\bar{Y}_N = N^{-1}\sum_{i\in U} y_i$
- We assume that $(\boldsymbol{x}_i, y_i)$ is observed in nonprobability sample and only $\boldsymbol{x}_i$ is observed in probability sample

Notations (3)

- In practice, $\boldsymbol{x}_i$ can be "bridge" variable such as age, gender, race, etc. which is usually collected in public-use probability sample
- y_i is the outcome variable which is designed for the nonprobability sample such as disease status, blood measurement, etc.
- Naïve estimator only based on nonprobability sample ($\hat{\theta}_{NA} = n_B^{-1} \sum_{i \in S_B} y_i$) might be biased due to selection bias unless the selection probability for nonprobability sample is missing completely at random (MCAR)

Nonprobability sample (S_B)

Age	Gender	Race	Education	BMI
20	M	White	< HS	25
40	F	Non-White	>= HS	18
50	F	White	< HS	35
70	M	White	>= HS	23
80	M	Non-White	< HS	32
30	F	Non-White	>= HS	19

Probability sample (S_A)

Age	Gender	Race	Education	Final weight
30	F	White	< HS	3
40	M	Non-White	< HS	5
50	F	Non-White	>= HS	10
60	M	White	>= HS	8
90	F	Non-White	< HS	3
20	F	Non-White	>= HS	9
25	M	White	< HS	9

2. Calibration weighting approach

Calibration weighting approach (Valliant, 2020; Tsung et al., 2018)

- Idea: Obtain calibrated weights by minimizing the distance function

$$\sum_{i \in S_B} \frac{(w_{i,B} - w_{i,B}^{(0)})^2}{2w_{i,B}^{(0)}}$$

subject to $\sum_{i \in S_B} w_{i,B}\, \boldsymbol{x}_i = \sum_{i \in S_A} w_i\, \boldsymbol{x}_i$ and $w_{i,B} > 0$, where $w_{i,B}^{(0)}$ is the initial weight such that $w_{i,B}^{(0)} = 1$

- Calibrated estimator of θ_N can be written as $\hat{\theta}_C = \hat{N}^{-1} \sum_{i \in S_B} w_{i,B}\, y_i$ such that $\hat{N} = \sum_{i \in S_B} w_{i,B}$

Calibration weighting approach (Cont'd)

- Raking (e.g. iterative proportional fitting algorithm) proposed by Deming and Stephan (1940) can be used for calibration
- To avoid negative weights, other distance functions in Deville and Särndal (1992) can be considered
- The performance of calibration approach depends on the association between outcome variable and covariate vector. If $E(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, then the calibration estimator is consistent

Model Calibration (Rao and Sitter, 2001)

- Assume $y_i = m(\boldsymbol{x}_i; \boldsymbol{\beta}) + \epsilon_i$, where $m(\boldsymbol{x}_i; \boldsymbol{\beta})$ is a working model and ϵ_i satisfies $E(\epsilon_i | \boldsymbol{x}_i) = 0$
- Strategy: Minimize the distance function subject to constraint

$$\sum_{i \in S_B} w_{i,B}\, m(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}}) = \sum_{i \in S_A} w_i\, m(\boldsymbol{x}_i; \widehat{\boldsymbol{\beta}})$$

where $\widehat{\boldsymbol{\beta}}$ is the estimator of $\boldsymbol{\beta}$

- If $m(\boldsymbol{x}_i; \boldsymbol{\beta}) = \sum_{k=1}^{L} c_k b_k(\boldsymbol{x}_i)$, then one can use

$$\sum_{i \in S_B} w_{i,B} \{b_1(\boldsymbol{x}_i), \ldots b_L(\boldsymbol{x}_i)\} = \sum_{i \in S_A} w_i \{b_1(\boldsymbol{x}_i), \ldots b_L(\boldsymbol{x}_i)\}$$

Nonprobability sample (S_B) with calibrated weight

Age	Gender	Race	Education	BMI	C_Weight
20	M	White	< HS	25	5
40	F	Non-White	>= HS	18	8
50	F	White	< HS	35	6
70	M	White	>= HS	23	2
80	M	Non-White	< HS	32	3
30	F	Non-White	>= HS	19	9

3. Inverse propensity score weighting approaches

Inverse Propensity score weighting (IPW) approaches

- Rescaled design weight (RDW) method (Valliant and Dever, 2011)
- Log-likelihood estimating equation (LEE) method (Chen, Li, and Wu, 2019)
- Adjusted logistic propensity weighting (ALP) methods (Wang et al, 2021)

Rescaled design weight (RDW) method

- Assume a logistic regression model for the selection probability of nonprobability sample

$$\log\left\{\frac{p(\boldsymbol{x}_i;\boldsymbol{\alpha})}{1-p(\boldsymbol{x}_i;\boldsymbol{\alpha})}\right\} = \boldsymbol{\alpha}^T\boldsymbol{x}_i, \qquad \mathrm{i} \in U$$

- Let $p_i = p(\boldsymbol{x}_i;\boldsymbol{\alpha})$. The population level log-likelihood can be written as

$$l(\boldsymbol{\alpha}) = \sum_{i\in U} r_i \log(p_i) + \sum_{i\in U}(1-r_i)\log(1-p_i)$$

$$= \sum_{i\in S_B} \log(p_i) + \sum_{i\in U/S_B} \log(1-p_i)$$

Rescaled design weight (RDW) method (2)

- Idea: use the following weighted log-likelihood and the weight to approximate the population level log-likelihood

$$\tilde{l}^{RDW}(\boldsymbol{\alpha}) = \sum_{i \in S_B} w_i^* \log(p_i) + \sum_{i \in S_A} w_i^* \log(1 - p_i)$$

$$w_i^* = \begin{cases} 1, & for\ i \in S_B \\ \dfrac{w_i(\widehat{N} - n_B)}{\widehat{N}}, & for\ i \in S_A \end{cases}$$

where $\widehat{N} = \sum_{i \in S_A} w_i$

- RDW estimator: $\hat{\theta}_{RDW} = \sum_{i \in S_B} \hat{p}_{i,RDW}^{-1} y_i / \sum_{i \in S_B} \hat{p}_{i,RDW}^{-1}$

Rescaled design weight (RDW) method (3)

- RDW estimator might be biased unless
  - The response mechanism of nonprobability sample is missing completely at random (MCAR)
  - The nonprobability sample units have small response rates such that n_B/N and p_i are close to 0 for all $i \in U$
- In many practice applications, the second condition will hold since N might be much larger than n_B
- RDW method is easy to implement in practice since one can use existing software to perform weighted logistic regression

Log-likelihood estimating equation (LEE) method

- Idea: The population level log-likelihood can also be written as

$$l(\boldsymbol{\alpha}) = \sum_{i \in U} r_i \log(p_i) + \sum_{i \in U} (1 - r_i) \log(1 - p_i)$$

$$= \sum_{i \in S_B} \log(p_i) + \sum_{i \in U / S_B} \log(1 - p_i)$$

$$= \sum_{i \in S_B} \log(p_i) + \sum_{i \in U} \log(1 - p_i) - \sum_{i \in S_B} \log(1 - p_i)$$

$$= \sum_{i \in S_B} \log\left(\frac{p_i}{1 - p_i}\right) + \sum_{i \in U} \log(1 - p_i)$$

Log-likelihood estimating equation (LEE) method (2)

- The population level log-likelihood can be estimated by the following pseudo log-likelihood

$$\tilde{l}^{CLW}(\boldsymbol{\alpha}) = \sum_{i \in S_B} \log\left(\frac{p_i}{1-p_i}\right) + \sum_{i \in S_A} w_i \log(1-p_i)$$

- Under the same logistic regression model, the corresponding pseudo estimation equation is

$$\tilde{S}(\boldsymbol{\alpha}) = \frac{1}{N}\left(\sum_{i \in S_B} \boldsymbol{x}_i - \sum_{i \in S_A} w_i\, p_i \boldsymbol{x}_i\right) = \mathbf{0}$$

- LEE estimator: $\hat{\theta}_{LEE} = \sum_{i \in S_B} \hat{p}_{i,LEE}^{-1}\, y_i / \sum_{i \in S_B} \hat{p}_{i,LEE}^{-1}$

Log-likelihood estimating equation (LEE) method (3)

- LEE estimator is consistent under the correct response model of nonprobability sample. Different from RDW method, It does not need other extra conditions for consistency
- Drawback: no existing software can be used directly. One needs to write computational code for solving the pseudo estimation equation. When there is large number of covariate variables, the computational time can be very large and the convergence of the algorithm may not be guaranteed

Adjusted logistic propensity weighting (ALP) methods

- (Step 1): Search for covariates x available in both probability sample S_A and nonprobability sample S_B and combine the two samples. Assign $\delta_i = 1$ for $i \in S_B$ and $\delta_i = 0$ for $i \in S_A$ in the combined sample
- (Step 2): Fit a logistic regression model for $q_i = \Pr(\delta_i = 1|x_i)$ in the combined unweighted S_B and weighted S_A, with the survey sample weights $\{w_i, i \in S_A\}$, and obtain the estimate $\hat{q}_i$
- (Step 3): Estimate θ_N by using $\hat{\theta}_{ALP} = \sum_{i \in S_B} \hat{p}_i^{-1} y_i / \sum_{i \in S_B} \hat{p}_i^{-1}$ such that $\hat{p}_i = \hat{q}_i/(1- \hat{q}_i)$

Adjusted logistic propensity weighting (ALP) methods (2)

- A scaled version of ALP estimator can be used to improve the efficiency of ALP estimator, see Wang et al. (2021)
- ALP method is based on the parametric model assumption in the combined sample, which is different from the parametric model assumption for the nonprobability sample in other methods
- ALP method has computational advantage than LEE method since one can directly use weighted logistic regression model in existing computational software

Nonprobability sample (S_B) with inverse propensity score weight

Age	Gender	Race	Education	BMI	I_Weight
20	M	White	< HS	25	9
40	F	Non-White	>= HS	18	7
50	F	White	< HS	35	2
70	M	White	>= HS	23	12
80	M	Non-White	< HS	32	13
30	F	Non-White	>= HS	19	8

Questions

- Did you use propensity score weighting in practice including missing data analysis and causal inference?

4. Mass imputation approaches

Mass imputation approaches

- Assume the following outcome regression model holds in both nonprobability sample and the population

$$y_i = m(\boldsymbol{x}_i; \boldsymbol{\beta}) + \epsilon_i,$$

where $\mathrm{E}(\epsilon_i | \boldsymbol{x}_i) = 0$ and ϵ_i and ϵ_j are independent for $i \neq j$

- Idea: One can first use nonprobability sample to fit regression model $\hat{m}(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}})$, then generate imputed values of y for all units in the probability sample by using $\hat{y}_i = \hat{m}(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}})$. At last, the mass imputed estimator of θ_N can be obtained by using probability sample as

$$\hat{\theta}_{MI} = \frac{\sum_{i \in S_A} w_i \hat{y}_i}{\sum_{i \in S_A} w_i}$$

Mass imputation approaches (2)

- Parametric mass imputation approach: Chen, Li, and Wu (2019), Kim et al. (2021)
- Nonparametric mass imputation approach: Chen, Yang, and Kim (2022)
- Machine learning based mass imputation approach: Chen, Xu, and Cutler (2023)
- Multivariate mass imputation approaches: Chen et al. (2023)

Probability sample (S_A) with imputed outcome variable

Age	Gender	Race	Education	Final weight	Imputed BMI
30	F	White	< HS	3	20
40	M	Non-White	< HS	5	25
50	F	Non-White	>= HS	10	35
60	M	White	>= HS	8	32
90	F	Non-White	< HS	3	29
20	F	Non-White	>= HS	9	18
25	M	White	< HS	9	12

Questions

- Did you use any imputation methods in practice?
- What are the commonly used imputation methods?

5. Variance estimation

Variance estimation

- Taylor linearization: Tedious and Case by Case
- Approximation by using probability proportional to size with replacement (PPSWR) design for weighting methods
- Bootstrap: General and Algorithm based, but can be time-consuming

Bootstrap (Kim et al., 2021)

- (Step 1): We first treat nonprobability sample S_B as a simple random sample to draw a large number (K) bootstrap samples $S_B^{*(k)}$ for $k = 1,2, \dots K$
- (Step 2): Obtain a large number (K) bootstrap weights $w_i^{(k)}$ for $i \in S_A$ and $k = 1,2, \dots K$
- (Step 3): For each pair of $S_B^{*(k)}$ and S_A with $w_i^{(k)}$, obtain the k-th bootstrap estimator $\hat{\theta}^{*(k)}$ for $k = 1,2, \dots K$
- (Step 4): The bootstrap variance estimator can be calculated by $\hat{V}_{boot} = K^{-1} \sum_{k=1}^{K} \left(\hat{\theta}^{*(k)} - \hat{\theta}\right)^2$

PPSWR method for weighting approaches

```
PROC SURVEYMEANS DATA=INDAT;
VAR V1 V2 V3;
STRATUM ST;
CLUSTER PSU;
WEIGHT WT; /*Calibration/Propensity Score/Pseudo weight*/
RUN;
```

6. Real data applications and in-class exercise

Real data applications – Data Harmonization and Cleaning

- Common variables have the same name, same format, same categories, same column numbers in the probability sample file and nonprobability sample file
- Categorical variables need to be dummy coded as 1 or 0 for each category
- Collapsing needs to be done if there are sparse cells
- Missing values need to be imputed

Real data applications – Plan

- The Korea National Health and Nutrition Examination Survey (KNHANES) and the National Health Insurance Sharing Services (NHISS)

KNHANES and NHISS - Data

- The Korea National Health and Nutrition Examination Survey (KNHANES) is a national survey that studies the health and nutritional status of Koreans and has been conducted annually since 1998
- A nationally representative cross-sectional survey that includes approximately 10, 000 individuals each year and collects information on social-economic status, health-related behaviors, quality of life, healthcare utilization, anthropometric measures, biochemical and clinical profiles for non-communicable diseases, and dietary intakes with three component surveys: health interview, health examination, and nutrition survey

KNHANES and NHISS – Data (2)

- The nonprobability sample from NHISS provides health-related information collected from National Health Screening Program (NHSP) in South Korea
- The data that we used in the present study are from the subset corresponding to the blood test results that are associated with metabolic syndrome from the 2014 program
- The data set is made publicly available after anonymization and random selection of 1 million observations (National Health Insurance Data Sharing Service, https://nhiss.nhis.or.kr/bd/ab/bdabf006cv.do)

KNHANES and NHISS – Data (3)

- In our real-world application, we treat the KNHANES sub-sample data for blood test as the probability sample S_A with sample size 4,929 after removing the missing values for key items. We treat the sampling design of KNHANES as probability proportional to size without replacement
- In order to reduce the computational burden, we first draw a simple random sample with size 5, 000 from the original NHISS data and treat the subsample as the nonprobability sample S_B for our analysis

KNHANES and NHISS – Variables

- Predictors: Sex (1 for male, 2 for female), Age, Hemoglobin (HGB), Triglyceride (TG), Anemia (ANE) (1 for yes, 0 for no), and High-density Lipoprotein Cholesterol (HDL, mg/dL)
- Age has 27 categories as following: 1 for ‘20 to 24’, 2 for ‘25 to 26’, 3 for ‘27 to 28’,...,27 for ‘75 or higher’
- Outcome variable: Total Cholesterol (TCHOL)
- TCHOL is a variable in both KNHANES and NHISS, so we can evaluate the performance of different approaches by comparing the estimates with weighted estimates calculated from KNHANES

Data integration methods

- Calibration weighting approaches
- Inverse propensity score weighting approaches (IPW)
- Mass imputation approaches (MI)

Comparison of covariate distributions

Var	Mean A	Mean B
SEX(Male)	0.427	0.518
ANE(Yes)	0.075	0.083
AGE	14.431	13.843
HGB	14.052	14.034
TG	136.396	129.404
HDL	51.134	55.538

Point Estimation Result

	Bias_Mean	Bias_mean (Male)
Mean B	7.522	7.455
Calibration	6.265	5.823
LM (MI)	6.265	7.608
GAM (MI)	4.427	4.846
RDW (IPW)	5.787	5.679
LEE (IPW)	5.786	5.679
ALP (IPW)	5.787	5.679

Variance Estimation Result

	eV	LB	UB	eV_D	LB_D	UB_D
Calibration	0.4	192.4	194.9	0.68	190.7	194
LM (MI)	0.4	192.4	194.9	0.7	192.5	195.8
GAM (MI)	0.4	190.6	193	0.71	189.7	193
RDW (IPW)	0.5	191.9	194.5	0.65	190.6	193.8
LEE (IPW)	0.5	191.9	194.5	0.65	190.6	193.8
ALP (IPW)	0.5	191.9	194.5	0.65	190.6	193.8

Discussion of R code and implementation

7. Discussion

Nonprobability samples

- Nonprobability sample may suffer from selection bias
- Naïve estimates by only using nonprobability sample can be misleading
- Data integration methods show promising performance for handling nonprobability samples
- Data integration methods can be categorized into three methods:
  - Weighting approaches: Calibration, Propensity Score
  - Mass imputation approaches
  - Hybrid approaches

Data integration methods

- Calibration: Calibration variables are linearly associated with study vars
- Propensity score: Propensity score models should be correctly specified
- Mass imputation: The same imputation models should hold in both samples; The imputation model should be correctly specified
- Hybrid: Provide further protection for model misspecification

Comparison of propensity score methods

- RDW method was not recommended due to theoretical weakness
- LEE method may not produce stable convergence numerical results if the dimensional of covariate variables is large
- ALP method had comparable performance with LEE and it is computationally more efficient

In all, I would recommend ALP method in practice

Weighting vs Mass imputation

- If the survey is general purpose and there are many study variables of interest, it is recommended to use weighting methods
- If the purpose of the survey is very specific and there are only limited number of study variables, mass imputation methods can be used if the model fitting is good. In addition, Hybrid methods can be used to improve the performance

Variance Estimation

- Bootstrap methods provide valid tools for statistical inference
- Bootstrap methods may require large computational time
- Taylor methods are specific for each method
- Approximation by using PPSWR design provides practical solution
- Variance estimation methods for machine learning based data integration methods have not been developed

Future Research

- Develop customized R package for data integration
- Develop model diagnostic tools for different methods
- Develop indicator measure for selection bias of nonprobability sample
- Develop statistical inference tools for machine learning based data integration methods
- Develop data integration methods when the selection probability is nonignorable
- Compare data integration methods by both simulation and real application

Questions?

- Contact: Sixia-Chen@ouhsc.edu

Open Q & A

Reference

- Griswold, W., and N. Wright. 2004.“Cowbirds, Locals, and the Dynamic Endurance of Regionalism.” American Journal of Sociology109 (6): 1411–51.
- Martin, K.A. 2009.“Normalizing Heterosexuality: Mothers’ Assumptions, Talk, and Strategies with Young Children.” American Sociological Review 74: 190–207.
- O'Brien, R. 2017.“Redistribution and the New Fiscal Sociology: Race and the Progressivity of State andLocal Taxes.”American Journal of Sociology122 (4): 1015–49.
- Sagar, T., D. Jones, K. Symons, J. Tyrie, and R. Roberts. 2016.“Student Involvement in the UK Sex Industry: Motivations and Experiences.” British Journal of Sociology67 (4): 697–718

Reference (2)

- Willems, P., R. van Ossenbruggen, and T. Vonk. 2006. The Effects of Panel Recruitment andManagement on Research Results: A Study Across 19 Online Panels. ESOMAR Panel Research
- Bethlehem, J.G. 2016.“Solving the Nonresponse Problem With Sample Matching?” Social Science Computer Review34 (1): 59–77
- Curtin, R., S. Presser, and E. Singer. 2005.“Changes in Telephone Survey Nonresponse Over the Past Quarter Century.” Public Opinion Quarterly69: 87–98
- Baker, R., Brick, J.M., Bates, N.A., Battaglia, M., Couper, M.P., Dever, J.A., Gile, K.J. and Tourangeau, R., 2013. Summary report of the AAPOR task force on non-probability sampling. *Journal of survey statistics and methodology*, *1*(2), pp.90-143.

Reference (3)

- Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427-444.
- Valliant, R., (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology*, *8*(2), pp.231-263.
- Tsung, C., Kuang, J., Valliant, R.L. and Elliott, M.R., (2018). Model-assisted calibration of non-probability sample survey data using adaptive LASSO. *Survey Methodology*, *44*(1), pp.117-145.
- Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, *87*(418), 376-382.

Reference (4)

- Valliant R, Dever JA. Estimating propensity adjustments for volunteer web surveys.Sociol Methods Res. 2011;40(1):105-137.
- Chen Y, Li P, Wu C. Doubly robust inference with nonprobability survey samples.J Am Stat Assoc. 2019;115(532):2011-2021.
- 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: Statistics in Society*, *184*(3), 941-963.
- Chen, S., Yang, S., & Kim, J. K. (2022). Nonparametric mass imputation for data integration. *Journal of survey statistics and methodology*, *10*(1), 1-24.

Reference (5)

- Chen, S., Xu, C., and Cutler, J. (2023). Integrating probability and non-probability samples through deep learning based mass imputation. Revision invited by *Survey Methodology*
- Chen, S., Campbell, J., Spain, E., Woodruff, A., & Snider, C. (2023). Improving the representativeness of the tribal behavioral risk factor surveillance system through data integration. *BMC public health*, *23*(1), 273.
- Wu, C., & Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, *96*(453), 185-193.
- 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: Statistics in Society*, *184*(3), 941-963.

Reference (6)

- Wang, L., Valliant, R., & Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in medicine*, *40*(24), 5237-5250.
- Wang, L., Graubard, B. I., Katki, H. A., & Li, Y. (2022). Efficient and robust propensity-score-based methods for population inference using epidemiologic cohorts. *International Statistical Review*, *90*(1), 146-164.
- Elliott, M. R., & Valliant, R. (2017). Inference for nonprobability samples.
- Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review/Revue Internationale de Statistique*, 317-337.

Reference (7)

- Chen, S., Xu, C., and Cutler, J. (2023). Integrating probability and non-probability samples through deep learning based mass imputation. Revision invited by *Survey Methodology*
- Cochran, W. G. (1977). *Sampling techniques*. John Wiley & Sons.
- Lohr, S. L. (2021). *Sampling: design and analysis*. CRC press.
- Wu, C., & Thompson, M. E. (2020). *Sampling theory and practice*. Cham: Springer International Publishing.
- DiSogra, C., Cobb, C., Chan, E., & Dennis, J. M. (2011, August). Calibrating non-probability internet samples with probability samples using early adopter characteristics. In *Joint Statistical Meetings (JSM), Survey Research Methods* (pp. 4501-4515).