Survey Data Integration for Cumulative Distribution Function and Quantile Estimation

Jeremy R.A. Flood[1]

Applied Science & Technology – Data Science & Analytics
North Carolina A&T State University

ITSEW 2024

The Need for Data Integration

- **Probability samples** ensure every possible sample from a finite population possess some chance of selection [8] and hence is the gold standard for population-based inference
  - **Cons:** Costly, prone to non-response, and generally of small size [6, 11, 10]
- **Nonprobability samples** (i.e., **convenience samples**) are flexible, rich, and cheap sources of data
  - **Cons:** No probabilistic design; no way to control for sampling bias [12, 6]
- **Research Goal:** "Integrate" data from large convenience samples with that from smaller probability samples, to leverage the strengths of both

Why Data Integration?

1. **Vulnerable populations**, where new probability samples can't be obtained, but old ones only include common demographic variables
   - Can be combined with cheap, current, and colossal data
2. **Political polls**, which are abundant but prone to error
   - These can be used to predict the outcome of a probabilistic, pre-election poll

Data Integration Setting

- $\boldsymbol{A}$ denotes a *probability sample* of size n_{A} from a finite population $\mathcal{U}$ of size N
  - Has covariates $X_1, X_2, \cdots, X_p$
  - Has $\pi_i = \Pr\left(i \in \mathcal{U} \cap i \in \boldsymbol{A}\right)$
- $\boldsymbol{B}$ denotes a *convenience sample* of size n_{B} from the same finite population
  - Has covariates $X_1, X_2, \cdots, X_p$
  - Has Y, which is the variable of interest
- **Goal**: Use data from $\boldsymbol{A}$ and $\boldsymbol{B}$ to estimate finite population quantities

Data Integration Setting (cont.)

Sample	π	X_1	X_2	$\cdots$	X_p	Y
Probability ($\boldsymbol{A}$)	✓	✓	✓	✓	✓	×
Nonprobability ($\boldsymbol{B}$)	×	✓	✓	✓	✓	✓

Table 1: *The data integration sample setup.*

Estimating Distribution Functions

- Sometimes we are interested in estimating **distribution functions**, as well as **quantiles**:

$$F_{\mathsf{N}}(t) = \frac{1}{N} \sum_{u \in \mathcal{U}} \mathbb{1}\left(Y_u \leq t\right)$$

$$t_{\mathsf{N}}(\alpha) = \inf_{t} \left\{t : F_{\mathsf{N}}(t) \geq \alpha\right\} \quad ; \quad \alpha \in (0,1),$$

- Some examples:
  1. Estimating % of individuals in a food desert with income at or below poverty
  2. Estimating 80^{th} percentile of BMI after conditioning on age, sex, and race

- **Research Question**: How can we do this using $\boldsymbol{A}$ and $\boldsymbol{B}$?

Semiparametric Regression

Assume distribution of Y in finite population follows

$$Y = m(\boldsymbol{X}; \boldsymbol{\beta}_{\mathsf{N}}) + \nu(\boldsymbol{X})\epsilon, \tag{1}$$

where

- $m(\boldsymbol{X}; \boldsymbol{\beta}_{\mathsf{N}}) = \mathbb{E}\left(Y|\boldsymbol{X}\right)$: *known* function of $\boldsymbol{X}$, parameterized by unknown $\boldsymbol{\beta}_{\mathsf{N}}$
- $\boldsymbol{\beta}_{\mathsf{N}}$: $\mathcal{U}$'s estimate of the true $\boldsymbol{\beta}$ in the superpopulation model
- $\nu(\cdot)$: a known, strictly positive variance function
- ϵ: a random error term satisfying $\mathbb{E}\left(\epsilon|\boldsymbol{X}\right) = 0$ and $\mathbb{E}\left(\epsilon^2|\boldsymbol{X}\right) = \sigma_\epsilon^2$

Semiparametric Regression (cont.)

- Let $\widehat{\boldsymbol{\beta}}$ denote a sample-based estimate of $\boldsymbol{\beta}_{\mathsf{N}}$ that solves

$$\widehat{U}(\boldsymbol{\beta}) = \frac{1}{n_{\mathsf{B}}} \sum_{j \in \boldsymbol{B}} \Big(Y_j - m(\boldsymbol{X}_j; \boldsymbol{\beta}) \Big) \boldsymbol{W} \left(\boldsymbol{X}_j; \boldsymbol{\beta} \right) = 0$$

for some p-dimensional function $\boldsymbol{W}$ [4]

Ex: Simple Linear Regression w/ OLS ($p = 1$)

$$\begin{aligned} \hat{\beta} &= \min_{\beta} [\mathsf{RSS}] \\ &= \min_{\beta} \left[\frac{1}{n_{\mathsf{B}}} \sum_{j \in \boldsymbol{B}} \Big(Y_j - \beta X_j \Big)^2 \right] \\ &= \frac{\sum_{j \in \boldsymbol{B}} Y_j X_j}{\sum_{j \in \boldsymbol{B}} X_j^2}. \end{aligned}$$

Residual eCDF Estimator

- Our **residual, eCDF-based estimate of the finite population CDF**:

$$
\begin{aligned}
\hat{F}_{\mathsf{R}}(t) &= \frac{1}{N} \sum_{i \in \boldsymbol{A}} \pi_i^{-1} \hat{G}_i \\
&= \frac{1}{N n_{\mathsf{B}}} \sum_{i \in \boldsymbol{A}} \sum_{j \in \boldsymbol{B}} \pi_i^{-1} \mathbb{1}\left(\hat{\epsilon}_j \leq \frac{t - m(\boldsymbol{X}_i; \widehat{\boldsymbol{\beta}})}{\nu(\boldsymbol{X}_i)} \right) \qquad (2)
\end{aligned}
$$

- Corresponding quantile estimator:

$$
\hat{t}_{\mathsf{R}}(\alpha) = \inf_t \left\{ t : \hat{F}_{\mathsf{R}}(t) \geq \alpha \right\}
$$

Asymptotic Results: Summary

1. Under Assumptions 1 - 7,

$$\frac{\hat{F}_R(t) - F_N(t)}{AV\{\hat{F}_R(t)\}} \xrightarrow{\mathcal{L}} N(0,1),$$

where

$$AV\{\hat{F}_R(t)\} = \frac{1}{N^2} \sum_{u \in \mathcal{U}} \sum_{v \in \mathcal{U}} \left(\frac{\pi_{uv}}{\pi_u \pi_v} - 1 \right) G_u G_v$$

2. An asymptotically unbiased estimate of $AV\{\hat{F}_R(t)\}$ is

$$\widehat{AV}\{\hat{F}_R(t)\} = \frac{1}{N^2} \sum_{h \in \boldsymbol{A}} \sum_{i \in \boldsymbol{A}} \left(\frac{\pi_{hi}}{\pi_h \pi_i} - 1 \right) \frac{1}{\pi_{hi}} \hat{G}_h \hat{G}_i$$

Simulation Overview

- We conducted a two-phase Monte-Carlo simulation study to contrast the performance of our proposed distribution estimators to that using $\boldsymbol{B}$ alone

- Performance metric: **relative root mean squared error** (RRMSE), defined generically for some estimator $\hat{\theta}$ as

$$\mathrm{RRMSE}(\hat{\theta}) = \sqrt{\frac{\mathrm{MSE}(\hat{\theta})}{\mathrm{MSE}(\hat{\theta}_\pi)}},$$

where

1. $\hat{\theta}_\pi$ for CDF: $\hat{F}_\pi(t) = \frac{1}{N} \sum_{i \in \boldsymbol{A}} \pi_i^{-1} \mathbb{1}\big(Y_i \le t\big)$

2. $\hat{\theta}_\pi$ for quantile: $\hat{t}_\pi(\alpha) = \inf_t \{t : \hat{F}_\pi(t) \ge \alpha\}\ ; \quad \alpha \in (0,1)$

Simulation Setup

- $\mathcal{U}$: Simple random sample without replacement (SRSWOR) of size $N = 100,000$ from four superpopulation models
- $\boldsymbol{A}$: SRSWOR of size $n_{\mathsf{A}} = 500$ from $\mathcal{U}$
- $\boldsymbol{B}$: Distratified SRSWOR of size $n_{\mathsf{B}} = 10000$ from $\mathcal{U}$
  1. **Missing at random** (MAR): binary stratification based on the covariate with the highest Pearson correlation to Y
  2. **Missing not at random** (MNAR): binary stratification based on the population mean
- $n_{\mathsf{I}} = .85 n_{\mathsf{B}}$; $n_{\mathsf{II}} = .15 n_{\mathsf{B}}$
- $\alpha = \begin{bmatrix} .01 & .10 & .25 & .50 & .75 & .90 & .99 \end{bmatrix}$

Superpopulation Models

- Model f_1 [3]: $Y = .3 + 2X_1 + 2X_2 + \epsilon$, where

$$
\begin{aligned}
X &\sim \mathrm{N}\left(\mu = 2, \sigma = 1\right) \\
\epsilon &\sim \mathrm{N}\left(\mu = 0, \sigma = 1\right)
\end{aligned}
$$

- Model f_2 [3]: $Y = .3 + .5X_1^2 + .5X_2^2 + \epsilon$, where

$$
\begin{aligned}
X &\sim \mathrm{N}\left(\mu = 2, \sigma = 1\right) \\
\epsilon &\sim \mathrm{N}\left(\mu = 0, \sigma = 1\right)
\end{aligned}
$$

Superpopulation Models (cont.)

- Model f_3 [9, 5]: $Y = -\sin(X_1) + X_2^2 + X_3 - e^{-X_4^2} + \epsilon$, where

$$X_1, \cdots, X_6 \sim \mathrm{Unif}(-1, 1)$$

$$\epsilon \sim \mathrm{N}(\mu = 0, \sigma = \sqrt{.5})$$

- Model f_4 [7, 5]:

$$Y = X_1 + .707X_2^2 + 2\mathbb{1}(X_3 > 0) + .873\ln(|X_1|)|X_3| \\ + .894X_2X_4 + 2\mathbb{1}(X_5 > 0) + .464e^{X_6} + \epsilon,$$

where

$$X_1, \cdots, X_6 \sim \mathrm{Unif}(-1, 1)$$

$$\epsilon \sim \mathrm{N}(\mu = 0, \sigma = \sqrt{.5})$$

Estimators Under Comparison

- CDF Estimators
  - $\hat{F}_{\mathsf{B}}(t)$: The näive CDF of $\boldsymbol{B}$
  - $\hat{F}_{\mathsf{P}}(t)$: Plug-in CDF estimator, $\hat{F}_{\mathsf{P}}(t) = \frac{1}{N} \sum_{i \in \boldsymbol{A}} \pi_i^{-1} \mathbb{1}\big(\hat{Y}_i \leq t\big)$
  - $\hat{F}_{\mathsf{R}}(t)$: Our residual eCDF estimator
- Quantile Estimators
  - $\hat{t}_{\mathsf{B}}(\alpha)$: The näive quantile function of $\boldsymbol{B}$
  - $\hat{t}_{\mathsf{P}}(\alpha)$: The estimated quantile function associated with our plug-in CDF estimator
  - $\hat{t}_{\mathsf{R}}(\alpha)$: The estimated quantile function associated with our residual eCDF estimator

Name Shortening

Estimator names have been shortened to ‘B’, ‘P’, and ‘R’, respectively, to preserve readability.

Figure 1: *RRMSE Values for MAR Missingness at $n_B = 10,000$*

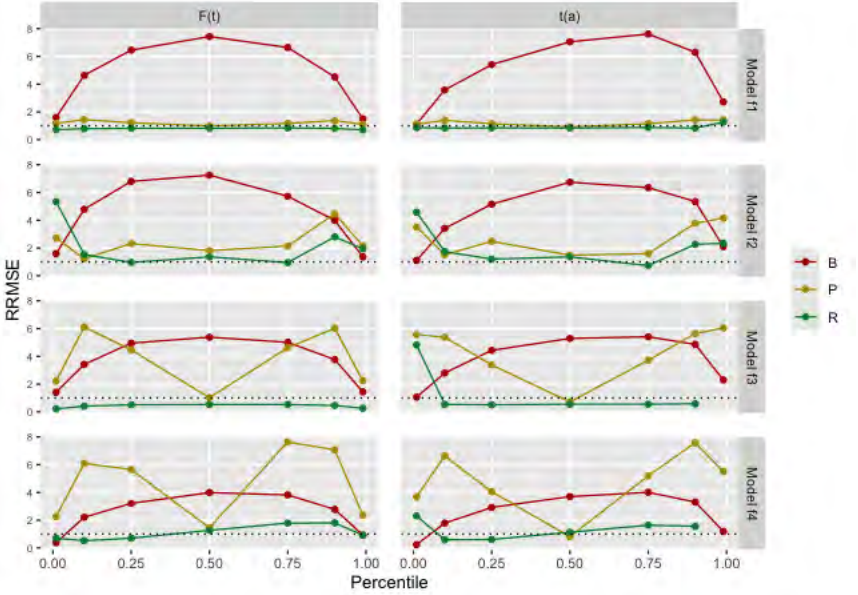
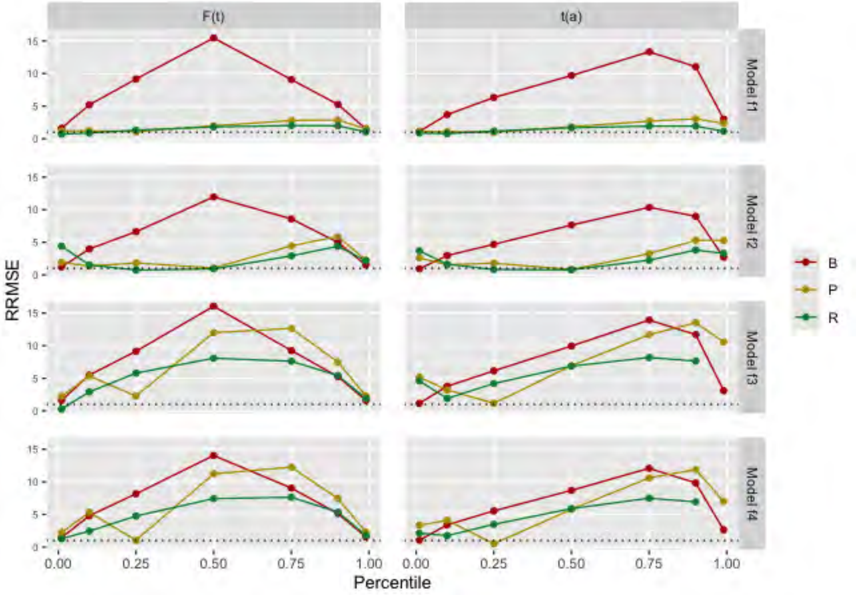


Figure 2: *RRMSE Values for MNAR Missingness at* $n_B = 10,000$



Real Data: Description

- Using NHANES [1] data, we sought to estimate the CDF / quantile function of total cholesterol (in mg/dL) using the following seven covariates:
  - X_1: Biological Sex
  - X_2: Age
  - X_3: Glycohemoglobin (i.e., hemoglobin A1c, in %)
  - X_4: Triglycerides (in mg/dL)
  - X_5: Direct high-density lipoprotein cholesterol (HDL, in mg/dL)
  - X_6: Body mass index (BMI, X_6, kg/m^2)
  - X_7: Pulse

Sampling Setup

- $\mathcal{U}$: Population of U.S. adults
- $\boldsymbol{A}$: 2015-2016 NHANES cohort ($n_{\mathsf{A}} = 2,474$)
- $\boldsymbol{B}$: 2017-2020 cohort ($n_{\mathsf{B}} = 3,770$)
- Performance metric: **percent absolute relative bias**, defined generically for some $\hat{\theta}$ as

$$\mathrm{RB}\left(\hat{\theta}\right) = \frac{\left|\hat{\theta}_\pi - \hat{\theta}\right|}{\hat{\theta}_\pi} \times 100$$

Table 2: *Percent absolute relative bias of $\hat{F}_B(t)$, $\hat{F}_P(t)$, and $\hat{F}_R(t)$, as well as their respective quantile estimators, relative to HT equivalents using the 2015-2016 NHANES dataset ($\boldsymbol{A}$).*

			$\mathrm{RB}(\hat{F})$			$\mathrm{RB}(\hat{t})$		
α	$\hat{F}_\pi(t)$	$\hat{t}_\pi(\alpha)$	B	P	R	B	P	R
1%	0.01	107.00	99.00	100.00	**52.49**	**6.54**	38.71	25.51
10%	0.10	138.00	50.99	99.49	**17.67**	5.80	15.87	**0.37**
25%	0.25	158.00	30.34	70.55	**10.17**	5.06	7.07	**2.03**
50%	0.51	184.00	16.59	16.92	**6.95**	4.89	**2.38**	2.40
75%	0.75	212.00	7.53	21.61	**4.53**	3.77	8.21	**2.41**
90%	0.90	244.00	3.56	9.85	**2.68**	4.51	14.24	**3.60**
99%	0.99	295.00	0.12	0.77	**0.04**	0.68	18.69	**0.50**

Concluding Summary

- **Research Question:** How to extend the field of data integration to distribution function estimation?
- **Idea:** Substitute $\mathbb{1}(Y_i \leq t)$ in $\hat{F}_\pi(t)$ with $\hat{G}_i$, the eCDF of estimated residuals from a regression model built on $\boldsymbol{B}$
- **Empirical Results:** $\hat{F}_{\mathsf{R}}(t)$ seemed robust to model misspecification if ignorability held, and robust to ignorability if the model was correctly specified
- **Next Steps:** Replacing semiparametric regression with a nonparametric alternative

Contact Information:

jrfloodusc@gmail.com

Asymptotic Assumptions

Asymptotic Assumptions

1. The sampling design of $\boldsymbol{B}$ is ignorable; that is, $\Pr(\delta_j|\boldsymbol{X}, Y) = \Pr(\delta_j|\boldsymbol{X})$ for all $j \in \boldsymbol{B}$.

2. The sampling fraction $\frac{n_\mathsf{s}}{N} = \frac{n_\mathsf{A}+n_\mathsf{B}}{N}$ converges to a limit in $(0, 1]$ as both n_s and N tend to infinity [2].

3. There exist some positive real constants c_1, c_2 such that $c_1 \leq \frac{N\pi_i}{\mathbb{E}_\mathcal{D}(n_\mathsf{A})} \leq c_2$ for all $i \in \boldsymbol{A}$, where $\mathbb{E}_\mathcal{D}(\cdot)$ denotes the design-based expectation. Furthermore,

$$\lim_{N\to\infty}\left[\left(\frac{\mathbb{E}_\mathcal{D}(n_\mathsf{A})}{n_\mathsf{B}}\right)^{1/2}\right] = 0,$$

implying $n_\mathsf{B}^{-1/2} = o\left(\mathbb{E}_\mathcal{D}^{-1/2}(n_\mathsf{A})\right)$.

Asymptotic Assumptions (cont.)

4. For any random variable z with finite $2+\delta$ population moments and arbitrarily small $\delta > 0$,

$$\mathrm{Var}_{\mathcal{D}}\left(\frac{1}{N}\sum_{i\in \boldsymbol{A}} \pi_i^{-1} z_i\right) \leq \frac{c_3}{\mathbb{E}_{\mathcal{D}}\left(n_{\mathsf{A}}\right)(N-1)} \sum_{u\in\mathcal{U}_{\mathsf{N}}} \left(z_u - \bar{z}_{\mathsf{N}}\right)^2,$$

where $\bar{z}_{\mathsf{N}} = \frac{1}{N}\sum_{u\in\mathcal{U}_{\mathsf{N}}} z_u$ is the finite population mean of z.

5. For any random variable z with a finite fourth population moment,

$$\mathrm{Var}_{\mathcal{D}}\left(\bar{z}_\pi\right)^{-1/2}\left(\bar{z}_\pi - \bar{z}_{\mathsf{N}}\right) \xrightarrow{\mathcal{L}} \mathrm{N}\left(0,1\right)$$
$$\mathrm{Var}_{\mathcal{D}}\left(\bar{z}_\pi\right)^{-1/2}\widehat{\mathrm{Var}}_\pi\left(\bar{z}_\pi\right) - 1 = O_{\mathsf{P}}\left(\mathbb{E}_{\mathcal{D}}\left(n_{\mathsf{A}}^{-1/2}\right)\right),$$

where $\bar{z}_\pi = \frac{1}{N}\sum_{i\in\boldsymbol{A}} \pi_i^{-1} z_i$ denotes the HT mean estimate of $\bar{z}_{\mathsf{N}}$ and $\widehat{\mathrm{Var}}_\pi\left(\bar{z}_\pi\right)$ denotes the HT estimate of $\mathrm{Var}_{\mathcal{D}}\left(\bar{z}_\pi\right)$.

Asymptotic Assumptions (cont.)

6. $F_{\mathsf{N}}(t)$ converges to a smooth function $F^*(t)$ as N goes to infinity; that is,

$$\lim_{N\to\infty} F_{\mathsf{N}}(t) = F^*(t),$$

where the limiting function $F^*(t)$ is uniformly continuous with finite first and second derivatives.

7. There exists some positive real constants c_3, c_4, c_5 such that $\boldsymbol{X}_i \leq c_3$, $\nu(\boldsymbol{X}_i) \leq c_4$, and $\boldsymbol{X}_j \leq c_5$ for all $i \in \boldsymbol{A}$ and $j \in \boldsymbol{B}$.

Works Cited

References I

[1] Centers for Disease Control and Prevention (CDC). *NHANES - National Health and Nutrition Examination Survey*. https://www.cdc.gov/nchs/nhanes/index.htm (visited: 2023-10-11). 2015-2020.

[2] Richard L Chambers and R Dunstan. "Estimating distribution functions from survey data". In: *Biometrika* 73.3 (1986), pp. 597–604.

[3] Sixia Chen, Shu Yang, and Jae Kwang Kim. "Nonparametric mass imputation for data integration". In: *Journal of Survey Statistics and Methodology* 10.1 (2022), pp. 1–24.

[4] Jae Kwang Kim et al. "Combining non-probability and probability survey samples through mass imputation". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 184.3 (2021), pp. 941–963.

References II

[5] Mateus Maia, Arthur R Azevedo, and Anderson Ara. "Predictive comparison between random machines and random forests". In: *Journal of Data Science* 19.4 (2021), pp. 593–614.

[6] National Academies of Sciences, Engineering, and Medicine. *Federal statistics, multiple data sources, and privacy protection: next steps*. National Academies Press, 2018.

[7] Marie-Hélène Roy and Denis Larocque. "Robustness of random forests for regression". In: *Journal of Nonparametric Statistics* 24.4 (2012), pp. 993–1006.

[8] Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer Science & Business Media, 2003.

[9] Erwan Scornet. "Random forests and kernel methods". In: *IEEE Transactions on Information Theory* 62.3 (2016), pp. 1485–1500.

References III

[10] Arkadiusz Wiśniowski et al. "Integrating probability and nonprobability samples for survey inference". In: *Journal of Survey Statistics and Methodology* 8.1 (2020), pp. 120–147.

[11] Shu Yang and Jae Kwang Kim. "Statistical data integration in survey sampling: A review". In: *Japanese Journal of Statistics and Data Science* 3 (2020), pp. 625–650.

[12] Shu Yang, Jae Kwang Kim, and Youngdeok Hwang. "Integration of data from probability surveys and big found data for finite population inference using mass imputation". In: *Survey Methodology* 47.1 (2021), pp. 29–58.