Large sample theory for merged data from multiple sources

Takumi Saegusa

University of Maryland
Division of Statistics

August 22 2018

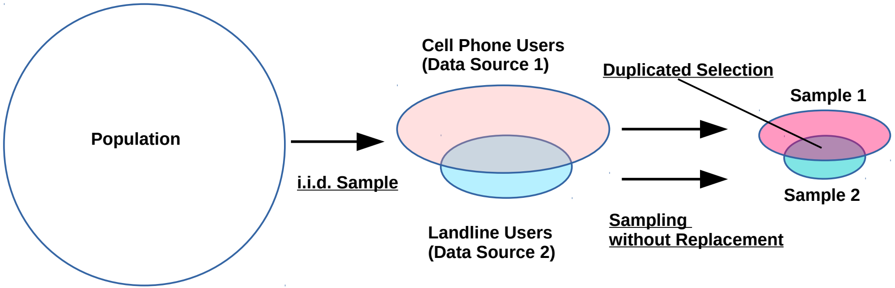
Section 1

Introduction

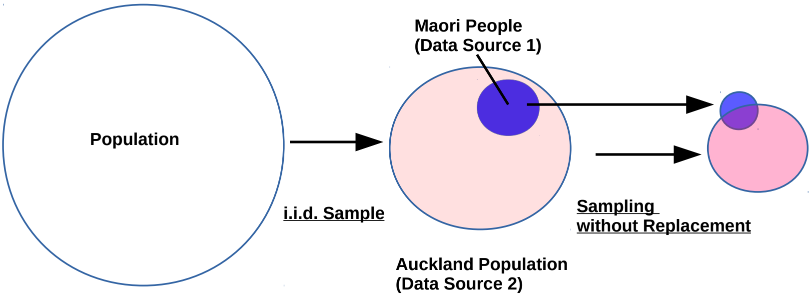
Problem: Data Integration

- Massive data are collected from various sources:
    - online surveys
    - social networks
    - business transactions
    - sensor networks
    - scientific research
    - (in a public health setting) disease registry, clinical trials, epidemiological studies, health surveys, hospital records, healthcare databases, etc.
    - etc.
- The representativeness of a sample critically depends on technology for data collection
- A remedy: to merge multiple data sets with different coverage

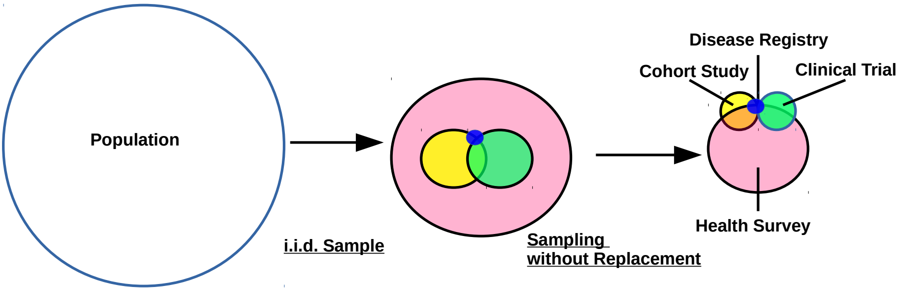
Motivating Examples: Telephone Surveys



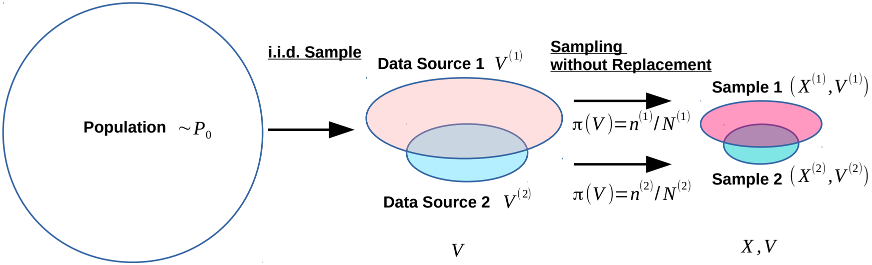
Motivating Examples: Study on Rare Populations



Motivating Examples: Combining Medical Studies



Two-Stage Formulation



The issue: Biased and Dependent Sample with Duplicated Selection

- Biasedness
  - Data Sources of Different Sizes
  - Overlapping Data Sources
- Dependence
  - (Across Samples) Duplicated Selection
  - (Within Samples) Sampling without Replacement
- Lack of Identification of Duplicated Items
  - Independent Data Collection across Data Sources

Resemblance to Other Frameworks

- Stratified Sampling with Overlapping “Strata”
  - (Practice) Single entity designs the entire sampling so that duplicated items can be identified
  - (Method) Naive method produces bias from overlaps
  - (Theory) The quantity of interest is decomposed into stratum means, and they are asymptotically independent due to the disjoint nature of strata.

Resemblance to Other Frameworks

- Stratified Sampling with Overlapping "Strata"
    - (Practice) Single entity designs the entire sampling so that duplicated items can be identified
    - (Method) Naive method produces bias from overlaps
    - (Theory) The quantity of interest is decomposed into stratum means, and they are asymptotically independent due to the disjoint nature of strata.
- Multiple-frame surveys: sampling from overlapping sampling frames
    - (Practice) Applications are limited to survey sampling
    - (Method) Hartley's estimator works well
    - (Theory) Finite population framework
    - (Theory) No empirical process theory

Resemblance to Other Frameworks

- Stratified Sampling with Overlapping "Strata"
  - (Practice) Single entity designs the entire sampling so that duplicated items can be identified
  - (Method) Naive method produces bias from overlaps
  - (Theory) The quantity of interest is decomposed into stratum means, and they are asymptotically independent due to the disjoint nature of strata.
- Multiple-frame surveys: sampling from overlapping sampling frames
  - (Practice) Applications are limited to survey sampling
  - (Method) Hartley's estimator works well
  - (Theory) Finite population framework
  - (Theory) No empirical process theory
- Meta-analysis
  - (Practice) Does not cover overlaps in samples

Resemblance to Other Frameworks

- Stratified Sampling with Overlapping "Strata"
  - (Practice) Single entity designs the entire sampling so that duplicated items can be identified
  - (Method) Naive method produces bias from overlaps
  - (Theory) The quantity of interest is decomposed into stratum means, and they are asymptotically independent due to the disjoint nature of strata.
- Multiple-frame surveys: sampling from overlapping sampling frames
  - (Practice) Applications are limited to survey sampling
  - (Method) Hartley's estimator works well
  - (Theory) Finite population framework
  - (Theory) No empirical process theory
- Meta-analysis
  - (Practice) Does not cover overlaps in samples
- Record Linkage: Identification of duplications
  - (Issue) Produces bias of wrong links and non-links
  - (Theory) Requires a correctly specified model of linking errors

Comparison with Approaches in Sampling Theory

	Finite Population	Super Population	Ours
Randomness			
sampling from data sources	✓	✓	✓
distribution on variables		✓	✓
Model on Variables		✓	✓
Parameter			
finite population parameter	✓		
parameter in the model		✓	✓
Dependence			
within samples	✓	✓	✓
across samples		✓	✓
Applications			
sample mean	✓		
generalized linear model	?	✓	✓
semiparametric model		?	✓
Asymptotics			
LLN	✓	✓(?)	✓
CLT	✓	✓(?)	✓
U-LLN for a class of functions			✓
U-CLT for a class of functions			✓
Conditions			
super population		✓	✓
design	✓	✓	

Section 2

Empirical Process

Empirical Process Approach

- Empirical process is a stochastic process
- very useful in semiparametric and nonparametric models.
- Major tools for statistical theory
    - Uniform LLN and Uniform CLT
    - Rate of convergence
    - Concentration inequalities, etc.

Empirical Process Approach

- Empirical process is a stochastic process
- very useful in semiparametric and nonparametric models.
- Major tools for statistical theory
  - Uniform LLN and Uniform CLT
  - Rate of convergence
  - Concentration inequalities, etc.
- Let $X_1, \ldots, X_n$ i.i.d. P taking values in $\mathcal{X}$. The empirical measure is defined as
$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$
where δ_x is a Dirac measure putting a unit mass at x.
- The empirical measure is a probability measure. The probability of the event $A \subset \mathcal{X}$ under $\mathbb{P}_n$ is
$$\mathbb{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i) = \frac{\#\{X_i : X_i \in A\}}{n},$$
and the expectation of $f(X)$ under $\mathbb{P}_n$ is
$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

- The empirical process indexed by the class $\mathcal{F}$ of functions on $\mathcal{X}$ is defined as
$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P).$$

- The empirical process indexed by the class $\mathcal{F}$ of functions on $\mathcal{X}$ is defined as

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P).$$

- This is a stochastic process indexed by $\mathcal{F}$, i.e., given $f \in \mathcal{F}$, the following random variable is obtained:

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - Pf\right).$$

  Here $Pf = E_P f(X)$ is the expectation of $f(X)$ under P.

- The empirical process indexed by the class $\mathcal{F}$ of functions on $\mathcal{X}$ is defined as

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P).$$

- This is a stochastic process indexed by $\mathcal{F}$, i.e., given $f \in \mathcal{F}$, the following random variable is obtained:

$$\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i) - Pf\right).$$

  Here $Pf = E_P f(X)$ is the expectation of $f(X)$ under P.
- Examples of index sets are
  - $\mathcal{F} = \{t \mapsto 1_{(-\infty,t]}(s) : t \in \mathbb{R}\}$ yields $\mathbb{P}_n 1_{(-\infty,t]} = \mathbb{F}_n(t)$
  - $\mathcal{F} = \{x \mapsto \log p_\theta(x) : \theta \in \Theta\}$

- An important goal of modern empirical process theory is to provide a uniform control of the sample average over the class of functions.

- An important goal of modern empirical process theory is to provide a uniform control of the sample average over the class of functions.

  The class $\mathcal{F}$ of functions on $\mathcal{X}$ is called *P*-Glivenko-Cantelli if

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \rightarrow_P \text{ or }_{a.s.} 0.$$

- An important goal of modern empirical process theory is to provide a uniform control of the sample average over the class of functions.

  The class $\mathcal{F}$ of functions on $\mathcal{X}$ is called P-Glivenko-Cantelli if

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \to_P \text{ or }_{a.s.} 0.$$

- The class $\mathcal{F}$ of functions on $\mathcal{X}$ is called P-Donsker if

$$\mathbb{G}_n \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{F}),$$

  where $\mathbb{G}$ is the P-Brownian bridge, a Gaussian process with covariance function

$$\rho_P(f,g) = \mathrm{Cov}_P(f(X), g(X)) = Pfg - (Pf)(Pg) \ \text{ for } f, g \in \mathcal{F}.$$

- An important goal of modern empirical process theory is to provide a uniform control of the sample average over the class of functions.

  The class $\mathcal{F}$ of functions on $\mathcal{X}$ is called P-Glivenko-Cantelli if

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \to_P \text{ or }_{a.s.} 0.$$

- The class $\mathcal{F}$ of functions on $\mathcal{X}$ is called P-Donsker if

$$\mathbb{G}_n \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty(\mathcal{F}),$$

  where $\mathbb{G}$ is the P-Brownian bridge, a Gaussian process with covariance function

$$\rho_P(f,g) = \mathrm{Cov}_P(f(X), g(X)) = Pfg - (Pf)(Pg) \ \text{ for } f, g \in \mathcal{F}.$$

  - At $f, g \in \mathcal{F}$, this implies

$$\begin{pmatrix} \mathbb{G}_n f \\ \mathbb{G}_n g \end{pmatrix} \to_d \begin{pmatrix} \mathbb{G} f \\ \mathbb{G} g \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \rho_P(f,f) & \rho_P(f,g) \\ \rho_P(f,g) & \rho_P(g,g) \end{pmatrix} \right).$$

  - There exits a version of a Gaussian process with sample continuity. Here we further have asymptotic equicontinuity:

$$\sup_{\rho_P(f,g) < \delta} |\mathbb{G}_n(f-g)| = o_P(1), \quad \text{as } \delta \downarrow 0.$$

Why Empirical Process Theory?

We have enough tools already?

- “Regularity conditions”
- Calculus
- Law of Large Numbers (LLN)
- Central Limit Theorem (CLT)
- Martingale theory if you like

Motivating Example: Uniform LLN

- M-estimator $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} \mathbb{P}_n m(\theta)$
    - Condition for Consistency (van der Vaart 1998, Theorem 5.7)

$$\sup_{\theta \in \Theta} |\mathbb{P}_n m(\theta) - Pm(\theta)| \rightarrow_P 0.$$

Motivating Example: Uniform LLN

- M-estimator $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} \mathbb{P}_n m(\theta)$
  - Condition for Consistency (van der Vaart 1998, Theorem 5.7)

$$\sup_{\theta \in \Theta} |\mathbb{P}_n m(\theta) - Pm(\theta)| \to_P 0.$$

- In some literature, it is claimed that under "regularity conditions", the law of large numbers yields

$$\frac{1}{n} \sum_{i=1}^{n} \log p_{\hat{\theta}_n}(X_i) \to_{a.s.} E \log p_{\theta_0}(X) \quad (?)$$

Motivating Example: Uniform LLN

- M-estimator $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} \mathbb{P}_n m(\theta)$
  - Condition for Consistency (van der Vaart 1998, Theorem 5.7)

$$\sup_{\theta \in \Theta} |\mathbb{P}_n m(\theta) - Pm(\theta)| \to_P 0.$$

- In some literature, it is claimed that under "regularity conditions", the law of large numbers yields

$$\frac{1}{n} \sum_{i=1}^{n} \log p_{\hat{\theta}_n}(X_i) \to_{a.s.} E \log p_{\theta_0}(X) \quad (?)$$

  - The law of large numbers requires the independent summand:

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{\log p_{\hat{\theta}_n}(X_i)}_{Independent?}$$

Motivating Example: Uniform LLN

- M-estimator $\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} \mathbb{P}_n m(\theta)$
  - Condition for Consistency (van der Vaart 1998, Theorem 5.7)

$$\sup_{\theta \in \Theta} |\mathbb{P}_n m(\theta) - Pm(\theta)| \to_P 0.$$

- In some literature, it is claimed that under "regularity conditions", the law of large numbers yields

$$\frac{1}{n} \sum_{i=1}^{n} \log p_{\hat{\theta}_n}(X_i) \to_{a.s.} E \log p_{\theta_0}(X) \quad (?)$$

  - The law of large numbers requires the independent summand:

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{\log p_{\hat{\theta}_n}(X_i)}_{Independent?}$$

  - The sample $X_1, \ldots, X_n$ is independent but $\hat{\theta}_n$ depends on $X_1, \ldots, X_n$. Hence $\log p_{\hat{\theta}_n}(X_1), \ldots, \log p_{\hat{\theta}_n}(X_n)$ are dependent.

- The Glivenko-Cantelli theorem and the dominated convergence theorem yield

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\log p_{\hat{\theta}_n}(X_i) &= \frac{1}{n}\sum_{i=1}^{n}\log p_{\hat{\theta}_n}(X_i) - E\log p_{\hat{\theta}_n}(X) + E\log p_{\hat{\theta}_n}(X) \\
&\leq \sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}\log p_{\theta}(X_i) - E\log p_{\theta}(X)\right| + E\log p_{\hat{\theta}_n}(X) \\
&\to 0 + E\log p_{\theta_0}(X)
\end{aligned}
$$

Motivating Example: Asymptotic Equicontinuity

- The MLE solves the likelihood equation $(1/n)\sum_{i=1}^{n} \dot{\ell}_{\hat{\theta}_n}(X_i) = 0$. For asymptotic normality, Taylor's theorem from Calculus yields

$$0 = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_{\hat{\theta}_n}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + \frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{\theta_n^*}(X_i)(\hat{\theta}_n - \theta_0).$$

  Hence we can apply LLN and CLT to obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left(\frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{\theta_n^*}(X_i)\right)^{-1} \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) \to_d X \sim N(0, I^{-1}).$$

Motivating Example: Asymptotic Equicontinuity

- The MLE solves the likelihood equation $(1/n)\sum_{i=1}^{n} \dot{\ell}_{\hat{\theta}_n}(X_i) = 0$. For asymptotic normality, Taylor's theorem from Calculus yields

$$0 = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_{\hat{\theta}_n}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) + \frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{\theta_n^*}(X_i)(\hat{\theta}_n - \theta_0).$$

Hence we can apply LLN and CLT to obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left(\frac{1}{n}\sum_{i=1}^{n} \ddot{\ell}_{\theta_n^*}(X_i)\right)^{-1} \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) \rightarrow_d X \sim N(0, I^{-1}).$$

- Rubin-Bleuer and Kratina (Annals of Statistics, 2005) adopted a two-phase framework for estimating a Euclidean parameter from estimating equations:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = \underbrace{\sqrt{N}(\hat{\theta}_N - \theta_N)}_{\text{Asymptotic Normality from Design Conditions}} + \underbrace{\sqrt{N}(\theta_N - \theta_0)}_{\text{Asymptotic Normality from Superpopulation Condition}}$$

where $\hat{\theta}_N$ is a solution to weighted estimating equations and θ_N is a solution of unweighted estimating equations.

The previous argument works for many parametric models. In survival analysis, however, semiparametric models play a pivotal role in determining effects of treatments and risk factors. A semiparametric model is a collection of probability measures indexed by

- a finite-dimensional parameter, and
- a infinite-dimensional parameter.

An example is the Cox proportional hazards model with regression parameter $\beta \in \mathbb{R}^d$ and the cumulative hazard function Λ in the class of positive increasing functions. The conditional hazard function given covariates $X = x$ is

$$\lambda(t|x) = \lambda_0(t) \exp(x^T \beta).$$

The previous argument works for many parametric models. In survival analysis, however, semiparametric models play a pivotal role in determining effects of treatments and risk factors. A semiparametric model is a collection of probability measures indexed by

- a finite-dimensional parameter, and
- a infinite-dimensional parameter.

An example is the Cox proportional hazards model with regression parameter $\beta \in \mathbb{R}^d$ and the cumulative hazard function Λ in the class of positive increasing functions. The conditional hazard function given covariates $X = x$ is

$$\lambda(t|x) = \lambda_0(t)\exp(x^T\beta).$$

The following is the likelihood for the Cox model with current status data. Can you use the Taylor expansion around $\theta_0 = (\beta_0, \Lambda_0)$ as usual?

$$\dot{\ell}_\beta(\theta) = \frac{1}{n}\sum_{i=1}^{n} X_i e^{\beta^T X_i}\Lambda(Y_i)\left(\Delta_i \frac{1 - e^{-e^{\beta^T X_i}\Lambda(Y_i)}}{e^{-e^{\beta^T X_i}\Lambda(Y_i)}} - (1 - \Delta_i)\right)$$

Suppose the asymptotic equicontinuity condition holds:

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\dot\ell_{\hat\theta_n}(X_i) - E\log\dot\ell_{\hat\theta_n}(X)\right\} - \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\dot\ell_{\theta_0}(X_i) - E\log\dot\ell_{\theta_0}(X)\right\} = o_P(1+\sqrt{n}\|\hat\theta_n-\theta_0\|).$$

Since $(1/n)\sum_{i=1}^{n}\dot\ell_{\hat\theta_n}(X_i) = 0$ and $E\dot\ell_{\theta_0}(X) = 0$, it follows

$$\begin{aligned}
&\sqrt{n}(E\dot\ell_{\hat\theta_n}(X) - E\dot\ell_{\theta_0}(X)) \\
&= -\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\dot\ell_{\hat\theta_n}(X_i) - E\log\dot\ell_{\hat\theta_n}(X)\right\} \\
&= \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\dot\ell_{\theta_0}(X_i) - E\log\dot\ell_{\theta_0}(X)\right\} + o_P(1+\sqrt{n}\|\hat\theta_n-\theta_0\|)
\end{aligned}$$

If $\theta \to E\dot\ell_\theta(X)$ is differentiable at θ_0 and $\sqrt{n}\|\hat\theta_n-\theta_0\| = O_P(1)$, we obtain

$$\sqrt{n}E\ddot\ell_{\theta_0}(X)(\hat\theta_n-\theta_0) \quad = \quad \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\dot\ell_{\theta_0}(X_i) - E\log\dot\ell_{\theta_0}(X)\right\} + o_P(1).$$

For semiparametric models,

- the derivative $E\ddot\ell_{\theta_0}$ is replaced by the functional derivative.
- the number of the likelihood equations becomes infinitely many.

Motivating Example: Martingale

- The Cox's partial likelihood score can be written as

$$\sum_{i=1}^{n} \int_0^{\tau} \underbrace{H_i(s)}_{\text{Predictable Process}} \; d \; \underbrace{M_i(s)}_{\text{Martingale}}$$

to which Martingale Central Limit Theorem applies.

Motivating Example: Martingale

- The Cox's partial likelihood score can be written as

$$\sum_{i=1}^{n} \int_0^{\tau} \underbrace{H_i(s)}_{\text{Predictable Process}} \, d \underbrace{M_i(s)}_{\text{Martingale}}$$

  to which Martingale Central Limit Theorem applies.

- In the analysis of complex samling data where sampling depends on the event, we analyze inverse probability weighted partial likelihood score

$$\sum_{i=1}^{n} \int_0^{\tau} \underbrace{W_i}_{\text{Not Predictable}} \underbrace{H(s)}_{\text{Predictable Process}} \, d \underbrace{M(s)}_{\text{Martingale}}$$

  so that the Martingale CLT does not apply.

- the Martingale CLT and Empirical process approaches must address dependence issues from complex sampling but the former approach intrinsically fails to address sampling that depends on events even if dependence can be addressed.

Some Literature on the Cox Model in Sampling Theory

- D.Y. Lin, On fitting Cox's proportional hazards models to survey data, Biometrika 87 (2000) 37-47.
  - The paper simply cited Andersen and Gill (Annals of Statistics 10(4) 1982 1100-1120) for consistency but there are too many difficulties left to the reader (martingale, LLN, etc.)
  - The paper simply assumes the existence of the U-CLT a priori.
- S. Rubin-Bleuer, "The proportional hazards model for survey data from independent and clustered super-populations," Journal of Multivariate Analysis 102 (2011), 884-895
  - Most parts assumes sampling does not depend on the event so that the martingale CLT can be used
  - Consistency results counts on K.H. Yuan an R. Jennrich (J. Multivariate Anal. 65 ,1998, 245-260) where the uniform LLN are assumed that this condition is not verified in the paper.
  - The last part where sampling depend on the event counts on Lin (2000).

Some Literature on Empirical Process Theory on Complex Surveys

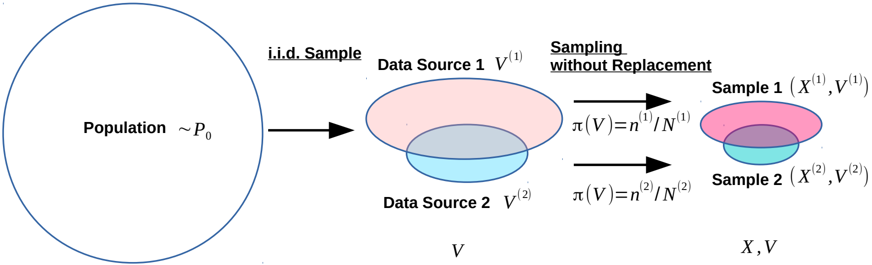
- Bertail, P., Chautru, E. and Clémençon, S. (2017). Empirical processes in survey sampling with (conditional) Poisson designs. Scand. J. Stat. 44 97-111.
  - Rejective Sampling
  - U-LLN is not obtained
- Boistard, H., Lopuhaä , H. P. and Ruiz-Gazen, A. (2017). Functional central limit theorems for single-stage sampling designs. Ann. Statist. 45 1728-1758.
  - Single-stage sampling in a general way
  - finite dimensional CLT is assumed
  - U-LLN is not obtained
  - A class of functions is restricted to a indicator function of variables less than some number
- Daniel Bonnéry, F. Jay Breidt, and François Coquet, Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population
  - A class of functions is restricted to a indicator function of variables less than some number

Section 3

Our Approach

Setting (2 Data Sources)

- V: auxiliary variables available for all items
  - $V_1, \ldots, V_N$ i.i.d.
  - $\mathcal{V}^{(j)}, j = 1, 2$: sampling frames of size $N^{(j)}$ $(\mathcal{V}^{(1)} \cap \mathcal{V}^{(2)} \neq \emptyset)$
  - $V \in \mathcal{V}^{(j)}$ means the item belong to source j
- $n^{(j)}$ items are selected without replacement from source j
- $R^{(j)}, j = 1, 2$: sampling indicator from source j ($R_i^{(j)}$'s with $V_i \in \mathcal{V}^{(j)}$ are dependent)
- $\pi^{(j)}(v)$: sampling probability from source j (e.g. $\pi^{(1)}(v) = n^{(1)}/N^{(1)} I_{\mathcal{V}^{(1)}}(v)$)
- $n^{(j)}/N^{(j)} \to p^{(j)} > 0$.
- X: only available on selected items



Canonical Estimator: Hartley's Estimator

Solution to Duplicated Selection: Hartleys' estimator (1962, 1974)

- reweighing function $\rho : \mathcal{V} \to \mathbb{R}^2$

$$\rho(v) = \left(\rho^{(1)}(v), \rho^{(2)}(v)\right) = \begin{cases} (1,0) & \text{if} \quad v \in \mathcal{V}^{(1)} \cap \left(\mathcal{V}^{(2)}\right)^c \\ (0,1) & \text{if} \quad v \in \left(\mathcal{V}^{(1)}\right)^c \cap \mathcal{V}^{(2)} \\ (c_1, c_2) & \text{if} \quad v \in \mathcal{V}^{(1)} \cap \mathcal{V}^{(2)} \end{cases}$$

where $c_1 + c_2 = 1$.

Canonical Estimator: Hartley's Estimator

Solution to Duplicated Selection: Hartleys' estimator (1962, 1974)

- reweighing function $\rho : \mathcal{V} \to \mathbb{R}^2$

$$\rho(v) = \left(\rho^{(1)}(v), \rho^{(2)}(v)\right) = \begin{cases} (1,0) & \text{if} \quad v \in \mathcal{V}^{(1)} \cap \left(\mathcal{V}^{(2)}\right)^c \\ (0,1) & \text{if} \quad v \in \left(\mathcal{V}^{(1)}\right)^c \cap \mathcal{V}^{(2)} \\ (c_1, c_2) & \text{if} \quad v \in \mathcal{V}^{(1)} \cap \mathcal{V}^{(2)} \end{cases}$$

  where $c_1 + c_2 = 1$.

- Hartley's estimator of $\overline{X} = (1/N) \sum_{i=1}^{N} X_i$ is

$$\frac{1}{N} \sum_{i=1}^{N} \left(\underbrace{\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}}_{\substack{\text{Inverse probability} \\ \text{weighting for source 1}}} \underbrace{\rho^{(1)}(V_i)}_{\substack{\text{Reweighing} \\ \text{for Source 1}}} + \underbrace{\frac{R_i^{(2)}}{\pi^{(2)}(V_i)}}_{\substack{\text{Inverse probability} \\ \text{weighting for source 2}}} \underbrace{\rho^{(2)}(V_i)}_{\substack{\text{Reweighing} \\ \text{for Source 2}}} \right) X_i$$

- Unbiasedness:
  - Biased Sampling: $E[R^{(j)}|V,X] = \pi^{(j)}(V)$.
  - $\rho^{(1)}(v) + \rho^{(2)}(v) = 1$ for every v.
- No identification of duplicated items:

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i)X_i}_{\text{computed from sample from source 1}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i)X_i}_{\text{computed from sample from source 2}}$$

Hartley-Type Empirical Process

- The empirical measure is replaced by Hartley's estimator of the distribution function. Define Hartley-type inverse probability weighted (H-IPW) empirical measure by

$$\mathbb{P}_N^H = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i)\right)\delta_{X_i}$$

- Note that this is NOT a probability measure measure:

$$\mathbb{P}_N^H 1 \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(\frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i)\right) \neq 1$$

  in general in contrast to $\mathbb{P}_n 1 = 1$.
- Define H-IPW empirical process by

$$\mathbb{G}_N^H = \sqrt{N}(\mathbb{P}_N^H - P).$$

Decomposition of H-Empirical Process

Key Idea 1: Decompose H-Empirical Process into different sampling:

- Stage 1 + Stage 2:

$$
\begin{aligned}
\mathbb{G}_N^H f &= \sqrt{N}(\mathbb{P}_N^H - P)f + \sqrt{N}(\mathbb{P}_N - \mathbb{P}_N)f \\
&= \sqrt{N}(\mathbb{P}_N - P)f + \sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f \\
&= \underbrace{\mathbb{G}_N f}_{\text{Sampling from Population}} + \underbrace{\sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f}_{\text{Sampling from Data Sources}}
\end{aligned}
$$

Decomposition of H-Empirical Process

Key Idea 1: Decompose H-Empirical Process into different sampling:

- Stage 1 + Stage 2:

$$
\begin{aligned}
\mathbb{G}_N^H f &= \sqrt{N}(\mathbb{P}_N^H - P)f + \sqrt{N}(\mathbb{P}_N - \mathbb{P}_N)f \\
&= \sqrt{N}(\mathbb{P}_N - P)f + \sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f \\
&= \underbrace{\mathbb{G}_N f}_{\text{Sampling from Population}} + \underbrace{\sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f}_{\text{Sampling from Data Sources}}
\end{aligned}
$$

- It can be shown that $\mathbb{G}_N f$ and $\sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f$ are uncorrelated. If the latter processes converge to Gaussian process, the limiting process of $\mathbb{G}_N^H$ is a sum of independent Gaussian processes.

Decomposition of H-Empirical Process

Key Idea 1: Decompose H-Empirical Process into different sampling:

- Stage 1 + Stage 2:

$$
\begin{aligned}
\mathbb{G}_N^H f &= \sqrt{N}(\mathbb{P}_N^H - P)f + \sqrt{N}(\mathbb{P}_N - \mathbb{P}_N)f \\
&= \sqrt{N}(\mathbb{P}_N - P)f + \sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f \\
&= \underbrace{\mathbb{G}_N f}_{\text{Sampling from Population}} + \underbrace{\sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f}_{\text{Sampling from Data Sources}}
\end{aligned}
$$

- It can be shown that $\mathbb{G}_N f$ and $\sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f$ are uncorrelated. If the latter processes converge to Gaussian process, the limiting process of $\mathbb{G}_N^H$ is a sum of independent Gaussian processes.
- It follows by Donsker theorem,

$$\mathbb{G}_N \rightsquigarrow \mathbb{G}.$$

Key Idea 2: View sampling from sources as a single realization of *m out of n without-replacement bootstrap* with $m = n^{(j)}$ and $n = N^{(j)}$.

- The average within data source j before sampling

$$\mathbb{P}^{(j)}_{N^{(j)}} f = \frac{1}{N^{(j)}} \sum_{i: V_i \in \mathcal{V}^{(j)}} f(X_i)$$

plays a role of sample average in a bootstrap framework.

- The average within data source j after sampling

$$\hat{\mathbb{P}}^{(j)}_{n^{(j)}} f = \frac{1}{n^{(j)}} \sum_{i: V_i \in \mathcal{V}^{(j)}} R^{(j)}_i f(X_{(j),i})$$

plays a role of bootstrap sample average in a bootstrap framework.

Key Idea 2: View sampling from sources as a single realization of m out of n without-replacement bootstrap with $m = n^{(j)}$ and $n = N^{(j)}$.

- The average within data source j before sampling

$$\mathbb{P}^{(j)}_{N^{(j)}} f = \frac{1}{N^{(j)}} \sum_{i:V_i \in \mathcal{V}^{(j)}} f(X_i)$$

  plays a role of sample average in a bootstrap framework.

- The average within data source j after sampling

$$\hat{\mathbb{P}}^{(j)}_{n^{(j)}} f = \frac{1}{n^{(j)}} \sum_{i:V_i \in \mathcal{V}^{(j)}} R^{(j)}_i f(X_{(j),i})$$

  plays a role of bootstrap sample average in a bootstrap framework.

- Sampling from each source:

$$\sqrt{N}(\mathbb{P}^{H}_{N} - \mathbb{P}_N) f \quad = \quad \sum_{j=1}^{2} \sqrt{\frac{N^{(j)}}{N}} \sqrt{N^{(j)}} (\hat{\mathbb{P}}^{(j)}_{n^{(j)}} - \mathbb{P}^{(j)}_{N^{(j)}}) \rho^{(j)} f$$

  where with reindexing $X_{(j),i}, i = 1, \ldots, N^{(j)}, j = 1, 2$.

- It can be shown that $\mathbb{G}_N$ and $\sqrt{N^{(j)}/N}\sqrt{N^{(j)}}(\hat{\mathbb{P}}^{(j)}_{n^{(j)}} - \mathbb{P}^{(j)}_{N^{(j)}})$ are all uncorrelated.

Theorem (Uniform Law of Large Numbers)

Suppose the class $\mathcal{F}$ of measurable functions is P-Glivenko-Cantelli. Then

$$\|\mathbb{P}_N^H - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_N^H f - Pf| \to_P 0.$$

Theorem (Uniform Central Limit Theorem)

Suppose the class $\mathcal{F}$ of measurable functions is the P-Donsker. Then

$$\mathbb{G}_N^H \rightsquigarrow \mathbb{G} + \sum_{j=1}^{2} \sqrt{P(V \in \mathcal{V}^{(j)})} \sqrt{\frac{1 - p^{(j)}}{p^{(j)}}} \mathbb{G}^{(j)} \rho^{(j)} \cdot,$$

where P-Brownian bridge $\mathbb{G}$, and $P^{(j)}$-Brownian bridge $\mathbb{G}^{(j)}$ are all independent. Here $P^{(j)}$ is a conditional probability measure given membership in source j.

Implications

- Asymptotic distribution of $\mathbb{G}_N^H f$:

$$\mathbb{G}_N^H f \rightarrow_d Z^H \sim N(0, \Sigma^H)$$

with

$$\Sigma^H = \underbrace{\mathrm{Var}(f(X))}_{\text{Population Variance}} + \underbrace{\sum_{j=1}^{2} P(V \in \mathcal{V}^{(j)}) \frac{1 - p^{(j)}}{p^{(j)}} \mathrm{Var}\left\{\rho^{(j)}(V) f(X) | \mathcal{V}^{(j)}\right\}}_{\text{Design Variance}}.$$

- Optimal $\rho^{(j)}$ and Optimal Calibration (Deville and Sarndal, JASA, 1992) can be derived from this variance formula.
  - (Our Approach) Optimal based on the limiting distribution of $\mathbb{G}_N^H f = \sqrt{N}(\mathbb{P}_N^H - P)f$.
  - (Finite Population Approach, Lohr and Rao, "Estimation in multiple-frame surveys," JASA, 2006, 101, 1019-1030.) Optimality based on the variable $\mathbb{P}_N^H f$.

Calibration

- General Idea (Deville and Sarndal, JASA 1992): Improve estimation of Horvitz-Thompson estimator of $\mathbb{P}_N X = (1/N)\sum_{i=1}^N X_i$ by the following relationship

$$\mathbb{P}_N V = \underbrace{\frac{1}{N}\sum_{i=1}^N V_i}_{\text{Sample Average}} \approx \underbrace{\frac{1}{N}\sum_{i=1}^N \frac{R_i}{\pi(V_i)} V_i}_{\text{Horvitz-Thompson Estimator}}$$

Calibration

- General Idea (Deville and Sarndal, JASA 1992): Improve estimation of Horvitz-Thompson estimator of $\mathbb{P}_N X = (1/N)\sum_{i=1}^N X_i$ by the following relationship

$$\mathbb{P}_N V = \underbrace{\frac{1}{N}\sum_{i=1}^{N} V_i}_{\text{Sample Average}} \approx \underbrace{\frac{1}{N}\sum_{i=1}^{N} \frac{R_i}{\pi(V_i)} V_i}_{\text{Horvitz-Thompson Estimator}}$$

- For Multiple-frame sampling, Ranalli et al (2018) uses several different relationship to improve the estimation of $\mathbb{P}_N X$. One of them is to use the following idea:

$$\underbrace{\mathbb{P}_N V}_{\text{Sample Average}} \approx \underbrace{\mathbb{P}_N^H V}_{\text{Hartley's estimator}}$$

- Another method considered by Ranalli et al. (2018) uses the relationship given by

$$\underbrace{\mathbb{P}_N VI\{V \in \mathcal{V}^{(j)}\}}_{\text{Sample Average in Source } j} \approx \underbrace{\mathbb{P}_N^H VI\{V \in \mathcal{V}^{(j)}\}}_{\text{Hartley's estimator in Source } j}$$

- Another method considered by Ranalli et al. (2018) uses the relationship given by

$$\underbrace{\mathbb{P}_N VI\{V \in \mathcal{V}^{(j)}\}}_{\text{Sample Average in Source } j} \approx \underbrace{\mathbb{P}_N^H VI\{V \in \mathcal{V}^{(j)}\}}_{\text{Hartley's estimator in Source } j}$$

- Our approach:

$$\underbrace{\frac{1}{N^{(j)}} \sum_{i:V_i \in \mathcal{V}^{(j)}} \rho^{(j)}(V_i) V_i I\{V_i \in \mathcal{V}^{(j)}\}}_{\text{Sample Average in Source } j} \approx \underbrace{\frac{1}{N^{(j)}} \sum_{i:V_i \in \mathcal{V}^{(j)}} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \rho^{(j)}(V_i) V_i}_{\text{Horvitz-Thompson Estimator in Source} j}$$

- Another method considered by Ranalli et al. (2018) uses the relationship given by

$$\underbrace{\mathbb{P}_N VI\{V \in \mathcal{V}^{(j)}\}}_{\text{Sample Average in Source } j} \approx \underbrace{\mathbb{P}_N^H VI\{V \in \mathcal{V}^{(j)}\}}_{\text{Hartley's estimator in Source } j}$$

- Our approach:

$$\underbrace{\frac{1}{N^{(j)}} \sum_{i:V_i \in \mathcal{V}^{(j)}} \rho^{(j)}(V_i) V_i I\{V_i \in \mathcal{V}^{(j)}\}}_{\text{Sample Average in Source } j} \approx \underbrace{\frac{1}{N^{(j)}} \sum_{i:V_i \in \mathcal{V}^{(j)}} \frac{R_i^{(j)}}{\pi^{(j)}(V_i)} \rho^{(j)}(V_i) V_i}_{\text{Horvitz-Thompson Estimator in Source} j}$$

Method	Ours	Ranalli (2)
Which variables?	$\rho^{(1)}(V)V$	V in source 1
	$\rho^{(2)}(V)V$	V in source 2
Where variable come from?	Sampling from source 1	Both sampling
	Sampling from source 2	Both sampling
What is computed	Horvitz-Thompson	Hartley
	Horvitz-Thompson	Hartley

General semiparametric model

- Semiparametric model: $X \sim P_{\theta,\eta} \in \mathcal{P}$
  - (parametric) $\theta \in \Theta \subset \mathbb{R}^p$
  - (nonparametric) $\eta \in H \subset \mathcal{B}$, $\mathcal{B}$, a Banach space
  - Scores $\dot{\ell}_{\theta,\eta}$ for θ and $B_{\theta,\eta}h$ for η with $h \in \mathcal{H}$, Hilbert space
  - Efficient influence function $\tilde{\ell}_0$
  - Semiparametric efficiency bound $I_0^{-1} = E\tilde{\ell}_0^{\otimes 2}$
- Assumptions (for complete data)
  - Smoothness of the model
  - Asymptotic equicontinuity

Hartley-type Weighted Likelihood Estimator (H-WLE)

- The MLE $(\hat{\theta}, \hat{\eta})$ for complete data is obtained from the likelihood equations;

$$\frac{1}{N}\sum_{i=1}^{N} \dot{\ell}_{\hat{\theta},\hat{\eta}}(X_i) = o_P(N^{-1/2}),$$

$$\frac{1}{N}\sum_{i=1}^{N} B_{\hat{\theta},\hat{\eta}} h(X_i) = o_P(N^{-1/2}).$$

- The WLE $(\hat{\theta}_N, \hat{\eta}_N)$ is obtained from the weighted likelihood equations;

$$\frac{1}{N}\sum_{i=1}^{N} \left\{ \frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i) \right\} \dot{\ell}_{\hat{\theta}_N,\hat{\eta}_N}(X_i) = o_P(N^{-1/2}),$$

$$\frac{1}{N}\sum_{i=1}^{N} \left\{ \frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i) \right\} B_{\hat{\theta}_N,\hat{\eta}_N} h(X_i) = o_P(N^{-1/2}).$$

Hartley-type Weighted Likelihood Estimator (H-WLE)

- The MLE $(\hat{\theta}, \hat{\eta})$ for complete data is obtained from the likelihood equations;

$$\frac{1}{N}\sum_{i=1}^{N} \dot{\ell}_{\hat{\theta},\hat{\eta}}(X_i) = o_P(N^{-1/2}),$$
$$\frac{1}{N}\sum_{i=1}^{N} B_{\hat{\theta},\hat{\eta}} h(X_i) = o_P(N^{-1/2}).$$

- The WLE $(\hat{\theta}_N, \hat{\eta}_N)$ is obtained from the weighted likelihood equations;

$$\frac{1}{N}\sum_{i=1}^{N} \left\{ \frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i) \right\} \dot{\ell}_{\hat{\theta}_N,\hat{\eta}_N}(X_i) = o_P(N^{-1/2}),$$
$$\frac{1}{N}\sum_{i=1}^{N} \left\{ \frac{R_i^{(1)}}{\pi^{(1)}(V_i)}\rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)}\rho^{(2)}(V_i) \right\} B_{\hat{\theta}_N,\hat{\eta}_N} h(X_i) = o_P(N^{-1/2}).$$

Theorem (H-WLE for Data Integration)

Assume the WLE's are consistent, and $\|\hat{\eta}_N - \eta_0\| = O_P(N^{-\beta})$. *Then*

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \rightsquigarrow Z \sim N(0, \Sigma),$$

and

$$\Sigma = \underbrace{I_0^{-1}}_{\textit{Superpopulation Variance}} + \underbrace{\sum_{j=1}^{J} P(V \in \mathcal{V}^{(j)}) \frac{1-p_j}{p_j} \mathrm{Var}(\rho^{(j)}(V)\tilde{\ell}_0 | \mathcal{V}^{(j)})}_{\textit{Design Variances}}.$$

Variance Estimation: Plug-in Estimator

- For sampling fraction and data source probability

$$\hat{p}_j = \frac{n^{(j)}}{N^{(j)}}, \quad P(\widehat{V \in \mathcal{V}^{(j)}}) = \frac{N^{(j)}}{N}$$

- For the inverse of the efficient information $I_0^{-1} = E\tilde{\ell}_0^{\otimes 2}$,

$$\widehat{I_0^{-1}} = \mathbb{P}_N^H \tilde{\ell}_{\hat{\theta}_N, \hat{\eta}_N}^{\otimes 2}$$

- For variance from data sources,

$$\begin{aligned} Var(\widehat{\rho^{(j)}\tilde{\ell}_0 | \mathcal{V}^{(j)}}) \quad = \quad & \frac{1}{N^{(j)}} \sum_{i=1}^{N^{(j)}} \frac{R_{(j),i}^{(j)}}{\pi^{(j)}(V_{(j),i})} \left\{ \rho^{(j)}(V_{(j),i}) \tilde{\ell}_{\hat{\theta}_N, \hat{\eta}_N}(X_{(j),i}) \right\}^{\otimes 2} \\ & - \left\{ \sum_{i=1}^{N^{(j)}} \frac{R_{(j),i}^{(j)}}{\pi^{(j)}(V_{(j),i})} \rho^{(j)}(V_{(j),i}) \tilde{\ell}_{\hat{\theta}_N, \hat{\eta}_N}(X_{(j),i}) \right\}^{\otimes 2} \end{aligned}$$

Section 4

Numerical Study

Simulation for Cox Model with Right Censoring

- $T \sim \mathit{Weibull}(\alpha, \beta)$: time to event
- $C \sim \mathit{Uniform}(0, c)$: censoring variable
- $Y = \min\{T, C\}$, $\Delta = I\{T \leq C\}$.
- covariates $Z_1 \sim \mathit{Bernoulli}(1/2)$, $Z_2 \sim N(0, 1)$
- Z_1 is collected only in the final combined sample
- Data sources $\mathcal{V}^{(j)}$ are created from $V = (Y, \Delta, Z_2)$

	$\mathcal{V}^{(1)}$	$\mathcal{V}^{(2)}$	N	$N^{(1)}$	$N^{(2)}$	$n^{(1)}$	$n^{(2)}$	Duplication
Scenario 1	$Z_2 \geq -1$	$Z_2 \leq 1$	500	421	421	85	127	21
			10000	8413	8414	1683	2525	410
Scenario 2	$\mathcal{V}$	$Z_2 \leq 1$	500	500	421	100	127	25
			10000	10000	8413	2000	2524	505
Scenario 3	$\mathcal{V}$	$\Delta = 1$	500	500	76	100	76	15
			10000	10000	1529	2000	1529	305

								Duplication	
	N	$N^{(1)}$	$N^{(2)}$	$N^{(3)}$	$n^{(1)}$	$n^{(2)}$	$n^{(3)}$	twice	thrice
Scenario 4	500	76	423	278	76	43	28	13	1
	10000	8475	5564	1529	848	556	1529	258	9

Table: Sample sizes and the numbers of duplications based on 2000 simulated datasets. In Scenarion 4, $\mathcal{V}^{(1)} = \{\Delta = 1\}$, and membership in $\mathcal{V}^{(2)} \cap \{\mathcal{V}^{(3)}\}^C$, $\mathcal{V}^{(2)} \cap \mathcal{V}^{(3)}$, and $\{\mathcal{V}^{(2)}\}^C \cap \mathcal{V}^{(3)}$ are determined via multinomial logistic regression on Z_2

$\theta_1 = \theta_2$		$\log 2$		0					
N		500	10000	500	10000	500	10000	500	10000
		Scenario 1				Scenario 3			
θ_1	Bias	.024	.0061	.011	.0017	.005	.0009	.006	.0011
	SD	.482	.0985	.429	.0887	.330	.0733	.301	.0676
	SEE	.467	.0989	.419	.0899	.330	.0728	.305	.0668
θ_2	Bias	.005	.0031	.011	.0011	.023	.0003	.001	.0007
	SD	.251	.0526	.234	.0495	.181	.0378	.163	.0342
	SEE	.260	.0524	.244	.0507	.171	.0381	.156	.0334
		Scenario 2				Scenario 4			
θ_1	Bias	.062	.0005	.009	.0010	.010	.0019	.005	.0003
	SD	.479	.0967	.416	.0876	.368	.0789	.372	.0775
	SEE	.467	.0981	.412	.0871	.355	.0789	.347	.0765
θ_2	Bias	.016	.0000	.015	.0001	.023	.0018	.012	.0016
	SD	.250	.0526	.222	.0493	.192	.0407	.185	.0367
	SEE	.252	.0510	.232	.0480	.181	.0407	.169	.0367

Table: Bias, an absolute Monte Carlo sample bias; SD, a Monte Carlo sample standard deviation; SEE, average of a plug-in estimator of a standard error.

$(\alpha, \beta) = (.2, .5)$		$N = 500$				$N = 10000$		
$\theta_1 = \log(2)$	w/o	SC	C	DC	w/o	SC	C	DC
MLE	.246				.0534			
S	.368	.333	.370	.371	.0789	.0720	.0789	.0789
SF	.375	.341	.376	.376	.0809	.0740	.0809	.0804
B	.497	.474	.497	.497	.1060	.1005	.1060	.1060
$\theta_2 = \log(2)$	w/o	SC	C	DC	w/o	SC	C	DC
MLE	.121				.0270			
S	.192	.188	.193	.193	.0407	.0395	.0405	.0403
SF	.197	.192	.197	.196	.0414	.0401	.0412	.0409
B	.258	.253	.258	.258	.0530	.0517	.0530	.0530

Note: S, the proposed weights; SF, ρ for a single-frame estimator; B, a balanced weights; w/o, non-calibration; SC, the proposed calibration; C, the standard calibration; DC, the data-source-specific calibration. All calibrations use U and Y.

National Wilms Tumor Study

- Complete information is available for comparison of designs
- $N = 1957$
- Histology is determined in the final sample
- Event = Relapse of Wilms Tumor
- Data integration ($n = 506$ with 68 duplications)
  - Data Source 1: Death (all sampled)
  - Data Source 2: Unfavorable Histology (50% sampled)
  - Data Source 3: Entire Cohort (10% sampled)
- Stratified Sample ($n = 502$)
  - Stratum 1: Death (all sampled)
  - Stratum 2: Alive with Unfavorable Histology (50% sampled)
  - Stratum 3: the rest (14% sampled)

	Full cohort		Data integration				Stratified sampling	
ρ			Proposed		Balanced			
# subjects	1957		438 (506 with duplication)				502	
Duplication	0		64 (twice)		2 (thrice)		0	
Partial likelihood	-2458.8		-2464.7		-2463.2		-2467.2	
Variable	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE
Histology	1.430	0.125	1.243	0.236	1.383	0.268	1.419	0.190
Age	0.084	0.021	0.045	0.043	0.043	0.047	0.110	0.035
Stage (III/IV)	1.506	0.356	2.680	0.761	2.589	0.848	2.157	0.705
Tumor	0.064	0.020	0.082	0.046	0.076	0.052	0.106	0.041
Stage $\times$Tumor	-0.079	0.029	-0.156	0.061	-0.079	0.068	-0.134	0.055

Note: Histology is measured at a central laboratory.

Table: Point estimates and estimated standard errors in the analysis of the NWTS study with different sampling schemes. "Proposed" means results for the estimator with proposed $\rho^{(j)}$ and "Balanced" means results for the estimator with the value for $\rho^{(j)}$ across sources.

Section 5

Discussion

- Empirical process theory is shown to be extended to data integration problems.
- Empirical process theory is powerful tool to study semiparametric models under complex surveys.
- Discussion above for more than 2 sources and stratified sampling from each source as an alternative to sampling without replacement are straightforward.
- Other sampling designs to combine different data sources are to be investigated.
- Many methods proposed in the i.i.d. setting should be modified to accomodate sampling procedures.

Thank you!