Public Health Data Integration

Takumi Saegusa

University of Maryland
Department of Mathematics

Workshop on Statistical Data Integration 2019, August 06 2019
Section 1

Problem
Problem: Data Integration

- Massive data are collected from online surveys, social networks, business transactions, sensor networks, scientific research, etc.
Problem: Data Integration

- Massive data are collected from online surveys, social networks, business transactions, sensor networks, scientific research, etc.
- Benefit of Data Integration
  - Cost reduction
  - Larger sample size (which contains rare outcomes and exposure)
  - Accelerate scientific discovery
Problem: Data Integration

- Massive data are collected from online surveys, social networks, business transactions, sensor networks, scientific research, etc.
- Benefit of Data Integration
  - Cost reduction
  - Larger sample size (which contains rare outcomes and exposure)
  - Accelerate scientific discovery
- Challenges of Data Integration
  - Heterogeneity
  - Opportunistically collected data
  - Duplication
Heterogeneity in Standard Approach

- Typical Applications: Meta-Analysis
- Combining individual data:
  - Heterogeneity in multiple data sets are often captured by
    - fixed effects
    - random effects
- Combining statistics
  - test statistics
  - P-values
- Why do we need new methods?
Motivating Examples: Study on Rare Populations

- Population
- Korean / Vietnamese (Source 3)
- Landline (Source 1)
- Cell Phone (Source 2)
- Sample 1
- Sample 2
- Sample 3
- i.i.d. Sample

Sampling without Replacement
Motivating Examples: Combining Medical Studies

Challenges for Biostatistical Data Integration

- Duplication
  - Duplication naturally occurs in population-based clinical studies
  - A single random effect or multiple random effects for a multiply selected person?
Challenges for Biostatistical Data Integration

- Duplication
  - Duplication naturally occurs in population-based clinical studies
  - A single random effect or multiple random effects for a multiply selected person?

- Privacy preservation
  - Key identifiers such as names, addresses, and SSNs are not available
  - Identifying duplication by a statistical model necessarily induces bias
Challenges for Biostatistical Data Integration

• Duplication
  • Duplication naturally occurs in population-based clinical studies
  • A single random effect or multiple random effects for a multiply selected person?

• Privacy preservation
  • Key identifiers such as names, addresses, and SSNs are not available
  • Identifying duplication by a statistical model necessarily induces bias

• Target population
  • Well-defined inclusion and exclusion criteria for a single study
  • Unclear target population for meta-analysis
Challenges for Biostatistical Data Integration

- Duplication
  - Duplication naturally occurs in population-based clinical studies
  - A single random effect or multiple random effects for a multiply selected person?

- Privacy preservation
  - Key identifiers such as names, addresses, and SSNs are not available
  - Identifying duplication by a statistical model necessarily induces bias

- Target population
  - Well-defined inclusion and exclusion criteria for a single study
  - Unclear target population for meta-analysis

- Finite population sampling
  - Each data set is often opportunistically collected
  - Dependence may be induced by finite population sampling
The issue: Biased and Dependent Sample with Duplicated Selection

- Biasedness
  - Data Sources of Different Sizes
  - Overlapping Data Sources

- Dependence
  - (Across Samples) Duplicated Selection
  - (Within Sample) Sampling without Replacement

- Lack of Identification of Duplicated Items
  - Independent Data Collection across Data Sources
  - Record linkage?
Resemblance to Other Frameworks

• Stratified Sampling with Overlapping “Strata”
  • (Practice) Single entity designs all sampling procedures without duplication
  • (Method) Inverse probability weighting does not handle duplication
  • (Theory) Stratum-wise CLT to yields independent normals

• Public health data integration
  • (Practice) Different research entities independently conduct data collection
  • (Method) a different weighting handles duplication
  • (Theory) CLT for each data source yields dependent normals
Resemblance to Other Frameworks

• Multiple-frame surveys: sampling from overlapping sampling frames
  • (Practice) Applications are limited to survey sampling with emphasis on finite population parameters
  • (Method) Hartley’s estimator works well
  • (Theory) Finite population framework (≡ small variance)

\[
\frac{1}{N} \sum_{i=1}^{N} R_{i}^{(1)} X_i + \frac{1}{N} \sum_{i=1}^{N} R_{i}^{(2)} X_i \rightarrow_{d} \text{ (Independent Normals)}
\]

Average for Source 1    Average for Source 2

• Public health data integration
  • (Practice) Various statistical models such as semiparametric censored regression and nonparametric survival functions
  • (Method) borrow Hartley’s estimator
  • (Theory) Superpopulation framework (≡ larger variance)

\[
\frac{1}{N} \sum_{i=1}^{N} R_{i}^{(1)} X_i + \frac{1}{N} \sum_{i=1}^{N} R_{i}^{(2)} X_i = \text{ (Dependent Averages)}
\]

Average for Source 1    Average for Source 2
• Meta-analysis
  • (Practice) Does not cover overlaps in samples
  • (Practice) Does not consider target populations
  • (Theory) Standard CLT theory/Bayesian analysis

• Public Health Data Integration
  • (Practice) Duplication and target populations are explicitly considered
Resemblance to Other Frameworks

- Record Linkage: Identification of duplications
  - (Practice) Produces bias of wrong links and non-links but still useful
  - (Method) Correction of bias due to wrong links and non-links
  - (Reference) Lahiri and Larsen, “Regression analysis with linked data”, JASA 2005
  - (Method) Likelihood accounting for record linkage

- Public Health Data Integration
  - (Practice) Enough key identifiers are not available
  - (Practice) For large population studies, overlaps of data sources is plausible but actual duplications may be few
  - (Practice) Much more complicated statistical models are considered and general methods are needed
  - (Practice) Target populations are important for interpretations
Section 2

Estimator
Ideal Properties for Our Estimator

- Addresses duplication
- Clear target population
- Computable without identification of duplicated selection
- Theory available for semiparametric inference
Recall Typical Example: Telephone Surveys

Population

i.i.d. Sample

Cell Phone Users (Data Source 1)

Duplicated Selection

Landline Users (Data Source 2)

Sampling without Replacement

Sample 1

Sample 2
Canonical Estimator: Hartley’s Estimator

**Issues:** Duplicated Selection, Lack of Identification, Biasedness

**Solution:** Hartley’s estimator (1962, 1974)
Canonical Estimator: Hartley’s Estimator

Issues: Duplicated Selection, Lack of Identification, Biasedness
Solution: Hartleys’ estimator (1962, 1974)

• reweighing function \( \rho : \mathcal{V} \rightarrow \mathbb{R}^2 \)

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

where \( c_1 + c_2 = 1 \).
Canonical Estimator: Hartley’s Estimator

Issues: Duplicated Selection, Lack of Identification, Biasedness

Solution: Hartleys’ estimator (1962, 1974)

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

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

where $c_1 + c_2 = 1$.

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

$$\mathbb{P}_N^H \bar{X} \equiv \frac{1}{N} \sum_{i=1}^{N} \left( \frac{R^{(1)}_i}{\pi^{(1)}(V_i)} \rho^{(1)}(V_i) + \frac{R^{(2)}_i}{\pi^{(2)}(V_i)} \rho^{(2)}(V_i) \right) X_i$$

where $R^{(j)}$ is a sampling indicator from data source $j$. 
• **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:**

\[
\frac{1}{N} \sum_{i=1}^{N} \frac{R^{(1)}_i}{\pi^{(1)}(V_i)} \rho^{(1)}(V_i) X_i + \frac{1}{N} \sum_{i=1}^{N} \frac{R^{(2)}_i}{\pi^{(2)}(V_i)} \rho^{(2)}(V_i) X_i
\]

computed from sample from source 1  
computed from sample from source 2
Asymptotic Theory

Conditions:

- $N$ total sample size
- $V \in \mathcal{V} = \mathcal{V}^{(1)} \cup \mathcal{V}^{(2)}$
- $\pi^{(j)}(V) = n^{(j)}/N^{(j)} \to p^{(j)} > 0$ for $V \in \mathcal{V}^{(j)}$

Theorem (Law of Large Numbers)

Suppose $E|X| < \infty$. Then the Hartley's estimator $P^H_N X$ is consistent for $EX$:

$$
P^H_N X \equiv \frac{1}{N} \sum_{i=1}^N \left( \frac{R^{(1)}_i}{\pi^{(1)}(V_i)} \rho^{(1)}(V_i) + \frac{R^{(2)}_i}{\pi^{(2)}(V_i)} \rho^{(2)}(V_i) \right) X_i \to_P EX.
$$
Theorem (Central Limit Theorem)

Suppose $EX^2 < \infty$. Then

$$\sqrt{N} \left( \mathbb{P}_N^H X - EX \right) \to_d Z \sim N(0, \Sigma)$$

where

$$\Sigma = Var(X) + \sum_{j=1}^{2} P(V \in \mathcal{V}^{(j)}) \frac{1 - p^{(j)}}{p^{(j)}} Var(\rho^{(j)}(V)X | \mathcal{V}^{(j)}).$$

$\Sigma$ is large when

- Large data source: $P(V \in \mathcal{V}^{(j)})$ is large
- Small sampling fraction: $n^{(j)}/N^{(j)} \approx p^{(j)} \approx 0.$

$\Sigma$ is the same as in the i.i.d. case when

- Complete selection: $n^{(j)}/N^{(j)} = 1 = p^{(j)}, j = 1, 2.$
Variance Estimation

- For \( \text{Var}(X) = EX^2 - \{EX\}^2 \),
  \[
  \hat{\text{Var}}(X) = \mathbb{P}^H_N X^2 - \{\mathbb{P}^H_N X\}^2.
  \]

- For \( \text{Var}(\rho^{(j)}(V)X|V^{(j)}) \),
  \[
  \text{Var}(\rho^{(j)}(V)X|V^{(j)}) = \frac{1}{N^{(j)}} \sum_{i=1}^{N^{(j)}} \frac{R^{(j)}_{j,i}}{\pi^{(j)}(V_{j,i})} \{\rho^{(j)}(V_{j,i})X_{j,i}\}^2
  \]
  \[\quad - \left\{ \frac{1}{N^{(j)}} \sum_{i=1}^{N^{(j)}} \frac{R^{(j)}_{j,i}}{\pi^{(j)}(V_{j,i})} \rho^{(j)}(V_{j,i})X_{j,i} \right\}^2\]
  where \( X_{j,i} \) is the \( i \)th observation in data source \( j \).

- For \( P(V \in V^{(j)}) \),
  \[
  P(V \in V^{(j)}) = N^{(j)}/N.
  \]

- For \( p^{(j)} \),
  \[
  \hat{p}^{(j)} = n^{(j)}/N^{(j)}.
  \]
Section 3

Empirical Process Theory
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 $X$. The empirical measure is defined as $P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ where $\delta_x$ is a Dirac measure putting a unit mass at $x$. The empirical measure is a probability measure. The probability of the event $A \subset X$ under $P_n$ is $P_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_{A}(X_i) = \frac{\# \{X_i : X_i \in A\}}{n}$, and the expectation of $f(X)$ under $P_n$ is $P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.
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
  \[
P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},
\]
  where $\delta_x$ 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 $P_n$ is
  \[
P_n(A) = \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) = \frac{\# \{X_i : X_i \in A \}}{n},
\]
  and the expectation of $f(X)$ under $P_n$ is
  \[
P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i).
\]
• The empirical process indexed by the class $\mathcal{F}$ of functions on $\mathcal{X}$ is defined as

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

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

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

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

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

$$G_n = \sqrt{n}(P_n - P).$$

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

$$G_nf = \sqrt{n}(P_nf - Pf) = \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) - Pf\right).$$

Here $Pf = E_Pf(X)$ 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 $P_n1_{(-\infty,t]} = F_n(t)$
  - $\mathcal{F} = \{x \mapsto \log p_\theta(x) : \theta \in \Theta\}$
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

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

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

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

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

where $G$ is the $P$-Brownian bridge, a Gaussian process with covariance function $\rho(f, g) = \text{Cov}_P(f(X), g(X)) = Pf - (Pf)(Pg)$ for $f, g \in \mathcal{F}$. At $f, g \in \mathcal{F}$, this implies

$$\begin{pmatrix} G_n f \\ G_n g \end{pmatrix} \to_d \begin{pmatrix} G f \\ G g \end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \text{Var}(f(X)) & \text{Cov}_P(f(X), g(X)) \\ \text{Cov}_P(f(X), g(X)) & \text{Var}(g(X)) \end{pmatrix} \right).$$

Here $\rightsquigarrow$ denotes the weak convergence: $X_n \rightsquigarrow X$ in $\ell^\infty(\mathcal{T})$ if

$$EH(X_n) \to EH(X)$$

for every bounded continuous function $H : \ell^\infty(\mathcal{T}) \to \mathbb{R}$. 
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 1

Consider a parametric model $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$. Suppose $X_1, \ldots, X_n$ are i.i.d. $X$ and $\hat{\theta}_n$ is the MLE converging to $\theta_0$ almost surely. Under “regularity conditions,” the law of large numbers yields

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

Consider a parametric model $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$.
Suppose $X_1, \ldots, X_n$ are i.i.d. $X$ and $\hat{\theta}_n$ is the MLE converging to $\theta_0$ almost surely. Under “regularity conditions,” the law of large numbers yields

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

The law of large numbers requires the independent summand:

$$\frac{1}{n} \sum_{i=1}^{n} \log p_{\hat{\theta}_n}(X_i) \quad [\text{Independent}?]$$
Motivating Example 1

Consider a parametric model $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ with $\Theta \subset \mathbb{R}^d$. Suppose $X_1, \ldots, X_n$ are i.i.d. $X$ and $\hat{\theta}_n$ is the MLE converging to $\theta_0$ almost surely. Under “regularity conditions,” the law of large numbers yields

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

The law of large numbers requires the independent summand:

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

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.
Suppose

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i) - E \log p_\theta(X) \right| \rightarrow_{a.s.} 0 \quad \text{(Glivenko-Cantelli property)}.$$ 

and

$$| \log p_\theta(x) | \leq F(x) \quad \forall x, \theta, \quad EF(X) < \infty \quad \text{(integrable envelope)}.$$
Suppose

\[ \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \log p_{\theta}(X_i) - E \log p_{\theta}(X) \right| \rightarrow_{a.s.} 0 \quad \text{(Glivenko-Cantelli property)}. \]

and

\[ |\log p_{\theta}(x)| \leq F(x) \quad \forall x, \theta, \quad EF(X) < \infty \quad \text{(integrable envelope)}. \]

Then the dominated convergence theorem yields

\[ \frac{1}{n} \sum_{i=1}^{n} \log p_{\hat{\theta}_n}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \log p_{\hat{\theta}_n}(X_i) - E \log p_{\hat{\theta}_n}(X) + E \log p_{\hat{\theta}_n}(X) \]
\[ \leq \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \log p_{\theta}(X_i) - E \log p_{\theta}(X) \right| + E \log p_{\hat{\theta}_n}(X) \]
\[ \rightarrow 0 + E \log p_{\theta_0}(X) \]

Note that \( E \log p_{\hat{\theta}_n}(X) \) is understood as the expectation of \( \log p_{\theta}(X) \) at \( \theta = \hat{\theta}_n \) (i.e., \( \hat{\theta}_n \) is fixed in the expectation).
Motivating Example 2

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

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

Hence we can apply LLN and CLT to obtain

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = - \left(\frac{1}{n} \sum_{i=1}^{n} \ddot{\ell}_{\theta_0^*}(X_i)\right)^{-1} \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) \rightarrow_d X \sim N(0, I^{-1}).
\]
Motivating Example 2

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

\[
0 = \frac{1}{n} \sum_{i=1}^{n} \ell_{\hat{\theta}_n}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\theta_0}(X_i) + \frac{1}{n} \sum_{i=1}^{n} \ell_{\theta_0}^*(X_i)(\hat{\theta}_n - \theta_0).
\]

Hence we can apply LLN and CLT to obtain

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = - \left(\frac{1}{n} \sum_{i=1}^{n} \ell_{\theta_0}^*(X_i)\right)^{-1} \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \ell_{\theta_0}(X_i) \to_d X \sim N(0, I^{-1}).
\]

Semiparametric models play a pivotal role in Biostatistics. It is a collection of probability measures indexed by a finite-dimensional parameter and infinite-dimensional parameter. An example is the Cox proportional hazards model with regression parameter \(\beta \in \mathbb{R}^d\) and the cumulative hazard function \(\Lambda\) in the class of positive increasing functions.
Motivating Example 2

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

\[
0 = \frac{1}{n} \sum_{i=1}^{n} \ell_{\hat{\theta}_n}(X_i) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\hat{\theta}_0}(X_i) + \frac{1}{n} \sum_{i=1}^{n} \ell_{\hat{\theta}_n^*}(X_i)(\hat{\theta}_n - \theta_0).
\]

Hence we can apply LLN and CLT to obtain

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) = -\left(\frac{1}{n} \sum_{i=1}^{n} \ell_{\hat{\theta}_n^*}(X_i)\right)^{-1} \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \ell_{\hat{\theta}_0}(X_i) \to_d X \sim N(0, I^{-1}).
\]

Semiparametric models play a pivotal role in Biostatistics. It is a collection of probability measures indexed by a finite-dimensional parameter and infinite-dimensional parameter. An example is the Cox proportional hazards model with regression parameter \(\beta \in \mathbb{R}^d\) and the cumulative hazard function \(\Lambda\) in the class of positive increasing functions.

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?

\[
\ell_{\beta}(\theta) = \frac{1}{n} \sum_{i=1}^{n} X_i e^{\beta^T X_i} \Lambda(Y_i) \left(\Delta_i \frac{1 - e^{-e^{\beta^T X_i} \Lambda(Y_i)}}{e^{-e^{\beta^T X_i} \Lambda(Y_i)}} - (1 - \Delta_i)\right)
\]
Suppose the asymptotic equicontinuity condition holds:

\[
\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\hat{\theta}_n}(X_i) - E \log \hat{\ell}_{\hat{\theta}_n}(X) \right\} - \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\theta_0}(X_i) - E \log \hat{\ell}_{\theta_0}(X) \right\} = o_P(1 + \sqrt{n} \| \hat{\theta}_n - \theta_0 \|).
\]
Suppose the asymptotic equicontinuity condition holds:

\[
\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_{\hat{\theta}_n}(X_i) - E \log \hat{\ell}_{\hat{\theta}_n}(X) \right\} - \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) - E \log \dot{\ell}_{\theta_0}(X) \right\} = o_P(1 + \sqrt{n} \| \hat{\theta}_n - \theta_0 \|).
\]

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

\[
\sqrt{n}(E \dot{\ell}_{\hat{\theta}_n}(X) - E \dot{\ell}_{\theta_0}(X))
\]

\[
= -\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_{\hat{\theta}_n}(X_i) - E \log \dot{\ell}_{\hat{\theta}_n}(X) \right\}
\]

\[
= \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) - E \log \dot{\ell}_{\theta_0}(X) \right\} + o_P(1 + \sqrt{n} \| \hat{\theta}_n - \theta_0 \|)
\]
Suppose the asymptotic equicontinuity condition holds:

\[
\sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\hat{\theta}_n}(X_i) - E\log \hat{\ell}_{\hat{\theta}_n}(X) \right\} - \sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\theta_0}(X_i) - E\log \hat{\ell}_{\theta_0}(X) \right\} = o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|).
\]

Since \((1/n) \sum_{i=1}^{n} \hat{\ell}_{\hat{\theta}_n}(X_i) = 0\) and \(E\hat{\ell}_{\theta_0}(X) = 0\), it follows

\[
\sqrt{n}(E\hat{\ell}_{\hat{\theta}_n}(X) - E\hat{\ell}_{\theta_0}(X))
\]

\[
= -\sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\hat{\theta}_n}(X_i) - E\log \hat{\ell}_{\hat{\theta}_n}(X) \right\}
\]

\[
= \sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\theta_0}(X_i) - E\log \hat{\ell}_{\theta_0}(X) \right\} + o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|)
\]

If \(\theta \rightarrow E\hat{\ell}_{\theta}(X)\) is differentiable at \(\theta_0\) and \(\sqrt{n}\|\hat{\theta}_n - \theta_0\| = O_P(1)\), we obtain

\[
\sqrt{n}E\ddot{\ell}_{\theta_0}(X)(\hat{\theta}_n - \theta_0) = \sqrt{n}\left\{ \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_{\theta_0}(X_i) - E\log \hat{\ell}_{\theta_0}(X) \right\} + o_P(1).
\]
Suppose the asymptotic equicontinuity condition holds:

\[ \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\hat{\theta}_n}(X_i) - E \log \dot{\theta}_n(X) \right\} - \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) - E \log \dot{\theta}_0(X) \right\} = o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|). \]

Since \((1/n) \sum_{i=1}^{n} \dot{l}_{\hat{\theta}_n}(X_i) = 0\) and \(E \dot{l}_{\theta_0}(X) = 0\), it follows

\[ \sqrt{n}(E \dot{l}_{\hat{\theta}_n}(X) - E \dot{l}_{\theta_0}(X)) \]

\[ = -\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\hat{\theta}_n}(X_i) - E \log \dot{\theta}_n(X) \right\} \]

\[ = \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) - E \log \dot{\theta}_0(X) \right\} + o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|) \]

If \(\theta \rightarrow E \dot{l}_{\theta}(X)\) is differentiable at \(\theta_0\) and \(\sqrt{n}\|\hat{\theta}_n - \theta_0\| = O_P(1)\), we obtain

\[ \sqrt{n}E \ddot{l}_{\theta_0}(X)(\hat{\theta}_n - \theta_0) = \sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \dot{l}_{\theta_0}(X_i) - E \log \dot{\theta}_0(X) \right\} + o_P(1). \]

For semiparametric models,

- the derivative \(E \ddot{l}_{\theta_0}\) is replaced by the functional derivative.
- the number of the likelihood equations becomes infinitely many.
Motivating Example 3: Martingale Theory

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

$$\sum_{i=1}^{n} \int_{0}^{T} H_i(s) \, d \left\langle \underbrace{M_i(s)}_{\text{Martingale}} \right\rangle$$

Predictable Process

Martingale

to which Martingale Central Limit Theorem applies.
Motivating Example 3: Martingale Theory

• The Cox’s partial likelihood score can be written as

\[
\sum_{i=1}^{n} \int_{0}^{T} H_i(s) \, d \{M_i(s)\}
\]

Predictable Process \hspace{1cm} Martingale
to which Martingale Central Limit Theorem applies.

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

\[
\sum_{i=1}^{n} \int_{0}^{T} W_i \mathcal{H}(s) \, d \{M(s)\}
\]

Not Predictable \hspace{1cm} Predictable Process \hspace{1cm} 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

  - The paper 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 assumes the existence of the U-CLT a priori.

  - Most parts assume 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 is assumed but 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

  - Rejective Sampling
  - U-LLN was not obtained

  - Single-stage sampling in a general way
  - finite dimensional CLT is assumed
  - U-LLN was 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 (2012), Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. Bernoulli pp1361-1385
  - A class of functions is restricted to a indicator function of variables less than some number
Some Literature on Empirical Process Theory on Complex Surveys

  • Stratified sampling without replacement
  • The same conditions as in the i.i.d. setting
  • U-LLN was not obtained

  • Stratified sampling without replacement
  • U-LLN and other empirical process results were obtained
Extension of Empirical Process Theory
to Data Integration

In the i.i.d. case,
- Empirical process theory provides useful theoretical tools to study complex models such as semiparametric, nonparametric and high-dimensional models.
- An important assumption is an i.i.d. sample.

In the data integration problems,
- No theory of stochastic process indexed by a class of functions was available before.
- Most statistical models have never been studied in this context.
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

\[
P_H^N = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{R_i^{(1)}}{\pi^{(1)}(V_i)} \rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)} \rho^{(2)}(V_i) \right) \delta_{X_i}
\]

- Note that this is NOT a probability measure measure:

\[
P_H^N 1 = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{R_i^{(1)}}{\pi^{(1)}(V_i)} \rho^{(1)}(V_i) + \frac{R_i^{(2)}}{\pi^{(2)}(V_i)} \rho^{(2)}(V_i) \right) \neq 1
\]

in general in contrast to \(P_n 1 = 1\).

- Define H-IPW empirical process by

\[
G_H^N = \sqrt{N} (P_H^N - P).
\]
Decomposition of H-Empirical Process

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

- Stage 1 + Stage 2:
  \[ G_H N f = \sqrt{N} (p_N - p) f + \sqrt{N} (p_H^N - p_N) f \]
  \[ = G_N f + \sqrt{N} (p_H^N - p_N) f \]

  Sampling from Population  Sampling from Data Source
Decomposition of H-Empirical Process

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

- **Stage 1 + Stage 2:**
  \[
  G_N^H f = \sqrt{N}(\mathbb{P}_N - \mathbb{P})f + \sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f \\
  = \underbrace{G_N f}_\text{Sampling from Population} + \underbrace{\sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f}_\text{Sampling from Data Source}
  \]

- **Sampling from each source:**
  \[
  \sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N)f = \sum_{j=1}^{2} \sqrt{\frac{N(j)}{N}} \sqrt{N(j)}(\hat{\mathbb{P}}^{(j)}_{n(j)} - \mathbb{P}^{(j)}_{N(j)})\rho^{(j)} f
  \]
  where with reindexing \(X^{(j)}, \delta X^{(j)}, i = 1, \ldots, N^{(j)}, j = 1, 2\) for each source,

  \[
  \hat{\mathbb{P}}^{(j)}_{n(j)} = \frac{1}{n(j)} \sum_{i=1}^{N^{(j)}} R^{(j)}_{(j),i} \delta X^{(j)},i, \quad \mathbb{P}^{(j)}_{N(j)} = \frac{1}{N^{(j)}} \sum_{i=1}^{N^{(j)}} \delta X^{(j)},i.
  \]
Decomposition of H-Empirical Process

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

• Stage 1 + Stage 2:

\[
\mathcal{G}_N^H f = \sqrt{N}(\mathbb{P}_N - \mathbb{P}) f + \sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N) f \\
= \mathcal{G}_N f + \sqrt{N}(\mathbb{P}_N^H - \mathbb{P}_N) f
\]

Sampling from Population \hspace{1cm} Sampling from Data Source

• Sampling from each source:

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

where with reindexing \( X_{(j),i}, i = 1, \ldots, N(j), j = 1, 2 \) for each source,

\[
\hat{\mathbb{P}}(j)_{n(j)} = \frac{1}{n(j)} \sum_{i=1}^{N(j)} R_{(j),i} \delta X_{(j),i}, \quad \mathbb{P}(j)_{N(j)} = \frac{1}{N(j)} \sum_{i=1}^{N(j)} \delta X_{(j),i}.
\]

• It can be shown that \( \mathcal{G}_N \) and \( \sqrt{N(j)/N} \sqrt{N(j)} (\hat{\mathbb{P}}(j)_{n(j)} - \mathbb{P}(j)_{N(j)}) \) are all uncorrelated. If the latter processes converge to Gaussian process, the limiting process of \( \mathcal{G}_N^H \) is a sum of independent Gaussian processes.
Dependence and Bootstrap Asymptotics

Key Idea: View sampling from sources as a single realization of \( m \) out of \( n \) without-replacement bootstrap with \( m = n^{(j)} \), \( n = N^{(j)} \) and \( W_{ni} = R_{i}^{(j)} \).

- The average within data source \( j \) before sampling

\[
\mathbb{P}^{(j)}_{N^{(j)}} f = \frac{1}{N^{(j)}} \sum_{i=1}^{N^{(j)}} f(X^{(j)},i)
\]

plays a role of sample average in a bootstrap framework.

- The average within data source \( j \) after sampling

\[
\hat{\mathbb{P}}^{(j)}_{n^{(j)}} f = \frac{1}{n^{(j)}} \sum_{i=1}^{n^{(j)}} R_{i}^{(j)} f(X^{(j)},i)
\]

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

- Let $X_1, \ldots, X_n$ be i.i.d. sample
- Let $\hat{X}_1, \ldots, \hat{X}_n$ be a bootstrap sample.
Bootstrap Asymptotics

- Let $X_1, \ldots, X_n$ be i.i.d. sample
- Let $\hat{X}_1, \ldots, \hat{X}_n$ be a bootstrap sample.
- Let $(W_{n1}, \ldots, W_{nn})$ follows a multinomial distribution with $n$ and probabilities $(1/n, \ldots, 1/n)$.
Bootstrap Asymptotics

- Let $X_1, \ldots, X_n$ be i.i.d. sample
- Let $\hat{X}_1, \ldots, \hat{X}_n$ be a bootstrap sample.
- Let $(W_{n1}, \ldots, W_{nn})$ follows a multinomial distribution with $n$ and probabilities $(1/n, \ldots, 1/n)$.
- Bootstrap average is
  \[ \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i = \frac{1}{n} \sum_{i=1}^{n} W_{ni} X_i \]
Bootstrap Asymptotics

- Let $X_1, \ldots, X_n$ be i.i.d. sample
- Let $\hat{X}_1, \ldots, \hat{X}_n$ be a bootstrap sample.
- Let $(W_{n1}, \ldots, W_{nn})$ follows a multinomial distribution with $n$ and probabilities $(1/n, \ldots, 1/n)$.
- Bootstrap average is
  \[
  \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i = \frac{1}{n} \sum_{i=1}^{n} W_{ni} X_i
  \]
- Bootstrap weights $W_{ni}$’s are dependent.
Bootstrap Asymptotics

- Let $X_1, \ldots, X_n$ be i.i.d. sample
- Let $\hat{X}_1, \ldots, \hat{X}_n$ be a bootstrap sample.
- Let $(W_{n1}, \ldots, W_{nn})$ follows a multinomial distribution with $n$ and probabilities $(1/n, \ldots, 1/n)$.
- Bootstrap average is
  \[
  \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i = \frac{1}{n} \sum_{i=1}^{n} W_{ni} X_i
  \]
- Bootstrap weights $W_{ni}$'s are dependent.
- (Un)conditionally on $X_1, X_2, \ldots$,
  \[
  \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} W_{ni} X_i - \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} (W_{ni} - 1) X_i \rightarrow_d N(0, \text{Var}(X)).
  \]
Bootstrap Asymptotics

- Let $X_1, \ldots, X_n$ be i.i.d. sample
- Let $\hat{X}_1, \ldots, \hat{X}_n$ be a bootstrap sample.
- Let $(W_{n1}, \ldots, W_{nn})$ follows a multinomial distribution with $n$ and probabilities $(1/n, \ldots, 1/n)$.
- Bootstrap average is
  \[
  \frac{1}{n} \sum_{i=1}^{n} \hat{X}_i = \frac{1}{n} \sum_{i=1}^{n} W_{ni} X_i
  \]
- Bootstrap weights $W_{ni}$'s are dependent.
- (Un)conditionally on $X_1, X_2, \ldots,$
  \[
  \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} W_{ni} X_i - \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} (W_{ni} - 1)X_i \rightarrow_d N(0, \text{Var}(X)).
  \]
- In general for exchangeable bootstrap weights (Praestgaard and Wellner, 1993)
  \[
  \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} (W_{ni} - \overline{W}_n) X_i \rightarrow_d N(0, c^2 \text{Var}(X)).
  \]
  where $(1/n) \sum_{i=1}^{n} (W_{ni} - \overline{W}_n)^2 \rightarrow_P c^2.$
Theorem (Uniform Law of Large Numbers)

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

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

Proof.

Use decomposition and bound each term by the triangle inequality. Apply the Glivenko-Cantelli theorem and the bootstrap Glivenko-Cantelli theorem to $\|P_N - P\|_{\mathcal{F}}$ and

$$\left\| \frac{1}{N(j)} \sum_{i=1}^{N(j)} \left( R_{(j),i} - \frac{n(j)}{N(j)} \right) \rho(j)(V_{(j),i})X_{(j),i} \right\|_{\mathcal{F}},$$

respectively. Since $N(j)/N \to_{a.s.} P(V \in \mathcal{V}(j))$ and $n(j)/N(j) \to p(j)$, this completes the proof.
Theorem (Uniform Central Limit Theorem)

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

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

where $P$-Brownian bridge $G$, and $P^{(j)}$-Brownian bridge $G^{(j)}$ are all independent. Here $P^{(j)}$ is a conditional probability measure given membership in source $j$. 
Proof.
Decomposition for $G_N^H$:

\[ G_N^H = \sqrt{N}(P_N - P) + \sqrt{N}(P_N^H - P_N) = G_N + \sum_{j=1}^{2} \sqrt{\frac{N(j)}{N}} \frac{N(j)}{n(j)} \sqrt{N(j)} \frac{1}{N(j)} \sum_{i=1}^{(j)} \left( R(j) - \frac{n(j)}{N(j)} \right) \rho(j)(V(j,i)) \delta(x_{(j),i}) \]

Bootstrap empirical process for source $j$

Apply the Donsker theorem and bootstrap Donsker theorem to each term. □
Section 4

Applications
Nonparametric Estimation of Survival Functions

• Let $T$ be a time to event with distribution function $F = 1 - S$ and $C$ be a censoring variable.

• Under right censoring, we observe $Y = \min\{T, C\}$ and censoring indicator $\Delta = I\{T \leq C\}$.

• Let $X$ be the exposure of interest which is not observed for everyone.

• We are interested in survival curves for different levels of $X$:

$$S(t|x) = P(T \leq t|X = x).$$
Nonparametric Estimation of Survival Functions

- Let $T$ be a time to event with distribution function $F = 1 - S$ and $C$ be a censoring variable.
- Under right censoring, we observe $Y = \min\{T, C\}$ and censoring indicator $\Delta = I\{T \leq C\}$.
- Let $X$ be the exposure of interest which is not observed for everyone.
- We are interested in survival curves for different levels of $X$:
  $$S(t|x) = P(T \leq t | X = x).$$

- Let $V = (\Delta, U)$ where $U$ is a vector of auxiliary variables.
- We select $n^{(j)}$ subjects from source $j$ of size $N^{(j)}$ without replacement and $X$ is measured on selected subjects.
• Observation: survival function is the map of two functions

\[(N, Y) \mapsto \Lambda(t|x) = \int_0^t \frac{N(dt, x)}{Y(t, x)},\]

\[\Lambda(t|x) \mapsto S(t|x) = \prod_{0 < u \leq t} \{1 - \left(\Lambda(t|x) - \Lambda(t - |x))\right)\} \exp\{\Lambda^c(t|x)\}\]

where \(\Lambda^c\) is a continuous part of \(\Lambda\) and

\[N(t, x) = P(T \leq t, \Delta = 1, X = x), \quad Y(t, x) = P(Y \geq t, X = x).\]

These maps are Hadamard differentiable.
• Observation: survival function is the map of two functions

\[(N, Y) \mapsto \Lambda(t|x) = \int_0^t \frac{N(dt, x)}{Y(t, x)},\]

\[\Lambda(t|x) \mapsto S(t|x) = \prod_{0<u\leq t} \{1 - (\Lambda(t|x) - \Lambda(t - |x))\} \exp\{\Lambda^c(t|x)\}\]

where \(\Lambda^c\) is a continuous part of \(\Lambda\) and

\[N(t, x) = P(T \leq t, \Delta = 1, X = x), \quad Y(t, x) = P(Y \geq t, X = x).\]

These maps are Hadamard differentiable.

• The plug-in estimator is

\[\hat{N}_N(t, x) = P_N^H I\{T \leq t, \Delta = 1, X = x\},\]

\[\hat{Y}_N(t, x) = P_N^H I\{Y \geq t, X = x\},\]

\[\hat{\Lambda}_N(t|x) = \int_0^t \frac{\hat{N}_N(dt, x)}{\hat{Y}_N(t, x)},\]

\[\hat{S}_N(t|x) = \prod_{0<u\leq t} \{1 - (\hat{\Lambda}_N(t|x) - \hat{\Lambda}_N(t - |x))\}.\]
Applying the **Functional Delta Method**, we obtain

\[ \sqrt{N}(\hat{S}_N(x) - S(x)) \overset{D}{\rightarrow} \mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 \]

where \( \mathcal{G}_1 \) and \( \mathcal{G}_2 \) are independent Gaussian processes on \([0, \tau]\) with \( \tau < \infty \).
Construction of Confidence Bands: Method

- In the i.i.d. case, $\sqrt{N}(\hat{S}_N(t|x) - S(t|x))$ weakly converges to a Brownian motion process $G'$ with certain covariate function. Since quantiles of $\sup_{t \in [0, \tau]} |G'(t|x)|$ or some modification are analytically obtained,

$$P\left( \sup_{t \in [0, \tau]} \sqrt{N}|\hat{S}_N(t|x) - S(t|x)| \leq q_{1-\alpha} \right) \rightarrow P\left( \sup_{t \in [0, \tau]} |G'(t|x)| \leq q_{1-\alpha} \right) = 1 - \alpha,$$

so that a confidence band is obtained as

$$\hat{S}_N(t|x) - \frac{q_{1-\alpha}}{\sqrt{N}} \leq S(t|x) \leq \hat{S}_N(t|x) + \frac{q_{1-\alpha}}{\sqrt{N}}.$$

- Gillespie and Fisher (Annals, 1979), Hall and Wellner (Biometrika, 1980), Borgan and Liestol, (Scandinavian Journal of Statistics, 1990), Hollander and McKeague (JASA, 1997), etc.

- Another approach is to bootstrap Kaplan-Meier estimators (Akritas, JASA 1986).
Construction of Confidence Bands: Methods

- Observation: $G_1$ is exactly the same limiting process obtained from the analysis of i.i.d. data. The limiting process $G_2$ captures randomness only from sampling from data sources.
Construction of Confidence Bands: Methods

- Observation: $G_1$ is exactly the same limiting process obtained from the analysis of i.i.d. data. The limiting process $G_2$ captures randomness only from sampling from data sources.

- Key Idea: Simulate $G_1$ and bootstrap $G_2$ separately to estimate $G$. 
Construction of Confidence Bands: Methods

- **Observation:** $G_1$ is exactly the same limiting process obtained from the analysis of i.i.d. data. The limiting process $G_2$ captures randomness only form sampling from data sources.
- **Key Idea:** Simulate $G_1$ and bootstrap $G_2$ separately to estimate $G$.
- **For $G_1$, its covariance function can be estimated by celebrated Greenwood’s formula (with weighting). Thus generate a Gaussian process $\hat{G}_1$ based on the estimated covariance function.
Construction of Confidence Bands: Methods

• Observation: $G_1$ is exactly the same limiting process obtained from the analysis of i.i.d. data. The limiting process $G_2$ captures randomness only from sampling from data sources.

• Key Idea: Simulate $G_1$ and bootstrap $G_2$ separately to estimate $G$.

• For $G_1$, its covariance function can be estimated by celebrated Greenwood’s formula (with weighting). Thus generate a Gaussian process $\hat{G}_1$ based on the estimated covariance function.

• For selected items in the same sampling procedure, apply Gross’ bootstrap (If 20 out of 100 are selected from a data source, create 5 copies of selected items to obtain a bootstrap population of size 100 and then select 20 without replacement).
• Generate $\hat{G}_1$ based on the estimated covariance function of $G$.

• Apply Gross' bootstrap and compute the bootstrap estimator $S^*_N(|x)$. Compute $\hat{G}_2 = \sqrt{N}(S^*_N(|x) - \hat{S}_N(|x))$.

• Compute $\hat{G} = \hat{G}_1 + \hat{G}_2$. 
• Generate $\hat{G}_1$ based on the estimated covariance function of $G$.

• Apply Gross' bootstrap and compute the bootstrap estimator $S_N^*(|x|)$. Compute $\hat{G}_2 = \sqrt{N}(S_N^*(|x|) - \hat{S}_N(|x|))$.

• Compute $\hat{G} = \hat{G}_1 + \hat{G}_2$.

• Note that

$$P \left( \sup_{t \in [0, \tau]} |\sqrt{N} \hat{S}_N(t|x) - S(t|x)| \leq q \right) \rightarrow P \left( \sup_{t \in [0, \tau]} |G(t|x)| \leq q \right).$$
• Generate $\hat{G}_1$ based on the estimated covariance function of $G$.

• Apply Gross’ bootstrap and compute the bootstrap estimator $S^*_N(|x|)$. Compute $\hat{G}_2 = \sqrt{N}(S^*_N(|x|) - \hat{S}_N(|x|))$.

• Compute $\hat{G} = \hat{G}_1 + \hat{G}_2$.

• Note that

$$P \left( \sup_{t \in [0, \tau]} |\sqrt{N}|\hat{S}_N(t|x) - S(t|x)| \leq q \right) \rightarrow P \left( \sup_{t \in [0, \tau]} |\hat{G}(t|x)| \leq q \right).$$

• Repeat the above procedures $B$ times and compute the $1 - \alpha\%$ tile $\hat{q}_{1-\alpha}$ of $\sup_{t \in [0, \tau]} |\hat{G}(t|x)|$.

• The $1 - \alpha\%$ confidence band for $S(|x|)$ is

$$\hat{S}_N(t|x) - \frac{\hat{q}_{1-\alpha}}{\sqrt{N}} \leq S(t|x) \leq \hat{S}_N(t|x) + \frac{\hat{q}_{1-\alpha}}{\sqrt{N}}, \quad t \in [0, \tau].$$
• If we compute $1 - \alpha$%tile of $\sup |\hat{G}(t)|/f(x)$ for a positive $f$, then the resultant band is

$$\hat{S}_N(t|x) - \frac{\hat{q}_{1-\alpha}}{\sqrt{N}} f(t) \leq S(t|x) \leq \hat{S}_N(t|x) + \frac{\hat{q}_{1-\alpha}}{\sqrt{N}} f(t), \quad t \in [0, \tau].$$

• An increasing function $f$ mitigates larger uncertainty in the estimation of $S$ on the right tail. We use $f(t) = \exp(t)$ in a simulation.
Simulation Study

- Binary independent covariates $X$ and $Z$ with prevalence 30%.
- $T$ follows the Cox model with the baseline hazard function is based on Weibull($\alpha, \beta$) with $\alpha = .2$ and $\beta = 0.5$, $1$, $5$ and covariates $X$ and $Z$.
- $C \sim \text{Uniform}(0, c)$ s.t. censoring proportions are approximately 80%.
- $V^{(1)} = \{D = 1\}$, logistic regression on $X$ to determine $V^{(2)}$ and $V^{(3)}$.

<table>
<thead>
<tr>
<th>$N$</th>
<th>$N^{(1)}$ ($n^{(1)}$)</th>
<th>$N^{(2)}$ ($n^{(2)}$)</th>
<th>$N^{(3)}$ ($n^{(3)}$)</th>
<th>Dup 2</th>
<th>Dup 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>103 (103)</td>
<td>366 (74)</td>
<td>382 (39)</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>1000</td>
<td>206 (206)</td>
<td>732 (147)</td>
<td>766 (77)</td>
<td>51</td>
<td>2</td>
</tr>
</tbody>
</table>

Table: Sample sizes
**Table:** Simulated coverage probabilities in % of various 95% confidence bands of the survival functions with level $X$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$\beta$</th>
<th>$n$</th>
<th>$m$</th>
<th>equal</th>
<th>variable</th>
<th>Nair</th>
<th>H–W</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>0.5</td>
<td>500</td>
<td>150</td>
<td>94.0</td>
<td>94.4</td>
<td>91.4</td>
<td>95.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>301</td>
<td>94.0</td>
<td>94.0</td>
<td>94.0</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>500</td>
<td>150</td>
<td>95.1</td>
<td>95.0</td>
<td>89.8</td>
<td>95.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>301</td>
<td>94.4</td>
<td>94.8</td>
<td>91.6</td>
<td>95.2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>500</td>
<td>150</td>
<td>93.1</td>
<td>95.0</td>
<td>87.6</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>301</td>
<td>93.1</td>
<td>94.9</td>
<td>87.8</td>
<td>96.4</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>0.5</td>
<td>500</td>
<td>77</td>
<td>93.6</td>
<td>95.2</td>
<td>94.9</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>144</td>
<td>94.5</td>
<td>94.8</td>
<td>93.7</td>
<td>96.8</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>500</td>
<td>72</td>
<td>94.4</td>
<td>94.8</td>
<td>89.8</td>
<td>93.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>141</td>
<td>95.2</td>
<td>94.7</td>
<td>93.4</td>
<td>92.6</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>500</td>
<td>68</td>
<td>90.1</td>
<td>97.5</td>
<td>87.4</td>
<td>95.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1000</td>
<td>135</td>
<td>91.2</td>
<td>95.7</td>
<td>90.6</td>
<td>96.0</td>
</tr>
</tbody>
</table>
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)
Cox Proportional Hazards Model with Right Censoring

- The hazard function given the vector of covariates $X = x$ is
  \[ \lambda(t|x) = \lambda(t) \exp(x^T \theta) \]

  where $\lambda(t)$ is a baseline hazard function and $\theta$ is regression coefficients.

- Under right censoring, we observe $Y = \min\{T, C\}$ where $C$ is a censoring variable.

- The estimator maximizes the weighted nonparametric likelihood
  \[ \prod^H_N \Delta X^T \theta + \log \Lambda\{Y\} - e^{X^T \theta} \Lambda(Y) \]

  where $\Lambda$ is a baseline cumulative hazard function and $\Lambda\{t\}$ is a jump of $\Lambda$ at $t$. $\hat{\theta}_N$ is the weighted partial likelihood estimator and $\hat{\Lambda}_N$ is the weighted Breslow estimator.
Simulation for Cox Model with Right Censoring

- $\lambda(t)$ is based on Weibull($\alpha, \beta$)
- $C \sim \text{Uniform}(0, c)$: censoring variable
- $Y = \min\{T, C\}$, $\Delta = I\{T \leq C\}$.
- covariates $Z_1 \sim \text{Bernoulli}(1/2)$, $Z_2 \sim N(0, 1)$
- $Z_1$ is collected only in the final combined sample
- Data sources $\mathcal{V}^{(j)}$ are created from $\mathcal{V} = (Y, \Delta, Z_2)$
<table>
<thead>
<tr>
<th>Scenario</th>
<th>Property</th>
<th>$N$</th>
<th>$N^{(1)}$</th>
<th>$N^{(2)}$</th>
<th>$N^{(3)}$</th>
<th>$n^{(1)}$</th>
<th>$n^{(2)}$</th>
<th>$n^{(3)}$</th>
<th>Duplication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>$Z_2 \geq -1$</td>
<td>500</td>
<td>421</td>
<td>421</td>
<td>1683</td>
<td>85</td>
<td>127</td>
<td>410</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>$Z_2 \leq 1$</td>
<td>10000</td>
<td>8413</td>
<td>8414</td>
<td>2525</td>
<td>127</td>
<td>410</td>
<td>410</td>
<td>21</td>
</tr>
<tr>
<td>Scenario 2</td>
<td>$\mathcal{V}$</td>
<td>500</td>
<td>500</td>
<td>421</td>
<td>2000</td>
<td>127</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$Z_2 \leq 1$</td>
<td>10000</td>
<td>10000</td>
<td>8413</td>
<td>2524</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 3</td>
<td>$\mathcal{V}$</td>
<td>500</td>
<td>500</td>
<td>76</td>
<td>2000</td>
<td>76</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\Delta = 1$</td>
<td>10000</td>
<td>10000</td>
<td>1529</td>
<td>1529</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$N$</td>
<td>$N^{(1)}$</td>
<td>$N^{(2)}$</td>
<td>$N^{(3)}$</td>
<td>$n^{(1)}$</td>
<td>$n^{(2)}$</td>
<td>$n^{(3)}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario 4</td>
<td>500</td>
<td>76</td>
<td>423</td>
<td>278</td>
<td>13</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10000</td>
<td>8475</td>
<td>5564</td>
<td>1529</td>
<td>258</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table:** Sample sizes and the numbers of duplications based on 2000 simulated datasets. In Scenario 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 $Z_2$. 

...
\[ \theta_1 = \theta_2 \]

<table>
<thead>
<tr>
<th>( N )</th>
<th>Scenario 1</th>
<th>Scenario 3</th>
<th>Scenario 2</th>
<th>Scenario 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>10000</td>
<td>500</td>
<td>10000</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Bias</td>
<td>Bias</td>
<td>Bias</td>
<td>Bias</td>
</tr>
<tr>
<td></td>
<td>.024</td>
<td>.0061</td>
<td>.011</td>
<td>.0017</td>
</tr>
<tr>
<td></td>
<td>.482</td>
<td>.0985</td>
<td>.429</td>
<td>.0887</td>
</tr>
<tr>
<td></td>
<td>.467</td>
<td>.0989</td>
<td>.419</td>
<td>.0899</td>
</tr>
<tr>
<td></td>
<td>.005</td>
<td>.0031</td>
<td>.011</td>
<td>.0011</td>
</tr>
<tr>
<td></td>
<td>.251</td>
<td>.0526</td>
<td>.234</td>
<td>.0495</td>
</tr>
<tr>
<td></td>
<td>.260</td>
<td>.0524</td>
<td>.244</td>
<td>.0507</td>
</tr>
<tr>
<td></td>
<td>.062</td>
<td>.0005</td>
<td>.009</td>
<td>.0010</td>
</tr>
<tr>
<td></td>
<td>.479</td>
<td>.0967</td>
<td>.416</td>
<td>.0876</td>
</tr>
<tr>
<td></td>
<td>.467</td>
<td>.0981</td>
<td>.412</td>
<td>.0871</td>
</tr>
<tr>
<td></td>
<td>.016</td>
<td>.0000</td>
<td>.015</td>
<td>.0001</td>
</tr>
<tr>
<td></td>
<td>.250</td>
<td>.0526</td>
<td>.222</td>
<td>.0493</td>
</tr>
<tr>
<td></td>
<td>.252</td>
<td>.0510</td>
<td>.232</td>
<td>.0480</td>
</tr>
</tbody>
</table>

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.
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)
<table>
<thead>
<tr>
<th>( \rho )</th>
<th>Full cohort</th>
<th>Data integration</th>
<th>Stratified sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td># subjects</td>
<td>1957</td>
<td>438 (506 with duplication)</td>
<td>502</td>
</tr>
<tr>
<td>Duplication</td>
<td>0</td>
<td>64 (twice)</td>
<td>2 (thrice)</td>
</tr>
<tr>
<td>Partial likelihood</td>
<td>-2458.8</td>
<td>-2464.7</td>
<td>-2463.2</td>
</tr>
<tr>
<td>Variable</td>
<td>( \hat{\theta} )</td>
<td>SE</td>
<td>( \hat{\theta} )</td>
</tr>
<tr>
<td>Histology</td>
<td>1.430</td>
<td>0.125</td>
<td>1.243</td>
</tr>
<tr>
<td>Age</td>
<td>0.084</td>
<td>0.021</td>
<td>0.045</td>
</tr>
<tr>
<td>Stage (III/IV)</td>
<td>1.506</td>
<td>0.356</td>
<td>2.680</td>
</tr>
<tr>
<td>Tumor</td>
<td>0.064</td>
<td>0.020</td>
<td>0.082</td>
</tr>
<tr>
<td>Stage ( \times ) Tumor</td>
<td>-0.079</td>
<td>0.029</td>
<td>-0.156</td>
</tr>
</tbody>
</table>

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
• Large sample theory for survival analysis for merged censored data is developed through the extension of empirical process theory.

• Optimal calibration can be obtained from our theory.

• Optimal choice of $\rho$ can be obtained from our theory.

• (Problem) Bootstrap?

• (Problem) Testing?

• (Problem) Other sampling which is more realistic as opportunistically collected data?
Reference

Saegusa, Large Sample Theory for Merged Data from Multiple Sources, 2019, Annals of Statistics.
Thank you!