Day-2 Session-1: Incidence Weighting Estimation under BIG Sampling

Li-Chun Zhang[1,2,3] and ***Melike Oguz-Alper***[2]

[1] ***University of Southampton (L.Zhang@soton.ac.uk)***
[2] ***Statistisk sentralbyraa, Norway***
[3] ***Universitetet i Oslo***

Sample graph under BIGS

Sample graph by BIGS, or *sample BIG*, defined to be

$$\mathcal{B}_s = \big(s_0, \Omega_s; H_s\big)$$

given initial sample s_0 from F, where Ω_s consists of the nodes (in Ω) connected to s_0, and H_s contains the edges connecting s_0 to Ω_s denoted by $H_s = H \cap (s_0 \times \Omega)$

$$\begin{tikzcd}
\kappa_1 & \kappa_2 & \kappa_3 & \\
i_1 \arrow[u] & i_2 \arrow[u] \arrow[ul] & i_3 \arrow[u] & i_4
\end{tikzcd}$$

$F = \{i_1, i_2, i_3, i_4\}$ and $\Omega = \{\kappa_1, \kappa_2, \kappa_3\}$

s_0	Ω_s	H_s
$\{i_1, i_4\}$	$\{\kappa_1\}$	$\{(i_1\kappa_1)\}$
$\{i_2, i_3\}$	$\{\kappa_1, \kappa_2, \kappa_3\}$	$\{(i_2\kappa_1), (i_2\kappa_2), (i_3\kappa_3)\}$

Ancestry knowledge

Ancestry knowledge for sample graph $\mathcal{B}_s = (s_0, \Omega_s; H_s)$:

$$\{\beta_\kappa : \kappa \in \Omega_s\} \quad \text{and} \quad \beta(\Omega_s) = \bigcup_{\kappa \in \Omega_s} \beta_\kappa$$

In particular, out-of-sample nodes $\beta(\Omega_s) \setminus s_0$ needed

$$\begin{tikzcd}
\kappa_1 & \kappa_2 & \kappa_3 & \\
i_1 \arrow[u] & i_2 \arrow[u] \arrow[ul] & i_3 \arrow[u] & i_4
\end{tikzcd}$$

e.g. $s_0 = \{i_1, i_3\}$, $\Omega_s = \{\kappa_1, \kappa_3\}$, $H_s = \{(i_1\kappa_1), (i_3\kappa_3)\}$
then $\beta_{\kappa_1} = \{i_1, i_2\}$, $\beta_{\kappa_3} = \{i_3\}$ and $\beta(\Omega_s) \setminus s_0 = \{i_2\}$

$$\begin{tikzcd}
\kappa_1 & \kappa_2 & \cdots & \kappa_N \\
i_1 \arrow[u] & i_2 \arrow[u] & \cdots & i_N \arrow[u]
\end{tikzcd}$$

Element sampling

$$\begin{tikzcd}
\kappa_1 & \kappa_2 & \cdots & \kappa_N \\
 & i_1 \arrow[u] \arrow[ul] & \cdots & i_{|F|} \arrow[u] \arrow[ul]
\end{tikzcd}$$

Cluster sampling

Incidence weighting estimator (Patone and Zhang, 2020)

Total of interest: $\theta = \sum_{\kappa \in \Omega} y_\kappa$

IWE based on $\mathcal{B}_s = (s_0, \Omega_s; H_s)$ by BIGS:

$$\hat{\theta} = \sum_{(i\kappa) \in H_s} W_{i\kappa} \frac{y_\kappa}{\pi_i}$$

The IWE is unbiased for θ provided, for each $\kappa \in \Omega$,

$$\sum_{i \in \beta_\kappa} E(W_{i\kappa} | \delta_i = 1) = 1$$

Moreover,

$$V(\hat{\theta}) = \sum_{\kappa \in \Omega} \sum_{\ell \in \Omega} (\Delta_{\kappa\ell} - 1) y_\kappa y_\ell$$

$$\Delta_{\kappa\ell} = \sum_{i \in \beta_\kappa} \sum_{j \in \beta_\ell} \frac{\pi_{ij}}{\pi_i \pi_j} E\big(W_{i\kappa} W_{j\ell} | \delta_i \delta_j = 1\big)$$

HH-type estimator

So-called Hansen-Hurwitz (HH) type estimator uses weights $\omega_{i\kappa}$ that are constant of sampling, such that

$$\sum_{i\in\beta_\kappa} E(\omega_{i\kappa}|\delta_i = 1) = \sum_{i\in\beta_\kappa} \omega_{i\kappa} = 1$$

(Birnbaum and Sirken, 1965), where *multiplicity* weights

$$\omega_{i\kappa} \equiv d_\kappa^{-1} \quad \text{and} \quad d_\kappa = |\beta_\kappa|$$

are common for network sampling (e.g. Sirken, 2005), ACS (Thompson, 1990), indirect sampling (Birnbaum and Sirken, 1965; Lavalleè, 2007). Probability and inverse degree-adjusted (PIDA) weights (Patone and Zhang, 2020):

$$\omega_{i\kappa} \propto d_i^{-\gamma} \pi_i$$

where d_i = no. nodes connected to sampling unit i in s_0

HH-type estimator

- F = clinics, Ω = patients of a certain disease
  d_i = no. patients receiving treatment at hospital i
  d_κ = no. hospitals that treat patient κ
- F = parent (mother or father), Ω = children
  d_i = no. children of person i
  d_κ = no. parents in F of child κ
- F = Twitter accounts, Ω = followers (Twitter accounts)
  d_i = no. followers of account i
  d_κ = no. accounts κ follows
- F = products (online market), Ω = buying customers
  d_i = no. buyers of product i
  d_κ = no. products bought by κ
- $F = \Omega$ = individuals, $(i\kappa) \in H$ if in-contact, incl. $i = \kappa$

HH-type estimator

$$\hat{\theta}_z = \sum_{i \in s_0} \frac{z_i}{\pi_i} \qquad \text{and} \qquad z_i = \sum_{\kappa \in \alpha_i} \omega_{i\kappa} y_\kappa$$

where z_i is a constructed constant for each $i \in F$
and $\alpha_i = \{\kappa \in \Omega : (i\kappa) \in H\}$ its connected study units
Associated sampling variance

$$V(\hat{\theta}_z) = \sum_{i \in F} \sum_{j \in F} (\frac{\pi_{ij}}{\pi_i \pi_j} - 1) z_i z_j$$

PIDA weights (prop. to $d_i^{-\gamma} \pi_i$) aim to even out z_i / π_i
However, $d_i = |\alpha_i|$ for $i \in \beta(\Omega_s) \setminus s_0$ requires additional information beyond the ancestry knowledge
E.g. no. children to an out-of-s_0 parent in Birth Register

HT-estimator (HTE)

$$\hat{\theta}_y = \sum_{\kappa \in \Omega_s} y_\kappa \pi_{(\kappa)}^{-1} = \sum_{\kappa \in \Omega_s} y_\kappa \Big(\sum_{i \in s_0 \cap \beta_\kappa} W_{i\kappa} \pi_i^{-1} \Big)$$

is an IWE, where $W_{i\kappa}$ satisfy $\sum_{i \in s_0 \cap \beta_\kappa} W_{i\kappa} \pi_i^{-1} = \pi_{(\kappa)}^{-1}$
$W_{i\kappa}$ sample-dependent if $|\beta_\kappa| > 1$, e.g. $\beta_{\kappa_1} = \{i_1, i_2\}$ in

$$\begin{tikzcd}
\kappa_1 & \kappa_2 & \kappa_3 & \\
i_1 \arrow[u] & i_2 \arrow[u] \arrow[ul] & i_3 \arrow[u] & i_4
\end{tikzcd}$$

- $s_0 \cap \beta_{\kappa_1} = \{i_1\}$: $W_{i_1\kappa_1} = \pi_{i_1}/\pi_{(\kappa_1)}$
- $s_0 \cap \beta_{\kappa_1} = \{i_2\}$: $W_{i_2\kappa_1} = \pi_{i_2}/\pi_{(\kappa_1)}$
- $s_0 \cap \beta_{\kappa_1} = \{i_1, i_2\}$: $W_{i_1\kappa_1} = a\frac{\pi_{i_1}}{\pi_{(\kappa_1)}}$, $W_{i_2\kappa_1} = (1-a)\frac{\pi_{i_2}}{\pi_{(\kappa_1)}}$

HT-type estimator*

Let sample-dependent weights $W_{i\kappa}$ satisfy

$$\eta_{s_\kappa} = \pi_{(\kappa)} \sum_{i \in s_\kappa} \frac{W_{i\kappa}}{\pi_i}$$

$$\sum_{s_\kappa} \Pr(s_0 \cap \beta_\kappa = s_\kappa) \eta_{s_\kappa} = \pi_{(\kappa)}$$

HTE is the special case of $\eta_{s_\kappa} \equiv 1$
HT-type estimator given η_{s_κ} that differs for different sample intersects s_κ subject to the restriction above
But HTE = RB-estimator of such a HT-type estimator

$$E\Big(\sum_{\kappa \in \Omega_s} \sum_{i \in s_\kappa} \frac{W_{i\kappa}}{\pi_i} y_\kappa | \Omega_s \Big) = \sum_{\kappa \in \Omega_s} y_\kappa E\Big(\frac{\eta_{s_\kappa}}{\pi_{(\kappa)}} | \kappa \in \Omega_s \Big)$$

$$= \sum_{\kappa \in \Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}} \sum_{s_\kappa} \frac{\Pr(s_0 \cap \beta_\kappa = s_\kappa)}{\pi_{(\kappa)}} \eta_{s_\kappa} = \sum_{\kappa \in \Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}}$$

Priority-rule estimator (Birnbaum and Sirken, 1965)

Apply *priority rule* to the sample edges H_s:

$$I_{i\kappa} = \begin{cases} 1 & \text{if } i = \min(s_0 \cap \beta_\kappa) \\ 0 & \text{otherwise.} \end{cases}$$

Let $p_{(i\kappa)} = \Pr(I_{i\kappa} = 1 | \kappa \in \Omega_s)$ for prioritisation, and

$$\hat{\theta}_p = \sum_{(i\kappa) \in H_s} \Big(\frac{I_{i\kappa} \omega_{i\kappa}}{p_{(i\kappa)}} \Big) \frac{y_\kappa}{\pi_i}$$

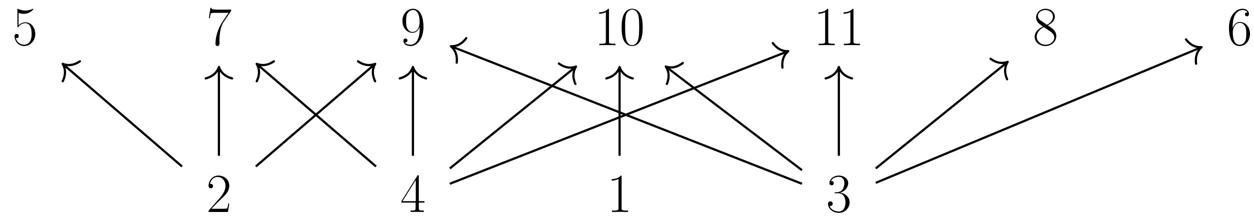
<u>Biased</u> if $p_{(i\kappa)}$ can be 0 for some $(i\kappa) \in H_s$, e.g. $\beta_\kappa = F$, or generally, if $\exists \kappa \in \Omega$ with $|\beta_\kappa| > 1$, where

$$\Pr(|s_0 \cap \beta_\kappa| > 1 \mid \kappa \in \Omega_s) = 1$$

then $p_{(i\kappa)} = 0$ for $i = \max(\beta_\kappa)$ — Patone and Zhang (2020)

Numerical example (Patone, 2020)

Consider BIGS from below, with SRS of s_0 and $|s_0| = 2$:



HH-type PIDA weights given γ

PIDA-γ	z_1	z_2	z_3	z_4	S_z^2
0	0.33	1.83	3.17	1.67	1.34
1	0.69	2.00	2.83	1.48	0.81
2	0.91	2.16	2.61	1.32	0.60
3	0.98	2.31	2.48	1.23	0.57

Variance of IWE

	$\hat{\theta}_{z0}$	$\hat{\theta}_{z1}$	$\hat{\theta}_{z2}$	$\hat{\theta}_{z3}$	$\hat{\theta}_p$	$\hat{\theta}_{pD}$	$\hat{\theta}_{pA}$	$\hat{\theta}_y$
Variance	5.37	3.25	2.41	2.28	3.06	2.55	6.32	3.98

$\hat{\theta}_{pD}$ given ordered $\tilde{F} = \{3, 4, 2, 1\}$ $\quad \hat{\theta}_{pA}$ given $\tilde{F} = \{1, 2, 4, 3\}$

[1] Birnbaum, Z.W. and Sirken, M.G. (1965). *Design of Sample Surveys to Estimate the Prevalence of IRareDiseases: Three Unbiased Estimates*. Vital and Health Statistics, Ser. 2, No.11. Washington:Government Printing Office.

[2] Lavalleè, P. (2007). *Indirect Sampling*. Springer.

[3] Patone, M. (2020) *Topics of Statistical Analysis with Social Media Data*. Unpublished PhD Thesis.

[4] Patone, M. and Zhang, L.-C. (2020). Incidence weighting estimation under bipartite incidence graph sampling. `https://arxiv.org/abs/2004.04257`

[5] Sirken, M.G. (2005). *Network Sampling*. In *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd. DOI: 10.1002/0470011815.b2a16043

[6] Thompson, S.K. (1990). Adaptive cluster sampling. *Journal of the American Statistical Association*, 85:1050–1059.