Day-2 Session-2: Strategy BIGS-IWE

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***

Observation links under graph sampling

Let Ω consist of the study units defined in G, such as nodes, case networks, triangles, whose total of interest is

$$\theta = \sum_{\kappa \in \Omega} y_\kappa$$

Let $F \subseteq U$ be the sampling frame for an *initial* sample s_0, following which Ω_s is observed from Ω by some specified *observation procedure*, such as under ACS

$\forall \kappa \in \Omega$, let $\beta_\kappa \subseteq F$ be such that, for any $i \in \beta_\kappa$, we have

$$\Pr(\kappa \in \Omega_s | i \in s_0) = 1 \tag{1}$$

i.e. under sampling from valued graph G with associated F and Ω, the study unit κ is observed in the sample Ω_s whenever node i in β_κ is included in the initial sample s_0

Observation links: $H = \bigcup_{\kappa \in \Omega} \beta_\kappa \times \kappa$ and $\mathcal{B} = (F, \Omega; H)$

Strategy BIGS-IWE using β_κ

$\boxed{\text{Theorem}}$ (Zhang and Oguz-Alper, 2020)*
Strategy BIGS-IWE defined for $\mathcal{B} = (F, \Omega; H)$ subjected to (1) and $\sum_{i \in \beta_\kappa} E(W_{i\kappa} | \delta_i = 1) = 1$ is unbiased for θ, under sampling from valued graph $G = (U, A)$, provided

(i) $\forall \kappa \in \Omega$, we have $\beta_\kappa \neq \emptyset$ such that $\bigcup_{i \in F} \alpha_i = \Omega$ in $\mathcal{B}$;

(ii) the observation procedure of sampling from G ensures the ancestry knowledge of Ω_s in $\mathcal{B}$.

$\boxed{\text{Proof}}$ Given (i), every κ in Ω has a positive probability of being included in Ω_s under BIGS from $\mathcal{B} = (F, \Omega; H)$. Given (ii), the IWE can be defined with respect to BIGS from $\mathcal{B}$ by virtue of (1). Given $\sum_{i \in \beta_\kappa} E(W_{i\kappa} | \delta_i = 1) = 1$ in addition, the IWE is unbiased for θ.

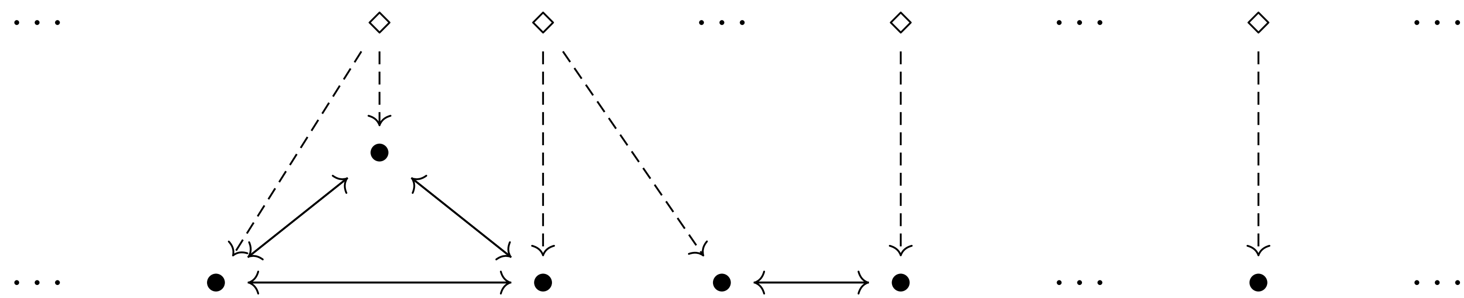
Indirect sampling: BIGS directly

- F = clinics, Ω = patients of a certain disease (e.g. Birnbaum and Sirken, 1965) (i) if excluding undiagnosed, (ii) if $\beta_\mathcal{K} \setminus s_0$ observed
- F = parent (mother or father), Ω = children (e.g. Lavalleè, 2007) (i) if excluding orphans, (ii) given Birth Register
- F = Twitter accounts, Ω = followers (Twitter accounts) (i) and (ii) guaranteed if BIGS by Twitter the company Depends on API provided by Twitter for the others

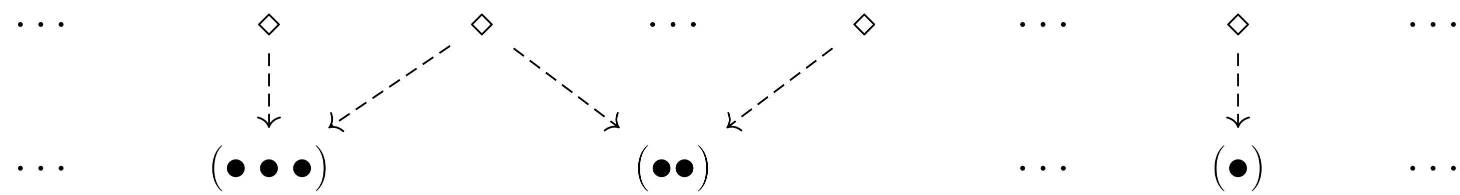
NB. as finite population sampling: potential non-sampling errors associated with frame, observation, operation, etc.

Network sampling (e.g. Sirken, 2005)

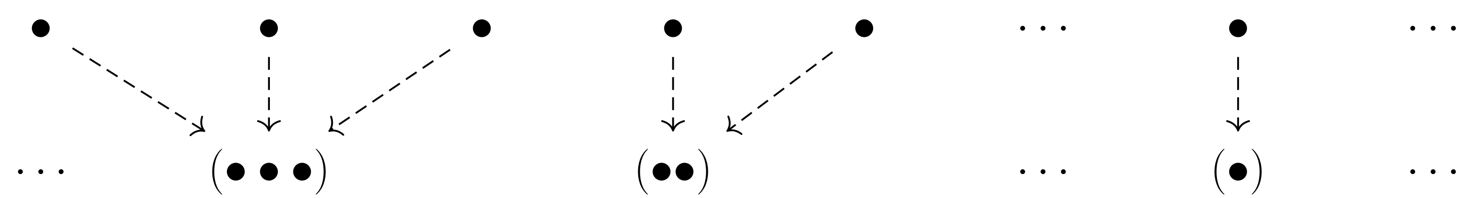
Sampling of siblings ($\bullet$) e.g. via households ($\diamond$): siblings reporting each other, sibling network exhausted



$$\Omega = \{\ \kappa : \text{individual}\ \} \qquad |\beta_\kappa| \geq 1$$



$$\Omega = \{\ \kappa : \text{sibling network}\ \} \qquad |\beta_\kappa| \geq 1$$



$$F = U(\Omega) \qquad \Omega = \{\ \kappa : \text{sibling network}\ \} \qquad |\beta_\kappa| \geq 1$$

ACS (e.g. Thomspon, 1990; Zhang, 2020)

$$
\begin{tikzcd}
1 & 0 & 2 & 10 & 1000 \\
1 \arrow[u] & 0 \arrow[u] & 2 \arrow[u] & 10 \arrow[u] \arrow[ul] \arrow[ur] & 1000 \arrow[u] \arrow[ul] \arrow[ulll]
\end{tikzcd}
$$

$F = \Omega = U$ $(H, \mathcal{B})$ with β_κ containing all ACS-links

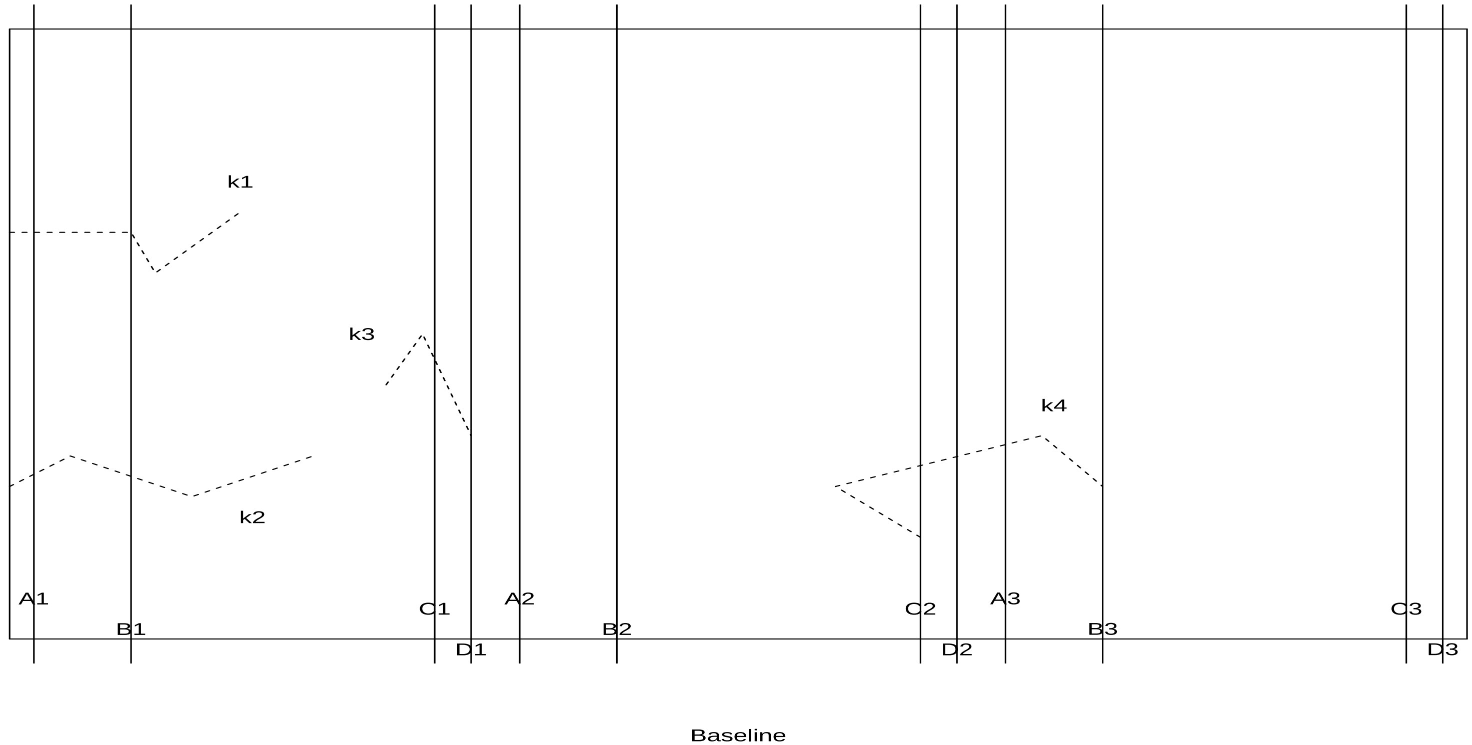
$$
\begin{tikzcd}
1 & 0 & 2 & 10 & 1000 \\
1 \arrow[u] & 0 \arrow[u] & 2 \arrow[u] & 10 \arrow[u] \arrow[ur] & 1000 \arrow[u] \arrow[ul]
\end{tikzcd}
$$

$F = \Omega = U$ $(H^*, \mathcal{B}^*)$ with β^*_κ of certain ACS-links

$$
\begin{tikzcd}
1 & 0 & 2 & & (10, 1000) & \\
1 \arrow[u] & 0 \arrow[u] & 2 \arrow[u] & 10 \arrow[ur] & & 1000 \arrow[ul]
\end{tikzcd}
$$

$F = U$ $\{\Omega : \text{cluster}\}$ $(H^*, \mathcal{B}^*)$ with β^*_κ of certain ACS-links

Line-intercept sampling (LIS, Becker, 1991)



Four systematic samples A, B, C and D, each containing 3 positions, are drawn on the *baseline* that is equally divided into 3 sections of length 12 miles each

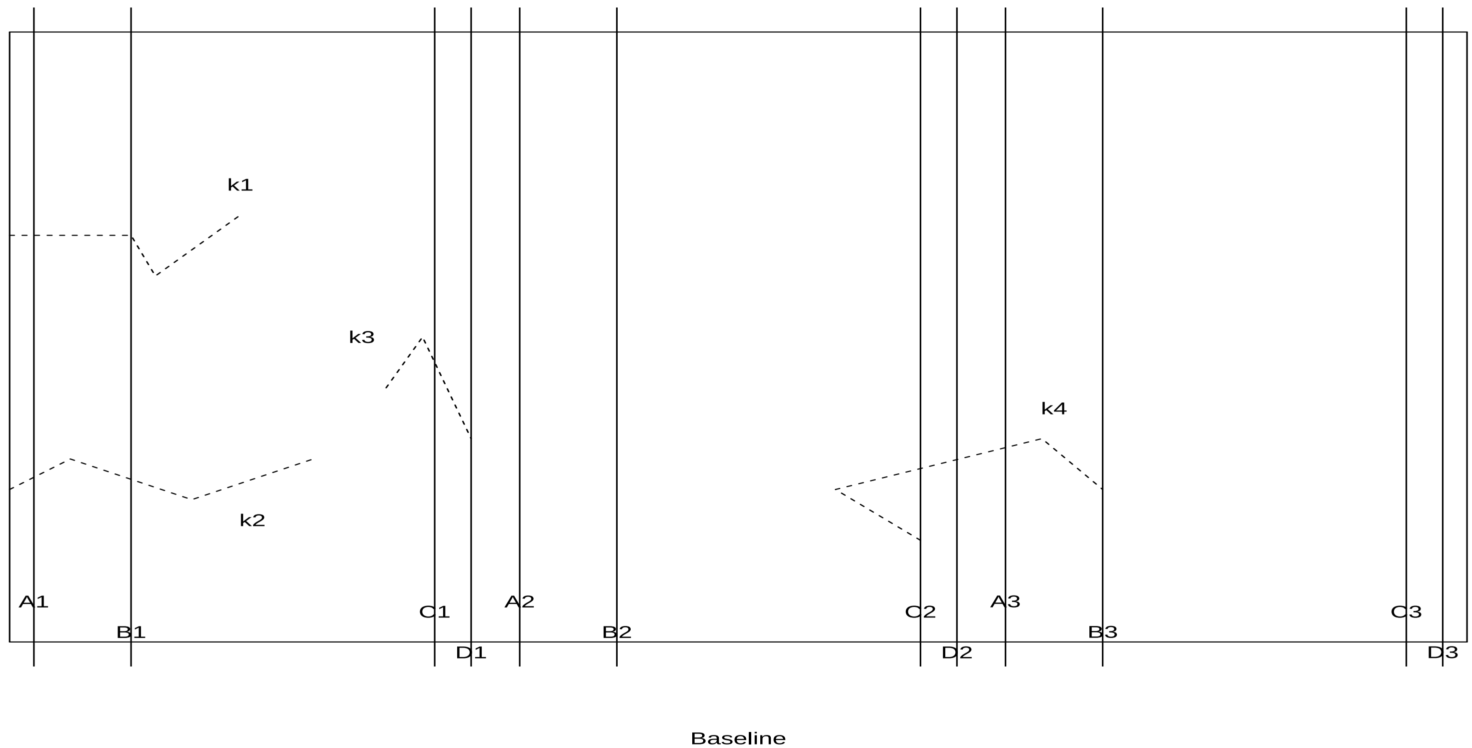
Follow the lines and any intercepting wolverine track (dashed) $\kappa = 1, ..., 4$

Let y_κ = no. wolverines, L_κ = length of *projection* on the baseline:

$(y_1, L_1) = (1, 5.25)$, $(y_2, L_2) = (2, 7.5)$, $(y_3, L_3) = (2, 2.4)$, $(y_4, L_4) = (1, 7.05)$

Interest of estimation: θ = total no. wolverines in the area

Line-intercept sampling (LIS, Becker, 1991)



$$\begin{tikzcd} \kappa_1 & \kappa_2 & & \kappa_3 & & \kappa_4 & \\ i_1 \arrow[u] \arrow[ur] & i_2 \arrow[u] & i_3 & i_4 \arrow[u] & i_5 & i_6 \arrow[u] & i_7 \end{tikzcd}$$

$\mathcal{B}^* = (F^*, \Omega_s; H^*)$ by (1) based on observed Ω_s under LIS

Partition baseline into *projection segments* $i_1, ..., i_7$ of length x_i and $p_i = \frac{x_i}{12}$:

$$i_1 \leftrightarrow \kappa_1 \quad ... \quad i_1 \cup i_2 \leftrightarrow \kappa_2 \quad ... \quad i_4 \leftrightarrow \kappa_3 \quad ... \quad i_6 \leftrightarrow \kappa_4 \quad ...$$

Line-intercept sampling (LIS, Becker, 1991)

Let $\Omega = \{1, ..., \kappa, ..., |\Omega|\}$ contain all the wolverine tracks in the area, $|\Omega| \geq 4$
Let $F = \{1, ..., i, ..., m_F\}$ consist of the corresponding projection segments
Let $H = \{(i\kappa); i \in F, \kappa \in \Omega\}$ and $\mathcal{B} = (F, \Omega; H)$

Only $\mathcal{B}^*$ can be constructed (given Ω_s) but not $\mathcal{B}$

In reality, field observation along a line has an actual width of *detectability*...
yielding known F' of *detectability partitions* and $\mathcal{B}' = (F', \Omega; H')$ by (1)

Given (i) and (ii), strategy BIGS-IWE applicable with $\mathcal{B}'$

As along as the unit of detectability is negligible in scale compared to the baseline, one can assume the elements of F' to be nested in those of F^* (or F), such that the selection probability of each observed track κ with respect to BIGS from $\mathcal{B}'$ can be correctly calculated using $\mathcal{B}^*$ (or $\mathcal{B}$)

Strategy BIGS-IWE for $\mathcal{B}'$ applicable using the observed $\mathcal{B}^*$, as well as $\mathcal{B}$ if it were known

Line-intercept sampling (LIS, Becker, 1991)

	$\hat{\theta}_y$	$\hat{\theta}_{z\beta}$	$\hat{\theta}_{z\alpha 0}$	$\hat{\theta}_{z\alpha .5}$
Estimate of θ	7.57	8.99	9.44	9.27
Variance Estimate	5.27	2.46	1.70	1.97

HTE $\hat{\theta}_y$ (Thompson, 2012), where

$$\begin{aligned}
\pi_{(\kappa)} &= 1 - (1 - p_{(\kappa)})^4 \\
\pi_{(\kappa\ell)} &= 1 - \Big(\Pr(\kappa \notin \Omega_s) + \Pr(\ell \notin \Omega_s) - \Pr(\kappa \notin \Omega_s, \ell \notin \Omega_s) \Big) \\
&= \pi_{(\kappa)} + \pi_{(\ell)} - 1 + \big(1 - p_{(\kappa \cup \ell)}\big)^4
\end{aligned}$$

Multiplicity estimator $\hat{\theta}_{z\beta}$ using equal weights $\omega_{i\kappa}$ where

$$\omega_{11} = \omega_{43} = \omega_{64} = 1 \qquad \text{and} \qquad \omega_{12} = \omega_{22} = 0.5$$

HH-type $\hat{\theta}_{z\alpha\gamma}$ with PIDA weights, where $\hat{\theta}_{z\alpha 0}$ with $\gamma = 0$ is the with-replacement Hansen-Hurwitz (HH) estimator used by Becker (1991):

$$\hat{\theta}_{HH} = \frac{1}{4} \sum_{r=1}^{4} \tau_r \qquad \text{and} \qquad \tau_r = \sum_{\kappa \in \Omega_r} \frac{y_\kappa}{p_{(\kappa)}}$$

[1] Becker, E.F. (1991). A terrestrial furbearer estimator based on probability sampling. *The Journal of Wildlife Management*, 55:730–737.

[2] 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.

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

[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.

[7] Thompson, S.K. (2012). *Sampling*. John Wiley & Sons, Inc.

[8] Zhang, L.-C. (2020). Sampling designs for epidemic prevalence estimation. https://arxiv.org/abs/2011.08669

[9] Zhang, L.-C. and Oguz-Alper, M. (2020). Bipartite incidence graph sampling. https://arxiv.org/abs/2003.09467