Day-1 Practical Session, 25 May 2021

Part 1: Epidemic Prevalence Estimation

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

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

Illustration I: Size-biased sampling and adaptive network tracing

In this illustration, the efficiency of the Horvitz-Thompson estimator (HTE) of epidemic prevalence under Adaptive Cluster Sampling (ACS) by adaptive network tracing will be compared with the HTE based on the initial sample s_0, selected with either Simple Random Sampling ($\eta = 1$) or *size-biased* sampling with $\eta \neq 1$, where η is the *odds-ratio* between case and non-case groups, which is defined by

$$\eta = \frac{Pr(i \in s_0 | y_i = 1)}{Pr(i \in s_0 | y_i = 0)}.$$

We will use the R-function **mainEpi** described below.

Description of the population and sampling strategies

- Population graph: $G = (U, A)$, $(ij) \in A$ if $i \in U$ and $j \in U$ are in-contact
- Population size is denoted by N
- The population consists of cases and non-cases: $y_i = 1$ if case, and $y_i = 0$ otherwise
- A *case network* in G contains all case-nodes connected to each other
- The set of case networks in G denoted by Ω
- The total number of cases in the population is denoted by $Y = \sum_{i \in U} y_i$ within the function below
- All the cases in the population divided into equal-size networks, K
- Initial sample s_0 of size n_0 is selected with either SRS or Poisson sampling with probabilities $\pi_i \propto \eta$
- For ACS, all case-networks that initially selected cases in s_0 belong to are taken into the sample
- Sampled case networks denoted by Ω_s

Description of R-function **mainEpi**

1. Function parameters

- **N**: the population size; deafult value 100 000
- **theta**: prevalence in the population, i.e. the proportion of the cases in the population, $\theta = Y/N$; default value 0.01
- **f**: sampling fraction, $f = n_0/N$; default value 0.01
- **M**: the number of case-networks in the population; default value 10
- **lift**: the odds-ratio between cases and non-cases, denoted by η; default value 1
- **B**: number of replications; default value 100

2. Main steps of the function

- Equal case-network size, **K**, is calculated based on **N** and **M**, and the total number of cases is calculated by **Y**=**MK**
- Index for case-networks, **kidx**, is created for all cases and set to 0 for all non-cases, such that the first **Y** units in the population take a network id number while the last $N - Y$ units take value 0, that is,
$$\underbrace{1, 1, \ldots, 1}_{K \text{ times}}, \underbrace{2, 2, \ldots, 2}_{K \text{ times}}, \ldots, \underbrace{M, M, \ldots, M}_{K \text{ times}}, \underbrace{0, 0, \ldots, 0}_{(N-Y) \text{ times}}$$
- Inclusion probabilities, $\pi_i = \Pr(i \in s_0)$, denoted by **p** inside the function, are calculated as proportional to **lift** if $y_i = 1$, and 1 otherwise
- Inclusion probabilites of case-networks are calculated by
    - $\pi_{(\kappa)} = 1 - (1 - p)^K$, denoted by **pr.k** inside the function, under Poisson sampling of s_0,
    - $\pi_{(\kappa)} = 1 - \binom{N-K}{n_0}/\binom{N}{n_0}$, denoted by **p.k** inside the function, under SRS of s_0
- An $Y \times 2$ array is created, called **nidx** inside the function, the first and the second columns of which are replaced with the case-network sizes and their inclusion probabilities, respectively
- The HTE of the prevalence the population:

$\hat{\theta}_{HT} = \frac{1}{N} \sum_{\kappa \in \Omega_s} \frac{y_\kappa}{\pi_{(\kappa)}}$, where y_κ the number of cases in network κ

- Sampling variances of the HTE based on s_0 are calculated under SRS and Poisson sampling: **v.srs** and **v.pois**

$$V_{srs}(\hat{\theta}_{HT;s_0}) = \left(1 - \frac{n_0}{N}\right)\frac{\theta(1-\theta)}{n_0}, \quad V_{pois}(\hat{\theta}_{HT;s_0}) = \frac{1}{N^2} \sum_{i \in U} \left(\frac{1}{\pi_i} - 1\right) y_i$$

- Sampling variance of the HTE under ACS is calculated by

$$V_{acs}(\hat{\theta}_{HT}) = \frac{1}{N^2} \left\{ \sum_{\kappa \in \Omega} \left(\frac{1}{\pi_{(\kappa)}} - 1\right) K^2 + \sum_{\kappa \in \Omega} \sum_{\ell \neq \kappa \in \Omega} \left(\frac{\pi_{(\kappa\ell)}}{\pi_{(\kappa)}\pi_{(\ell)}} - 1\right) K^2 \right\}.$$

Here, we have $\pi_{(\kappa\ell)} = \pi_{(\kappa)} + \pi_{(\ell)} - (1 - \bar{\pi}_{(\kappa\cup\ell)})$, with $\bar{\pi}_{(\kappa\cup\ell)} = (1 - p)^{2K}$ if poisson sampling.

- Simulation study
    - **B** random samples selected with *Sequential Poisson Sampling* (SPS) from the population
    - For each replication, the population prevalence is estimated based on s_0 and the sample of case-networks observed by adaptive tracing

2. Main outputs of the function

- Relative efficiency of the HTE under ACS against the HTE based on the initial sample s_0 selected with either SRS or poisson sampling