Relative Efficiencies of Two-Stage Sampling schemes for Mean Estimation in Multilevel Populations when Cluster Size is Informative

Francesco Innocenti, Math Candel, Frans Tan, Gerard van Breukelen
Maastricht University
Department of Methodology & Statistics

7th Channel Network Conference
Rothamsted Research UK
10-12 July 2019

Outline

- Motivating Examples
- General Framework
  - Population meanS in multilevel populations
  - Informative cluster size
  - Competing sampling schemes
- Unbiased Estimation
- Relative Efficiencies of Competing Sampling Schemes
- Applications: Sampling Schemes Comparison for Planning a Survey
- Conclusion and Future Research

Motivating Examples

1. Average alcohol consumption among adolescents:

- Adolescents clustered in schools
- Schools vary in size (i.e. # of enrolled students)
- Adolescent's alcohol consumption can be related to school size

2. Average per capita government expenditure on health:

- Patients clustered in general practices (GPs)
- GPs vary in size (i.e. # of registered patients)
- Government expenditure on health per patient can be related to GP size

1. **two-level population**
2. **cluster size variation**
3. **informative cluster size**

Definitions of Population Means

Two alternative definitions of population means:

1. The **average of all individual scores** in the population, ignoring cluster membership.

2. The **average of all cluster-specific means.**

- They coincide only under special conditions.
- Focus is on the **average of all individual scores**.

Assumptions

1. The population is composed of K clusters and cluster j contains N_j individuals, i.e. **clusters vary in size**.
2. The sampling scheme is either a **Simple Random Sampling** (SRS) of $m < N_{pop} = \sum_{j=1}^{K} N_j$ individuals , or a **Two-Stage Sampling** (TSS) of $k < K$ clusters and n or $n_j < N_j$ individuals per sampled cluster. The population is very large relative to the sample size at each design level.
3. The outcome variable Y_{ij} (individual i, cluster j) is **quantitative**

$$y_{ij} = \beta_0 + u_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2), \quad u_j \perp \varepsilon_{ij}.$$

4. **Informative cluster size**, i.e. cluster effect u_j is related to cluster size N_j:

$$u_j = \gamma(N_j - \theta_N) + v_j, \quad v_j \sim N(0, \sigma_v^2), \quad v_j \perp N_j.$$

Informativeness parameter

Population mean cluster size

Thus,

$$u_j | N_j \sim N(\gamma(N_j - \theta_N), \sigma_v^2).$$

Why two Population Means?

The overall mean is defined as $E(Y_{ij})$. The distinction between the **average of all individual Ys** (μ) and the **average of all cluster-specific means** (β_0) comes from considering the distribution of cluster effect u_j over

- the population of clusters → β_0, or
- the population of individuals → μ

Average of all individual Ys = $\mu = \beta_0 + \gamma\theta_N\tau_N^2$

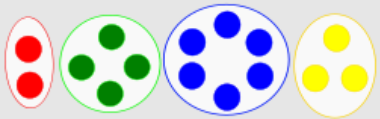
Population coefficient of variation of cluster size, i.e. $\tau_N = \frac{\sigma_N}{\theta_N}$, where σ_N st. dev. of cluster size

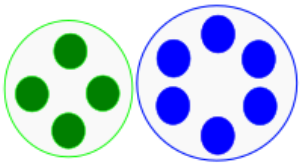
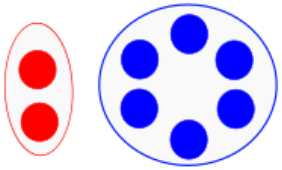
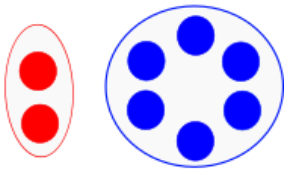
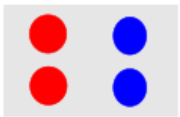
Informativeness parameter

Population mean cluster size

Three competing TSS schemes

Population



	TSS1	TSS2	TSS3
1st stage π_j	$\pi_j \propto N_j$	$\pi_j = k/K$	$\pi_j = k/K$
			
2nd stage $\pi_{i\|j}$	$\pi_{i\|j} = n/N_j$	$\pi_{i\|j} = n_j/N_j = p$	$\pi_{i\|j} = n/N_j$
			
$\pi_{ij} = \pi_j \pi_{i\|j}$	$\pi_{ij} \approx \frac{nk}{N_{pop}}$	$\pi_{ij} = \frac{k}{K} p$	$\pi_{ij} = \frac{k}{K} \frac{n}{N_j}$

Note that TSS1 requires prior knowledge of the whole cluster size distribution before sampling.

Unbiased Estimation of μ

❖ **Informative Cluster Size** (i.e. $\gamma \neq 0$)

- SRS and TSS1: unweighted average of cluster means ($\bar{y}_j$s)
- TSS2 and TSS3: weighting cluster means ($\bar{y}_j$s) by cluster size

(N_js) ← only asymptotically unbiased!

❖ **Non-informative Cluster Size** (i.e. $\gamma = 0 \rightarrow \mu = \beta_0$)

- SRS, TSS1, TSS3: unweighted average of cluster means ($\bar{y}_j$s)
- TSS2: weighting cluster means ($\bar{y}_j$s) by $\frac{1}{Var(\bar{y}_j)}$

Relative Efficiencies

To compare two competing sampling schemes (e.g. $D1$ and $D2$), define

$$RE(D1\ vs\ D2) = V_{D2}(\hat{\mu})/V_{D1}(\hat{\mu})$$

under the constraint of a **fixed total sample size** (i.e. same number of individuals).

❖ **Informative Cluster Size** ($\gamma \neq 0$)

- RE depends on: the average sample size per cluster $\bar{n}$, the intraclass correlation $\rho = \sigma_v^2/\left(\sigma_v^2 + \sigma_\varepsilon^2\right)$, and the cluster size distribution in the population (i.e. its coefficient of variation τ_N, skewness ζ_N, kurtosis η_N)
- SRS is more efficient than any TSS for $\rho > 0$, $\bar{n} > 1$, and $\tau_N > 0$
- TSS1 is the most efficient TSS for many cluster size distributions
- TSS3 is always the least efficient TSS

❖ **Non-informative Cluster Size** ($\gamma = 0$)

- RE depends on: $\bar{n}$, ρ, and τ_N
- SRS is more efficient than any TSS for $\rho > 0$, $\bar{n} > 1$, and $\tau_N > 0$
- TSS1 and TSS3 are equally efficient and outperform TSS2

Applications: Comparing TSS1 and TSS2

1. Average weekly alcohol consumption among adolescents in Italy

- In 2016/2017 in Italy: $6{,}235 = K$ public high schools, $2{,}515{,}060 = N_{pop}$ enrolled students
- **Informative School size**: the number of students per school can affect the degree of connection between students and school which, in turn, can be inversely related to alcohol consumption

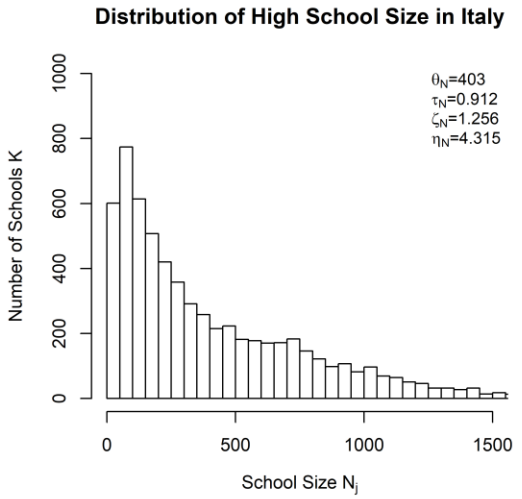
2. Average per capita government expenditure on health in England

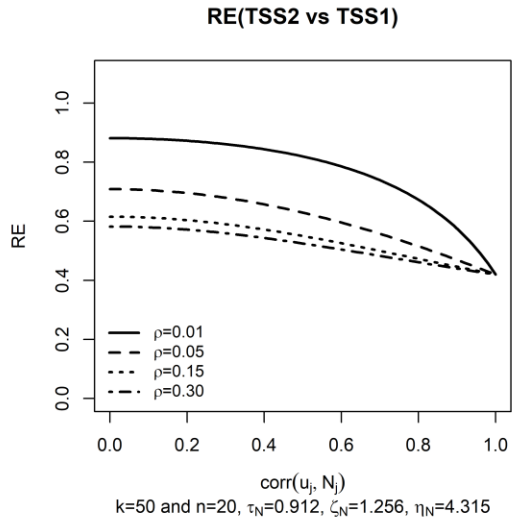
- In 2017 in England: $7{,}353 = K$ general practices, $58{,}719{,}921 = N_{pop}$ registered patients
- **Informative GP's Patient List size:** the number of patients per GP can affect GP's efficacy in preventing hospitalizations, thus affecting government expenditure on health

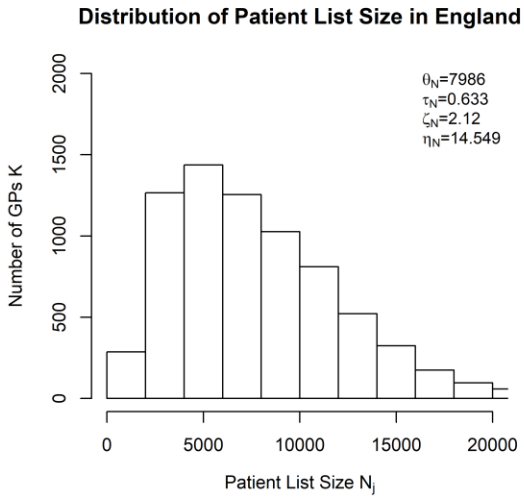
Applications: Comparing TSS1 and TSS2

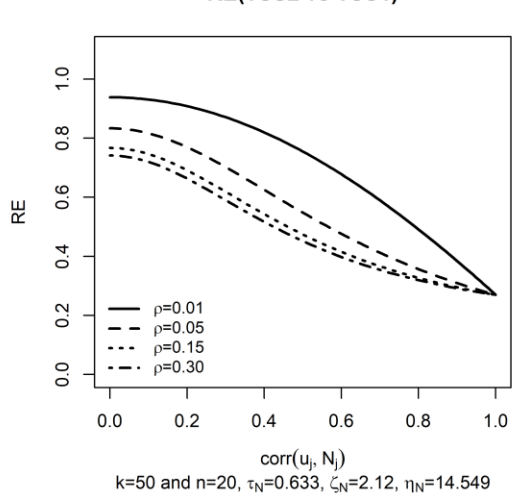
Aim: comparing TSS1 and TSS2 when planning a survey to estimate the population mean μ.

Sample size:
$k = 50$ clusters,
$\bar{n} = 20$ individuals per cluster.









Conclusion and Future Research

- The **average of all individual outcomes** and the **average of all cluster means** coincide only if cluster size does not vary or is not related to the outcome variable of interest
- The **relative efficiency** of two sampling schemes **depends on** the **cluster size distribution** in the population when cluster size is informative
- To estimate the average of all individual outcomes the **best strategy** (in terms of unbiasedness and efficiency) is
  - **Informative cluster size**: TSS1 + unweighted average of cluster means
  - **Non-informative cluster size**: TSS1/TSS3 + unweighted average of cluster means, but TSS1 requires prior knowledge of the whole cluster size distribution
- Model-based and Design-based inference yielded almost the same results (i.e. same estimators and relative efficiencies), if the model assumptions hold
- **Future Research**
  - Optimal Design (i.e. design that maximizes statistical power for a given budget)
  - Binary Outcomes (i.e. prevalence estimation)

Thanks for your attention!

APPENDIX

Marginal Expectations and Variances of u_j

The overall mean and marginal variance of Y_{ij} are defined as

$$E(Y_{ij}) = E(\beta_0 + u_j + \varepsilon_{ij}) = \beta_0 + E(u_j)$$

and

$$V(Y_{ij}) = V(\beta_0 + u_j + \varepsilon_{ij}) = V(u_j) + \sigma_\varepsilon^2$$

- If clusters are sampled with **equal probabilities** (e.g. with TSS2 or TSS3), the sampling distribution of u_j, for $n = 1$, is $\int f(u_j|N_j)f(N_j)dN_j$. Thus,

$$E(u_j) = 0 \text{ and } V(u_j) = \sigma_v^2 + \gamma^2\sigma_N^2$$

$$\rightarrow E(Y_{ij}) = \beta_0 \text{ and } V(Y_{ij}) = \sigma_v^2 + \gamma^2\sigma_N^2 + \sigma_\varepsilon^2$$

- If clusters are sampled with **probabilities proportional to their size** (e.g. with SRS or TSS1), the sampling distribution of u_j, for $n = 1$, is $\int \left(\frac{N_j}{\theta_N}\right) f(u_j|N_j)f(N_j)dN_j$. Thus,

$$E(u_j) = \gamma\theta_N\tau_N^2 \text{ and } V(u_j) = \sigma_v^2 + \gamma^2\sigma_N^2[\tau_N(\zeta_N - \tau_N) + 1]$$

$$\rightarrow E(Y_{ij}) = \beta_0 + \gamma\theta_N\tau_N^2 = \mu \text{ and } V(Y_{ij}) = \sigma_v^2 + \gamma^2\sigma_N^2[\tau_N(\zeta_N - \tau_N) + 1] + \sigma_\varepsilon^2$$

Mean Estimators and Sampling Variances

- SRS: $\hat{\mu} = \sum_{i=1}^{m} \frac{y_i}{m},$

$$V(\hat{\mu}) = \frac{\sigma_\varepsilon^2 + \sigma_v^2 + \gamma^2 \sigma_N^2 [\tau_N (\zeta_N - \tau_N) + 1]}{m}$$

- TSS1: $\hat{\mu} = \sum_{j=1}^{k} \frac{\bar{y}_j}{k},$

$$V(\hat{\mu}) = \frac{\sigma_\varepsilon^2 + n\{\sigma_v^2 + \gamma^2 \sigma_N^2 [\tau_N (\zeta_N - \tau_N) + 1]\}}{nk}$$

Mean Estimators and Sampling Variances

- TSS2:

$$\hat{\mu} = \frac{\sum_{j=1}^{k} pN_j \bar{y}_j}{\sum_{j=1}^{k} pN_j},$$

$$V(\hat{\mu}) = \frac{\sigma_\varepsilon^2 + \bar{n}\left\{\left(\frac{\tau_N^2 + 1}{\frac{\tau_N^2}{k} + 1}\right)\sigma_v^2 + \gamma^2\sigma_N^2\left[\left(\frac{k-1}{k}\right)^2 \tau_N^2\left(\eta_N - \frac{k-3}{k-1} + \tau_N(\tau_N - 2\zeta_N)\right) + 2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N - \tau_N) + 1\right]\right\}}{\bar{n}k}$$

- TSS3:

$$\hat{\mu} = \frac{\sum_{j=1}^{k} N_j \bar{y}_j}{\sum_{j=1}^{k} N_j},$$

$$V(\hat{\mu}) = \frac{(\sigma_\varepsilon^2 + n\sigma_v^2)\left(\frac{\tau_N^2 + 1}{\frac{\tau_N^2}{k} + 1}\right) + n\gamma^2\sigma_N^2\left[\left(\frac{k-1}{k}\right)^2 \tau_N^2\left(\eta_N - \frac{k-3}{k-1} + \tau_N(\tau_N - 2\zeta_N)\right) + 2\left(\frac{k-1}{k}\right)\tau_N(\zeta_N - \tau_N) + 1\right]}{nk}$$

Sufficient conditions under which TSS1 is more efficient than TSS2 and TSS3 when cluster size is informative

- TSS1 is more efficient than TSS2 and TSS3 if the cluster size distribution in the population meets one of the following conditions:
  - The cluster size distribution is positively skewed (i.e. $\zeta_N > 0$) with $\tau_N \in [0, \zeta_N]$
  - The cluster size distribution is Normal
  - The cluster size distribution is symmetric (i.e. $\zeta_N = 0$) with $\tau_N \in [0,1]$ and $k \in$

$$\left[1, \frac{(2-\tau_N^2)+\sqrt{2-\tau_N^2}}{1-\tau_N^2}\right]$$

- If the cluster size distribution does not meet any of the conditions given above, then compute the relative efficiency for that specific distribution to see whether TSS1 is more efficient than TSS2 and TSS3.

Model-based vs Design-based inference

Model-based approach:

- Y_{ij} is random
- Inference based on the stochastic model for Y_{ij}
- PRO: It simplifies sample size planning and sampling schemes comparisons

Design-based approach:

- Y_{ij} is fixed but unknown. The *inclusion indicator* I_{ij} is random (i.e. $I_{ij} = 1$ with π_{ij}, and $I_{ij} = 0$ otherwise)
- Inference based on the distribution of I_{ij} over repeated samples with a probability sampling design
- PRO: robustness

In the considered setting, the two approaches yield almost the same results (if the model assumptions are met)

- **same estimators** of the population mean
- approximately the **same relative efficiencies** (i.e. for k sufficiently large)