Evaluating asymptotic intervals of weighted and FH estimates using simulated data mirroring Nigeria height-for-age

Alana McGovern

2025-10-30

The goal of this simulation study to assess the properties of the asymptotic intervals of weighted and FH estimates, and to isolate the various factors (misspecification of sampling model, oversmoothing) which affect these interval properties under the FH model. We simulate data meant to mirror the height-for-age z-score in the 2018 Nigeria DHS.

Simulation Procedure

1. Obtain weighted mean estimates at the admin1 and admin2 level, $(\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)})$, using the DHS data.

2. For each admin2 area, j, which doesn't have a valid weighted estimate, draw $\mu_j^{(2)} \sim \mathcal{N}(\mu_i^{(1)}, s_i)$, where i is the admin1 area which admin2 area j is contained in and s_i is the standard deviation of the valid weighted estimates for the admin2 areas contained in admin1 area i (i.e. the between admin2 variation within that admin1).

3. Then we use the Nigeria under-five population from worldpop, and draw a continuous response for each individual, k, in cluster, c, in admin2 area, j,

$$y_{jck} = \mu_j^{(2)} + \epsilon_c + \gamma_k$$

and $\epsilon_c \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\gamma_c \sim \mathcal{N}(0, \sigma_\gamma^2)$.

The parameter σ_ϵ represents between-cluster variation, while σ_γ represents individual variation, and should be varied across settings.

Now for each simulation, k,

4. Take a two-stage stratified random sample where the first stage is at the admin1 level and the second is at the cluster level. Sampling design and sizes are chosen to mirror the Nigeria 2018 DHS.

5. For this sample dataset, calculate the weighted estimates and corresponding variance estimates, $(\hat{\boldsymbol{\theta}}^{(1,k)}, \hat{\mathbf{V}}^{(1,k)}, \hat{\boldsymbol{\theta}}^{(2,k)}, \hat{\mathbf{V}}^{(2,k)})$

After completing these simulations,

6. Use $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$ from the simulations to calculate the true variance of the weighted estimates for each admin1 and admin2 area, $(\mathbf{V}^{(1)}, \mathbf{V}^{(2)})$.

Now for each simulation, k,

7. Obtain admin1 and admin2 estimates under an IID FH model using $(\hat{\boldsymbol{\theta}}^{(1,k)}, \hat{\mathbf{V}}^{(1,k)})$ and $(\hat{\boldsymbol{\theta}}^{(2,k)}, \hat{\mathbf{V}}^{(2,k)})$, respectively. We will refer to this as using the 'naive sampling model' because it makes assumptions about normality and variance.

8. Obtain admin1 and admin2 estimates under an IID FH model using $(\hat{\boldsymbol{\theta}}^{(1,k)}, \mathbf{V}^{(1)})$ and $(\hat{\boldsymbol{\theta}}^{(2,k)}, \mathbf{V}^{(2)})$, respectively. We will refer to this as using the 'correct-variance sampling model' because it assumes normality, but uses the correct variance.

9. For each admin2 area, j, draw $\theta_j^{*(2,k)} \sim \mathcal{N}(\mu_j^{(2)}, V_j^{(2)})$, and analogously for each admin1 area.

10. Obtain admin1 and admin2 estimates under an IID FH model using $(\boldsymbol{\theta}^{*(1,k)}, \mathbf{V}^{(1)})$ and $(\boldsymbol{\theta}^{*(2,k)}, \mathbf{V}^{(2)})$, respectively. We will refer to this as using the 'correct sampling model' because we have ensured the sampling model is exactly correctly.

Simulation Settings

The only parameters we will vary here are $(\sigma_\epsilon, \sigma_\gamma)$. We set $\sigma_\epsilon^2 + \sigma_\gamma^2 = 2$, which is close to what we observe in the data. Then we specify which proportion of that variation is at the individual level to be ρ, so that $\rho = \frac{\sigma_\gamma^2}{\sigma_\epsilon^2 + \sigma_\gamma^2}$. In the first setting we define $\rho = 0.9$ and in the second we set $\rho = 0.5$,

We could also vary the summation of the two variances, but will leave it at one level for now.

Simulation Metrics

We will look at the coverage and widths of the 90% intervals for the weighted estimates and each of the three types of FH estimates.