Small Area Estimation: Weighted Estimation and Summarizing Uncertainty

Jon Wakefield

Department of Statistics,
University of British Columbia

Departments of Statistics and Biostatistics
University of Washington

CEMA SAE Short Course
November 10–12th, 2026

Outline

Weighted Estimation

Summary Measures

SAE for MCV1 Coverage in Kenya

Discussion

Weighted Estimation

Weighted Estimates

- **Prevalence** p_i is the empirical proportion in a finite population with a certain condition.
- In area i, if N_i is the population size and T_i is the number of events, then the prevalence is

$$p_i = \frac{T_i}{N_i}.$$

- The distinction between risk and prevalence is important in the survey sampling literature – the target in SAE is generally the prevalence.
- If we were to take a complete census in area i, i.e., sample all N_i individuals, then we would know the prevalence, and no statistics required!
- But we would not know the risk, as this is a summary from a hypothetical infinite population.

Weighted Estimates

- Let $Y_{ik} = 0/1$ again be the indicator for the event k in area i, with $w_{ik} = 1/\pi_{ik}$ being the design weight, that accounts for the (probabilistic) sample selection.

- w_{ik} can be thought of as the number of people represented by person k.

- For area i, the weighted estimator is

$$\widehat{p}_i^{\,w} \quad = \quad \frac{\widehat{T}_i}{\widehat{N}_i} \quad = \quad \frac{\sum_{k \in S_i} w_{ik} Y_{ik}}{\sum_{k \in S_i} w_{ik}}. \tag{1}$$

  where S_i is the set of sampled births.

- Known as a **direct** estimate since based on data from the area in question only.

- Standard errors can be obtained using design-based theory.

The weighted mean is the direct estimator that is the first choice for producing subnational estimates

Direct Estimates

Advantages:

- Estimates and uncertainty measures account for design via weighting.
- Inference is based on minimal assumptions, since there is no explicit model for the data.
- Often produces reasonable inference for Admin1 areas (unless outcome is very rare). Sometimes OK for Admin2 also.

Disadvantages:

- When data are sparse in an area, then estimator will have uncertainty that is high, or not possible to estimate (variance formula breaks).
- For example, for the SRS estimator in a model-based framework, the standard error is $\sqrt{\widehat{r}_i^{\text{SRS}}(1-\widehat{r}_i^{\text{SRS}})/n_i}$ which is 0 if $\widehat{r}_i^{\text{SRS}}$ is equal to 0 or 1.
- If no data in an area, then no estimate available.

Summary Measures

Summarizing Information

- A map of the prevalence estimates is visually appealing, but in isolation is impossible to interpret.
- To make interpret and take action, it is vital to understand the inherent uncertainty in the estimate, with possibilities:
  - The standard error – useful if unceratinty distribution is symmetric.
  - An uncertainty interval.
  - The width of an uncertainty interval.
  - The coefficient of variation.
  - An examination of ridgeplots.
  - Calculating exceedence probabilities.
- These measures are of generally interest, not just for weighted estimates.

Uncertainty Measures

- The variance of $\widehat{p}_i^{\mathrm{w}}$ can be obtained using survey sampling techniques which are tuned to the specific type of survey – call this $\widehat{V}_i$.
- The standard error is simply $\sqrt{\widehat{V}_i}$.
- If the sample size n_i is sufficiently large, an accurate 95% confidence interval is

$$\left[\widehat{p}_i^{\mathrm{w}} - 1.96 \times \sqrt{\widehat{V}_i} \, , \, \widehat{p}_i^{\mathrm{w}} + 1.96 \times \sqrt{\widehat{V}_i} \right]$$

Uncertainty Measures

- The width of the confidence interval (CI) is

$$2 \times 1.96 \times \sqrt{\widehat{V}_i} \approx 4 \times \sqrt{\widehat{V}_i}$$

- Notice that the width of the interval does not depend directly on the point estimate, which may be deceptive since, for example, an interval of width 0.05 can mean very different things:
  - $\widehat{p}_i^{\mathrm{W}} = 0.50$, interval: (0.475,0.525), versus
  - $\widehat{p}_i^{\mathrm{W}} = 0.05$, interval: (0.025,0.075).

  When prevalence is close to 0 or 1, we might prefer an alternative measure.

Uncertainty Measures

- The coefficient of variation (CV) is popular for SAE, particularly when the prevalence is close to zero.
- The CV is often expressed as a percentage and is defined as

$$\text{CV} = 100 \times \frac{\text{Standard Error}}{\text{Point Estimate}} = 100 \times \frac{\sqrt{V_i}}{\widehat{p}_i^{\text{w}}}.$$

- Examples:
  - if the point estimate is $\widehat{p}_i^{\text{w}} = 0.05$ and the standard error is 0.025, then we have a CV of 50% which would be viewed as large uncertainty.
  - By contrast, if the point estimate is $\widehat{p}_i^{\text{w}} = 0.25$ and the standard error is again 0.025, then we have a CV of 10% which may be viewed as reasonable uncertainty.

Uncertainty Measures

- Some agencies have criteria for reporting, for example, Statistics Canada have guidelines:
  - reliable if coefficient of variation $< 17\%$,
  - view with caution if in the range $(17\%, 33\%)$ and
  - conclude unreliable if $> 33\%$.
- Since the CV is an explicit function of $\widehat{p}_i^{\mathrm{w}}$ it may be preferable as a summary to the standard error or width of a CI – I find it most useful when $\widehat{p}_i^{\mathrm{w}}$ is small.
- Notice the CI depends on a level (95%) while the CV does not.

Ridgeplots

- Ridgeplots, as we use them in SAE, show the complete range of uncertainty for a collection of areas.
- We may examine, for example,
  - All Admin1 areas stacked on top of each other
  - Admin2 areas within a particular Admin1
- The ordering may be via the point estimate.

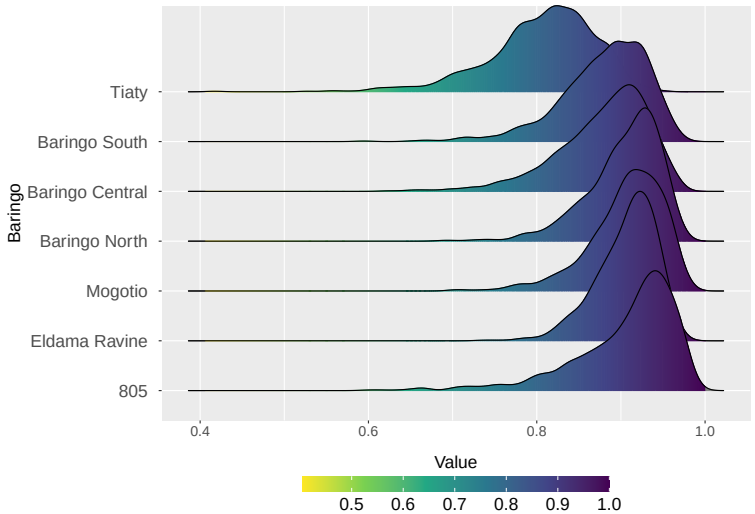


Figure 1: Ridgeplots for MCV1 prevalence for sub (Admin-2) areas of Baringo.

SAE for MCV1 Coverage in Kenya

Subnational MCV1 Coverage in Kenya

- Initial Aim: Estimate prevalence of MVC1 in 12–23 month old children.
- Results from the 2022 Kenya DHS, analyzed using the DHS R shinyApp.
- The national weighted estimate is:
  - 88.7% with 95% confidence interval of (87.2%, 90.2%)

- We see Mandera and Garissa have lower coverage than other counties.
- Look at the overlap in the remaining counties, even at county level quite a lot of uncertainty.
- This is sobering: we often have to deal with large levels of uncertainty.

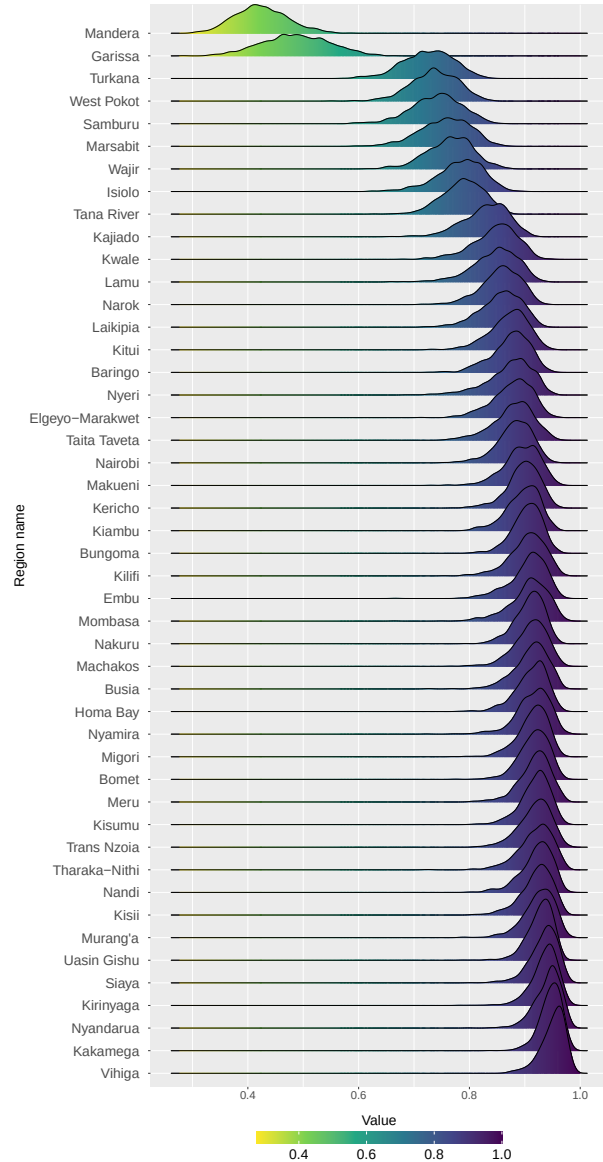


FIGURE 2: Ridgeplots for unit-level estimates of MCV1 for counties of Kenya.

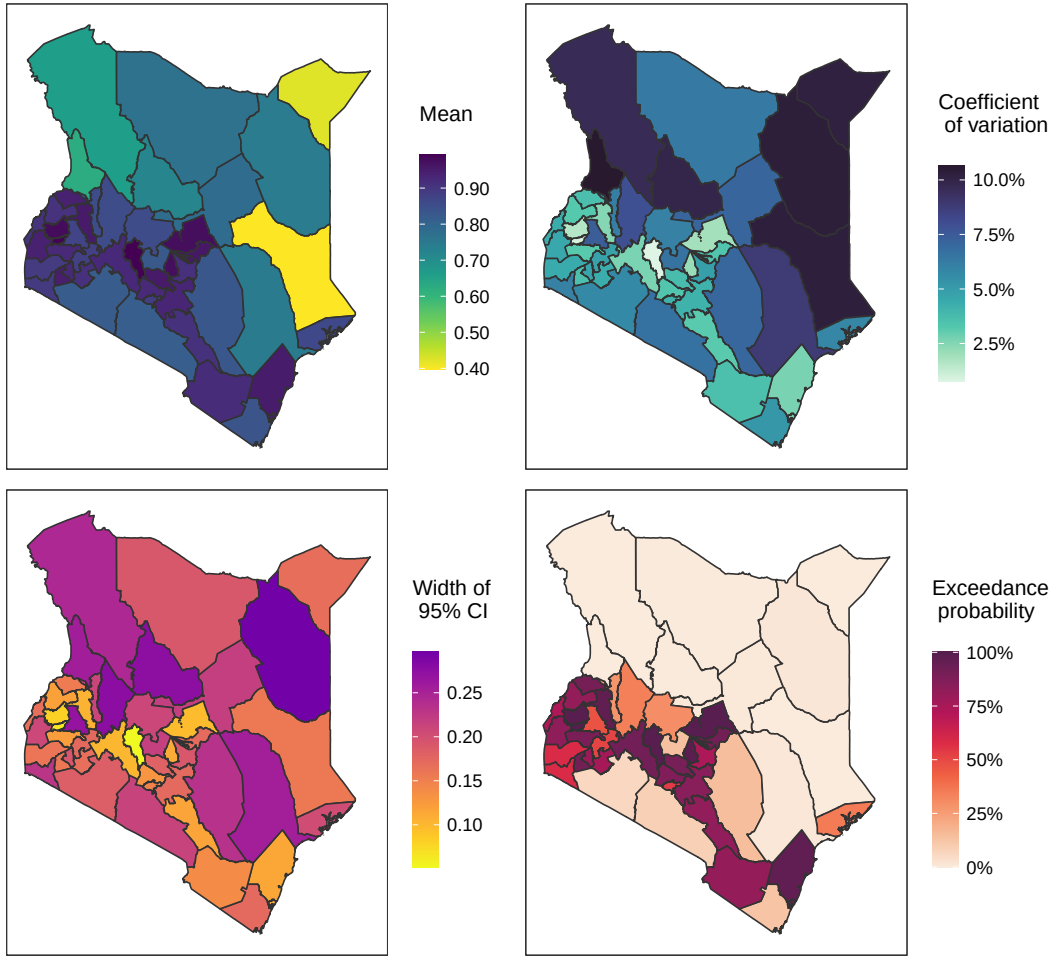


FIGURE 3: Weighted estimates summaries for MVC1 in counties of Kenya.

- Range of estimates is 40% to 99% – large between-area variation.
- But uncertainty in each estimate is needed for interpretatation.

Numerical Summaries for MCV1 Weighted Estimates in Kenya

- We can summarize uncertainty via the standard error or the width of a 95% confidence interval.
- Coefficient of variation is less easy to interpret here (easier for a rare outcome).

KE2022_CH_VACC_C_MSL_Admin_1_Direct

Region_Name	Mean	Median	Standard_Error	Lower_CI	Upper_CI	Width_95_CI	Coefficient_of_Variation
Baringo	0.8566170221546640	0.8566170221546640	0.0677254357450500	0.6696813020669020	0.9462518467614520	0.2765705446945500	0.07906163644069750
Bomet	0.9100241187744040	0.9100241187744040	0.03975497510244470	0.796135853726681	0.9632276257246590	0.16709177199797800	0.0436856279765262
Bungoma	0.908090699141394	0.908090699141394	0.029362611103571900	0.8321595817114740	0.9516657510279510	0.11950616931647700	0.03233444757372190
Busia	0.9054176266099540	0.9054176266099540	0.039708856468605500	0.7941473547151710	0.9596021386832340	0.1654547839680630	0.043856950982148000
Elgeyo-Marakwet	0.8527526619934150	0.8527526619934150	0.052681918129595300	0.7178940029088760	0.9294763038143310	0.21158230090545400	0.061778661595080500
Embu	0.9112783367047510	0.9112783367047510	0.044619466246464100	0.7768983239710410	0.9680467568784060	0.19114843290736500	0.04896359811186930
Garissa	0.39669479953496000	0.39669479953496000	0.04058599402282830	0.3204656288190520	0.47829385455638400	0.1578282257373320	0.10231037581134600
Homa Bay	0.8903533697673630	0.8903533697673630	0.03937843795801670	0.786468804633037	0.94709698906773	0.16062818443469300	0.04422787546511530
Isiolo	0.7840763598430860	0.7840763598430860	0.055842430541625900	0.6554546593442810	0.8739186252593550	0.218463965915074	0.07122065324454030
Kajiado	0.815976491162315	0.815976491162315	0.054122898480232700	0.6862999084556540	0.8998687190224800	0.21356881056682600	0.06632899239920190

Discussion

Discussion

- Design-based inference is very appealing since it directly considers a finite population and males few modeling assumptions, being based only on the randomization distribution.
- The reliance on large samples for inference (for example, via asymptotic confidence intervals) is a weakness, since by definition SAE methods were developed for small samples.
- In the next set of slides we consider stratified sampling and cluster sampling.

References