EU Twinning Project on Statistics in Jordan

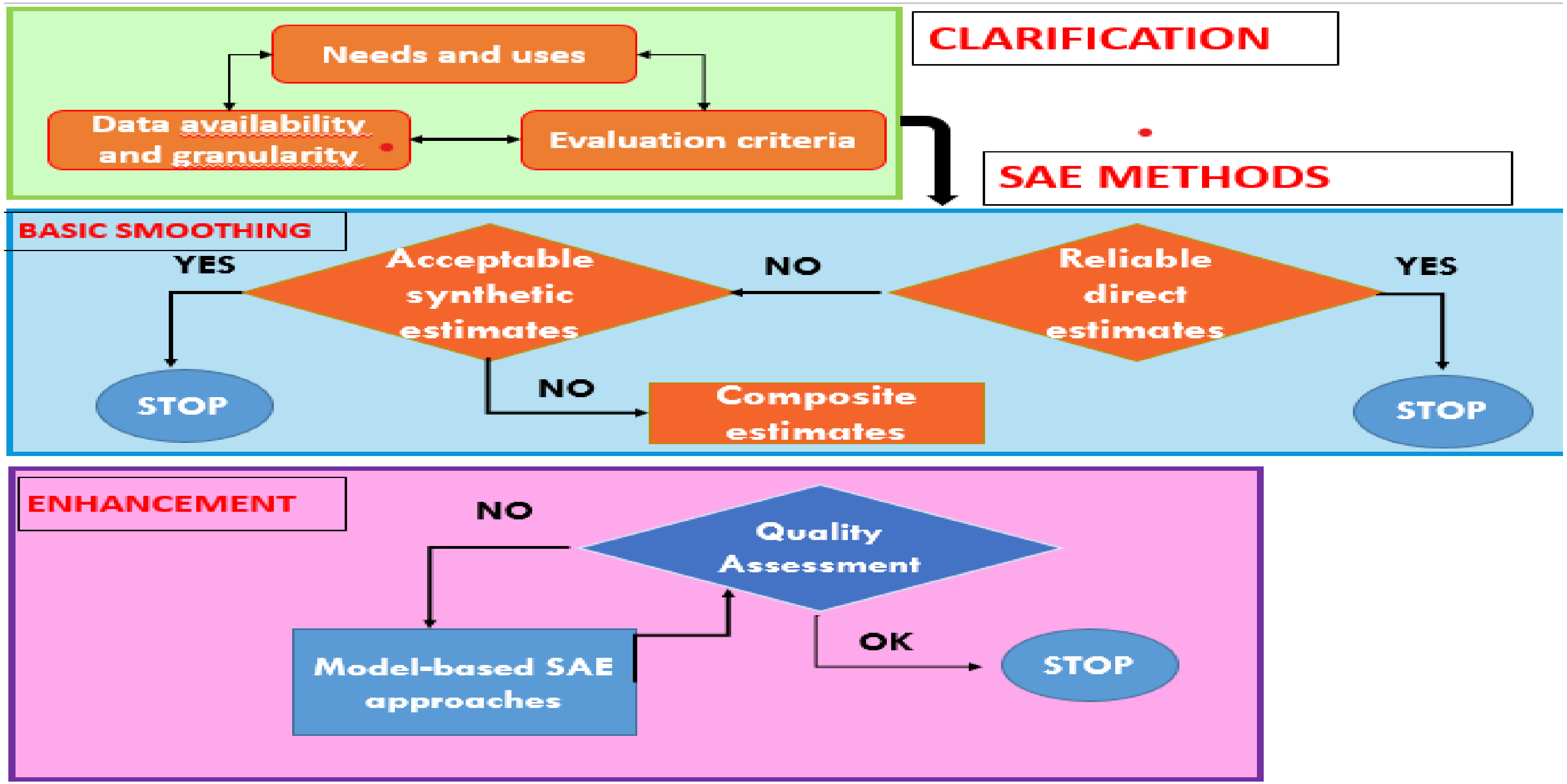
دائرة الإحصاءات العامة
Department of Statistics

Theory and best practice of Small area estimations

Component 2: Methodology for producing Small Area Statistics

26 February - 2024

STEP TO COMPUTE SAS



DESIGN-BASED INFERENCE

Element	Sample Design
Population	$U = \{1, \ldots, N\}$
	$Y = \{y_1 \ldots, y_N\}$
Sample	$s = \{i_1 \ldots., i_n\} \in S_\pi$
	$y = (y_{i1} \ldots, y_{in})$
Probability Distribution	$P(s)$
Parameter	$\theta = h(y_1, \ldots, y_N)$
Estimator	$\hat{\theta}(s)$

DIRECT ESTIMATOR

- U finite population of size N
- $(y_1 \ldots, y_N)$ measurements at the target variable on population units
- Target quantity: example population mean

$$h\ (y_1 \ldots, y_N) = \sum_{i=1}^{N} y_i \,/N$$

- s random sample of size n drawn from the population U
- r = U-s non-sample unit of size (N-n)

BASIC DIRECT ESTIMATOR

- π_j probability of inclusion of unit *j* in the sample

- $d_j = 1/\pi_j$ sampling weight for unit *j*

- Horvitz-Thompson (HT) estimator of mean

$$\hat{\bar{Y}}_{DIR} = \frac{1}{N}\sum_{i=1}^{n} d_i y_i$$

- Design-unbiased variance estimator (under approximation):

$$\widehat{V}(\hat{\bar{Y}}_{DIR}) \cong \frac{1}{N}\sum_{i=1}^{n} d_i(1-d_i)y_i^2$$

DOMAIN ESTIMATION

- U partitioned into D domains $U_1, \ldots, U_D$ of sizes $N_1, \ldots, N_D$
- s_d sub-sample of size n_d drawn from U_d $\left(n = \sum_{d=1}^{D} n_d\right)$
- $r_d = U_d - s_d$ sample complement, of size $N_d - n_d$.
- Target parameter: domain mean

$$\bar{Y}_d = \sum_{i=1}^{N_d} y_i \,/ N_d$$

BASIC DIRECT ESTIMATOR (DOMAIN)

- Horvitz-Thompson (HT) estimator of domain mean

$$\hat{\bar{Y}}_{DIR}^{d} = \frac{1}{N_d} \sum_{i \in s_d} d_i y_i$$

- Design-unbiased variance estimator (under approximation)

$$\widehat{var}\left(\hat{\bar{Y}}_{DIR}^{d}\right) \cong \frac{1}{N_d} \sum_{i \in s_d} d_i (1 - d_i) y_i^2$$

- HT uses only target variable and area-specific sample data

Let’s go to R

Delegation of the European Union to Jordan

ADJUSTMENTS BASIC ESTIMATOR USING AUXILIARY VARIABLES

- *p* auxiliary variables X_{ik}, $k = 1, \ldots, p$ *and* $i = 1, \ldots, n$

- Known population totals of the p variables in the domains d

$$\boldsymbol{X_d} = (X_{1d}, \ldots, X_{pd}), \quad \mathrm{d} = 1, \ldots, \mathrm{D}$$

ADJUSTMENTS BASIC ESTIMATOR USING AUXILIARY VARIABLES

- Attempts to improve the precision of the traditional HT estimator by using **correlation between target variable and covariates** through an adjustment of the initial sampling weights.

- This estimator is still **approximatively design-unbiased**, and should allow decreasing the design-variance.

- It is still a **direct estimator**, because it makes use of just the domain information

EXAMPLE 1. RATIO ESTIMATOR

- HT estimator of $\bar{X}_d$ $\quad \hat{\bar{X}}_{DIR}^{d} = \frac{1}{N_d} \sum_{i \in s_d} d_i x_i$

- Adjustment factor: $g_d = \frac{\bar{X}_d}{\hat{\bar{X}}_{DIR}^{d}}$

- Ratio estimator with auxiliary variable X:

$$\hat{\bar{Y}}_{R}^{d} = \frac{\bar{X}_d}{\hat{\bar{X}}_{DIR}^{d}} \hat{\bar{Y}}_{DIR}^{d} = \frac{1}{N_d} \sum_{i \in s_d} d_i g_d y_i$$

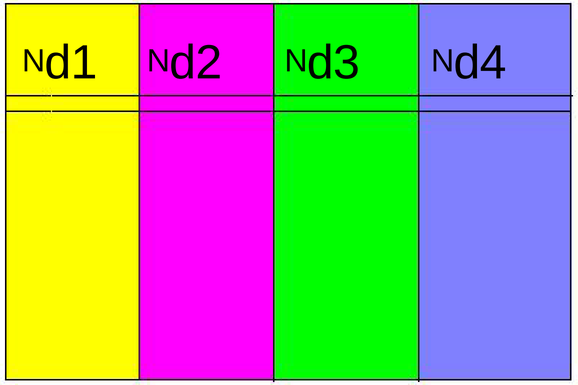
EXAMPLE 2. POST STRATIFIED ESTIMATOR

- J post-strata ($j = 1, \ldots, J$) cut across the domains.

- N_{dj} known count in the intersection of domain d and post-stratum j.

- Mean of domain d : $\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{J} N_{dj} \bar{Y}_{dj}$

$$\hat{\bar{Y}}_{PST}^{d} = \sum_{j=1}^{J} \frac{N_{dj}}{N_d} \hat{\bar{Y}}_{DIR}^{dj}$$



area d

$N_d = N_{d1} + N_{d2} + N_{d3} + N_{d4}$

EXAMPLE 3.GENARALIZED REGRESSION ESTIMATOR

- Linear regression model $y_j = x_j^T \beta + e_j \;\; E(e_j) = 0, \;\; E(e_j^2) = \sigma^2, j = 1, \dots, N$
- Generalized regression (GREG) estimator

$$\hat{\bar{Y}}^d_{GREG} = \hat{\bar{Y}}^d_{DIR} - (\boldsymbol{X_d} - \hat{\bar{\boldsymbol{X}}}^d_{DIR})\hat{\boldsymbol{B}}_d = \frac{1}{N_d} \sum_{i \in s_d} w_i y_i$$

Let's go to R

SUMMING UP DIRECT ESTIMATOR

Data Requirements:

- Design weights assigned to sample units across the specified area
- Horvitz-Thompson (HT) estimator: total domain population count (Nd)
- Generalized Regression Estimator (GREG): Population totals of auxiliary variables within the specified domains.
- Post-stratified estimator: Population totals of auxiliary variables within the specified domains and post-strata.

SUMMING UP DIRECT ESTIMATOR

ADVANTAGES:

- Nonparametric approach: Free from reliance on specific model assumptions
- Incorporation of sampling weights: Allows for approximate design-unbiasedness and design consistency with increasing sample size (n)
- Additivity (Benchmarking property): Demonstrates efficacy in benchmarking comparisons.

DRAWBACKS :

- Increased variance of the estimator (V(Y)) as sample size (n) decreases, rendering it highly inefficient for small domains.
- They cannot be calculated for non-sampled areas (nd = 0).

INDIRECT ESTIMATOR

SYNTHETIC ESTIMATORS:

- A reliable direct estimator for a broad area, covering several small areas, is used to derive an indirect estimator for a small area.
- *Produced under the assumption that the small areas have the same characteristics as the broad area*.

COMPOSITE ESTIMATORS:

- A linear combination between a direct estimator and a synthetic one *using a design-based approach or by assuming an explicit area or unit-level model*
- Represents a good compromise in terms of efficiency between the characteristics of the two components

SYNTHETIC ESTIMATORS

SIMPLE EXAMPLE:

- Target: $\bar{Y}_d = \sum_{i=1}^{N_d} y_i \,/ N_d$
- Assumption: $\bar{Y}_d = \bar{Y}$
- Synthetic estimator of $\bar{Y}_d$:

$$\hat{\bar{Y}}_{SYN}^{d} = \frac{1}{N} \sum_{i \in s} d_i y_i$$

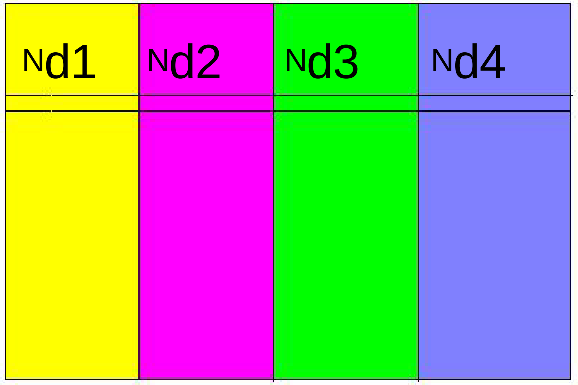
POST-STRATIFIED SYNTHETIC ESTIMATOR

- J post-strata ($j = 1, \ldots, J$) cut across the domains.

- N_{dj} known count in the intersection of domain d and post-stratum j.

- Mean of domain d: $\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{J} N_{dj} \bar{Y}_{dj}$

- Implicit model $\bar{Y}_{dj} = \bar{Y}_j$ for all d and j



area d

$N_d = N_{d1} + N_{d2} + N_{d3} + N_{d4}$

POST-STRATIFIED SYNTHETIC ESTIMATOR

- Post-stratified synthetic estimator:

$$\hat{\hat{Y}}^{d}_{SYN} = \sum_{j=1}^{J} \frac{N_{dj}}{N_d} \hat{\hat{Y}}^{j}_{DIR}$$

Need:

- reliable direct estimators of $\hat{\hat{Y}}^{j}_{DIR}$.
- homogeneity within each post-stratum.

MSE SYNTHETIC ESTIMATOR

- The variance of synthetic estimators depends upon the variance of $\hat{\bar{Y}}^{j}_{DIR}$ being relatively smaller compared to that of the direct estimator in the domain.

- Synthetic estimators are reliant on robust assumptions and may exhibit bias when these assumptions are violated.

- Therefore, needs to estimate the Mean Squared Error (MSE), accounting for both bias and variance.

Let’s go to R

SUMMING UP SYNTETIC ESTIMATOR

ADVANTAGES:

- They facilitate the production of estimates even in non-sampled regions.
- They can reduce the variance of direct estimates.
- They are straightforward to implement.

DRAWBACKS :

- They do not account for between-area heterogeneity, introducing significant bias.
- The assumption necessitates validation
- Stable and area-specific design MSE estimators are unavailable.
- Adjustments for benchmarking are indispensable.

COMPOSITE ESTIMATORS

Defined as a **linear combination** of a direct estimator and a synthetic estimator. This approach aims to balance the bias of the synthetic estimator with the variance of the direct estimator within a given domain.

$$\hat{\bar{Y}}_{CE}^{d} = \phi_d \hat{\bar{Y}}_{DIR}^{d} + (1 - \phi_d)\hat{\bar{Y}}_{SYN}^{d}$$

where:

- $\hat{\bar{Y}}_{DIR}^{d}$ is the direct estimator for the d small area
- $\hat{\bar{Y}}_{SYN}^{d}$ is a synthetic estimator for the d small area
- ϕ_i is a suitably chosen weight, with $0 \leq \phi_d \leq 1$

Let’s go to R

SUMMING UP COMPOSITE ESTIMATOR

ADVANTAGES:

- They cannot exhibit a higher design variance than the direct estimator or a greater bias than the synthetic one.

DRAWBACKS :

- They cannot be computed for non-sampled domains.
- Stable and domain-specific design Mean Squared Error estimators are unavailable.
- Adjustment for benchmarking is necessary.

DESIGN-BASED vs MODEL-BASED INFERENCE

Element	Under Design	Under Model
Population	$U = \{1, \ldots, N\}$	y$\sim P_\theta$
	$Y = \{y_1 \ldots, y_N\}$	
Sample	$\mathrm{s} = \{i_1 \ldots, i_n\} \in S_\pi$	$\boldsymbol{y} = (y_1 \ldots, y_n)$
	$y = (y_{i1} \ldots, y_{in})$	$y_i\ iid$
Probability Distribution	$P(s)$	$P_\theta(\boldsymbol{y})$
Parameter	$\theta = h(y_1, \ldots, y_N)$	$\theta\ i.e. \quad E_{P_\theta}(y)$
Estimator	$\hat{\theta}(s)$	$\hat{\theta}(\boldsymbol{y})$

SMALL AREA ESTIMATION - MODEL-BASED METHODS

AREA-LEVEL MODELS

- Models are specified at area level.
- Rely on area-level data obtained from surveys, both direct estimates and relative precision, as well as covariates.
- Accessing data is less complex compared to acquiring unit-level data.

UNIT-LEVEL MODELS

- Models are specified at the unit level.
- Utilize unit-level data, such as survey data, for model fitting purposes.
- Incorporate area-level covariates as predictor variables.
- Accessing unit-level data may be difficult due to potential confidentiality concerns.

AREA-LEVEL MODELS: THE FAY-HERRIOT MODEL

1. Sampling model

$$\hat{\theta}_d = \theta_d + e_d \quad d = 1, \dots, D$$

$\hat{\theta}_d$ is a direct design-unbiased estimator (e.g HT)
e_d is the known sampling error of the direct estimator

2. Linking model

$$\theta_d = \boldsymbol{X}^T\beta + u_d \quad d = 1, \dots, D$$

$u_d \sim N(0, \sigma_u^2)$ with σ_u^2 unknown

3. Combined model: Linear mixed model

$$\hat{\theta}_d = \boldsymbol{X}^T\beta + u_d + e_d$$

AREA-LEVEL MODELS: THE FAY-HERRIOT MODEL

- The EBLUP under the Fay-Herriot (FH) model is obtained by

$$\hat{\theta}_d^{FH} = \boldsymbol{X}^T\hat{\beta} + \hat{u}_d = \gamma\hat{\theta}_d^{DIR} + (1-\gamma)\,\boldsymbol{X}^T\hat{\beta}$$

- An MSE estimator of the small area estimator of the mean

$$MSE\left(\hat{\theta}_d^{FH}\right) = g_1 + g_2 + g_3$$

g_1 and g_2 uncertainty of BLUP, treating variance components as known
g_3 uncertainty due to estimation of the variance components

Let's go to R

SUMMING UP THE FAY-HERRIOT MODEL

ADVANTAGES:

- Relies only on area-level auxiliary data
- Automatically allocates greater weight to the regression estimator in areas with limited sample sizes and use direct estimator as the domain sample size increases
- Often exhibits superior efficiency compared to the direct estimator
- Addresses unexplained between-area heterogeneity

SUMMING UP THE FAY-HERRIOT MODEL

DRAWBACKS:

- There is a loss of information with the aggregation of auxiliary variables
- The model is fitted with only D observations
- Model checking is essential, introducing potential linearity issues for non-linear parameters.
- Preliminary estimation of sampling variances is necessary
- Cannot be disaggregated for subdomains
- The estimates needs Benchmarking adjustment

UNIT-LEVEL MODELS: BATTESE-HARTER-FULLER MODEL

Random effects model

Notation: (i =individual, d =domain)

$$y_{id} = x_{id}^T \beta + u_d + e_{id} \quad i = 1, \dots, n\ , \quad d = 1, \dots, D$$

Random effects $\quad u_d \sim N(0, \sigma_u^2)$

Error term $\quad e_{id} \sim N(0, \sigma_e^2)$

Basic concept: This linear mixed model is referred to as a random intercept model: the intercepts are allowed to differ among the small domains, whereas the effects of the covariates is equal for all domains.

UNIT-LEVEL MODELS: BATTESE-HARTER-FULLER MODEL

- The EBLUP under the Fay-Herriot (FH) model is obtained by

$$\hat{\theta}_d^{BHF} = \frac{1}{N_d}\left(\sum_{i\in s} y_{id} + \sum_{i\in r} \hat{y}_{id}\right) = \frac{1}{N_d}\left(\sum_{i\in s} y_{id} + \sum_{i\in r} (x_{id}^T\hat{\beta} + \hat{u}_d)\right)$$

- An MSE estimator of the small areia estimator of the mean

$$MSE\left(\hat{\theta}_d^{BHF}\right) = g_1 + g_2 + g_3$$

g_1 and g_2 uncertainty of BLUP, treating variance components as known
g_3 uncertainty due to estimation of the variance components

Let's go to R

SUMMING UP BATTESE-HARTER-FULLER MODEL

ADVANTAGES

- Unit-level auxiliary information, which exploit the correlation with the target variable more effectively than area-level data.

- The total sample size is typically big

- It dynamically assigns higher weight to the regression estimator in areas with smaller sample sizes, transitioning to the direct estimator as the domain sample size grows

- Estimates can be disaggregated for subareas, providing detailed insights

- The synthetic component can be applied to non-sampled areas, enhancing coverage and comprehensiveness.

SUMMING UP BATTESE-HARTER-FULLER MODEL

DRAWBACKS:

- Unit-level auxiliary information is often challenging to obtain
- Limited to linear parameters
- Does not incorporate sampling weights
- Susceptible to outliers and/or deviations from normality
- Rigorous model checking
- Estimates require benchmarking adjustment to ensure comparability and reliability.

EU Twinning Project on Statistics in Jordan

DOS
1949
دائرة الإحصاءات العامة
Department of Statistics

Thank you