# Analysis of Sample Design for Estimation

## Sample Design Specification (Task 1 Analysis)

The first step in formulating the estimation plan involves defining the structure for sampling error calculations based on the provided sample data ('final\_sample.csv'). This requires specifying the stratum codes and the Sampling Error Computation Units.  
For this analysis, the 'Region' variable from the dataset is designated as the primary stratum code, denoted by 'h'. Furthermore, each unique school or building, identified by its 'BCODE' in the dataset, is defined as a Sampling Error Computation Unit. The identifier 'SECU\_ID' used in subsequent references corresponds directly to the 'BCODE'. This structure suggests a cluster design where schools likely served as Primary Sampling Units (Primary Sampling Unit). Variance is therefore estimated based on the variation between these schools (treated as Sampling Error Computation Units) within the same region.  
The number of Sampling Error Computation Units within each stratum (Region 'h') was calculated based on the sample data, as detailed in Appendix Table 1. A key finding from this calculation is the substantial variation in the number of Sampling Error Computation Units per stratum. Specifically, Regions 1 through 4 each contain only one Sampling Error Computation Unit. This poses a challenge for standard variance estimation techniques, which typically require at least two units per stratum to estimate within-stratum variance directly. Consequently, collapsing these strata may be necessary during variance computation, potentially impacting the accuracy of the variance estimate and reducing the effective degrees of freedom. Regions 5 and 6 also contain relatively few Sampling Error Computation Units (2 and 4, respectively), which might lead to less stable variance estimates for these specific strata. In contrast, Regions 7, 8, and 9 contain a more substantial number of Sampling Error Computation Units (21, 21, and 44, respectively), providing a better foundation for stable variance estimation within those strata.  
Finally, regarding the size or expected sample size per Sampling Error Computation Unit, the 'tot\_all' column in the dataset represents the total number of students in each school. This value, denoted as 'SECU\_Size', serves as the measure of size for each Sampling Error Computation Unit. This considerable variation in Sampling Error Computation Unit size, viewed alongside the Sampling Error Computation Unit definition, strongly implies that the sample design is not an equal probability of selection mechanism. This size variability can be considered the expected sample size if the entire school is included in measurements, or it can serve as a size measure if a probability proportional to size sampling scheme was employed.

## Variance Estimation Framework (Task 2 Analysis)

Task 2 focuses on describing the variance estimation procedures for this sample design, determining the available degrees of freedom, formulating an example estimator, and discussing the necessity of sample weights, all building upon the structure defined in Task 1. Standard variance estimation techniques for stratified cluster samples, such as Taylor Series Linearization for complex statistics like means or proportions, are applicable. The general approach involves summing within-stratum variance estimates (conceptualized in Appendix Formula 1 and 2) to obtain the overall variance estimate for statistics like totals or means.

Confidence Intervals are constructed using the point estimate, its estimated standard error (derived from the variance estimate), and a critical value from the t-distribution (as shown structurally in Appendix Formula 3). The degrees of freedom for this t-distribution are calculated as the total number of Sampling Error Computation Units minus the total number of strata. Based on the 'final\_sample.csv' data, this yields 87 degrees of freedom (Appendix Formula 4). However, it's crucial to note the practical implication from the Task 1 analysis: strata containing only one Sampling Error Computation Unit will require collapsing for variance estimation, which will reduce the effective degrees of freedom available in practice.

Regarding estimation, an appropriate example for this data is the weighted stratified mean (Appendix Formula 5), necessary because the design is clearly not an equal probability of selection mechanism due to significant variation in school sizes ('tot\_all'). Therefore, using sampling weights, reflecting inverse selection probabilities and potentially adjustments, is essential for producing unbiased and consistent estimates from this dataset. Compensating for unequal selection probabilities, whether arising from probability proportional to size sampling or varying sampling rates, is mandatory for valid analysis.

## Subclass Analysis Considerations (Task 3 Analysis)

Task 3 addresses the specific requirement to provide estimates and inference for a population subclass constituting approximately 20 percent of the total population. This involves considering how Confidence Intervals for subclass estimates are formed and evaluating whether the current Sampling Error Computation Unit structure is adequate.

The formation of Confidence Intervals for estimates related to the 20 percent subclass (e.g., θ̂\_sub) generally follows the same methodology outlined in Task 2. The variance estimation procedures (e.g., Taylor Series Linearization), accounting for the stratification and clustering (Sampling Error Computation Units), remain applicable. The primary difference lies in restricting the calculations to only those sample units belonging to the specified subclass. The degrees of freedom (87) determined previously would typically still be used for the t-distribution critical value in the Confidence Interval formula (referencing the structure of Appendix Formula 3).

A critical consideration is whether the sample design, specifically the number (96) and characteristics of the Sampling Error Computation Units, provides sufficient data to support reliable analysis for a 20 percent subclass. While the overall sample might be adequate for population estimates, focusing on a smaller subgroup presents challenges. The effective sample size for the subclass is significantly reduced. Sampling Error Computation Units with smaller total enrollments ('SECU\_Size') may contain very few, or even zero, members of the target subclass. This data sparsity can lead to instability in variance estimates, as these estimates rely on observing variation between Sampling Error Computation Units within strata. If many Sampling Error Computation Units contribute little information about the subclass, variance estimates can become unreliable, standard errors might be inflated, and the nominal degrees of freedom might overstate the true precision. Therefore, the current design could pose risks for achieving stable estimates for this 20 percent subclass, particularly if subclass members are unevenly distributed. If high precision for this subclass is essential, further investigation or alternative strategies (e.g., increasing sample size, oversampling) might be warranted.

## Conclusion

This report has detailed the development of an estimation plan tailored to the provided sample design ('final\_sample.csv'). Key components established include the definition of strata based on 'Region' ('h') and Sampling Error Computation Units based on school 'BCODE' ('SECU\_ID'). The analysis framework provides 87 nominal degrees of freedom for variance estimation, although practical adjustments may be needed. Standard procedures for variance estimation, likely using Taylor Series Linearization, were described. The rationale for the essential use of sampling weights due to the design likely not being an equal probability of selection mechanism was highlighted. Finally, potential challenges related to the reliable estimation for a 20 percent population subclass were discussed, focusing on estimate stability given the current Sampling Error Computation Unit structure. These elements form the basis for calculating sampling errors and constructing Confidence Intervals from this sample data.

# Appendix

Appendix Table 1: Number of Sampling Error Computation Units (SECUs) per Stratum (h)

|  |  |
| --- | --- |
| **Stratum (h)** | **Number of Sampling Error Computation Units (nh)** |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 2 |
| 6 | 4 |
| 7 | 21 |
| 8 | 21 |
| 9 | 44 |

## Appendix Formulas

Note: These formulas are represented using standard mathematical notation text. They may need to be converted using Word's Equation Editor for direct editing within Word.

### Formula 1: Variance of Estimated Total (Ŷ)

### Formula 2: Within-Stratum Variance Estimate (Approximate, WR or small fraction)

where:

• H is the total number of strata

• nh is the number of sampled Sampling Error Computation Units in stratum h

• y\_hi is the (weighted) value for Sampling Error Computation Unit i in stratum h

• ȳh is the sample mean of y\_hi values in stratum h

### Formula 3: Confidence Interval for an Estimate (θ̂)

where:

• t(df, 1-α/2) is the critical value from t-distribution with degrees of freedom

• V̂(θ̂) is the estimated variance of the estimate θ̂

### Formula 4: degrees of freedom

df = (Total number of Sampling Error Computation Units) - (Total number of strata)

df = 96 - 9 = 87

### Formula 5: Weighted Stratified Mean Estimator (Ȳ̂w)

where:

• w\_hi is the final sampling weight for Sampling Error Computation Unit i in stratum h