Stratified simple random sampling

Aibat Kossumov

Abstract

This report investigates the effect of stratified simple random sampling without replacement on the estimation of the population mean of Ca (calcium) content in soil, using a real dataset from Barro Colorado Island (BCI), Panama. The dataset contains measurements of mineral concentrations across a rectangular 1000×500 meter grid, with no missing pixels. We compare three sampling methods: cum-root-f stratification based on the covariate Fe (iron), geographical stratification, and simple random sampling without replacement. For the stratified designs, proportional allocation was applied. A simulation study was conducted to estimate the population mean, sampling variance, and design effect for different sample sizes n and numbers of strata H.

1 Introduction

Spatial sampling is a design-based approach to estimation, and the goal is to estimate population characteristics such as means or totals. Unlike model-based approaches, all estimates here are based on the inclusion probabilities of the sampling units, which are determined by the sampling design. Design-based approaches are especially useful when working with a finite (true) population whose units have some spatial structure. In the case of finite populations, we are usually interested in estimating the population mean of a study variable z:

$$m(z) = \frac{1}{N} \sum_{k=1}^{N} z_k, \tag{1}$$

where $\{z_k : k = 1, \ldots, N\}$ are the values of the study variable for the entire population of size N. In our case, the BCI dataset is considered as the full population. Specifically, BCI contains 1250 grid cells, with each row representing one cell. The columns 'x' and 'y' correspond to the coordinates of the cell centers, while the other columns contain measurements of various mineral concentrations within those cells. In this report, we focus on Ca (calcium) as our study variable z, and Fe (iron) as the covariate used for stratification. The main goal of this study is to compare stratification based on the covariate Fe with geographical stratification. Specifically, we aim to determine which of these stratification strategies yields a more precise estimate of population mean $m(z) = 1697.3$.

2 Methods

2.1 Estimation of the population mean

A reasonable estimator of $m(z)$, in the sense of design-unbiasedness, is the Horvitz-Thompson estimator:

$$\widehat{m}_\pi(z) = \frac{1}{N} \sum_{k \in \mathcal{S}} \frac{z_k}{\pi_k}, \tag{2}$$

where $\mathcal{S}$ is the selected (unordered) sample of size n, and π_k denotes the inclusion probability of unit k. Note that in equation (2), the only source of randomness lies in how the sample $\mathcal{S}$ is selected. To select this sample, we must first specify the sampling design. In general, we aim to choose a sampling strategy that yields a more precise estimate of the population mean, meaning that the **sampling variance** $\text{var}\,[\widehat{m}_\pi(z)]$ should be as small as possible.

Two of the most commonly used variants of the Horvitz–Thompson estimator are:

- Simple random sampling without replacement, where all units have the same inclusion probability $\frac{n}{N}$:

$$\widehat{m}_{\text{srswor}}(z) = \frac{1}{n}\sum_{k\in\mathcal{S}} z_k. \tag{3}$$

- Stratified simple random sampling without replacement, where the population is divided into H strata, and simple random sampling without replacement is applied independently within each stratum:

$$\widehat{m}_{\text{st.srswor}}(z) = \sum_{h=1}^{H} \frac{N_h}{N}\widehat{m}_h(z), \text{ where } \widehat{m}_h(z) = \frac{1}{n_h}\sum_{k\in\mathcal{S}_h} z_k^h. \tag{4}$$

  For the notation used in (4), see Chapter 4.1 of [1].

2.2 Stratification techniques

Stratified random sampling is often used because it can improve the precision of population mean estimates. To perform stratified random sampling, we first need to know which stratum each population unit belongs to. If the strata are not specified apriori, then there are several techniques for constructing them. In this study, we consider cum-root-f stratification based on a stratification variable, and geographical stratification. In both approaches, the total number of strata H must be specified in advance.

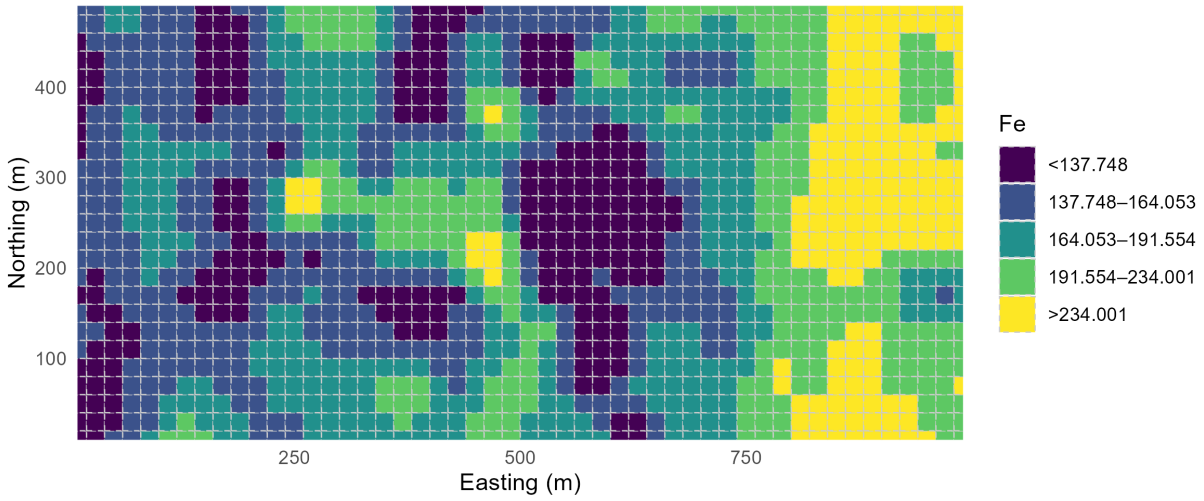
The idea behind cum-root-f stratification is to select a quantitative covariate v (stratification variable) that is correlated with the study variable z and is known for the entire population. Units with similar values of v are then grouped into the same stratum. The steps of cum-root-f stratification are as follows:

- Specify the total number of strata H, and choose a sufficiently large number of bins B for constructing a frequency histogram of the stratification variable.
- Calculate the cumulative sum of the square-rooted frequencies: $\sum_{b=1}^{B}\sqrt{f_b}$.
- The dividing points that define the strata are given by $c, 2c, \ldots, (H-1)c$, where $c = \frac{1}{H}\sum_{b=1}^{B}\sqrt{f_b}$.

In contrast, geographical stratification assigns units to strata based only on their spatial coordinates, which is useful when no auxiliary covariates are available. One way to implement this is by applying the k-means algorithm using the spatial coordinates of the grid cell centers as clustering variables.

Figure 1 shows examples of different stratifications into $H = 5$ strata. The first panel illustrate cum-root-f stratification based on the stratification variable Fe, using $B = 500$ bins. The second panel shows geographical stratification based on the k-means algorithm with 100 random initial configurations.

Cum-root-f stratification of BCI, using Fe as a stratification variable.



Geographical stratification.

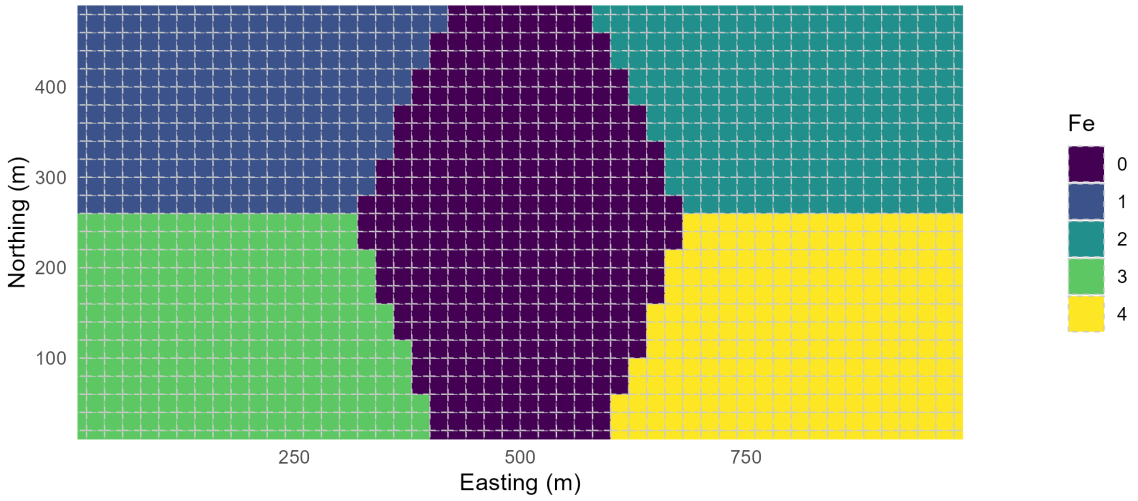


Figure 1: Stratification of the BCI dataset.

2.3 Allocation of sample size to strata

Once the entire population has been stratified and the total sample size n is known, the next step is to decide how many units to sample from each stratum. A natural approach is to allocate the sample proportionally to the size of each stratum in the population, which is known as proportional allocation:

$$n_h = n \cdot \frac{N_h}{\sum_{l=1}^{H} N_l} = n \cdot \frac{N_h}{N}, \tag{5}$$

with N_h the total number of population units in stratum h.

2.4 Sampling standard error

In the following text, the term **sampling standard error** refers to the square root of the sampling variance of the corresponding estimator of the population mean. Sampling standard errors can be computed as follows:

- Sampling standard error of the population mean estimator under simple random sampling without replacement:

$$\text{se}\left[\widehat{m}_{\text{srswor}}(z)\right] = \sqrt{\left(1 - \frac{n}{N}\right)\frac{\sigma^2(z)}{n}}, \tag{6}$$

  where $\sigma^2(z) = \frac{1}{N-1}\sum_{k=1}^{N}(z_k - m(z))^2$ is the population variance of the study variable z.

- Sampling standard error of the population mean estimator under stratified simple random sampling without replacement:

$$\text{se}\left[\widehat{m}_{\text{st.srswor}}(z)\right] = \sqrt{\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2(z)}{n_h}}, \tag{7}$$

  where $\sigma_h^2(z) = \frac{1}{N_h-1}\sum_{k=1}^{N_h}(z_k^h - m_h(z))^2$ is the population variance of the study variable z in stratum h.

Note that the sampling standard errors above can be computed only if the population variances $\sigma^2(z)$ and $\sigma_h^2(z)$ are known, which is typically not the case in practical applications. Therefore, the following estimators of the sampling standard error are used:

- Estimator of the sampling standard error under simple random sampling without replacement:

$$\widehat{\text{se}}_{\text{srswor}}(z) = \sqrt{\left(1 - \frac{n}{N}\right)\frac{\widehat{\sigma}^2(z)}{n}}, \tag{8}$$

  where $\widehat{\sigma}^2(z) = \frac{1}{n-1}\sum_{k\in\mathcal{S}}(z_k - \widehat{m}_{\text{srswor}}(z))^2$ is the sample-based estimator of the population variance of the study variable z.

- Estimator of the sampling standard error under stratified simple random sampling without replacement:

$$\widehat{\text{se}}_{\text{st.srswor}}(z) = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\widehat{\sigma}_h^2(z)}{n_h}}, \tag{9}$$

  where $\widehat{\sigma}_h^2(z) = \frac{1}{n_h - 1} \sum_{k \in \mathcal{S}_h} (z_k^h - \widehat{m}_h(z))^2$ is the sample-based estimator of the population variance of the study variable z in stratum h.

These estimators are motivated by the fact that their second powers are design-unbiased estimators of the corresponding sampling variances. Figures 2 and 3 display the true sampling standard errors along with boxplots of their simulated estimates. More details about the simulation procedure are provided in the Section 3.

2.5 Design effect

It can also be useful to consider not just the sampling variance under stratified simple random sampling without replacement, but rather its ratio to the sampling variance under simple random sampling without replacement. This ratio is known as the **design effect**:

$$\text{DE} = \frac{\text{var}[\widehat{m}_{\text{st.srswor}}(z)]}{\text{var}[\widehat{m}_{\text{srswor}}(z)]} = \frac{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2(z)}{n_h}}{\left(1 - \frac{n}{N}\right) \frac{\sigma^2(z)}{n}}. \tag{10}$$

The smaller the design effect, the more precise the estimator of the population mean under stratified simple random sampling without replacement, compared to simple random sampling without replacement, assuming equal sample sizes in both designs. However, the computation of the design effect in (10) depends on the population variances $\sigma^2(z)$ and $\sigma_h^2(z)$, which are typically unknown in practice. Therefore, the design effect is estimated using the following expression:

$$\widehat{\text{DE}} = \frac{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\widehat{\sigma}_h^2(z)}{n_h}}{\left(1 - \frac{n}{N}\right) \frac{\widehat{\sigma}_w^2(z)}{n}}, \tag{11}$$

where $\widehat{\sigma}_w^2(z)$ denotes a weighted sample-based estimator of the population variance of the study variable z. Figures 4 and 5 display the true design effects along with boxplots of their simulated estimates. More details about the simulation procedure are provided in the next Section 3.

3 Simulations

Recall that our goal is to investigate the sampling variance of estimators of the population mean, where the estimators are obtained using different sampling methods (cum-root-f stratification based on the covariate Fe, geographical stratification, and simple random sampling without replacement). To achieve this, we conducted a simulation study where we estimated sampling standard errors and design effects under different conditions:

- Sample sizes ranged from 30 to 250, which corresponds to approximately 2% to 20% of the full population of size 1250.
- For each simulation scenario:
  1. We first specified the number of strata, H, to be 3, 5, 7, or 10.

2. Based on the value H, we stratified the entire population either by cum-root-f or geographical stratification.
3. After stratification, we fixed a sample size n and calculated the corresponding true sampling standard errors (see Equations 6 and 7) and true design effect (see Equation 10). Note that these computations can be performed because we have access to the entire population, which allows us to compute the exact values of $\sigma^2(z)$ and $\sigma_h^2(z)$.
4. We are also interested in the behavior of the estimated sampling standard errors (see 8 and 9) and the estimated design effect (see 11). Therefore, we conducted 1000 simulation runs for each specified sampling design and sample size. In each run, we performed the following steps:
   - Drew a sample according to the specified design,
   - Computed the estimated sampling standard errors (see Equations 8 and 9),
   - Computed the estimated design effect (see Equation 11).

- Finally, we aggregated all simulation results and visualized them using boxplots, as shown in the figures below.

From Figures 2 and 3, we observe that the smallest sampling standard error is achieved by geographical stratification, while the largest corresponds to simple random sampling without replacement. Even for larger sample sizes, there is a noticeable difference between the sampling standard errors under simple random sampling without replacement and those under stratified sampling. It is also apparent from the figures that all sampling standard errors decrease as the sample size increases. Additionally, when comparing the top and bottom panels in Figure 2, we see that increasing the number of strata from 3 to 5 caused the sampling standard errors of the two stratified sampling methods to become closer to each other. This effect is not so noticeable when increasing the number of strata further to 7 and 10, as shown in Figure 3.

Regarding the boxplots, for smaller sample sizes, the longest boxplots correspond to the estimated standard errors under simple random sampling without replacement, indicating that the estimator in Equation 8 exhibits greater variability than the estimator in Equation 9. As for the boxplots corresponding to cum-root-f and geographical stratification, both methods appear to produce boxplots of nearly equal length.

Figures 4 and 5 show the true design effects, as well as boxplots of the estimated design effects obtained from simulations. From these figures, we can see that the design effects for geographical stratification are smaller than those for cum-root-f stratification based on the covariate Fe. This confirms that the estimator of the population mean is more precise under geographical stratification. Moreover, the boxplots corresponding to geographical stratification are shorter than those for cum-root-f stratification, indicating that the estimator in Equation 11 is less variable under geographical stratification.

In the lower plot of Figure 4, we can see that for all considered sample sizes, the design effect of geographical stratification is close to 0.5. This means that the estimator of the population mean under geographical stratification with 5 strata is about twice as precise as the estimator under simple random sampling without replacement.

Figures 6 and 7 show boxplots of the estimated population means obtained from simulations. The results indicate that the estimates of population mean are least variable under geographical stratification and most variable under simple random sampling without replacement.

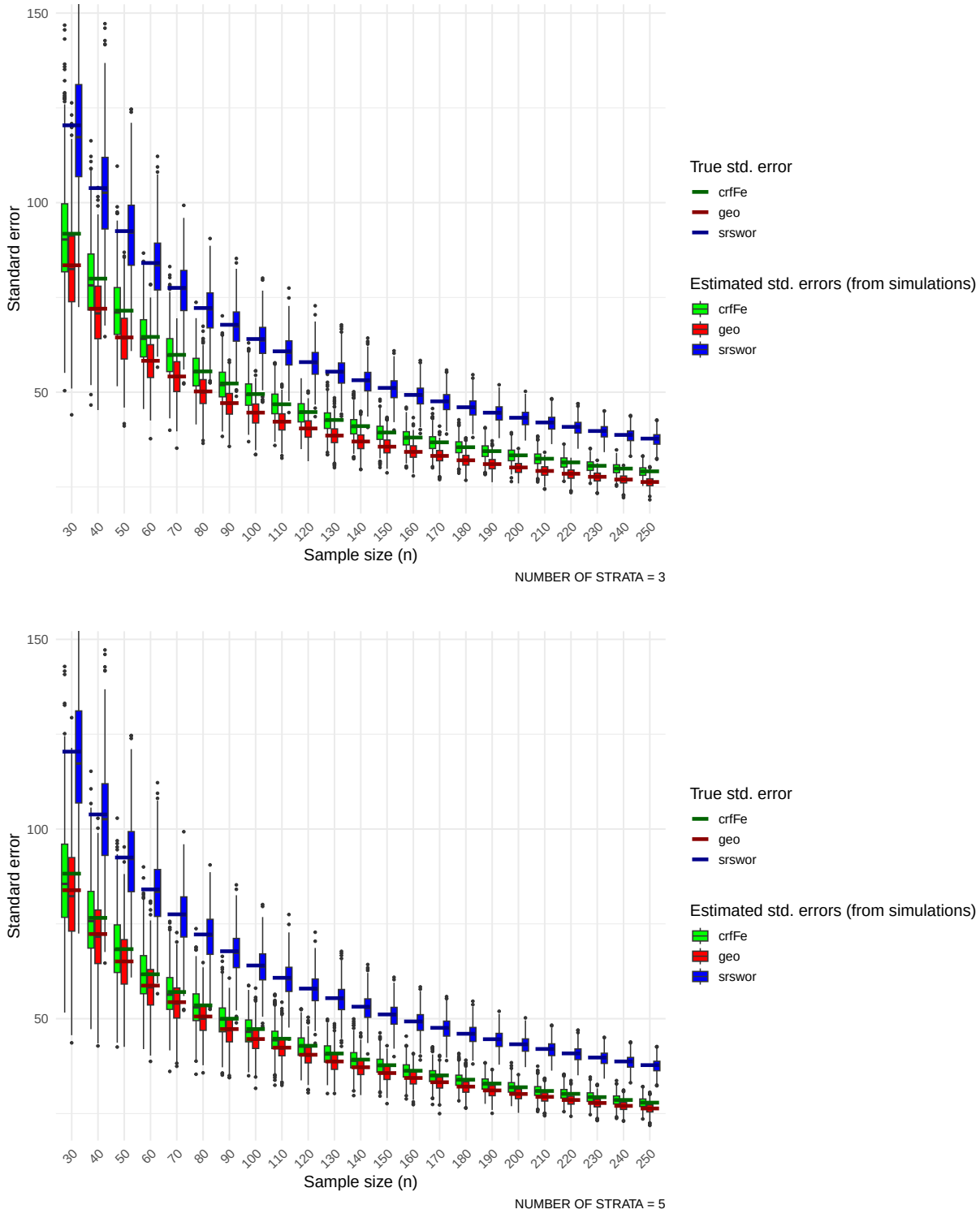


Figure 2: The figure shows the true sampling standard errors, computed using Equations 6 and 7, represented by small colored horizontal lines. Each boxplot corresponds to a collection of estimated sampling standard errors, obtained from 1000 simulation runs and calculated using Equations 8 and 9. The results are shown for stratification with 3 and 5 strata.

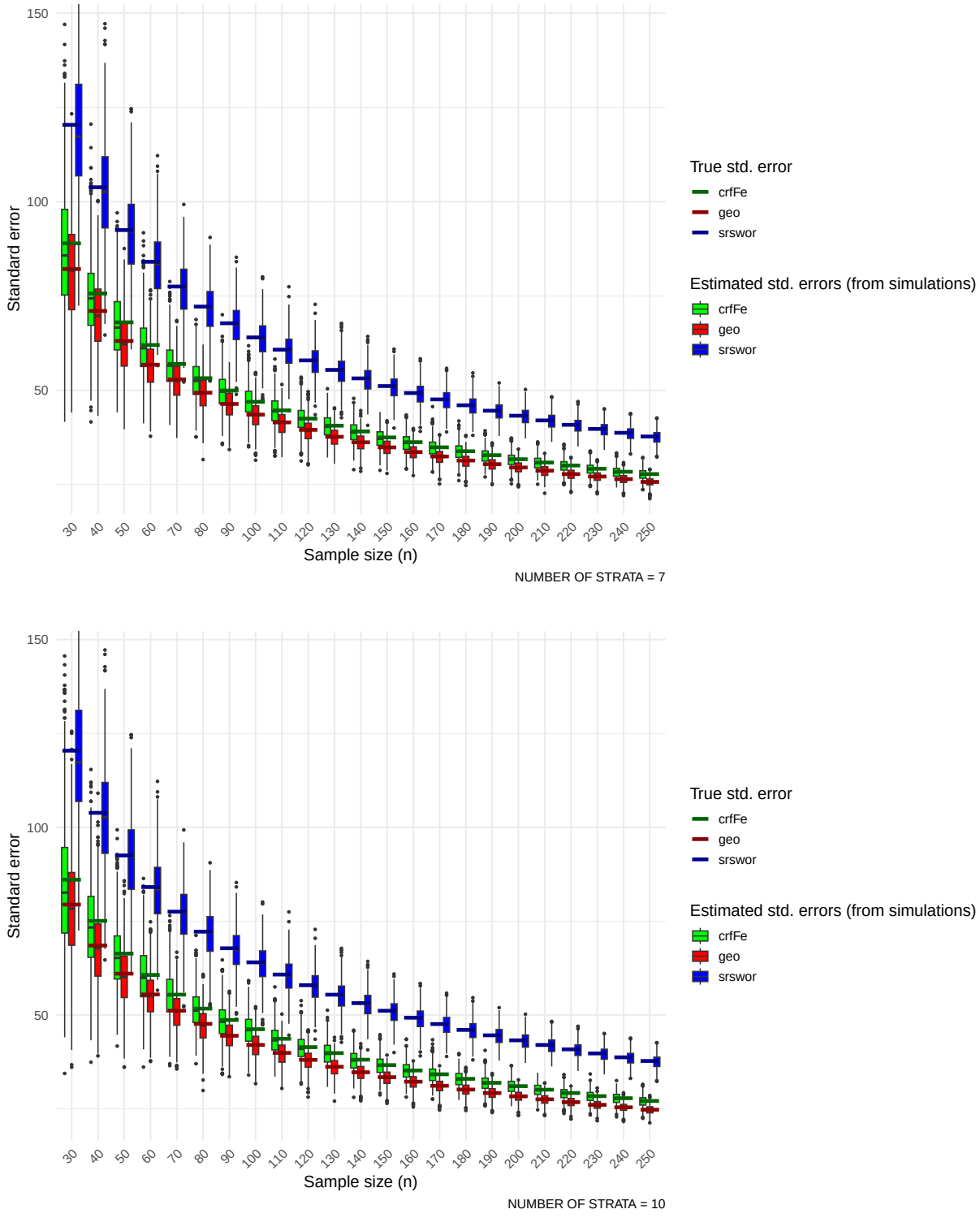


Figure 3: The figure shows the true sampling standard errors, computed using Equations 6 and 7, represented by small colored horizontal lines. Each boxplot corresponds to a collection of estimated sampling standard errors, obtained from 1000 simulation runs and calculated using Equations 8 and 9. The results are shown for stratification with 7 and 10 strata.

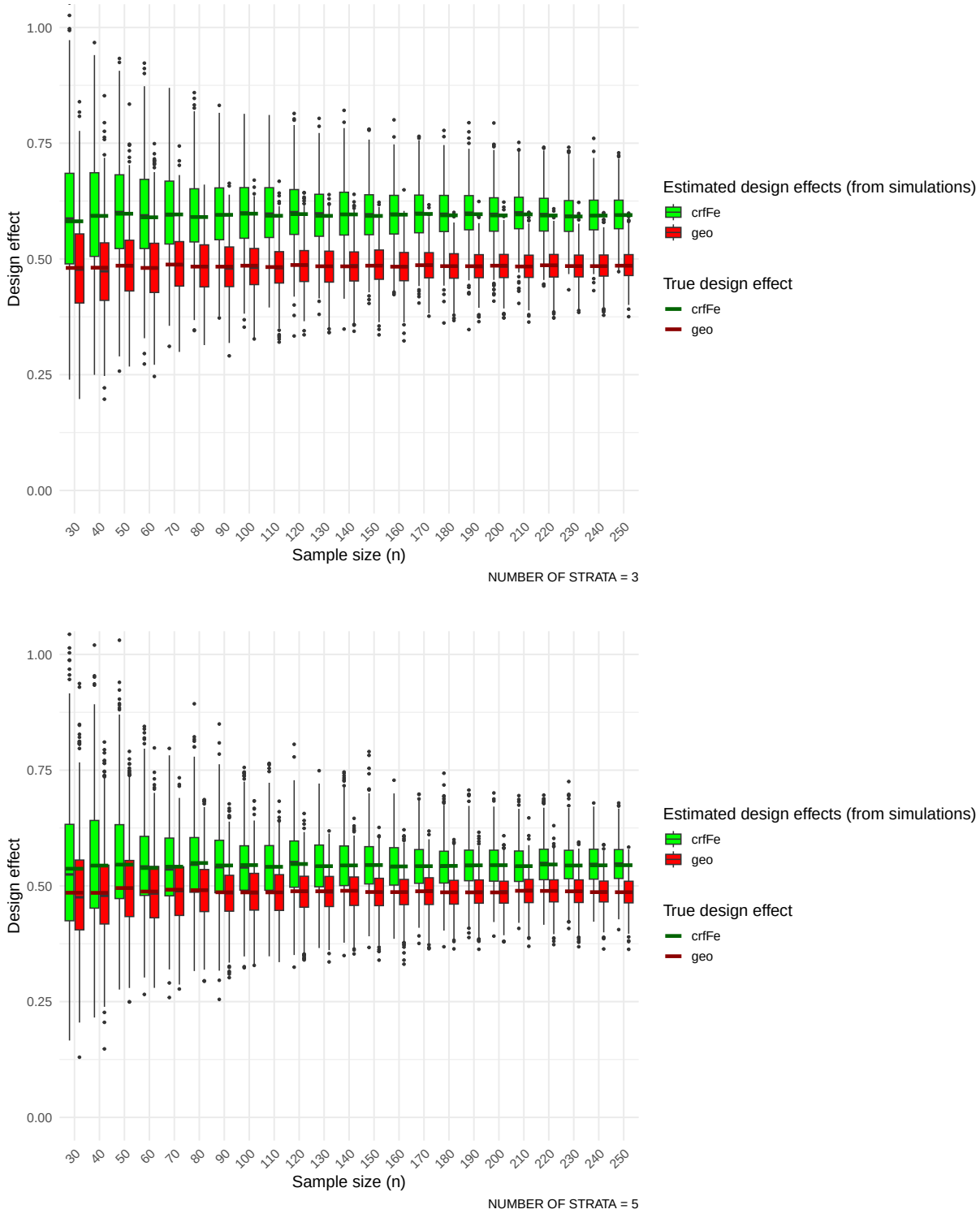


Figure 4: The figure shows the true design effects, computed using Equation 10, represented by small colored horizontal lines. Each boxplot corresponds to a collection of estimated design effects, obtained from 1000 simulation runs and calculated using Equation 11. The results are shown for stratification with 3 and 5 strata.

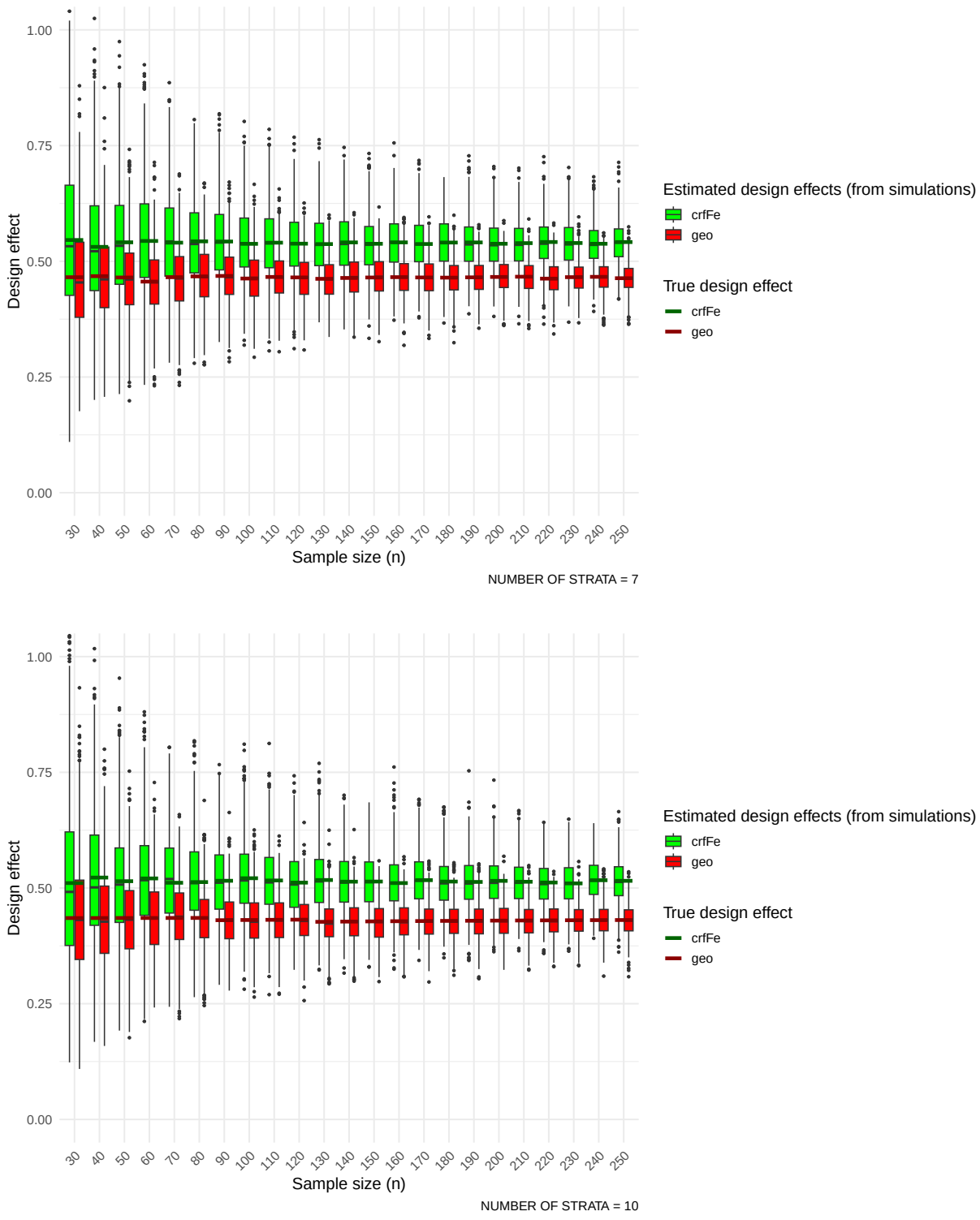


Figure 5: The figure shows the true design effects, computed using Equation 10, represented by small colored horizontal lines. Each boxplot corresponds to a collection of estimated design effects, obtained from 1000 simulation runs and calculated using Equation 11. The results are shown for stratification with 7 and 10 strata.

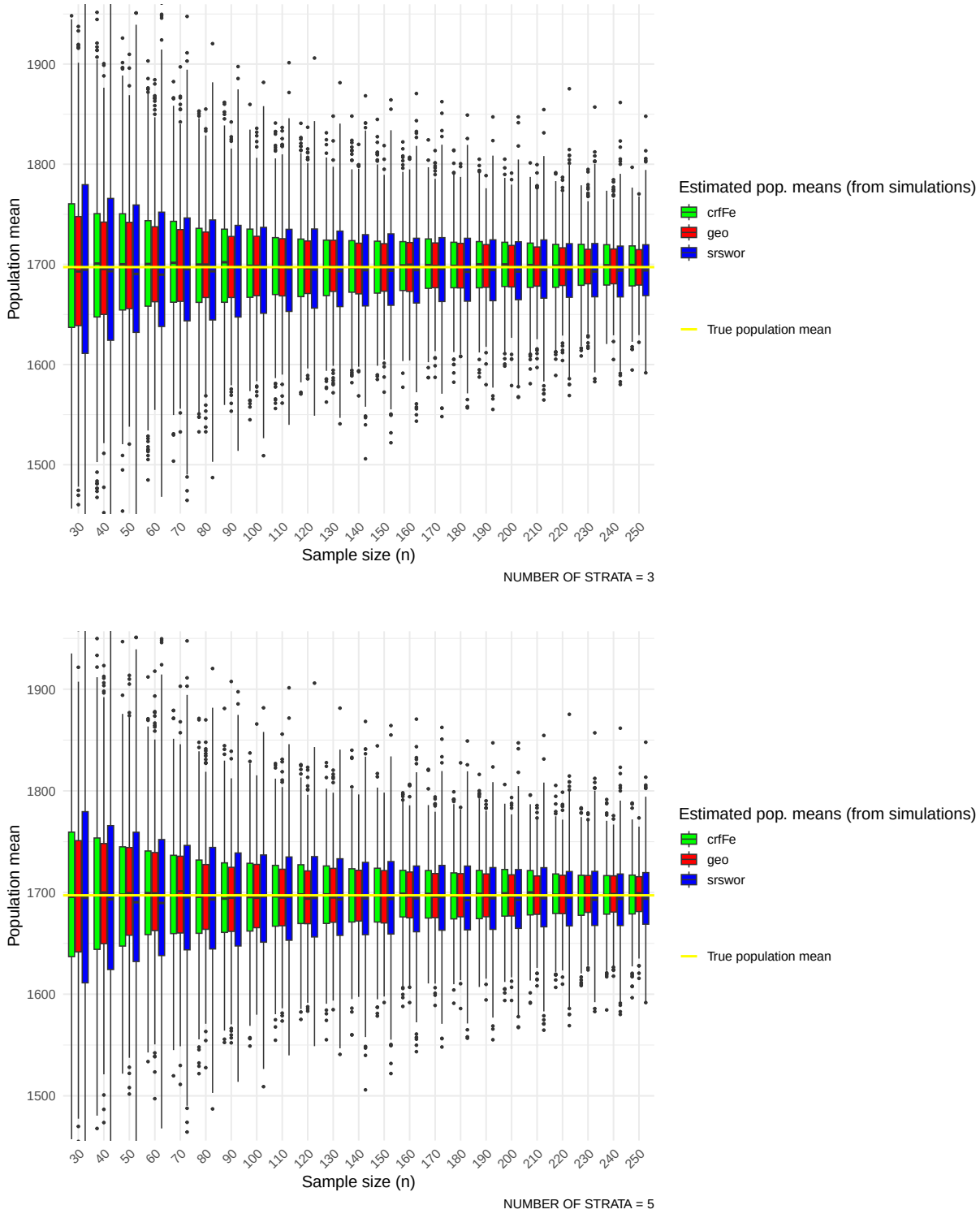


Figure 6: The figure shows the true population mean, computed using Equation 1, represented by a continuous yellow horizontal line. Each boxplot corresponds to a collection of estimated population means, obtained from 1000 simulation runs and calculated using Equations 3 and 4. The results are shown for stratification with 3 and 5 strata.

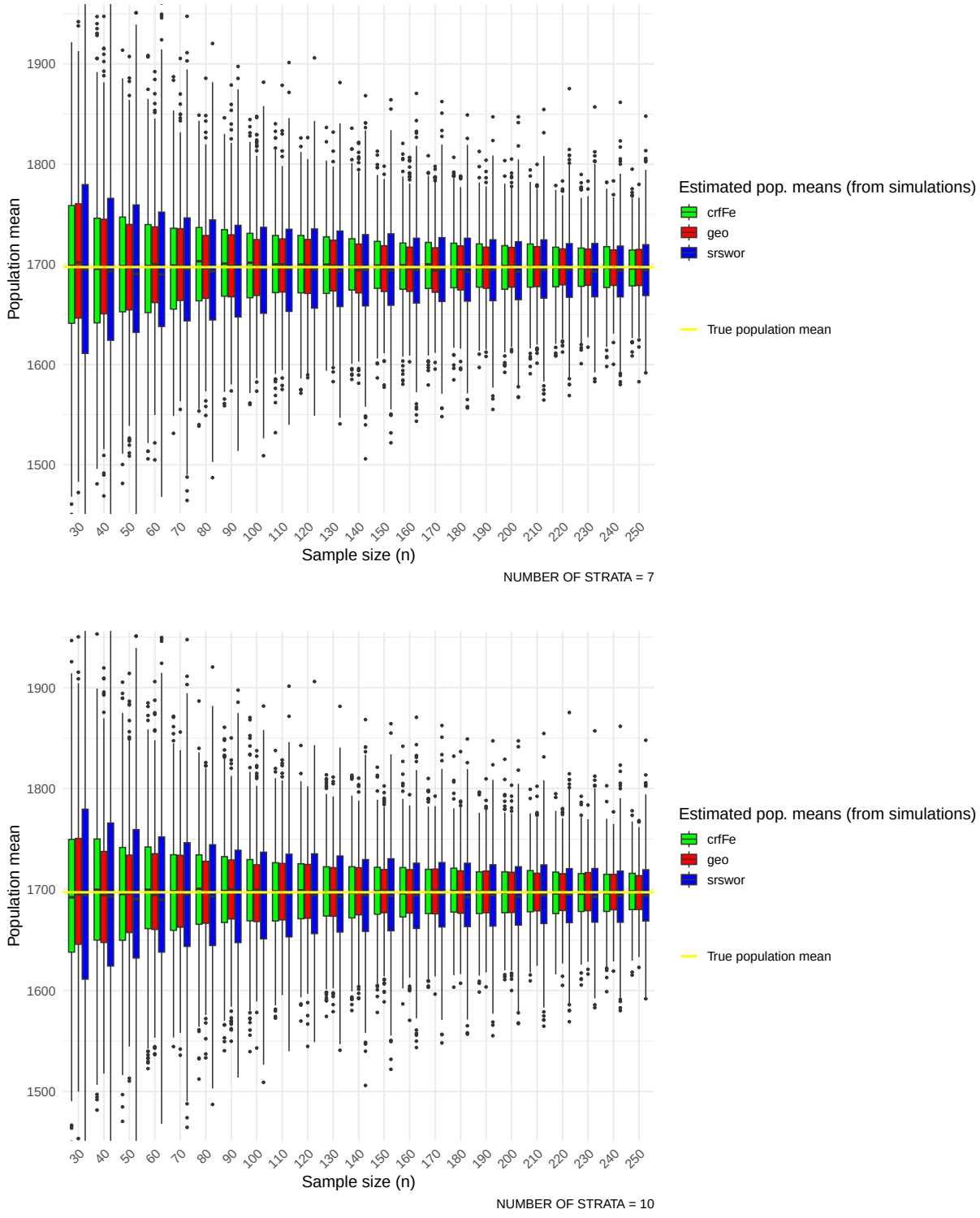


Figure 7: The figure shows the true population mean, computed using Equation 1, represented by a continuous yellow horizontal line. Each boxplot corresponds to a collection of estimated population means, obtained from 1000 simulation runs and calculated using Equations 3 and 4. The results are shown for stratification with 7 and 10 strata.

4 Conclusion

In this simulation study, we compared three sampling methods: cum-root-f stratification based on the covariate Fe, geographical stratification, and simple random sampling without replacement. The results showed that geographical stratification tends to have the smallest sampling variance, while simple random sampling without replacement exhibits the largest.

When comparing the two stratified sampling methods, cum-root-f and geographical stratification, we found that geographical stratification performs slightly better in terms of both sampling standard error and design effect. For example, the design effect for cum-root-f stratification was slightly above 0.5, whereas for geographical stratification it was slightly below 0.5. This suggests that the better performance of geographical stratification may be due to the fact that the spatial structure of the study variable Ca explains more of its variability than the stratification variable Fe.

An extension of this project could involve investigating how the above-mentioned methods perform when using Neyman allocation, or when the cum-root-f stratification is based on a better predictor than the stratification variable Fe.

References

[1] Dick J. Brus. *Spatial Sampling with R*. 2023. URL: `https://dickbrus.github.io/SpatialSamplingwithR/`.