

Item: Original

Efficiency of four probability sampling designs: a Monte Carlo simulation study

Eficiencia de cuatro diseños de muestreo probabilístico: estudio de simulación Monte Carlo

¹Marco Antonio Ramos Espinal  .

¹National Autonomous University of Honduras (UNAH), University City, Tegucigalpa, Honduras

*Corresponding author: Marco Antonio Ramos Espinal 

Submitted: 07-13-2026

Revised: 08-19-2026

Accepted: 09-03-2026

Published: 10-12-2026

How to cite:

Ramos Espinal, M. A. (2026). Efficiency of four probabilistic sampling designs: Monte Carlo simulation study. *Sapiens International Multidisciplinary Journal*, 3(5), 1-23. <https://doi.org/10.71068/SIMJ0005>.

ABSTRACT

Introduction: the choice of sampling design determines the precision and validity of survey inferences, yet differences between designs are rarely quantified on a population with fully known parameters. We compared four probability sampling designs (simple random, stratified, cluster and systematic) for estimating a mean, a standard deviation and a proportion.

Method: under specified assumptions, we generated a synthetic population of $N = 1,000,000$ individuals (age, weight and sex) in 40,000 clusters of 25. For each design we drew 1,000 samples of $n = 2,000$ and computed estimates and 90%, 95% and 99% confidence intervals. We assessed bias, root mean squared error, relative efficiency, design effect, coverage and interval width, with a sensitivity analysis of standard errors.

Results: all designs were practically unbiased (maximum relative bias 0.197%). Age stratification gained only for age, the stratification variable (relative efficiency 10.37 for the mean and 2.94 for the standard deviation), not for weight or the proportion of men; its mean relative efficiency (3.29) falls to 1.52 without mean age. Systematic sampling on a frame ordered by an age-related variable gained moderately (2.01 for mean age). Approximate standard errors produced conservative intervals (99.9% and 99.1% coverage at 95%); design-based standard errors brought them to 93.6% and 95.4%. Cluster sampling was the least efficient (design effects up to 2.85).

Conclusions: no design dominated; efficiency depended on how the design relates to the parameter. Because all designs used the same sample size but were not evaluated at equal operational cost, the comparison concerns statistical efficiency, not economic superiority. As the results come from one synthetic population with age-defined strata, the magnitudes should be generalized cautiously.

Keywords: probability sampling; stratified sampling; cluster sampling; systematic sampling; Monte Carlo simulation; design effect

RESUMEN

Introducción: el diseño de muestreo determina la precisión de las inferencias, pero sus diferencias rara vez se cuantifican con parámetros conocidos. Comparamos cuatro diseños probabilísticos (aleatorio simple, estratificado, por conglomerados y sistemático) para estimar una media, una desviación típica y una proporción.

Método: bajo supuestos especificados, generamos una población sintética de $N = 1\,000\,000$ de personas (edad, peso y sexo) en 40 000 racimos de 25. Por diseño extrajimos 1 000 muestras de $n = 2\,000$ con estimaciones e intervalos de confianza al 90 %, 95 % y 99 %. Evaluamos sesgo, error cuadrático medio, eficiencia relativa,

efecto de diseño, cobertura y amplitud de intervalos.

Resultados: los cuatro diseños fueron prácticamente insesgados (sesgo relativo $\leq 0,197\%$). Estratificar por edad ganó solo en la edad, variable de estratificación (eficiencia relativa: 10,37 en la media; 2,94 en la desviación típica), no en el peso ni en la proporción de hombres; su eficiencia media (3,29) baja a 1,52 sin la media de la edad. El sistemático con marco ordenado por una variable asociada a la edad ganó moderadamente (2,01). Los errores estándar aproximados dieron coberturas excesivas (99,9% y 99,1% al 95%), corregidas a 93,6% y 95,4% con errores coherentes con el diseño. Los conglomerados fueron los menos eficientes (efecto de diseño $\leq 2,85$).

Conclusiones: ningún diseño dominó; la eficiencia dependió de la relación entre el diseño y el parámetro. Como los diseños compartieron tamaño muestral, pero no se compararon a igual costo operativo, los resultados se refieren a eficiencia estadística, no a superioridad económica. La ventaja del estratificado depende de los estratos definidos por edad y de los supuestos de esta población sintética.

Palabras clave: muestreo probabilístico; muestreo estratificado; muestreo por conglomerados; muestreo sistemático; simulación Monte Carlo; efecto de diseño

INTRODUCTION

Probabilistic sampling has supported inference about finite populations since Neyman formalized the representative method in 1934 (1), and Horvitz and Thompson's estimator (2) completed the base. If the probability with which each unit enters the sample is known, the parameter and also the precision of that estimate can be estimated without making assumptions about the shape of the population (3-6). This property explains why it continues to be the reference compared to non-probabilistic samples, which are increasingly frequent (7,8).

Survey designers usually have at least four classic options: simple random sampling (MAS), stratified, cluster or cluster sampling, and systematic sampling (5,9). Each organizes the population in a different way. Stratified uses homogeneous subpopulations defined in advance; cluster saves travel by grouping neighboring units; systematic sampling follows an ordered frame at fixed intervals and often introduces implicit stratification (10-12). Choosing one or the other changes the variance of the estimators, the amplitude of the intervals, and thus the cost of achieving a given precision (13,14).

The theory of these differences is mature, but their concrete magnitude is rarely observed in practice, because in a real survey the true parameter is unknown and the estimation error cannot be measured. Monte Carlo simulation on a fully known artificial population solves the problem (15,16). With the actual mean, standard deviation and proportion in view, the bias, mean square error and coverage of the intervals are calculated directly and can be compared between designs (17). Synthetic populations and "digital twins" are increasingly used for this purpose, for example to evaluate forest inventory designs (18).

The available comparisons usually deal with one design or a family of estimators at a time: new estimators of the mean or variance under a fixed design (19-22), variants of the systematic for frames with a linear trend (12), estimators of the variance of the systematic (23) or optimal allocation rules in the stratified (24,25). Less frequent is a simultaneous comparison of the four basic designs on the same known population and with the same sample size, which evaluates both the point estimation and the validity of the intervals for parameters of different nature: a mean, a standard deviation and

a proportion. This is the gap addressed by this study. In addition to ordering the designs by efficiency, we show for which parameters the gain of each design occurs, when that gain is not transferred to the confidence intervals because the standard error used does not capture it, and how much the coverage changes when using a standard error consistent with the design.

In this work, we simulated a population of 1,000,000 people of age, weight, and sex, and compared the four designs in the estimation of a mean, a standard deviation, and a proportion. The general objective was to determine which design is more efficient and under what conditions. The specific objectives were (i) to quantify the bias and accuracy of each point estimator, (ii) to evaluate the empirical coverage and the amplitude of the 90%, 95%, and 99% confidence intervals, (iii) to order the designs according to their overall performance, and (iv) to assess the sensitivity of coverage to the standard error calculation method. We start from the hypothesis, derived from the theory, that no design dominates in all parameters and that efficiency depends on the relationship between the structure of the design and what is estimated (3,4).

The contribution has a methodological and an applied aspect. The study follows the ADEMP scheme for simulations (15,17) and addresses the recommendations to document the procedures to favor their replication (26) and to avoid comparisons that favor any method in advance (27). Verification scripts independent of the production code are accompanied; their exact correspondence with the published results must be contrasted before editorial delivery. Guidelines are also offered for planning community and population health surveys, where age, weight, and sex are commonly used variables.

LITERATURE REVIEW

We consider a finite population $U = \{1, 2, \dots, N\}$ in which each unit i has a value y_i of the variable of interest and, for a dichotomous attribute, an indicator $a_i \in \{0, 1\}$. The target parameters are the mean μ , the standard deviation σ , and the proportion P of equation (1) (2,3).

$$\mu = \frac{1}{N} \sum_{i=1}^N y_i, \quad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2}, \quad P = \frac{1}{N} \sum_{i=1}^N a_i \quad (1)$$

A sampling design is a probability distribution over the set of possible samples, which assigns to each unit a known and positive probability of inclusion π_i . The sampling fraction is $f = n/N$ and the correction factor per finite population is $1 - f$ (4).

2.1 Simple random sampling

In the MAS without replacement, all samples of size n have the same probability and $\pi_i = n/N$. The sample mean of equation (2) is unbiased for μ (3.5).

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (2)$$

Equations (3) and (4) give their variance and that of the estimator of the proportion p , with S^2 as the quasi-variance of the population. Since it does not use auxiliary information or any structure of the population, the MAS serves as a reference to measure the efficiency of the other designs (5).

$$\text{Var}(\bar{y}) = (1 - f) \frac{S^2}{n} \quad (3)$$

$$\text{Var}(p) = (1 - f) \frac{P(1-P)}{n-1} (4)$$

2.2 Stratified random sampling

The stratified design divides the population into L-sized strata of N_h size and extracts from each one an independent subsample by MAS. The mean estimator is the weighted mean of equation (5) and its variance is the weighted mean of equation (6) (1,3).

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h, \quad W_h = \frac{N_h}{N} (5)$$

$$\text{Var}(\bar{y}_{st}) = \sum_{h=1}^L W_h^2 (1 - f_h) \frac{S_h^2}{n_h} (6)$$

Variance depends only on the variability within the strata, so that the more homogeneous they are with respect to the variable of interest, the lower it will be (4). With proportional allocation ($n_h = n \cdot W_h$) the design is never less efficient than the MAS for the mean, and gains more when the stratification variable is closely associated with the study variable (13,14). Neyman's optimal allocation still refines this gain (1,24).

2.3 Cluster sampling

In this design, the population is organized into G clusters, g is selected from them and all their units are observed (one-stage sampling). For clusters of equal size M chosen by MAS, the estimator of the mean and its variance, which depends on the variability between clusters S_b^2 , are those of equations (7) and (8) (3,13).

$$\bar{y}_{cl} = \frac{1}{g} \sum_{i=1}^g \bar{y}_i (7)$$

$$\text{Var}(\bar{y}_{cl}) = \left(1 - \frac{g}{G}\right) \frac{S_b^2}{g} (8)$$

Kish summarized its efficiency in the design effect, which for clusters of equal size approximates equation (9), where ρ is the intraclass correlation coefficient (13).

$$\text{Deff} \approx 1 + (M - 1) \rho (9)$$

The units of the same cluster tend to look alike ($\rho > 0$) and therefore, with the same number of units, the design tends to be less precise than the MAS. It is chosen because of the savings in fieldwork and because there are often only cluster frames (13,28).

2.4 Systematic sampling

On an ordered frame of N units, a random start $r \in \{1, \dots, k\}$ is chosen, and then each k-th unit, with the sampling interval of equation (10).

$$k = \frac{N}{n} (10)$$

The result depends on the order of the frame (10,11). If the frame follows a variable associated with the variable of interest, the design behaves as an implicit stratification and reduces the variance

(11,12); if the order is random with respect to that variable, it is equivalent to the MAS. With a single systematic sample (a single random start) there is no unbiased estimator of variance. In practice, the MAS formula is used, which overestimates variance when the frame has a trend, or approximate estimators based on the differences between contiguous units of the frame (23,29), such as the one in equation (14) in section 2.5.

2.5 Spot and interval estimation

The point estimate gives a value $\hat{\theta}$ for the parameter θ . The confidence interval of level $1 - \alpha$ of equation (11) adds the uncertainty, with $EE(\hat{\theta})$ as the estimated standard error and $z_{1-\alpha/2}$ as the normal quantile (1.645, 1.960 and 2.576 for 90%, 95% and 99%).

$$\hat{\theta} \pm z_{1-\alpha/2} EE(\hat{\theta}) \quad (11)$$

For means and proportions, the standard error comes from the variance formulas of each design (equations 3 to 8). In the proportion we use the Wald interval, although the Wilson (30) and Agresti and Coull (31) intervals cover better in small samples or with extreme proportions (32,33). For the standard deviation, the main analysis uses the normal approximation of equation (12), valid with large n (29). This formula does not depend on the design: it assumes a simple random sample of a variable with the kurtosis of the normal distribution, so that it can overestimate or underestimate the error when the design gains precision or the variable is not normal (34,35).

$$EE(s) \approx \frac{s}{\sqrt{2(n-1)}} \quad (12)$$

As an alternative consistent with the design, we use Taylor's linearization (29). The standard deviation is approximated by the mean of the influencing variable of equation (13), and its standard error is obtained by applying the variance formula of the design to this mean: that of MAS (equation 3), stratified (equation 6) or that of clusters (equation 8). In the systematics, instead of the MAS formula, the estimator of successive differences of equation (14) can be used, which compares each unit with the previous one in the order of the frame and thus collects part of the implicit stratification (23,29).

$$u_i = \frac{(y_i - \bar{y})^2 - s^2}{2s} \quad (13)$$

$$v_{DS}(\bar{y}) = \frac{1-f}{2n(n-1)} \sum_{i=2}^n (y_i - y_{i-1})^2 \quad (14)$$

2.6 Efficiency and goodness-of-fit measures

As θ is known, the performance of each estimator is measured on the B replicates of the simulation (15,17). Equation (15) defines the bias and (16), the mean square error (ECM) and its root (RECM), which expresses the total error in the units of the parameter.

$$\text{Sesgo}(\hat{\theta}) = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b - \theta \quad (15)$$

$$\text{ECM} = \text{Sesgo}^2 + \text{Var}(\hat{\theta}), \quad \text{RECM} = \sqrt{\text{ECM}} \quad (16)$$

The relative efficiency against MAS (equation 17) compares mean square errors, and the design effect (equation 18) compares variances. An efficiency greater than 1 or a Deff less than 1 indicates an advantage over MAS (13).

$$Ef_{rel} = \frac{ECM_{MAS}}{ECM} \quad (17)$$

$$Deff = \frac{Var_{dis}(\hat{\theta})}{Var_{MAS}(\hat{\theta})} \quad (18)$$

At intervals, the empirical coverage of equation (19) is the proportion of replicates in which the interval contains θ , ideally equal to the nominal level, and the mean amplitude measures its precision. With equal coverage we prefer the narrowest interval (5,17).

$$\hat{C} = \frac{1}{B} \sum_{b=1}^B \mathbf{1}\{\theta \in IC_b\} \quad (19)$$

METHOD

We designed a Monte Carlo simulation experiment in which the population is fully known and the four designs are applied repeatedly under identical conditions. We follow the scheme of Morris et al. (15), known as ADEMP and developed in recent templates for planning and communicating simulations (17): objectives, data-generating mechanism, estimates, methods and performance measures.

3.1 Synthetic population

We generated $N = 1\,000\,000$ individuals grouped into $G = 40\,000$ clusters of $M = 25$ units, which represent territorial aggregates such as blocks or census tracts. To mimic the resemblance between neighbors, we added random cluster effects normal to age and weight (standard deviation of 3 years and 3 kg), which produces a positive intraclass correlation. Sex is a Bernoulli variable whose probability of being male is 0.49 plus a normal cluster effect with a standard deviation of 0.05 (bounded between 0.25 and 0.75). Age combines a mean of 42 years, the cluster effect and a normal individual term with a standard deviation of 14 years, and was truncated at the interval [18, 95]; truncation accumulates 4.7% of the population in 18 years. The weight is 62 kg plus 13 kg in men, 0.30 kg for each year of difference from 42 years, the cluster effect and an individual term with standard deviation of 9 kg, and was truncated to [40, 140] kg. With the complete population we calculate the true parameters that serve as a standard (Table 1).

The parameters were chosen to represent the adult population of a community or health survey, not to reproduce a specific country. The mean of 42 years and the individual standard deviation of 14 years, with limits of 18 and 95 years, give a population aged 18 years or older with a broad age structure, and truncation avoids values outside the domain of an adult survey. The 13 kg difference between sexes and the slope of 0.30 kg per year generate moderate associations between weight, sex, and age (in the population, $r = 0.53$ between sex and weight and $r = 0.34$ between age and weight), while age and sex were generated independently ($r = 0.001$); the 40 kg limit affects only 0.8% of the population. Cluster effects produce intraclass correlations of 0.044 in age, 0.067 in weight, and 0.010 in sex, low or moderate values such as those recorded in cluster household surveys (13,28).

These relationships determine which parameters can benefit from age stratification and frame order, and therefore we treat them as conditions of the experiment and not as general traits of any population.

Table 1.

Simulation Design Parameters and True Population Values

Parameter	Value
Population size (N)	1 000 000
Sample size per replica (n)	2 000 ($f = n/N = 0.2\%$)
Number of Monte Carlo samples (B)	1,000 per design
Cluster structure	40,000 clusters of size $M = 25$
Systematic sampling interval (k)	500
Age group strata (L)	4 ([18,30), [30,45), [45,60), [60,95])
Layer sizes (Nh)	199 966; 381 477; 313 221; 105 336
Proportional allocation (nh)	400; 763; 626; 211
Actual Mean Age (μ)	42.34 years
Actual standard deviation for age (σ)	13.75 years
Actual Average Weight (μ)	68.49 kg
Actual standard deviation of weight (σ)	12.14 kg
Actual proportion of men (P)	49,0 % (0,490)
Correlation of the auxiliary ordering variable (Z) with age	0.70 (realized: 0.7002)
Random Seed (Reproducibility)	20260627

Note: own elaboration based on the simulation (seed 20260627).

3.2 Implemented sampling designs

We applied the four designs to the same population with a common sample size of $n = 2\,000$. (a) MAS: selection of 2 000 units without replacement. (b) Stratified: four strata per age group, a variable available in most community frameworks, with proportional allocation and MAS within each stratum. (c) Clusters: $g = 80$ clusters of 25 people selected by MAS, with observation of all its units. (d) Systematic: the framework was ordered from lowest to highest according to an auxiliary variable $Z = 0.70 \cdot \text{zedad} + (1 - 0.702)1/2 \cdot \varepsilon$, where zedad is the standardized age and $\varepsilon \sim N(0, 1)$ an independent noise, so that $\text{corr}(Z, \text{age}) = 0.70$ per construction (0.7002 in the population generated). Z mimics an administrative record associated with age without coinciding with it. A random start r between 1 and $k = 500$ is chosen and the units $r, r + k, r + 2k, \dots$; since the design only admits 500 different samples, the replicates choose the replacement start. In each design we use the estimators and variance formulas from the previous section. In the systematic analysis, the main analysis estimates the variance with the MAS formula, as is usual (29); this formula is an approximation and not an unbiased estimator, which we treat as a limitation and examine in the sensitivity analysis. The equality of n controls the number of observations, but it does not homogenize displacements, frame development, personnel, time, or other execution costs; these were not measured or modeled.

3.3 Monte Carlo simulation procedure

For each design we extracted $B = 1\,000$ independent samples. In each one we calculated five estimates (mean and standard deviation of age, mean and standard deviation of weight and proportion of men), with their standard errors and intervals at 90 %, 95 % and 99 %. We thus obtained, by design and estimation, an empirical sample distribution of 1 000 values. The analysis of the manuscript is described as developed in Python 3 with NumPy and a fixed seed (20260627). To facilitate independent verification, the reference script S1 (Suplementario_S1_simulacion_muestreo.py), which implements the documented methodological decisions, and the script S2 (Suplementario_S2_figuras.py), which generates auxiliary graphs from the output of S1, are attached. These files are not the original production code and are not claimed to exactly reproduce the values in Tables 1 through 6 or Figures 1 through 6.

For each estimand we calculated two standard errors. The main one (EE1) was used in Tables 2 to 5: in the means and proportion, the variance formula of each design (equations 3 to 8, with that of the MAS in the systematic); in the standard deviation, $s/[2(n - 1)]^{1/2}$ in the four designs (equation 12), where s is the sample quasi-deviation in the MAS, the systematic, and the clusters, and in the stratified the root of $\sum W_h [s_h^2 + (\bar{y}_h - \bar{y}_{st})^2]$. The alternative (EE2) was used in the sensitivity analysis (Table 6): in the standard deviation, the linearization of equation (13) with the variance of the MAS, the stratified (with u_i calculated with respect to $\bar{y}_{st}$ and the stratified standard deviation) or that of clusters (on the means of u_i per cluster); in the systematic analysis, in addition, the estimator of successive differences (equation 14) for the means, the proportion and the influencing variable. In the other cases, EE2 coincides with EE1. The calculation of EE2 does not consume random numbers, so the point estimates and EE1 are identical with or without the sensitivity analysis.

3.4 Performance measures

We used seven measures: bias, relative percentage error, RECM, relative efficiency against MAS (ECMMAS/ECM), design effect (Deff), empirical coverage at all three levels, and mean amplitude of the intervals. We consider the design that combines a low ECM with coverage close to nominal coverage to be more efficient. For the overall balance, we averaged, in each design, the relative efficiency of the five estimands and the absolute deviation of coverage from 95%. Since the arithmetic mean is sensitive to extreme values and depends on the estimates chosen, we also reported the median and the mean without the age mean. The term "efficiency" is used only in its statistical meaning (comparison of estimation errors); the relationship between accuracy, budget, and field cost was not estimated.

3.5 Data analysis

The analysis took the form of a factorial experiment with three factors: design (four levels), estimating (five), and confidence level (three). This results in 20 point estimators and 60 intervals, each evaluated on 1,000 replicates. Normal critical values were fixed in the code and standard errors were calculated as described in section 3.3. As validity controls, we found that all four estimators were virtually unbiased (the highest relative bias was 0.197% in the proportion of men under MAS) and that MAS reproduced coverages close to nominal coverages (15). We also estimated the intraclass correlation of age, weight, and sex with the analysis of variance estimator, to contrast equation (9) with the

observed design effects. By running the code again with the same seed, we obtained values identical to those reported in the tables. The figures were made with Matplotlib.

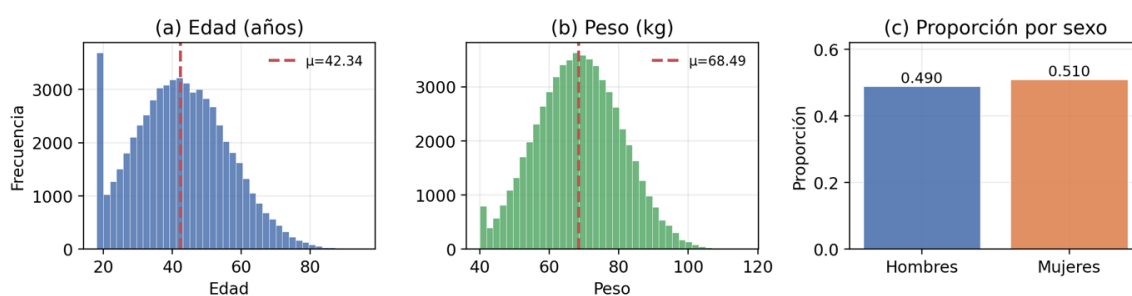
RESULTS

4.1 Population and True Parameters

The simulated population had a mean age of 42.34 years, with a standard deviation of 13.75 years; the mean weight was 68.49 kg, with a standard deviation of 12.14 kg, and 49.0% were men (Figure 1). The weight distribution is almost symmetrical and includes the difference between sexes incorporated in the model. The age distribution is unimodal, with a tail towards older ages and an isolated bar at 18 years that corresponds to truncation.

Figure 1.

Age (a) and weight distribution (b) and sex ratio (c) in the synthetic population (N = 1 000 000). The dashed line indicates the real mean



Note: Own elaboration based on the simulation (seed 20260627).

4.2 Point estimation and intervals in an illustrative sample

Table 2 lists the estimates and intervals of a single sample per design for the mean and standard deviation of weight and for the proportion of males. Its purpose is only to illustrate how the estimates are derived and read: the result of an isolated sample depends on the randomness of that extraction and is not used to compare designs, so efficiency comparisons are based on the 1,000 replicates summarized in Tables 3 to 6. In this sample, the intervals widen with the confidence level, as appropriate, and the cluster intervals are the widest. At 95%, the mean weight is 1.55 kg (from 67.48 to 69.03) versus 1.06 kg with MAS. The 36 intervals in the table contain the true value.

Table 2.

Point estimation and confidence intervals in an illustrative sample (n = 2 000) by sampling design

Method	Parameter	Est. puntual	90% CI	95% CI	99% CI
MORE	Average weight (kg)	68,56	[68,11; 69,01]	[68,03; 69,09]	[67,86; 69,26]
MORE	Desv. Typical weight (kg)	12,20	[11,88; 12,52]	[11,82; 12,58]	[11,70; 12,69]
BUT	Male Ratio	0,492	[0,474; 0,511]	[0,471; 0,514]	[0,464; 0,521]

Method	Parameter	Est. puntual	90% CI	95% CI	99% CI
Stratified	Average weight (kg)	68,65	[68,21; 69,08]	[68,13; 69,16]	[67,96; 69,33]
Stratified	Desv. Typical weight (kg)	12,45	[12,12; 12,77]	[12,06; 12,83]	[11,94; 12,95]
Stratified	Male Ratio	0,499	[0,481; 0,517]	[0,477; 0,521]	[0,470; 0,528]
By clusters	Average weight (kg)	68,25	[67,60; 68,90]	[67,48; 69,03]	[67,24; 69,27]
By clusters	Desv. Typical weight (kg)	12,27	[11,96; 12,59]	[11,89; 12,66]	[11,77; 12,77]
By clusters	Male Ratio	0,480	[0,459; 0,500]	[0,455; 0,504]	[0,447; 0,512]
Systematic	Average weight (kg)	68,49	[68,04; 68,93]	[67,96; 69,02]	[67,79; 69,18]
Systematic	Desv. Typical weight (kg)	12,12	[11,80; 12,43]	[11,74; 12,49]	[11,62; 12,61]
Systematic	Male Ratio	0,481	[0,463; 0,500]	[0,460; 0,503]	[0,453; 0,510]

Note: results from a single sample per design, for illustrative purposes; they should not be used to compare the efficiency of the designs (see Tables 3 to 6). Actual values: mean weight = 68.49 kg; standard deviation weight = 12.14 kg; proportion of men = 0.490. CI: confidence interval. Own elaboration based on the simulation.

4.3 Bias and efficiency in point estimation

Table 3 summarizes the point estimate over 1 000 replicates. No design showed appreciable bias. The highest relative bias was 0.197 %, in the proportion of men with MAS: the unrounded bias is +0.000965 over $P = 0.48998$, although Table 3 shows it rounded to +0.0010, which would give 0.204 % if calculated with the rounded figures. In the remaining 19 cases, the relative bias did not exceed 0.08 %, so the differences in performance come from variance. Efficiency, on the other hand, varied greatly depending on the parameter. In the mean age, which is the variable that defines the strata, the stratified reached a relative efficiency of 10.37 and reduced the RECM from 0.313 years with MAS to 0.097, with $Deff = 0.10$. The gain fell to 2.94 in the standard deviation of age and almost disappeared in weight (1.04 and 1.10) and in the proportion of men (1.00). The reason is that age strata explain 89.7% of the variance in age, but only 10.1% of that of weight and practically none of that of sex (3.4).

The systematic followed the same pattern with less intensity, thanks to the order of the frame, with an efficiency of 2.01 and a $Deff$ of 0.50 in the mean age. The clusters were at the opposite extreme, with $Deff$ of 2.85 in the mean weight and 1.94 in the mean of age, and efficiencies of 0.35 and 0.52 (13). In Figure 2, the distributions of stratified and systematic are concentrated around the actual age, while the cluster distribution is the most dispersed in both variables. Figure 3 shows the relative efficiency of all the designs and parameters.

Table 3.

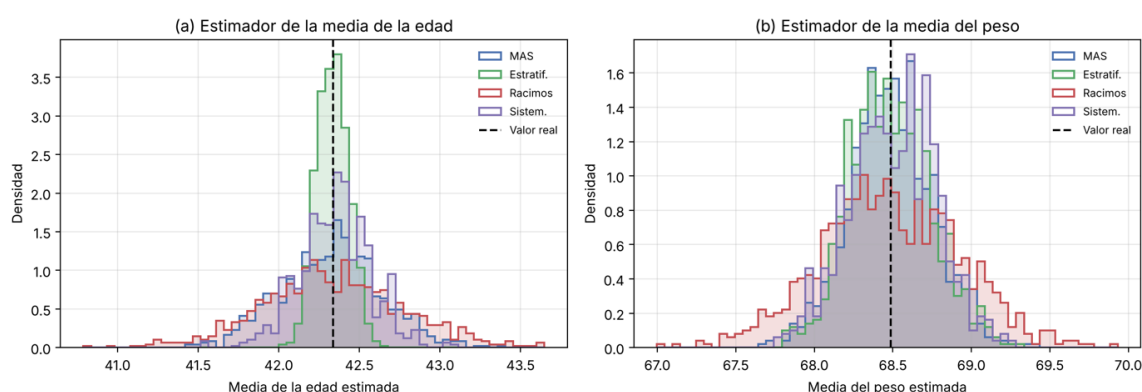
Performance of point estimation over 1,000 replicates: bias, relative bias, root mean square error (RECM), relative efficiency with respect to MAS, and design effect

Method	Estimating	Bias	Rel. bias (%)	RECM	Ef. relative	Deff
BUT	Mid-age	-0,0079	-0,019	0,3132	1,00	1,00
BUT	DT age	+0,0025	+0,018	0,2077	1,00	1,00
BUT	Medium weight	+0,0061	+0,009	0,2611	1,00	1,00
BUT	DT weight	-0,0059	-0,048	0,1862	1,00	1,00
BUT	Men's Prop.	+0,0010	+0,197	0,0110	1,00	1,00
Stratified	Mid-age	-0,0021	-0,005	0,0972	10,37	0,10
Stratified	DT age	+0,0033	+0,024	0,1211	2,94	0,34
Stratified	Medium weight	-0,0001	0,000	0,2562	1,04	0,96
Stratified	DT weight	+0,0030	+0,024	0,1776	1,10	0,91
Stratified	Men's Prop.	-0,0001	-0,027	0,0110	1,00	1,01
By clusters	Mid-age	-0,0024	-0,006	0,4361	0,52	1,94
By clusters	DT age	-0,0110	-0,080	0,2012	1,07	0,94
By clusters	Medium weight	-0,0050	-0,007	0,4410	0,35	2,85
By clusters	DT weight	-0,0021	-0,017	0,1873	0,99	1,01
By clusters	Men's Prop.	-0,0003	-0,053	0,0125	0,77	1,31
Systematic	Mid-age	+0,0141	+0,033	0,2212	2,01	0,50
Systematic	DT age	+0,0051	+0,037	0,1666	1,55	0,64
Systematic	Medium weight	+0,0097	+0,014	0,2770	0,89	1,12
Systematic	DT weight	-0,0011	-0,009	0,1722	1,17	0,86
Systematic	Men's Prop.	-0,0002	-0,045	0,0114	0,93	1,08

Note: the relative bias is $100 \cdot \text{bias} / \theta$, calculated with unrounded values. The relative efficiency is $\text{ECMMAS} / \text{ECM}$; values greater than 1 indicate greater efficiency than simple random sampling. The design effect (Deff) is the ratio of variances to MAS; values less than 1 indicate a gain in precision. Own elaboration based on the simulation.

Figure 2.

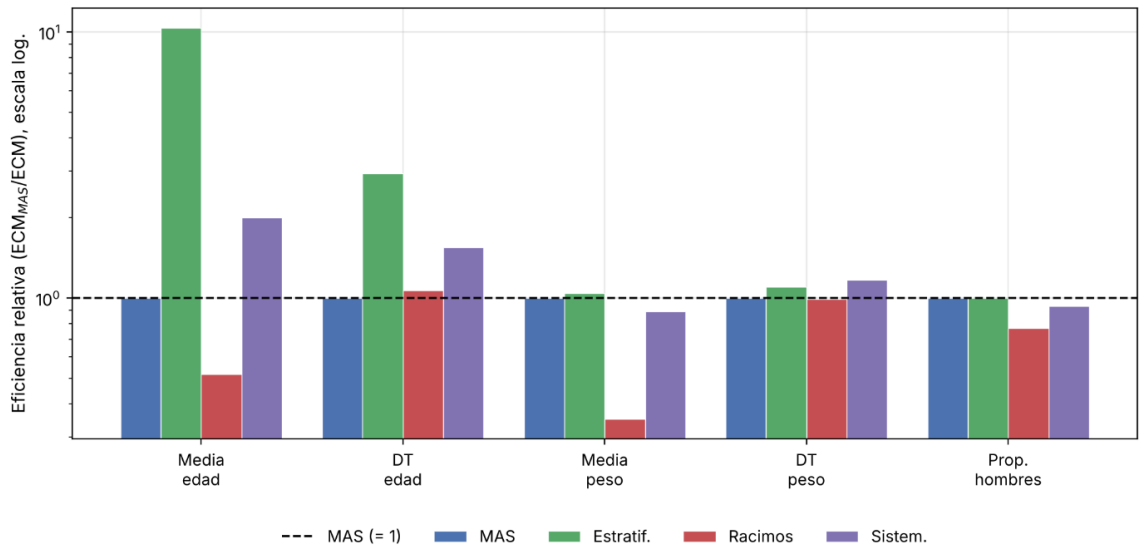
Sample distributions of the mean age (a) and weight estimator (b) under the four designs, over 1 000 replicates. The dashed line marks the real value



Note: Own elaboration based on the simulation (seed 20260627).

Figure 3.

Relative efficiency (ECMMAS/ECM, logarithmic scale) by design and parameter. The dashed line (= 1) corresponds to simple random sampling



Note: Own elaboration based on the simulation (seed 20260627).

4.4 Coverage and Breadth of Confidence Intervals

Table 4 assesses the validity and accuracy of the intervals. With MAS, the empirical coverages were between 89.0% and 91.7% at the 90% level, between 95.1% and 96.6% at 95%, and between 98.8% and 99.4% at 99%, confirming that the interval construction procedure works (Figure 4). The clusters also covered well, between 94.7% and 97.1% at 95%, a sign that the standard error calculated with cluster variability captures the loss of precision. The cost appears in the amplitude, as the interval at 95% of the mean weight averaged 1.72 kg, compared to 1.06 kg with MAS (Figure 5). With the observed Deff of 2.85, matching the accuracy of the MAS would require almost tripling the sample size (13).

The stratified and systematic were conservative just where they are most efficient. The coverage at 95% of the standard deviation of age reached 99.9% with the stratified, and that of the mean of age, at 99.1% with the systematic. These coverages are excessive and not a virtue of design: the mean standard error overestimated the real dispersion of the estimator by 80% and 39%, respectively (the ratio between the mean standard error and the empirical standard deviation of the estimator of 1.80 and 1.39), so that the intervals were wider than necessary and hide the gain in precision. In stratification, the reason is that the formula of equation (12) ignores stratification; in systematic, that the MAS formula does not reflect the implicit stratification of the ordered framework (23,29). Therefore, in those cases, the amplitude of the intervals in Table 4 does not measure the accuracy of the design, which should be read in the RECM in Table 3. Section 4.6 shows how coverage changes with standard errors consistent with the design.

Table 4.

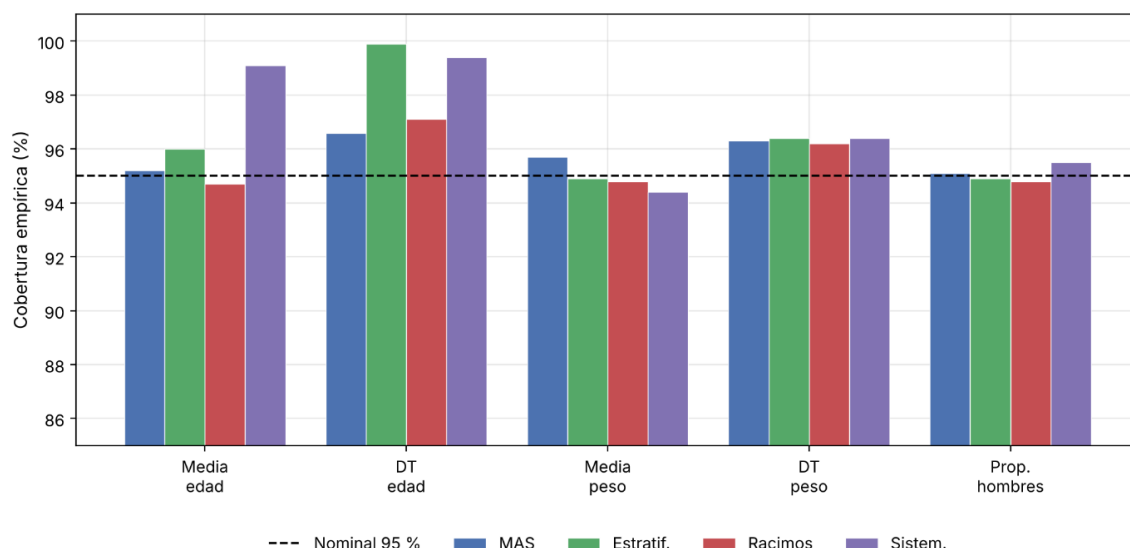
Empirical coverage of confidence intervals (90%, 95%, and 99%) and mean amplitude of the interval at 95%, by design and estimating

Method	Estimating	Cob. 90 %	Cob. 95 %	Cob. 99 %	95% CI amplitude
MORE	Mid-age	89,0	95,2	98,8	1,2039
BUT	DT age	91,7	96,6	99,4	0,8523
BUT	Medium weight	90,4	95,7	99,1	1,0626
BUT	DT weight	91,3	96,3	99,4	0,7523
BUT	Men's Prop.	90,4	95,1	99,0	0,0438
Stratified	Mid-age	91,4	96,0	99,4	0,3871
Stratified	DT age	99,5	99,9	100,0	0,8524
Stratified	Medium weight	90,8	94,9	99,4	1,0075
Stratified	DT weight	92,5	96,4	99,2	0,7529
Stratified	Men's Prop.	90,3	94,9	98,8	0,0437
By clusters	Mid-age	89,3	94,7	98,1	1,7165
By clusters	DT age	92,4	97,1	99,2	0,8515
By clusters	Medium weight	89,9	94,8	99,0	1,7151
By clusters	DT weight	90,4	96,2	99,1	0,7526
By clusters	Men's Prop.	89,7	94,8	99,0	0,0486
Systematic	Mid-age	98,3	99,1	99,6	1,2041
Systematic	DT age	96,1	99,4	99,8	0,8525
Systematic	Medium weight	89,6	94,4	99,3	1,0630
Systematic	DT weight	93,5	96,4	98,8	0,7526
Systematic	Men's Prop.	88,6	95,5	99,1	0,0438

Note: coverage in percentage, calculated with the main standard error (EE1); the amplitude is expressed in the units of each parameter (years, kg or proportion). The coverage of 99.9 % (stratified, standard deviation of age) and 99.1 % (systematic, mean of age) reflect overestimated standard errors and not greater validity; see Table 6. Prepared by the author based on the simulation.

Figure 4.

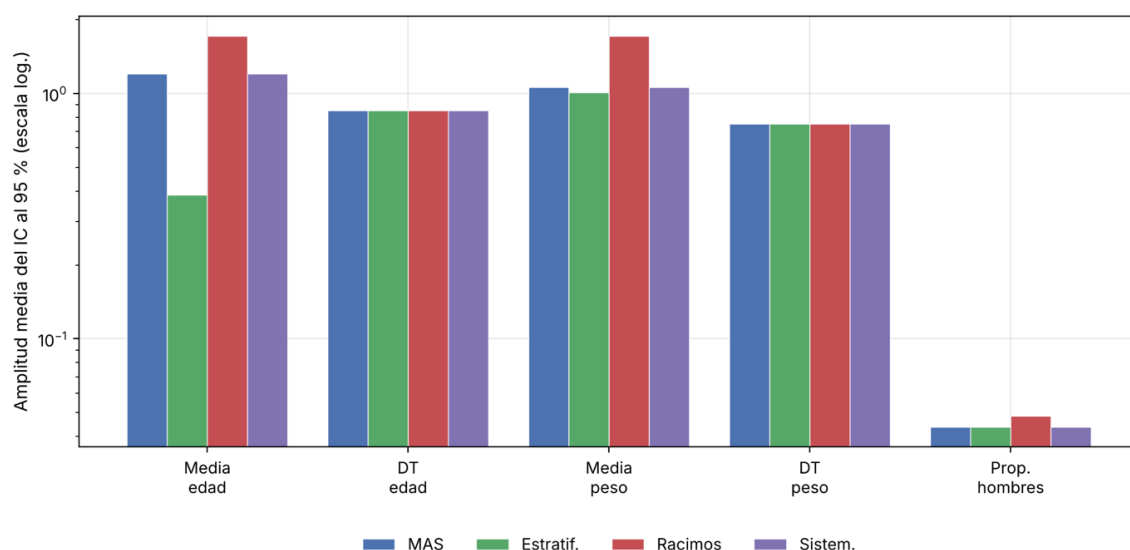
Empirical coverage of 95% confidence intervals by design and parameter. Dashed line marks 95% nominal level



Note: Own elaboration based on the simulation (seed 20260627).

Figure 5.

Average amplitude of 95% confidence intervals by design and parameter (logarithmic scale). With equal coverage, a narrower range indicates greater accuracy



Note: Own elaboration based on the simulation (seed 20260627).

4.5 Effect of overall design and ordering of efficiency

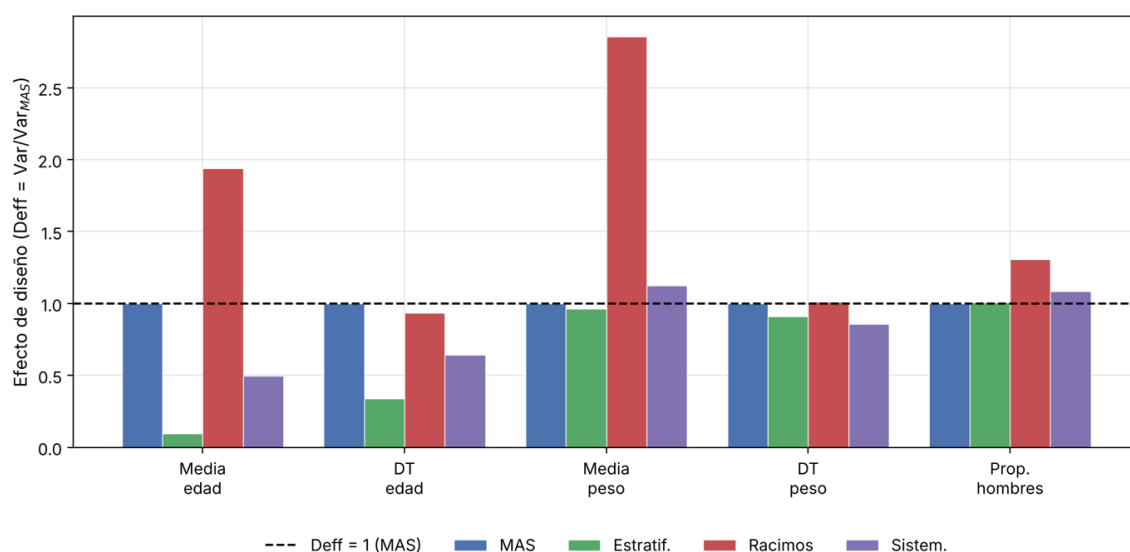
Figure 6 shows the design effects: clearly less than 1 for stratified and systematic in the age parameters, close to 1 in most other cases, and greater than 1 for clusters in the two means and in proportion (1.94, 2.85, and 1.31). By averaging the relative efficiency of the five estimands (Table 5), the stratified one ranks first (3.29), followed by the systematic (1.31), the MAS (1.00), and the clusters (0.74). This order should be read with caution, because it depends on the five estimates chosen, and the average of the stratified is dominated by the efficiency of 10.37 in the age mean. Without this estimate, its mean efficiency drops to 1.52, and its median is 1.10, close to that of the systematic

<https://doi.org/10.71068/SIMJ0005>

(1.14 and 1.17). The ordering does not express, therefore, a general superiority of the stratified, but its advantage in the parameters associated with the stratification variable, and could change with other estimates or with another stratification variable. The average should also be read together with the coverage: with EE1, the advantage of the systematic is accompanied by the largest mean deviation of the coverage, 2.20 percentage points, which is reduced to 0.50 with EE2 (section 4.6).

Figure 6.

Design effect ($Deff = Var/Var_{MAS}$) by design and parameter. Values below the line (= 1) indicate precision gain over simple random sampling



Note: Own elaboration based on the simulation (seed 20260627).

Table 5.

Overall ranking of efficiency of the four designs

Position	Method	Eph. Relative Mean	Median of Ef. relative	Ef. relative mean without mean of age	Desv. Mean Cob. 95% with EE1 (pp)	Desv. Mean cob. 95 % with EE2 (pp)
1	Stratified	3,29	1,10	1,52	1,50	0,56
2	Systematic	1,31	1,17	1,14	2,20	0,50
3	MORE	1,00	1,00	1,00	0,78	0,58
4	By clusters	0,74	0,77	0,79	0,80	0,24

Note: Relative efficiency averages or summarizes ECMMAS/ECM over the five estimates; the column "no age mean" excludes this estimate. The mean deviation of coverage is the average of $|empirical\ coverage - 95|$ in percentage points (pp), with the main standard error (EE1) and with the alternative standard error (EE2). Prepared by the author based on the simulation. All indicators summarize statistical efficiency at the same n ; they do not incorporate operating costs.

4.6 Coverage sensitivity to standard error

<https://doi.org/10.71068/SIMJ0005>

Table 6 compares the 95% coverage obtained with EE1 and EE2 in cases where the main standard error is not design-dependent or approximate: the standard deviations in the four designs and the means and proportion in the systematic. With linearization, the coverage of the standard deviation of age in the stratified decreased from 99.9% to 93.6% and the mean amplitude of the interval from 0.852 to 0.453 years; with successive differences, the mean age in the systematic went from 99.1% to 95.4% and the amplitude from 1.204 to 0.857 years. With EE2, the ratio between the mean standard error and the empirical standard deviation was between 0.95 and 1.04 in all cases, and the mean deviation of coverage from 95% decreased from 1.50 to 0.56 points in the stratified and from 2.20 to 0.50 in the systematic (Table 5). With a standard error consistent with the design, the intervals of both designs reflect their gain in precision.

Linearization also corrected for the slight overcoverage of the standard deviation with MAS (from 96.6 % to 93.9 % at age), which is explained by the fact that age, truncated, has a lower kurtosis than normal (excess of -0.34) and equation (12) then overestimates the error (34). At $n = 2\,000$, however, linearization was somewhat below the nominal level in MAS and stratified (between 93.6 % and 94.8 %), so it should not be taken as a definitive solution: replication methods, such as jackknife or bootstrap, could better calibrate these intervals (29).

Table 6.

Sensitivity analysis: 95% coverage and mean range amplitude with primary standard error (EE1) and alternate standard error (EE2)

Method	Estimating	EE/DT (EE1)	Cob. 95 % (EE1)	Amplitude (EE1)	EE/DT (EE2)	Cob. 95 % (EE2)	Amplitude (EE2)
BUT	DT age	1,05	96,6	0,8523	0,95	93,9	0,7745
BUT	DT weight	1,03	96,3	0,7523	0,95	94,2	0,6967
Stratified	DT age	1,80	99,9	0,8524	0,95	93,6	0,4526
Stratified	DT weight	1,08	96,4	0,7529	1,00	94,8	0,6944
By clusters	DT age	1,08	97,1	0,8515	1,01	95,1	0,7922
By clusters	DT weight	1,02	96,2	0,7526	1,00	94,6	0,7335
Systematic	DT age	1,31	99,4	0,8525	1,04	95,5	0,6790
Systematic	DT weight	1,11	96,4	0,7526	1,03	95,6	0,6961
Systematic	Mid-age	1,39	99,1	1,2041	0,99	95,4	0,8566
Systematic	Medium weight	0,98	94,4	1,0630	0,95	94,0	1,0333
Systematic	Men's Prop.	0,98	95,5	0,0438	0,98	95,0	0,0438

Note: EE/SD is the quotient between the mean standard error and the empirical standard deviation of the estimator in the 1 000 replicates (1 = standard error without bias). EE1: main standard error

(equation 12 for standard deviation; MAS formula in systematic). EE2: linearization (equation 13) for standard deviation and successive differences (equation 14) in systematic. Coverage in percentage; amplitude in parameter units. Own elaboration based on the simulation.

DISCUSSION

In a population that is fully known, the results reproduce a central idea of sampling theory: no design is superior in all parameters, and efficiency depends on how the design relates to what is estimated (3,4). The stratified obtained the highest average efficiency, but almost all of its advantage was in age, the variable that defines the strata; for weight and proportion of men, it behaved like the MAS. This is anticipated by the Neyman assignment (1) and the reference texts (5,6,9), according to which the reduction in variance depends on the fraction of the variability explained by the strata: here, 89.7% in age, 10.1% in weight, and practically nothing in sex.

The practical implication is that stratifying by a variable weakly related to the parameter of interest does not impair proportional allocation accuracy, but neither does it improve it, as shown by the efficiencies of weight (1.04) and the proportion of men (1.00). In a multi-objective survey, the stratification variable should be chosen based on the priority parameter or combine several auxiliary variables; if strata are defined by only one variable outside that parameter, the design adds operational complexity without benefit, and an optimal allocation for one variable may be inefficient for the others (24,25). When the relationship to auxiliary information is weak or non-linear, model-assisted estimators and calibration, which use that information in the estimation phase, offer a complementary pathway (36,37).

With the clusters, we were able to contrast Kish's approach with what was observed. The estimated intraclass correlation in the population was 0.044 for age, 0.067 for weight, and 0.010 for sex; with $M = 25$, equation (9) predicts design effects of 2.04, 2.62, and 1.24, close to those observed (1.94, 2.85, and 1.31). The result coincides with the experience in real surveys, where the effect of cluster design is rarely negligible (13,28). The lower statistical efficiency does not necessarily imply that the design should be discarded: its practical appeal may lie in the territorial concentration of the observations and the availability of area frames, operational aspects that this experiment did not quantify. Therefore, no result supports the conclusion that clusters are more or less economically efficient. Research on spatially balanced sampling shows methodological alternatives to reduce variability under certain conditions (38-40). The validity of the intervals was maintained because we used standard errors based on cluster design; with the MAS formula, intraclass dependence would be ignored (5,14).

The systematic study showed both sides of the implicit stratification. It was more efficient than the MAS in the parameters of age (2.01 in the mean and 1.55 in the standard deviation), but its variance cannot be estimated without bias with a single sample, and the MAS formula we used in the main analysis is a conservative approach (10,11,29). The successive differences estimator corrected for overcoverage (Table 6), in line with the studies that propose approximate estimators for ordered frames (23). It is advisable to use this design with frames whose order is known and, if possible, with several random starts or with an estimator of successive differences. In the proportion, the Wald

interval worked well with $n = 2\,000$ and $P = 0.49$ (95% coverage between 94.8% and 95.5%), although the literature documents its irregular coverage in small samples or with extreme proportions, where the Wilson or Agresti and Coull intervals (30-33) are appropriate.

We limit ourselves to designs with equal probabilities. The theory offers alternatives that take advantage of auxiliary information: sampling with unequal probabilities and the Horvitz and Thompson estimator (2,9); ratio estimators, regression estimators, and others that incorporate auxiliary variables (19,20); calibration to known totals (37); estimators assisted by machine learning models (36); balanced and spatially dispersed selection algorithms (39,40), and adaptive cluster sampling for rare or aggregated populations (41). Ordered set sampling has also shown gains over MAS when a cheap ordering variable is available, even in multistage designs (21,22). Our comparison serves as a baseline for these designs, and their qualitative pattern coincides with what is stated in the reference texts (5,6,14).

The study has limitations that restrict the generalization of the magnitudes obtained. First, the conclusions come from a single synthetic population, constructed under specific assumptions (section 3.1), and not from real surveys. The relative efficiency and design effect figures depend on this structure, especially on the intraclass correlation and the association of each variable with the stratification or ordering criteria; another population would give other figures, although the qualitative pattern is the one anticipated by the theory (3). Second, we use a single sample size ($n = 2\,000$), 1 000 replicates per design, and a single correlation structure between age, weight, and sex. Third, we do not model non-sampling errors (non-response, measurement, frame coverage), which in real surveys can dominate total error (42); non-response, for example, alters the relative performance of the designs (22). In addition, the comparison kept the number of units observed constant and not the total cost of applying each design, so the results do not identify an economically superior design. The favorable effect of stratification is conditioned by the role of age in the construction of strata and should not be extrapolated to other variables or populations.

We compared only single-stage designs with equal probabilities, without the multistage sampling, proportional probability sampling and model-assisted estimators, which are common in practice (4,9). We stratified by a single variable with proportional allocation; Neyman assignment or multivariate stratification could extend the observed advantage (1,24). In the main analysis, the standard deviation interval was based on a normal approximation of its standard error that does not depend on the design, and that of the systematic error on the MAS formula. The sensitivity analysis shows that linearization and successive differences correct for overcoverage, although linearization is somewhat below the nominal level and other methods could better calibrate the intervals (29). Finally, we do not yet know whether the ordering is maintained with small samples, with other cluster sizes or with other correlations, and this analysis is pending.

CONCLUSIONS

On a synthetic population of 1,000,000 people, fully known and generated under specific assumptions, the four designs gave practically unbiased estimators, so that the choice between them depends on the precision of the estimator and the validity of the intervals. No design was superior in

all parameters. Stratification achieved the greatest gains in age, the variable used to form the strata; in weight and in the proportion of men, its advantage practically disappeared, and its average relative efficiency of 3.29 depends on the estimates chosen. The MAS maintained coverages close to the nominal ones and the systematic improved the accuracy of age when the frame was ordered with an informative auxiliary variable, although the MAS variance formula was conservative. The clusters were the least statistically efficient at the same sample size, with design effects of up to 2.85. The coverage of the intervals depended on the standard error estimator, and the design-consistent alternatives attenuated the observed overcoverages.

These conclusions, obtained through a single simulation configuration, can guide statistical planning, but they do not replace the analysis of real surveys. It is advisable to choose the design according to the priority parameter, the available auxiliary information, and the stratification or ordering assumptions. Stratification is recommended when there is an informational auxiliary variable with respect to the parameter and an appropriate assignment is applied; here its advantage was strongly linked to age. The systematic can be useful with informative frameworks, provided that its uncertainty is appropriately estimated. Clusters need to evaluate their design effect and could respond to logistical needs, but this work does not estimate costs or identify economic superiority. To extend the comparison, it will be necessary to analyze other populations, sample sizes, correlation structures, multistage designs, unequal probabilities, costs, non-response, measurement errors, and alternative methods of interval construction.

BIBLIOGRAPHIC REFERENCES

1. Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*. 1934; 97(4):558-625. <https://doi.org/10.2307/2342192>.
2. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47(260):663-685. <https://doi.org/10.1080/01621459.1952.10483446>.
3. Cochran WG. *Sampling techniques*. 3rd ed. New York: John Wiley & Sons; 1977.
4. Särndal CE, Swensson B, Wretman J. *Model assisted survey sampling*. New York: Springer-Verlag; 1992. <https://doi.org/10.1007/978-1-4612-4378-6>
5. Lohr SL. *Sampling: design and analysis*. 3rd ed. Boca Raton: Chapman and Hall/CRC; 2022. <https://doi.org/10.1201/9780429298899>
6. Kim JK. *Statistics in survey sampling*. Boca Raton: Chapman and Hall/CRC; 2025. <https://doi.org/10.1201/9781003606147>
7. Kalton G. Probability vs. nonprobability sampling: from the birth of survey sampling to the present day. *Statistics in Transition New Series*. 2023; 24(3):1-22. <https://doi.org/10.59170/stattrans-2023-029>.
8. Wu C. Statistical inference with non-probability survey samples. *Survey Methodology*. 2022; 48(2):283-311. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2022002/article/00002-eng.htm>.
9. Tillé Y. *Sampling and estimation from finite populations*. Hoboken: John Wiley & Sons; 2020. <https://doi.org/10.1002/9781119071259>
<https://doi.org/10.71068/SIMJ0005>

10. Madow WG, Madow LH. On the theory of systematic sampling, I. The Annals of Mathematical Statistics. 1944; 15(1):1-24. <https://doi.org/10.1214/aoms/1177731312>.
11. Iachan R. Systematic sampling: a critical review. International Statistical Review. 1982; 50(3):293-303. <https://doi.org/10.2307/1402499>.
12. Azeem M, Hussain S, Ijaz M, Salahuddin N, Salam A. An improved version of systematic sampling design for use with linear trend data. Heliyon. 2023; 9(6):e17121. <https://doi.org/10.1016/j.heliyon.2023.e17121>.
13. Kish L. Survey sampling. New York: John Wiley & Sons; 1965.
14. Valliant R, Dever JA, Kreuter F. Practical tools for designing and weighting survey samples. 2nd ed. Cham: Springer; 2018. <https://doi.org/10.1007/978-3-319-93632-1>
15. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Statistics in Medicine. 2019; 38(11):2074-2102. <https://doi.org/10.1002/sim.8086>.
16. Chapuis K, Taillandier P, Drogoul A. Generation of synthetic populations in social simulations: a review of methods and practices. Journal of Artificial Societies and Social Simulation. 2022; 25(2):6. <https://doi.org/10.18564/jasss.4762>.
17. Siepe BS, Bartoš F, Morris TP, Boulesteix AL, Heck DW, Pawel S. Simulation studies for methodological research in psychology: a standardized template for planning, preregistration, and reporting. Psychological Methods. 2024. Publicación anticipada en línea. <https://doi.org/10.1037/met0000695>
18. Kondratev M, Corrao MV, Armstrong R, Smith AMS. Assessing the uncertainty of traditional sample-based forest inventories in mixed and single species conifer systems using a digital forest twin. Forests. 2025; 16(11):1617. <https://doi.org/10.3390/f16111617>.
19. Daraz U, Wu J, Agustiana D, Emam W. Finite population variance estimation using Monte Carlo simulation and real life application. Symmetry. 2025; 17(1):84. <https://doi.org/10.3390/sym17010084>.
20. Alghamdi AS, Almulhim FA. Optimizing finite population mean estimation using simulation and empirical data. Mathematics. 2025; 13(10):1635. <https://doi.org/10.3390/math13101635>.
21. Abbas M, Shehzad MA, Rabia M, Khurram H, Ijaz M. Estimation of finite population mean in a complex survey sampling. PLoS ONE. 2025; 20(5):e0324559. <https://doi.org/10.1371/journal.pone.0324559>.
22. Rehman SA, Al-Essa LA, Shabbir J, Khan Z. On the efficiency of paired ranked set sampling for estimating the population mean in the presence of non-response. Scientific Reports. 2024;14:25334. <https://doi.org/10.1038/s41598-024-76072-y>.
23. Franceschi S, Fattorini L, Gregoire TG. Exploiting nearest-neighbour maps for estimating the variance of sample mean in equal-probability systematic sampling of spatial populations. Spatial Statistics. 2024;64:100865. <https://doi.org/10.1016/j.spasta.2024.100865>.
24. Wesołowski J, Wieczorkowski R, Wójciak W. Recursive Neyman algorithm for optimum sample allocation under box constraints on sample sizes in strata. Survey Methodology. 2024; 50(2):487-511. <https://www150.statcan.gc.ca/n1/pub/12-001-x/2024002/article/00003-eng.htm>.
25. Sauer S, Hedt-Gauthier B, Haneuse S. Practical strategies for operationalizing optimal allocation in stratified cluster-based outcome-dependent sampling designs. Statistics in Medicine. 2023; 42(7):917-935. <https://doi.org/10.1002/sim.9650>.

26. Lohmann A, Astivia OLO, Morris TP, Groenwold RHH. It's time! Ten reasons to start replicating simulation studies. *Frontiers in Epidemiology*. 2022;2:973470. <https://doi.org/10.3389/fepid.2022.973470>.
27. Pawel S, Kook L, Reeve K. Pitfalls and potentials in simulation studies: questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*. 2024; 66(1):2200091. <https://doi.org/10.1002/bimj.202200091>.
28. Rhoda DA, Soofi SB, Hussain I, Clary CB, Yu M, Khanyile S, et al. Sample size parameters for vaccination coverage cluster surveys: insights from Pakistan. *Journal of Survey Statistics and Methodology*. 2025; 13(5):587-614. <https://doi.org/10.1093/jssam/smaf033>.
29. Wolter KM. Introduction to variance estimation. 2nd ed. New York: Springer; 2007. <https://doi.org/10.1007/978-0-387-35099-8>
30. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*. 1927; 22(158):209-212. <https://doi.org/10.1080/01621459.1927.10502953>.
31. Agresti A, Coull BA. Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*. 1998; 52(2):119-126. <https://doi.org/10.1080/00031305.1998.10480550>.
32. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science*. 2001; 16(2):101-133. <https://doi.org/10.1214/ss/1009213286>.
33. Andersson PG. The Wald confidence interval for a binomial p as an illuminating “bad” example. *The American Statistician*. 2023; 77(4):443-448. <https://doi.org/10.1080/00031305.2023.2183257>.
34. Curto J. Confidence intervals for means and variances of nonnormal distributions. *Communications in Statistics - Simulation and Computation*. 2023; 52(9):4414-4430. <https://doi.org/10.1080/03610918.2021.1963448>.
35. Bhushan S, Pandey S. Optimality of novel estimators for variance estimation under successive sampling. *Communications in Statistics - Theory and Methods*. 2025; 54(14):4635-4653. <https://doi.org/10.1080/03610926.2024.2425746>.
36. Dagdou M, Goga C, Haziza D. Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*. 2023; 118(542):1234-1251. <https://doi.org/10.1080/01621459.2021.1987250>.
37. Deville JC, Särndal CE. Calibration estimators in survey sampling. *Journal of the American Statistical Association*. 1992; 87(418):376-382. <https://doi.org/10.1080/01621459.1992.10475217>.
38. Heikkinen J, Henttonen HM, Katila M, Tuominen S. Stratified, spatially balanced cluster sampling for cost-efficient environmental surveys. *Environmetrics*. 2025; 36(5):e70019. <https://doi.org/10.1002/env.70019>.
39. Prentius W, Grafström A. How to find the best sampling design: a new measure of spatial balance. *Environmetrics*. 2024; 35(7):e2878. <https://doi.org/10.1002/env.2878>.
40. Jauslin R, Panahbehagh B, Tillé Y. Sequential spatially balanced sampling. *Environmetrics*. 2022; 33(8):e2776. <https://doi.org/10.1002/env.2776>.

41. Yasmeeen U, Noor-ul-Amin M, Hanif M. Variance estimation in stratified adaptive cluster sampling. *Statistics in Transition New Series*. 2022; 23(1):173-184.
<https://doi.org/10.2478/stattrans-2022-0010>.
42. Groves RM, Fowler FJ, Couper MP, Lepkowski JM, Singer E, Tourangeau R. *Survey methodology*. 2nd ed. Hoboken: John Wiley & Sons; 2009.

DATA AND CODE AVAILABILITY

Primary data from participants were not used. Two independently verified Python scripts are attached: S1 (Suplementario_S1_simulacion_muestreo.py) implements the population generation scheme, sample selection, and performance measures described in the method; S2 (Suplementario_S2_figuras.py) generates auxiliary graphs from its results. Both files are executable and document their parameters. They do not replace the original scripts with which the tables and figures of the manuscript were produced, which were not available for this review; therefore, the exact numerical reproduction of the published results is still pending verification by the author.

ETHICAL CONSIDERATIONS

The study was developed through a literature review based exclusively on previously published documentary sources. It did not involve the direct participation of human or animal beings, the performance of interventions, the use of biological samples, or access to personally identifiable data. Consequently, it did not require approval from an ethics committee or the obtaining of informed consent. During the research process, the principles of scientific integrity, adequate attribution of authorship, rigorous citation, and responsible use of the sources consulted were observed.

ARTIFICIAL INTELLIGENCE (AI) USAGE STATEMENT

The authors state that the generative AI tool was used exclusively to support the organization of the article and the translation of the text. Its use did not determine the scientific content, study design, source selection, data analysis, interpretation of results, or conclusions. Full responsibility for the content remains with the authors.

FUNDING

The authors state that this study did not receive any external funding from public, private, or non-profit organizations. All research, analysis, and development activities were carried out with their own resources.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest related to this study and that all the procedures followed comply with the ethical standards established by the journal. They also confirm that this work is unpublished and has not been published, either partially or totally, in any other publication.

AUTHOR'S CONTRIBUTION

Authors' names and initials: Marco Antonio Ramos Espinal (MARE)

1. Conceptualization: MARE

<https://doi.org/10.71068/SIMJ0005>

- 2. Data Curation:** MARE
- 3. Formal analysis:** MARE
- 4. Acquisition of financing:** not applicable
- 5. Research:** MARE
- 6. Methodology:** MARE
- 7. Project Management:** MARE
- 8. Remedies:** Not applicable
- 9. Software:** Not applicable
- 10. Supervision:** MARE
- 11. Validation:** MARE
- 12. View:** SEA
- 13. Original Draft:** MOTHER
- 14. Writing, Proofreading, and Editing:** MARE