Submitted to the South African Statistical Journal, Vol. XX, No. X, 1–14

ESTIMATING THE PROPORTION OF URBAN RESIDENTS LIKELY TO ADOPT ELECTRIC VEHICLES (EVS) IN THE NEXT 5 YEARS.

Odey R. Mofokeng
School of Actuarial Science and Statistics, University of the Witwatersrand, Johannesburg, South Africa
e-mail: *2555438@students.wits.ac.za*

This study estimates electric vehicle (EV) adoption in a simulated population of 10 000 Green City residents, stratified by age (18–34, 35–54, 55+), targeting a margin of error (MOE) ≤ 0.05 and improved efficiency over simple random sampling (SRS). Using a stratified sample of 450, the design-adjusted estimate is 0.4024 (SE = 0.0234, 95% CI: 0.3574–0.4490), closely matching the true proportion of 0.3891. The Bayesian estimate, from a binomial logit model, is 0.4139 (SE = 0.0115, 95% CI: 0.3907–0.4361), offering higher precision. Across 200 simulation replications, the stratified design shows a bias of 0.0087, MSE of 0.0006, MOE of 0.0453, and coverage of 0.930, while the Bayesian approach achieves an MOE of 0.0225 and coverage of 0.950. The SRS estimate (0.4709, MOE = 0.0492) fails the precision target, with a bias of 0.0011, MSE of 0.0007, and coverage of 0.960. The stratified design's relative efficiency of 1.2375 indicates a 23.75% variance reduction over SRS, though the empirical DEFF (1.1385) exceeds the theoretical DEFF (0.7896). Adoption varies significantly by age (0.6225, 0.3183, 0.1148 for 18–34, 35–54, 55+) and income (0.2884 for Very Low to 0.6778 for Very High), with younger and wealthier groups leading. The design detects a 0.08 subgroup difference with 84% power, confirming sensitivity to demographic variations. Approximately 40% of residents may adopt EVs within five years, informing policies like subsidies for lower-income groups and charging infrastructure for older residents. Validation with real-world data and additional predictors like geography could enhance precision for Green City's sustainable transportation planning.

Key words: Electric vehicles, Stratified sampling, Urban residents.

1. Introduction

The global shift towards sustainable transportation has intensified due to climate change, air pollution, and energy security concerns. Electric vehicles (EVs) offer reduced emissions and lower long-term costs, yet adoption varies by income, age, and infrastructure access (Rahman and Thill, 2024; Pandita, Bhatt, Kumar, Fatma and Vapiwala, 2024). This mini-project simulates a sample survey to estimate EV adoption intentions among urban residents in Green City, a fictional city modeled after Johannesburg, South Africa, to inform policy, urban planning, and EV marketing strategies (Pandita et al., 2024). Using stratified sampling, the study aims to achieve precise and efficient estimates in a diverse population.

The research question is: What proportion of Green City's adult residents are likely to adopt an EV in the next five years, and how does this vary by demographic groups? The primary objective is to

MSC2010 subject classifications. 62GXX, 62GXX, 62GXX.

estimate this proportion (p), defined as the share of individuals "likely" or "very likely" to purchase or lease an EV (1) versus not (0), using a stratified sampling design on a population of 10 000. Secondary objectives include comparing this design's precision and efficiency to simpler methods and ensuring a margin of error ≤ 0.05 through simulations (DeYoreo, 2018). This proportion is critical for assessing public readiness for EVs, guiding infrastructure investments (e.g., charging stations) and incentives (e.g., tax rebates) to address barriers like cost or range anxiety (Pamidimukkala, Kermanshachi, Rosenberger and Hladik, 2024; Pei, Huang, Zhang, Wang and Ye, 2025).
The target population comprises 10 000 adult residents in Green City. It is a fictional urban area modeled after a major metropolitan region like Johannesburg, South Africa. The population is designed to reflect realistic urban demographics:

- **Age distribution**: 40% aged 18–34 (young adults, typically tech-savvy and environmentally conscious), 35% aged 35–54 (middle-aged, often with family and career commitments), and 25% aged 55+ (older adults, potentially more resistant to adopting new technologies like electric vehicles).
- **Gender**: 48% male, 50% female, and 2% non-binary, capturing a spectrum to reflect diverse gender identities and their potential influence on technology adoption.
- **Income levels**: Income follows a continuous log-normal distribution (mean = 10.8, SD =0.8), categorised into five groups: Very Low (< R30 000), Low (R30 000–R60 000), Middle (R60 000–R100 000), High (R100 000–R150 000), and Very High (> R150 000). Higher income groups are expected to have greater financial capacity for EV adoption.
- **Geography**: 40% urban (dense areas with better access to charging infrastructure), 35% suburban (moderately dense, reliant on personal vehicles), and 25% rural (sparse, with limited charging access). This reflects varied infrastructure availability.

The variable of interest—likelihood of EV adoption—is binary (1 for likely, 0 otherwise) and follows a binomial distribution. Baseline adoption probabilities are 60% for young (18–34), 30% for middle-aged (35–54), and 5% for older (55+), adjusted by income (scaled log-income effect of 0.15), gender (+3% non-binary, +1% female, -2% male), and geography (+8% urban, +2% suburban, -5% rural). These reflect higher EV interest among younger, wealthier, urban, or non-binary individuals, constrained between 0.01 and 0.99 for stability (Powell, Cezar and Rajagopal, 2022).
Stratified random sampling is used for its precision and efficiency over simple random sampling (Van Hoeven, Janssen, Roes and Koffijberg, 2015). It ensures representation, thereby reducing sampling error in populations with varied EV adoption rates. The design optimises costs (R25 young, R35 middle-aged, R50 older) within a R15 000 budget, which leads to minimal variance (Diaz Quijano, 2018). Age stratification lowers within-stratum variance to improve accuracy (Zhao, Klaassen, Lisovski and Klaassen, 2019). Using the `sampling` package in R, samples are drawn from a population frame (ID, age, income, gender, geography) with cost-constrained allocation, so as to target $n \approx 450$ for a margin of error ≤ 0.05. The survey accounts for missing-at-random nonresponse based on income, age, and geography, which leads to tighter confidence intervals than naive methods.

2. Literature review/ Review of Methods

Stratified sampling divides the population into H non-overlapping subgroups based on auxiliary variables, with independent sampling from each stratum (Kim, Oh, Park, Cho and Park, 2013). It enhances efficiency and precision over simple random sampling by leveraging stratum-specific characteristics (César and Carvalho, 2011). Sample size per stratum h for proportional allocation is $n_h = n \cdot \frac{N_h}{N}$, where n is the total sample size, N_h is the stratum size, and N is the total population. This ensures representativeness but may not account for varying stratum variances. Optimal allocation, which minimises variance, uses $n_h = n \cdot \frac{N_h \cdot \sigma_h}{\sum(N_k \cdot \sigma_k)}$, where $\sigma_h = \sqrt{p_h(1-p_h)}$ is the stratum standard deviation, adjusted to sum to n and satisfy $n_h \leq N_h$ (Triveni, Danish and Tawiah, 2024). This suits EV adoption surveys with age-based variance differences.

Sample size determination ensures precise estimates within a margin of error $E = \pm 0.05$. The stratified sample size balances precision and variability, informed by auxiliary data (Triveni et al., 2024). The general formula for a proportion is:

$$n = \frac{\sum_{h=1}^{H} \left(\frac{N_h^2 p_h (1-p_h)}{n_h} \right)}{\frac{E^2}{z^2} + \sum_{h=1}^{H} \left(\frac{N_h (N_h - n_h) p_h (1-p_h)}{N^2 (N_h - 1)} \right)}, \tag{2.1}$$

where p_h is the stratum proportion and z is the z-score for the confidence level. An approximation adjusts the SRS formula with the design effect (DEFF) (Chen and Lumley, 2022):

$$n = \frac{z^2 p(1-p)}{E^2} \times \text{DEFF}, \tag{2.2}$$

where DEFF, calculated as $1 - \frac{\sum w_h (p_h - \bar{p})^2}{(\bar{p}(1-\bar{p}))}$, is typically < 1 (e.g., ≈ 0.8 per Cao, Vilar, Vilar and López (2013)), reflecting reduced variance in stratified designs.
Stratified sampling yields a lower sample size ($n \approx 200$) than SRS ($n \approx 260$) due to its efficiency. Stratum-specific proportions p_h and variances are used with cost-constrained allocation to optimise n_h, minimising variance within a R15 000 budget (Lohr, 2021). In R, the `sampling` package's functions handle allocation, with finite population correction (fpc) for $N = 10000$. Variance of the stratified proportion estimator $\hat{p}_{st} = \sum_{h=1}^{H} w_h \hat{p}_h$, where $w_h = \frac{N_h}{N}$, is

$$Var(\hat{p}_{st}) = \sum_{h=1}^{H} w_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{p_h (1-p_h)}{n_h - 1}. \tag{2.3}$$

This accounts for fpc $(1 - \frac{n_h}{N_h})$ (Lohr, 2021). Jackknife or bootstrap methods enhance variance estimation in simulations, especially when theoretical assumptions falter (Cao et al., 2013). The `survey` package's `svydesign(id=~1, strata=~age_group, weights=~wt, data=sample_df, fpc=~fpc)` creates a survey object, with `svymean()` computing $\hat{p}_{st}$ and SE via Taylor linearisation (Cao et al., 2013). Empirical SEs from 200 simulation repetitions will validate theoretical SEs.

Confidence intervals (CIs) for $\hat{p}_{st}$ use Wilson score intervals for better coverage, particularly for small strata or extreme proportions (Franco, Little, Louis and Slud, 2019):

$$\hat{p}_{st} + \frac{z^2}{2n} \pm z \frac{\sqrt{\frac{\hat{p}_{st}(1-\hat{p}_{st}) + \frac{z^2}{4n}}{n}}}{\left(1 + \frac{z^2}{n}\right)}. \tag{2.4}$$

Bayesian intervals with Beta quantiles improve coverage for small strata (Franco and Lahiri, 2012). These are implemented using `prop.test()` for naive CIs and `svyciprop()` for design-adjusted CIs, targeting ±5% precision (Franco and Lahiri, 2012). Stratified sampling outperforms SRS in heterogeneous populations like Green City, reducing variance and CI width by 20–30% (Cao et al., 2013; Triveni et al., 2024). Relative efficiency, $\frac{Var_{SRS}}{Var_{st}}$, is assessed via simulations, with SRS variance:

$$Var_{SRS}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{p(1-p)}{n}. \tag{2.5}$$

Simulations quantify estimator performance and CI coverage, addressing nonresponse and auxiliary data challenges to optimise survey design (Triveni et al., 2024).

Power analysis for detecting subgroup differences in stratified designs uses minimum detectable effect calculations, with sample size based on stratum variances and allocation (Wang, Goldfeld, Taljaard and Li, 2024). For a target difference δ between strata, power is derived from the stratified variance formula, crucial for studies on subgroup variations like health equity (Wang et al., 2024). Bayesian methods complement design-based inference, handling data sparsity and complex structures via hierarchical modelling, with the "calibrated Bayes' ' approach ensuring robust estimates and nominal coverage (Wu and Stephenson, 2024). Using `rstanarm` in R, Bayesian generalised linear models incorporate survey weights, which provides a modern alternative to traditional variance estimation (Wu and Stephenson, 2024).

3. Methodology

The stratified proportion estimator is as defined in Section 2, but for H=3 (non-overlapping distinct subgroups) from Equation (2.1) to the equations that follow.

3.1 Handling Missing Values

To simulate realistic survey nonresponse, missingness is modeled as missing at random (MAR), with probabilities based on income (higher for Very Low), age (higher for older, 55+), and geography (higher for rural). This reflects real-world patterns where lower-income, older, or rural respondents are less likely to respond due to access or engagement barriers (Schouten, Lugtig and Vink, 2018; Westfall and Edgar, 2022). Nonresponse is handled using post-stratification adjustment, which reweights non-missing units within each stratum to minimise bias. The adjusted weight for non-missing unit i in stratum h is

$$w_{hi}^{adj} = w_{hi} \times \frac{\sum_{j=1}^{n_h} w_{hj}}{\sum_{j \in \text{non-missing}} w_{hj}}, \tag{3.1}$$

where $w_{hi} = N_h/n_h$ is the original sampling weight (Kolenikov, 2016). Missing responses receive zero weight, which ensures the total weight sum aligns with the population under MAR (Lewis, 2012). This is implemented in the `survey` package's `svydesign` object, with simulations evaluating impacts on precision and coverage.

3.2 Simulation Study Design and Metrics

The simulation study evaluates the performance of the stratified sampling design against simple random sampling (SRS) using Monte Carlo methods, with 200 repetitions to ensure stable empirical estimates of accuracy and coverage (Clement, Udofia and Enang, 2014). Each repetition follows these steps:

1. Generate a population of 10 000 with binomial adoption outcomes, adjusted by income (log-normal, scaled effect 0.15), gender (+3% non-binary, +1% female, -2% male), and geography (+8% urban, +2% suburban, -5% rural).
2. Draw a stratified sample ($n \approx 450$) using cost-constrained allocation (costs: R25 young, R35 middle-aged, R50 older; budget: R15 000) and an SRS sample for comparison.
3. Introduce missing-at-random (MAR) nonresponse based on income, and geography.
4. Compute the stratified proportion estimator $\hat{p}_{st}$, standard error (SE), and 95% Wilson confidence interval (CI) using the `survey` package's `svydesign` and `svyciprop` with adjusted weights.
5. Estimate the Bayesian proportion using `stan_glm` (binomial logit model with strata, income group, and geography predictors; normal priors, SD = 1.5; 2 chains, 1000 iterations), weighted by stratum sizes. (Wu and Stephenson, 2024).
6. Calculate the SRS proportion $\hat{p}_{SRS}$, SE, and Wilson CI using `prop.test`.
7. Record metrics for aggregation.

Performance is assessed across valid repetitions (excluding those with fewer than 5 non-missing responses per stratum for stability) using the following metrics:

- **Bias**: Average deviation of the estimator from the true proportion p (Lin, Bun, Gaboardi, Kolaczyk and Smith, 2024):
$$\text{Bias} = \frac{1}{R}\sum_{r=1}^{R}(\hat{p}_r - p), \tag{3.2}$$
where R is the number of valid repetitions, and $\hat{p}_r$ is the estimate (stratified or SRS) in repetition r.
- **Mean Squared Error (MSE)**: Combines bias and variance for overall accuracy (Franco et al., 2019):
$$\text{MSE} = \frac{1}{R}\sum_{r=1}^{R}(\hat{p}_r - p)^2 = \text{Bias}^2 + \text{Var}(\hat{p}). \tag{3.3}$$

- **Empirical Standard Error (SE)**: The standard deviation of estimates across repetitions captures the inherent sampling variability (Franco et al., 2019):

$$\text{Empirical SE} = \sqrt{\frac{1}{R-1}\sum_{r=1}^{R}(\hat{p}_r - \bar{\hat{p}})^2}, \tag{3.4}$$

where $\bar{\hat{p}} = \frac{1}{R}\sum_{r=1}^{R}\hat{p}_r$.

- **Average Theoretical SE**: Mean of per-repetition SEs for stratified estimates, calculated using the variance formula with sample $\hat{p}_h$ (Lin et al., 2024):

$$\text{Theoretical SE}_r = \sqrt{\sum_{h=1}^{3} w_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\hat{p}_{h,r}(1-\hat{p}_{h,r})}{n_h - 1}}, \tag{3.5}$$

averaged over R, where $w_h = N_h/N$.

- **Coverage Rate** is the proportion of 95% CIs containing the true p (Qing and Valliant, 2025):

$$\text{Coverage} = \frac{1}{R}\sum_{r=1}^{R} I(\text{CI}_{low,r} \le p \le \text{CI}_{high,r}), \tag{3.6}$$

where I is the indicator function, applied to both stratified and Bayesian CIs.

- **Average Margin of Error (MOE)**: Mean CI half-width, computed as $1.96 \times \text{SE}_r$ averaged over repetitions, to verify the ±5% target (Franco et al., 2019).

- **Relative Efficiency** is the ratio of SRS to stratified empirical variances (Sarkar, Shrestha, Rosenbaum and Oral, 2024):

$$\text{Relative Efficiency} = \frac{\text{Empirical Var}_{SRS}}{\text{Empirical Var}_{st}}. \tag{3.7}$$

- **Empirical DEFF**: Stratified design's variance relative to SRS (Franco et al., 2019):

$$\text{DEFF} = \frac{\text{Empirical Var}_{st}}{p(1-p)/n \times (1-n/N)}. \tag{3.8}$$

- **Power Analysis** evaluates the ability to detect subgroup differences (e.g., between strata) with 80% power, using:

$$\text{Power} = \Phi\left(\sqrt{\frac{\delta^2}{2\sum w_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{p_h(1-p_h)}{n_h}}} - z_{\alpha/2}\right), \tag{3.9}$$

where δ is the effect size, Φ is the standard normal CDF, and $z_{\alpha/2}$ is the z-score for $\alpha = 0.05$ (Wang et al., 2024; Kennedy Shaffer and Hughes, 2020).

These metrics assess the stratified design's precision, bias reduction, and robustness to MAR non-response, compared to SRS. Bayesian estimates provide an alternative approach, incorporating demographic predictors to improve accuracy (Wu and Stephenson, 2024). Power analysis quantifies the design's sensitivity to subgroup differences, which ensures reliable detection of variations in EV adoption across age groups (Wang et al., 2024). Results are aggregated to confirm the MOE meets the ±5% target and to evaluate nonresponse impacts.

4. Analysis

Table 1. Single-Sample Estimation Results for EV Adoption Proportion

Metric	Design-Adjusted	Bayesian	Naive	True Proportion
Estimate	0.4024	0.4139	0.4709	0.3891
Standard Error	0.0234	0.0115	0.0251	-
95% CI Lower	0.3574	0.3907	0.4211	-
95% CI Upper	0.4490	0.4361	0.5213	-
Margin of Error	0.0458	0.0225	0.0492	-
Precision Target Met (≤ 0.05)	TRUE	TRUE	FALSE	-
Bias	0.0133	0.0248	0.0818	-
Nonresponse Rate	0.1578	0.1578	0.1578	-

Table 1 compares estimation methods for the EV adoption proportion. The design-adjusted estimate (0.4024) is closer to the true proportion (0.3891) than the naive estimate (0.4709), with a bias of 0.0133 versus 0.0818, highlighting stratification's ability to reduce error in heterogeneous populations. Its standard error (0.0234) yields an MOE of 0.0458, which meets the ≤ 0.05 target, and its 95% confidence interval (0.3574–0.4490) contains the true proportion, showing reliable coverage. The Bayesian estimate (0.4139) has a slightly higher bias (0.0248) but a lower standard error (0.0115), giving a tighter MOE (0.0225) and a 95% CI (0.3907–0.4361) that also covers the true proportion. The naive estimate's larger standard error (0.0251) and MOE (0.0492) fail the precision target, and its wider 95% CI (0.4211–0.5213) shows lower efficiency. The 15.78% nonresponse rate, adjusted via post-stratification in the design-adjusted and Bayesian approaches, minimally impacts accuracy, as both estimates remain close to the true proportion. These results suggest that both stratified and Bayesian methods outperform the naive approach, and the Bayesian method provides superior precision because it incorporates demographic predictors.

Table 2. Design-Adjusted EV Adoption Proportions by Income Group

Income Group	Proportion	Standard Error	95% CI	Margin of Error	Effective Sample Size
Very Low	0.2884	0.0427	[0.2047, 0.3721]	0.0837	92.1
Low	0.3250	0.0422	[0.2423, 0.4078]	0.0827	94.8
Middle	0.4434	0.0559	[0.3338, 0.5530]	0.1096	71.0
High	0.5672	0.0735	[0.4232, 0.7112]	0.1441	45.8
Very High	0.6778	0.0915	[0.4986, 0.8570]	0.1793	29.4

Table 2 presents design-adjusted estimates of EV adoption proportions across five income groups, adjusted for a 15.78% missing-at-random nonresponse rate. The Very Low income group has the lowest proportion at 0.2884 (SE = 0.0427, 95% CI: 0.2047–0.3721, MOE = 0.0837), followed by Low at 0.3250 (SE = 0.0422, 95% CI: 0.2423–0.4078, MOE = 0.0827). The Middle group shows 0.4434 (SE = 0.0559, 95% CI: 0.3338–0.5530, MOE = 0.1096), while High and Very High groups have higher proportions at 0.5672 (SE = 0.0735, 95% CI: 0.4232–0.7112, MOE = 0.1441) and 0.6778 (SE = 0.0915, 95% CI: 0.4986–0.8570, MOE = 0.1793), respectively. Effective sample sizes range from 29.4 (Very High) to 94.8 (Low). The increasing adoption trend with income reflects the simulation's log-income effect (0.15). However, none of the income groups meet the ≤ 0.05 MOE target, as MOEs range from 0.0827 to 0.1793, which indicates reduced precision due to smaller effective sample sizes and within-group variability. The Very High group's wide CI and high MOE suggest caution when interpreting its estimate. These results highlight income as a key driver of EV adoption and underscore the need for subsidies to address barriers for lower-income groups.

Table 3. Simulation Study Performance Metrics for Stratified, Bayesian, and SRS Designs

Metric	Stratified Design	Bayesian	SRS Design
Bias	0.0087	0.0248	0.0011
Mean Squared Error (MSE)	0.0006	0.0007	0.0007
Empirical Standard Error (SE)	0.0240	0.0115	0.0267
Average Theoretical SE	0.0231	-	0.0266
Coverage (95% CI)	0.930	0.950	0.960
Average Margin of Error	0.0453	0.0225	0.0523
Precision Target Met (≤ 0.05)	TRUE	TRUE	FALSE
Relative Efficiency	1.2375	-	-
Empirical DEFF	1.1385	-	-
Power (Detect $\delta = 0.08$)	0.8409	-	-
Valid Replications	200	200	200
Nonresponse Rate	0.1578	0.1578	0.1578

Table 3 summarises the performance of stratified, Bayesian, and SRS designs over 200 replications for estimating the EV adoption proportion (true proportion: 0.3891). The stratified design has low bias (0.0087), with an MSE of 0.0006 and empirical SE (0.0240) aligning with its theoretical SE (0.0231). Its coverage (0.930) is slightly below 95%, possibly due to the 15.78% MAR nonresponse rate, but its MOE (0.0453) meets the ≤ 0.05 target. The Bayesian approach shows higher bias (0.0248) but a lower empirical SE (0.0115), which yields a tighter MOE (0.0225) and better coverage (0.950). No theoretical SE is reported for Bayesian estimates, as they rely on posterior distributions. The SRS design has the lowest bias (0.0011) but higher MSE (0.0007) and empirical SE (0.0267), with an MOE (0.0523) that fails the precision target. Its coverage (0.960) is robust but less precise. The stratified design's relative efficiency (1.2375) indicates a 23.75% variance reduction over SRS, though its empirical DEFF (1.1385) exceeds the theoretical DEFF (0.7896), likely due to nonresponse adjustments. The power to detect a 0.08 subgroup difference (84%) demonstrates the design's ability to identify variations across age strata. These results confirm that stratified and Bayesian methods outperform SRS in precision, with Bayesian estimates providing the tightest MOE, suitable for informing EV adoption policies targeting diverse groups.

Figure 1 illustrates the relationship between the required sample size n and the margin of error E for stratified sampling across four scenarios scaling the baseline adoption probabilities (25%, 50%, 75%, 100%). As E increases, the required sample size decreases sharply, consistent with the theoretical expectation that higher precision (smaller E) requires larger samples. At the study's target $E = 0.05$, sample sizes range from approximately 150 (25% scenario) to 350 (100% scenario), with the chosen $n = 450$ exceeding these to ensure precision under cost-constrained allocation and missing values. Beyond $E = 0.05$, the curves plateau, which indicates diminishing returns where larger samples yield minimal precision gains. The 25% scenario requires the smallest sample sizes across all E, reflecting lower variance from scaled-down adoption probabilities. In contrast, the 100% and 75% scenarios demand larger samples due to higher variance from increased proportions. These trends highlight the critical role of prior adoption probability estimates in optimising sample size for efficient survey design.

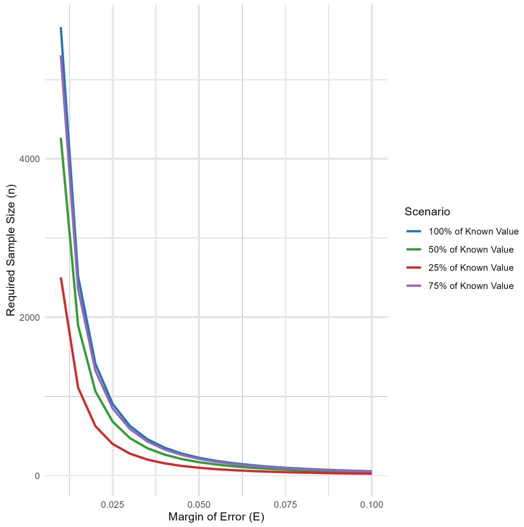


Figure 1. Sample Size vs Margin of Error for Stratified Sampling

Figure 2 illustrates the population adoption proportions of electric vehicles (EVs) across three age strata: stratum 1 (18–34 years), stratum 2 (35–54 years), and stratum 3 (55+ years). The y-axis shows the adoption proportion within each stratum, with values of 0.6225 for stratum 1, 0.3183 for stratum 2, and 0.1148 for stratum 3, based on the simulation results. The clear decreasing trend across age groups demonstrates a strong age-related gradient in EV adoption, with younger individuals (18–34) having a significantly higher likelihood of adoption compared to middle-aged (35–54) and older (55+) groups. This variation emphasises the importance of stratified sampling to ensure accurate representation of each age group and to prevent bias from the higher adoption rates in the younger stratum when estimating the population-level proportion.

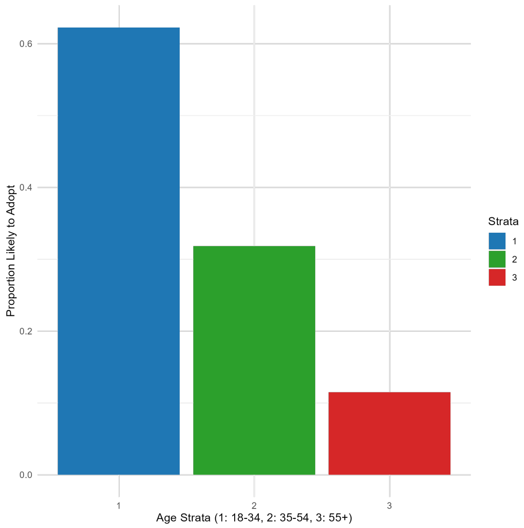
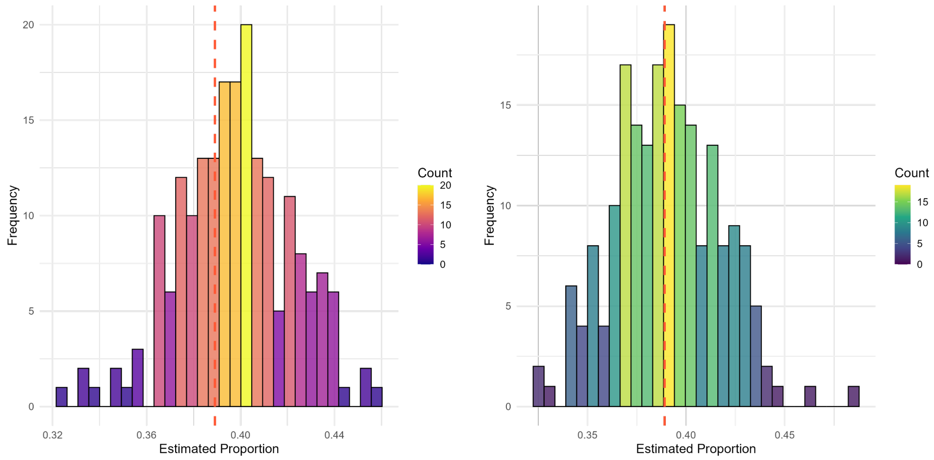


Figure 2. Population EV Adoption Proportion by Age Strata



(a) Design-Adjusted Estimates for Simulations
True p = 0.3891 | Bias = 0.00867

(b) SRS Estimates for Simulations
True p = 0.3891 | Bias = 0.00101

Figure 3. Comparison of Design-Adjusted and SRS Simulation Estimates

Figure 3(a) illustrates the distribution of design-adjusted estimates for EV adoption proportions from 200 simulation replications. The x-axis shows estimated proportions, and the y-axis indicates their frequency. The histogram forms a symmetric, bell-shaped curve, centered at approximately 0.4 (mean of estimates, given bias = 0.0087), with a red dashed line marking the true population proportion of 0.3891. A color gradient from purple (low density) to yellow (high density) highlights where estimates are concentrated, with most falling between 0.3738 and 0.4218 (based on empirical SE = 0.0240). The tight clustering shows low variability and high precision, with a coverage rate of 0.930

(see Table 3), confirming the stratified estimator's reliability despite a 15.78% missing-at-random nonresponse rate.

Figure 3(b) presents the distribution of SRS estimates for EV adoption proportions from 200 simulation replications. The x-axis displays estimated proportions, ranging approximately from 0.3357 to 0.4427 (based on empirical SE = 0.0267 around the mean estimate of 0.3902). The histogram forms a unimodal, nearly symmetric bell-shaped curve, peaking near 0.3902, with a red dashed line marking the true proportion of 0.3891. A color gradient from purple (low density) to yellow (high density) emphasises where estimates are concentrated. Compared to Figure 3(a), the SRS distribution shows a wider spread, which reflects higher variability (empirical SE = 0.0267) and lower precision, as evidenced by the relative efficiency of 1.2375 and higher MSE (0.0007 vs. 0.0006) in Table 3.

5. Conclusion and Discussion

The stratified sampling design effectively estimates the likelihood of electric vehicle (EV) adoption in Green City, with a single-sample design-adjusted proportion of 0.4024 (Table 1) closely approximating the true population proportion of 0.3891. Across 200 simulation replications, the stratified design achieves low bias (0.0087) and a mean squared error (MSE) of 0.0006, with an average margin of error (MOE) of 0.0453, which meets the precision target of ≤ 0.05 (Table 3). The Bayesian approach, with an estimate of 0.4139 and a tighter MOE of 0.0225, provides superior precision, while the simple random sampling (SRS) estimate (0.4709) fails the precision target (MOE = 0.0492). Significant demographic variations are evident, with adoption proportions decreasing across age strata (0.6225, 0.3183, 0.1148 for 18–34, 35–54, 55+ years, respectively, Figure 2) and increasing with income (0.2884 for Very Low to 0.6778 for Very High, Table 2). These trends provide actionable insights for targeted EV policies, such as subsidies for lower-income groups or infrastructure expansion for older residents.

The study meets its objectives of delivering precise (MOE ≤ 0.05) and efficient estimates compared to SRS. The stratified design's relative efficiency of 1.2375 reflects a 23.75% variance reduction over SRS, although the empirical design effect (DEFF) of 1.1385 exceeds the theoretical DEFF of 0.7896, likely due to missing-at-random nonresponse adjustments (Table 3). The power to detect a 0.08 subgroup difference with 84% power demonstrates the design's ability to identify meaningful variations across age strata. Graphical analyses (Figures 1 and 3) confirm stable estimate distributions, with design-adjusted estimates showing tighter clustering than SRS (empirical SE: 0.0240 vs. 0.0267). The sample size of $n = 450$, exceeding the precision-based requirement of 221 and budget-allowed 409 (Figure 1), ensures robustness under cost constraints.

However, the study's reliance on simulated data with assumed adoption probabilities (e.g., 60% for 18–34, 30% for 35–54, 5% for 55+) and income adjustments (scaled effect 0.15) may not fully capture real-world complexities. The high variability in the older stratum (proportion 0.1148) and Very High income group (SE = 0.0915) suggests potential overestimation of stratification benefits if actual heterogeneity is lower. The single-sample stratified CI (0.3574–0.4490) includes the true proportion, but high SEs for income groups (0.0422–0.0915, Table 2) indicate limited subgroup precision, particularly for higher-income groups with smaller effective sample sizes. Figure 3 shows stable distributions, while the Bayesian approach's tighter intervals indicate potential for improved modelling with additional predictors. The focus on age and income overlooks factors like geography, which influences adoption (e.g., +8% urban, -5% rural). Incorporating such variables could enhance estimate stability but may require larger samples, increasing costs beyond the R15 000 budget. Future studies should validate these findings with real survey data and explore additional demographic predictors to refine EV adoption strategies.

References

Cao, R., Vilar, J. A., Vilar, J. M., and López, A. K. (2013). Sampling Error Estimation in Stratified Surveys. *Open Journal of Statistics*, **3** (3), 200–212. doi:10.4236/ojs.2013.33023.
URL: *https://www.scirp.org/journal/paperinformation.aspx?paperid=35899*

César, C. C. and Carvalho, M. S. (2011). Stratified Sampling Design and Loss to Follow-Up in Survival Models: Evaluation of Efficiency and Bias. *BMC Medical Research Methodology*, **11** (1), 99. doi:10.1186/1471-2288-11-99.
URL: *https://doi.org/10.1186/1471-2288-11-99*

Chen, T. and Lumley, T. (2022). Optimal Sampling for Design-Based Estimators of Regression Models. *Statistics in Medicine*, **41** (8), 1482–1497. doi:10.1002/sim.9300.
URL: *https://doi.org/10.1002/sim.9300*

Clement, E. P., Udofia, G. A., and Enang, E. I. (2014). Estimation for Domains in Stratified Sampling Design in the Presence of Nonresponse. *American Journal of Mathematics and Statistics*, **4** (2), 65–71. doi:10.5923/j.ajms.20140402.03.
URL: *https://doi.org/10.5923/j.ajms.20140402.03*

DeYoreo, M. (2018). Stratified Random Sampling. *In* Frey, B. B. (Editor) *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*, volume 4. SAGE Publications, Inc., Thousand Oaks, CA, p. 1624. doi:10.4135/9781506326139.n671.
URL: *https://doi.org/10.4135/9781506326139.n671*

Diaz Quijano, F. A. (2018). Sample Allocation Balancing Overall Representativeness and Stratum Precision. *Annals of Epidemiology*, **28** (8), 570–575. doi:10.1016/j.annepidem.2018.04.011.
URL: *https://www.sciencedirect.com/science/article/pii/S1047279717310797*

Franco, C. and Lahiri, P. (2012). Interval Estimation for Small Area Proportions with Small True Proportions from Stratified Random Sampling Survey Data. *Proceedings of the Joint Statistical Meetings*.
URL: *https://www.asasrms.org/Proceedings/y2012/Files/305325_74654.pdf*

Franco, C., Little, R. J. A., Louis, T. A., and Slud, E. V. (2019). Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys. *Journal of Survey Statistics and Methodology*, **7** (3), 334–364. doi:10.1093/jssam/smy019.
URL: *https://doi.org/10.1093/jssam/smy019*

Kennedy Shaffer, L. and Hughes, M. D. (2020). Sample Size Estimation for Stratified Individual and Cluster Randomized Trials with Binary Outcomes. *Statistics in Medicine*, **39** (10), 1489–1513. doi:10.1002/sim.8492.
URL: *https://doi.org/10.1002/sim.8492*

Kim, Y. J., Oh, Y., Park, S., Cho, S., and Park, H. (2013). Stratified Sampling Design Based on Data Mining. *Healthcare Informatics Research*, **19** (3), 186–195. doi:10.4258/hir.2013.19.3.186.

URL: *https://doi.org/10.4258/hir.2013.19.3.186*

Kolenikov, S. (2016). Post-Stratification or Non-Response Adjustment? *Survey Practice*, **9** (3). doi:10.29115/SP-2016-0014.
URL: *https://doi.org/10.29115/SP-2016-0014*

Lewis, T. (2012). Weighting Adjustment Methods for Nonresponse in Surveys. *WUSS Proceedings*, **162**.
URL: *https://www.lexjansen.com/wuss/2012/162.pdf*

Lin, S., Bun, M., Gaboardi, M., Kolaczyk, E. D., and Smith, A. (2024). Differentially Private Confidence Intervals for Proportions under Stratified Random Sampling. *Electronic Journal of Statistics*, **18** (1), 1455–1494. doi:10.1214/24-EJS2234.
URL: *https://doi.org/10.1214/24-EJS2234*

Lohr, S. L. (2021). *Sampling: Design and Analysis*. 3rd edition. Chapman & Hall/CRC, New York. doi:10.1201/9780429298899.

Pamidimukkala, A., Kermanshachi, S., Rosenberger, J. M., and Hladik, G. (2024). Barriers and Motivators to the Adoption of Electric Vehicles: A Global Review. *Green Energy and Intelligent Transportation*, **3** (2), 100153. doi:10.1016/j.geits.2024.100153.
URL: *https://www.sciencedirect.com/science/article/pii/S2773153724000057*

Pandita, D., Bhatt, V., Kumar, V. V. R., Fatma, A., and Vapiwala, F. (2024). Electrifying the Future: Analysing the Determinants of Electric Vehicle Adoption. *International Journal of Energy Sector Management*, **18** (6), 1767–1786. doi:10.1108/IJESM-06-2023-0004.
URL: *https://doi.org/10.1108/IJESM-06-2023-0004*

Pei, M., Huang, Z., Zhang, Z., Wang, K., and Ye, X. (2025). Range Anxiety and Willingness to Pay: Psychological Insights for Electric Vehicle. *Journal of Renewable and Sustainable Energy*, **17** (1), 015902. doi:10.1063/5.0220237.
URL: *https://doi.org/10.1063/5.0220237*

Powell, S., Cezar, G. V., and Rajagopal, R. (2022). Scalable Probabilistic Estimates of Electric Vehicle Charging Given Observed Driver Behavior. *Applied Energy*, **309**, 118382. doi:10.1016/j.apenergy.2021.118382.
URL: *https://www.sciencedirect.com/science/article/pii/S0306261921016214*

Qing, S. and Valliant, R. (2025). Extending Cochran's Sample Size Rule to Stratified Simple Random Sampling with Applications to Audit Sampling. *Journal of Official Statistics*, **41** (1), 309–328. doi:10.1177/0282423X241277054.
URL: *https://doi.org/10.1177/0282423X241277054*

Rahman, M. M. and Thill, J.-C. (2024). A Comprehensive Survey of the Key Determinants of Electric Vehicle Adoption: Challenges and Opportunities in the Smart City Context. *World Electric Vehicle Journal*, **15** (12). doi:10.3390/wevj15120588.
URL: *https://www.mdpi.com/2032-6653/15/12/588*

Sarkar, I., Shrestha, P., Rosenbaum, C., and Oral, E. (2024). Comparing Different Sampling Schemes to Estimate the Population Mean: A Simulation Study. *School of Public Health Delta Omega Honors Day Poster Sessions*, **16**.
URL: *https://digitalscholar.lsuhsc.edu/dohd/2021/2021/16*

Schouten, R. M., Lugtig, P., and Vink, G. (2018). Generating Missing Values for Simulation Purposes: A Multivariate Amputation Procedure. *Journal of Statistical Computation and Simulation*, **88** (15), 2909–2930. doi:10.1080/00949655.2018.1491577.
URL: *https://doi.org/10.1080/00949655.2018.1491577*

Triveni, G. R. V., Danish, F., and Tawiah, K. (2024). Evolving Techniques for Enhanced Estimation: A Comprehensive Survey of Stratified Sampling and Post-Stratification Methods. *AIP Advances*, **14** (10). doi:10.1063/5.0193961.
URL: *https://doi.org/10.1063/5.0193961*

Van Hoeven, L. R., Janssen, M. P., Roes, K. C., and Koffijberg, H. (2015). Aiming for a Representative Sample: Simulating Random versus Purposive Strategies for Hospital Selection. *BMC Medical Research Methodology*, **15** (1), 90.

Wang, X., Goldfeld, K. S., Taljaard, M., and Li, F. (2024). Sample Size Requirements to Test Subgroup-Specific Treatment Effects in Cluster-Randomized Trials. *Prevention Science*, **25** (Suppl 3), 356–370. doi:10.1007/s11121-023-01590-6.
URL: *https://doi.org/10.1007/s11121-023-01590-6*

Westfall, J. A. and Edgar, C. B. (2022). Addressing Non-Response Bias in Urban Forest Inventories: An Estimation Approach. *Frontiers in Forests and Global Change*, **5**, 895969. doi: 10.3389/ffgc.2022.895969.
URL: *https://doi.org/10.3389/ffgc.2022.895969*

Wu, S. M. and Stephenson, B. J. K. (2024). Bayesian Estimation Methods for Survey Data with Potential Applications to Health Disparities Research. *Wiley Interdisciplinary Reviews: Computational Statistics*, **16** (1), e1633. doi:10.1002/wics.1633.
URL: *https://doi.org/10.1002/wics.1633*

Zhao, M., Klaassen, C. A. J., Lisovski, S., and Klaassen, M. (2019). The Adequacy of Aging Techniques in Vertebrates for Rapid Estimation of Population Mortality Rates from Age Distributions. *Ecology and Evolution*, **9** (3), 1394–1402. doi:10.1002/ece3.4854.
URL: *https://doi.org/10.1002/ece3.4854*

Manuscript received 2025-xx-xx, revised 2025-xx-xx, accepted 2025-xx-xx.