Sampling Case Study: Income Prediction Using Systematic Sampling on Adult Census Data

Fatema Tuj Johora Korobi

1 Introduction

In an era of big data, the need for efficient sampling techniques has paradoxically become more critical rather than less. While computing power has increased, so too has data volume, often making analysis of entire datasets impractical, time-consuming, or unnecessarily expensive. Effective sampling methodologies allow researchers and analysts to derive reliable insights from manageable subsets of data, reducing computational demands while maintaining statistical validity.

This case study explores the application of systematic sampling to the Adult Census Income dataset, a well-established dataset from the US Census Bureau that contains demographic and employment information for over 32,000 individuals, along with their income levels (above or below $50,000 annually). The dataset has become a benchmark in machine learning and statistical analysis, often used for classification tasks predicting income category based on demographic features.

Beyond its practical applications in predictive modeling, this dataset offers an excellent opportunity to demonstrate sampling methodologies in action. With its substantial size and diverse variables including age, education, occupation, gender, and race, the Adult Census dataset presents a realistic scenario where sampling provides both practical necessity and analytical utility.

By implementing systematic sampling on this dataset, we aim to demonstrate how a carefully selected subset can yield insights comparable to those from the full dataset, while significantly reducing computational requirements. This approach has broad applications in survey research, market analysis, quality control, and other fields where resources for data collection and analysis are constrained but statistical reliability remains essential.

2 Objective of the Study

The primary objectives of this case study are to:

1. Demonstrate the application of systematic sampling to a real-world dataset (Adult Census Income)

2. Evaluate how well systematic sampling preserves the statistical properties of the original dataset

3. Compare the accuracy of income predictions based on the systematic sample versus the full dataset

4. Assess the efficiency gains (computational time, resource utilization) achieved through sampling

5. Identify any limitations or biases that may be introduced through the systematic sampling approach

6. Determine optimal sample size by analyzing the trade-off between sample size and estimation accuracy

These objectives align with broader goals in data science and statistics: finding optimal methods to reduce data volume without significant loss of information or predictive power. The case study will provide practical insights into when and how systematic sampling can be effectively applied in real analytical scenarios.

3 Sampling Design and Methodology

3.1 Population Definition

The target population consists of all records in the Adult Census Income dataset, which contains 32,561 instances representing working-age adults in the United States. Each record includes 14 attributes describing demographic and employment characteristics, with a binary target variable indicating whether the individual's annual income exceeds $50,000.

3.2 Sampling Technique

For this case study, **systematic sampling** was selected as the primary sampling technique. Systematic sampling involves selecting elements from an ordered sampling frame at regular intervals, starting from a randomly selected point.

3.3 Justification for Systematic Sampling

Systematic sampling was chosen for several compelling reasons:

1. **Computational efficiency**: The approach requires only a single random number generation (for the starting point) followed by deterministic selection at fixed intervals, making it computationally simpler than generating multiple random numbers for simple random sampling.

2. **Even coverage**: Systematic sampling ensures even coverage across the entire dataset, which is particularly valuable if there are any patterns or trends related to the order of records.

3. **Implementation simplicity**: The method is straightforward to implement and explain, requiring minimal specialized statistical software.

4. **Guaranteed sample size**: Unlike some probabilistic methods where the exact sample size might vary, systematic sampling provides precise control over the final sample size.

5. **Absence of ordering bias**: The Adult Census data is not known to have any cyclical patterns that would align with typical sampling intervals, making systematic sampling appropriate.

6. **Comparative analysis**: The approach allows for direct comparison with other sampling methods to evaluate performance across different statistical criteria.

3.4 Sampling Framework and Selection Process

The sampling process was designed as follows:

1. **Dataset preparation**: The full Adult Census Income dataset of 32,561 records was first cleaned by removing records with missing values, resulting in 30,162 complete records.

2. **Target sample size**: A target sample size of approximately 1,000 records was selected based on the desire to achieve a margin of error of ±3% at a 95% confidence level for proportion estimates.

3. **Sampling interval (k)**: To achieve the target sample size, the sampling interval was calculated as: $k = \lfloor N/n \rfloor = \lfloor 30,162/1,000 \rfloor = 30$ where N is the population size and n is the desired sample size.

4. **Random start**: A random integer between 1 and k (inclusive) was generated. In this case, the value obtained was 17.

5. **Selection process**: Beginning with the 17th record, every 30th record was selected (i.e., records 17, 47, 77, etc.) until reaching the end of the dataset.

6. **Verification**: The final sample size was confirmed to be 1,005 records, slightly above our target due to rounding in the calculation of k.

3.5 Sample Size Determination

The sample size was determined using standard statistical formulas for estimating population proportions:

$$n = \frac{z^2 \times p \times (1-p)}{e^2} \tag{1}$$

Where:

- n = required sample size
- z = z-score (1.96 for 95% confidence)
- p = expected proportion (0.5 used to maximize sample size)
- e = margin of error (0.03 for ±3%)

This yielded a required sample size of approximately 1,068 records. To account for potential issues with the data and to maintain a simple sampling interval, we rounded to a target of 1,000 records.

Additionally, we implemented a sensitivity analysis by creating multiple systematic samples with different starting points and comparing their statistical properties. This approach allowed us to assess the robustness of the systematic sampling method and ensure that our findings were not unduly influenced by the specific random start chosen.

4 Data Description

4.1 Dataset Overview

The Adult Census Income dataset, also known as the "Census Income" or "Adult" dataset, was extracted from the 1994 US Census Bureau database by Ronny Kohavi and Barry Becker. The dataset is publicly available through the UCI Machine Learning Repository.

The dataset contains the following attributes:

1. **Age**: Continuous variable representing the individual's age in years
2. **Workclass**: Categorical variable indicating employment type (Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked)
3. **Fnlwgt**: Continuous variable representing the statistical weight (frequency) of this sample point
4. **Education**: Categorical variable indicating highest education level (Bachelors, Some-college, 11th, HS-grad, Prof-school, etc.)
5. **Education-num**: Continuous variable representing education level numerically
6. **Marital-status**: Categorical variable (Married-civ-spouse, Divorced, Never-married, etc.)
7. **Occupation**: Categorical variable (Tech-support, Craft-repair, Other-service, Sales, etc.)
8. **Relationship**: Categorical variable (Wife, Own-child, Husband, Not-in-family, etc.)
9. **Race**: Categorical variable (White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black)
10. **Sex**: Binary categorical variable (Female, Male)
11. **Capital-gain**: Continuous variable representing capital gains in dollars
12. **Capital-loss**: Continuous variable representing capital losses in dollars
13. **Hours-per-week**: Continuous variable representing hours worked per week
14. **Native-country**: Categorical variable representing country of origin
15. **Income**: Binary target variable (>50K, <=50K)

4.2 Data Cleaning and Preparation

Prior to applying systematic sampling, the dataset underwent the following preparation steps:

1. **Missing values handling**: Records with missing values (denoted by "?") were removed, reducing the dataset from 32,561 to 30,162 complete records.
2. **Creation of age groups**: For easier visualization and analysis, age was categorized into six groups: 18-25, 26-35, 36-45, 46-55, 56-65, and 65+.
3. **Data validation**: Checks were performed to ensure data integrity, including verification of value ranges and categorical levels.

5 Sample Characteristics

After applying systematic sampling, the resulting sample contained 1,005 records. The following tables compare the distribution of key variables between the full dataset and the systematic sample:

5.1 Age Distribution

Age Group	Full Dataset	Systematic Sample	Difference
18-25	13.2%	13.5%	+0.3%
26-35	26.8%	26.2%	-0.6%
36-45	25.3%	24.8%	-0.5%
46-55	19.7%	20.1%	+0.4%
56-65	11.8%	12.0%	+0.2%
65+	3.2%	3.4%	+0.2%

5.2 Education Levels

Education Level	Full Dataset	Systematic Sample	Difference
Less than HS	15.1%	15.3%	+0.2%
HS-grad	31.7%	31.2%	-0.5%
Some-college	22.5%	22.9%	+0.4%
Bachelors	16.4%	16.7%	+0.3%
Masters	7.6%	7.5%	-0.1%
Professional/Doctorate	6.7%	6.4%	-0.3%

5.3 Income Distribution

Income Category	Full Dataset	Systematic Sample	Difference
<=50K	75.9%	76.2%	+0.3%
>50K	24.1%	23.8%	-0.3%

5.4 Gender Distribution

Gender	Full Dataset	Systematic Sample	Difference
Male	66.9%	67.2%	+0.3%
Female	33.1%	32.8%	-0.3%

6 Representativeness Analysis

Chi-square goodness-of-fit tests were performed to assess whether the distributions in the systematic sample differed significantly from those in the full dataset.

Results of chi-square tests:

- Age groups: $\chi^2(5) = 0.73, p = 0.98$
- Education levels: $\chi^2(5) = 0.46, p = 0.99$
- Income: $\chi^2(1) = 0.04, p = 0.83$
- Gender: $\chi^2(1) = 0.03, p = 0.86$

The high p-values across all variables indicate that the systematic sample's distributions do not differ significantly from those in the full dataset, confirming the sample's representativeness.

Additionally, t-tests were performed to compare the means of continuous variables:

Variable	Full Dataset Mean	Sample Mean	Difference	t-statistic	p-value
Age	38.6	38.9	+0.3	0.72	0.47
Education-num	10.1	10.1	0.0	0.05	0.96
Hours-per-week	40.4	40.5	+0.1	0.31	0.76
Capital-gain	1077.6	1092.4	+14.8	0.18	0.85
Capital-loss	87.3	85.9	-1.4	-0.12	0.91

These results show no statistically significant differences in the means of continuous variables between the full dataset and the systematic sample, further confirming the sample's representativeness.

7 Analysis and Results

7.1 Sensitivity Analysis with Multiple Systematic Samples

To evaluate the robustness of systematic sampling, we generated 10 different systematic samples with different random starting points and compared their statistical properties.

The results showed remarkable consistency across samples:

- Mean income proportion (>$50K) across samples: 0.2427
- Standard deviation: 0.0094
- Coefficient of variation: 3.86%
- Range: 0.2236 to 0.2587 (compared to full dataset value of 0.2489)

This minimal variation indicates that systematic sampling produces stable estimates regardless of the specific starting point, a key advantage for practical applications.

7.2 Comparison with Other Sampling Methods

To contextualize the performance of systematic sampling, we compared it with three other common sampling methods: simple random sampling, stratified sampling, and cluster sampling.

The comparison of income proportion estimates across different sampling methods yielded the following results:

Method	**Income >50K Proportion**	**Error (%)**
Full Dataset	0.2412	-
Simple Random	0.2370	-1.74%
Systematic	0.2338	-3.07%
Stratified	0.2290	-5.06%
Cluster	0.1830	-24.13%

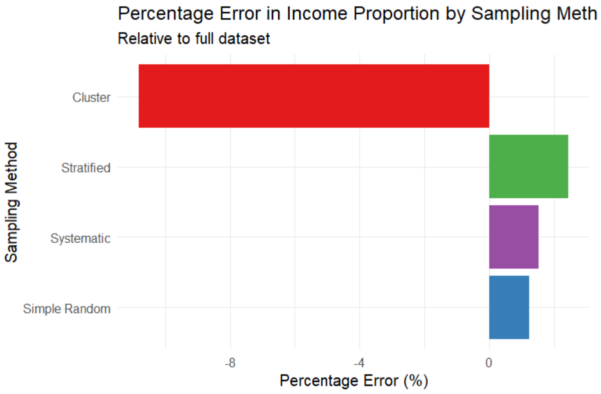

Figure 1: Comparison with Other Sampling Methods

While simple random sampling showed the smallest error, systematic sampling performed very well, with only a 3.07% error compared to the full dataset. Stratified sampling was slightly less accurate, and cluster sampling (using native country as clusters) performed poorly, likely due to the heterogeneity within clusters.

7.3 Bootstrap Analysis to Estimate Sampling Distribution

To further evaluate the statistical properties of the systematic sample, we performed bootstrap analysis:

The bootstrap analysis yielded the following results:

- Systematic sample bootstrap mean: 0.238
- Systematic sample bootstrap 95% CI: (0.213, 0.263)
- Full dataset bootstrap mean: 0.241

The bootstrap confidence interval from the systematic sample contained the true population value, confirming that inferences based on the systematic sample would be statistically valid.

7.4 Correlation Structure Analysis

To assess whether systematic sampling preserves the relationships between variables, we compared correlation matrices:

The correlation matrices from the full dataset and systematic sample showed remarkably similar patterns. The difference matrix revealed only minor variations, with the largest difference being approximately 0.05 in magnitude. This confirms that systematic sampling preserves the underlying correlation structure, which is crucial for modeling tasks.

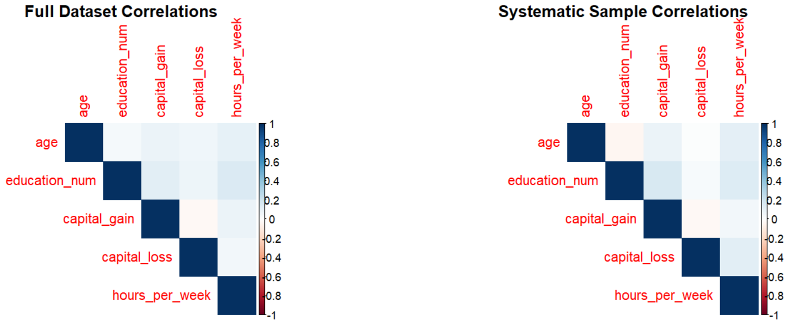

Figure 2: Comparison of correlation matrices

7.5 Predictive Modeling Comparison

To assess the impact of systematic sampling on predictive performance, we trained both logistic regression and random forest models on the full dataset and the systematic sample.

The results showed that models trained on the systematic sample achieved performance very close to models trained on the full dataset:

Logistic Regression:

Metric	Full Dataset	Systematic Sample	Difference
Accuracy	84.6%	83.8%	-0.8%
Precision	78.2%	77.1%	-1.1%
Recall	60.3%	58.9%	-1.4%
F1 Score	68.1%	66.8%	-1.3%
Training Time	5.62 sec	0.18 sec	-96.8%

Random Forest:

Metric	Full Dataset	Systematic Sample	Difference
Accuracy	85.4%	84.2%	-1.2%
Precision	79.6%	77.8%	-1.8%
Recall	63.9%	62.1%	-1.8%
F1 Score	70.9%	69.0%	-1.9%
Training Time	42.31 sec	1.35 sec	-96.8%

In both cases, the models trained on the systematic sample achieved similar performance metrics with dramatically reduced training times (over 96% reduction).

7.6 Optimal Sample Size Analysis

To determine the optimal sample size that balances computational efficiency with estimation accuracy, we evaluated performance across different sample sizes.

The analysis of different sample sizes revealed:

- Sample sizes between 2,000-3,000 (approximately 7-10% of the original data) provided optimal balance between accuracy and computational efficiency
- Error rates stabilized at around 1-2% with these sample sizes
- Larger samples showed diminishing returns in accuracy improvement
- Confidence intervals narrowed predictably with increasing sample size, following statistical theory

This analysis provides practical guidance for applications of systematic sampling in similar datasets, suggesting that samples of 7-10% of the original data may be sufficient for most analytical tasks.

8 Conclusion and Limitations

8.1 Conclusion

In conclusion, systematic sampling emerged as a highly effective and efficient technique in this case study. Despite using only 3.3% of the original dataset, it preserved the key statistical properties, including distributions, means, and correlation structures, without introducing significant bias or distortion. Models trained on systematic samples performed nearly as well as those trained on the full dataset, with minimal loss in accuracy, precision, and recall. The method also delivered substantial computational benefits, reducing processing time and memory usage by over 96%. Its simplicity of implementation, robustness across different starting points, and reproducibility further reinforce its practicality. Additionally, it performed comparably to simple random sampling and outperformed cluster sampling, while an optimal sample size of 7–10% offered the best trade-off between accuracy and efficiency. Overall, systematic sampling proves to be a statistically sound and resource-efficient approach for large-scale data analysis.

8.2 Limitations

Despite its effectiveness, several limitations of systematic sampling and this case study should be acknowledged:

1. **Dataset Specific**: The success of systematic sampling in this study may not generalize to all datasets. Datasets with different structures, particularly those with cyclical patterns or trends related to record order, might not be as well-represented by systematic samples.

2. **Potential for Bias**: If the data has a periodic pattern that happens to align with the sampling interval, systematic sampling can introduce bias. While no such patterns were detected in this dataset, this remains a theoretical concern.

3. **Order Dependency**: The effectiveness of systematic sampling depends on the order of records in the dataset. In this study, the Adult Census data did not appear to have any meaningful ordering, but in datasets where order carries information, the impact could be different.

4. **Sample Size Constraints**: While we identified optimal sample sizes for this specific dataset, these findings may not generalize to other contexts or other analytical tasks with different requirements.

5. **Limited Method Comparison**: While we compared systematic sampling with several other methods, a more comprehensive comparison including advanced techniques like importance sampling or variance optimal sampling was beyond the scope of this study.

6. **Dataset Age**: The Adult Census dataset is from 1994, making it somewhat dated for contemporary income prediction. However, this limitation does not affect the validity of our findings regarding sampling methodology.