Differentially Private Analysis of U.S. Household Income Statistics using Laplace and Gaussian Mechanisms

Muttaki I Bismoy
University of Michigan–Dearborn
Dearborn, Michigan, USA
mbismoy@umich.edu

Shomitro Ghosh
University of Michigan–Dearborn
Dearborn, Michigan, USA
shomitro@umich.edu

Farhan Tanvir
University of Michigan–Dearborn
Dearborn, Michigan, USA
farhanta@umich.edu

Ahmad Jadallah
University of Michigan–Dearborn
Dearborn, Michigan, USA
ahmadjd@umich.edu

Abstract

The widespread availability of open datasets enables powerful data-driven insights but also raises substantial privacy concerns. This project aims to perform a differentially private analysis on the *U.S. Household Income Statistics* [1] dataset released by Golden Oak Research Group on Kaggle. By incorporating both Laplace and Gaussian mechanisms, the project will demonstrate how accurate statistical summaries can be produced while providing formal privacy guarantees. In addition, a machine learning–based income predictor will be trained to illustrate how differentially private noise affects downstream model performance and fairness. The analysis focuses on state- and county-level queries such as mean and median income, standard deviation, and inequality ratios under differential privacy. The goal is to evaluate the trade-off between data utility and privacy loss (ε, δ) through empirical experiments, visualizations, and ML-based evaluations.

Keywords

Differential Privacy, Laplace Mechanism, Gaussian Mechanism, U.S. Income Statistics, Data Security, Privacy-Preserving Machine Learning

ACM Reference Format:
Muttaki I Bismoy, Shomitro Ghosh, Farhan Tanvir, and Ahmad Jadallah. 2025. Differentially Private Analysis of U.S. Household Income Statistics using Laplace and Gaussian Mechanisms. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Project Goals

Our primary goal is to quantify the privacy–utility trade-off when performing statistical and machine learning analysis on household income data using differential privacy (DP). Specifically, we aim to:

- Implement Laplace and Gaussian mechanisms to answer aggregate statistical queries under DP.
- Compare their sensitivity, noise calibration, and privacy budget allocation effects on analytical accuracy.
- Extend DP evaluation to a predictive machine learning task (income classification) to assess the downstream effect of privacy noise.
- Visualize, interpret, and compare how privacy levels impact both descriptive and predictive utility.

2 Dataset and Preprocessing

The **U.S. Household Income Statistics** dataset (Kaggle, 2017) contains socio-economic features derived from the U.S. Census Bureau's 2011–2015 ACS 5-Year reports. It includes 325,000+ rows of data containing over 12,232 U.S. cities and 1,957 counties, representing a rich cross-section of income. The dataset provides attributes such as:

- *State, County, City, Latitude, Longitude, Mean, Median, and Stdev of income, Household counts, Land/Water area, and demographic indicators.*

Preprocessing will include:

- Handling missing entries and normalizing numerical features.
- Converting income range data using midpoint interpolation from the official *Income Methodology Report.*
- Aggregating data at multiple levels (county, state and national) to support both macro- and micro-analysis.
- Partitioning the data into training/testing subsets for predictive ML experiments.

3 Proposed Approach

3.1 Queries and Statistical Analysis

The following statistical queries will be analyzed under both standard and DP settings:

- Mean and median household income per state.
- Standard deviation of income distribution per county.
- Gini-like inequality ratio $\frac{\text{Stdev}}{\text{Mean}}$ and regional disparity indices.
- Correlation analysis between income, population, and geographic latitude/longitude.

[1] Dataset: https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Each query will be evaluated with Laplace and Gaussian noise to observe differences in noise sensitivity, bias, and overall data utility.

3.2 Differential Privacy Mechanisms

Given the sensitivity Δf of each query f,

$$\text{Laplace: } M(x) = f(x) + \text{Lap}(\tfrac{\Delta f}{\varepsilon}),$$

$$\text{Gaussian: } M(x) = f(x) + \mathcal{N}(0, \sigma^2),\ \sigma \geq \frac{\sqrt{2\ln(1.25/\delta)}\,\Delta f}{\varepsilon}$$

We will vary ε in the range $[0.05, 5]$ and set $\delta = 10^{-5}$. Accuracy loss, bias, and privacy leakage will be measured at different scales of perturbation.

3.3 Machine Learning Extension

To go beyond class content, the project will train a regression-based and neural-network-based model to predict *median household income* from geographic and demographic attributes. We will employ:

- **Baseline model:** Linear Regression (no DP).
- **Pre-trained model:** Fine-tuned version of a neural regressor (e.g., DP-SGD implemented via TensorFlow Privacy) to demonstrate differential privacy in gradient updates.
- **Evaluation:** Comparison of model accuracy, mean absolute error (MAE), and fairness bias before and after DP training.

This ML extension will highlight the impact of DP noise not only on descriptive statistics but also on the predictive capacity of models trained on privatized data.

3.4 Experimental Workflow

The experimental phases will include:

1. **Data Preparation:** Cleaning, normalization, and feature selection.
2. **Statistical DP Analysis:** Apply Laplace and Gaussian mechanisms to selected queries.
3. **ML Training:** Train and fine-tune models on both original and DP-sanitized datasets using DP-SGD.
4. **Performance Evaluation:** Compare true and DP results using absolute/relative error, MAE, RMSE, and correlation scores.
5. **Visualization:** Generate histograms, scatter plots, and utility-vs-privacy trade-off curves.

4 Work Plan and Division

- **Muttaki Bismoy:** Algorithm and ML pipeline design, DP-SGD integration, technical writing.
- **Shomitro Ghosh:** Data cleaning, exploratory analysis, Laplace mechanism implementation.
- **Farhan Tanvir:** Gaussian mechanism, sensitivity computation, validation, and error analysis.
- **Ahmad Jadallah:** Comparative evaluation, visualization, presentation, and documentation.

5 Expected Outcomes

The project will produce:

- A complete Python-based pipeline integrating statistical DP mechanisms and ML models.
- Quantitative evaluation of Laplace vs. Gaussian mechanisms for real-world income data.
- ML-based demonstration of DP's effect on model utility and fairness.
- Visualizations summarizing privacy-utility trade-offs and per-state error variations.
- A final report and reproducible Jupyter notebook.

6 Conclusion

This project extends beyond the theoretical concepts taught in class by integrating differential privacy with a real-world socio-economic dataset and a pre-trained machine learning model. It aims to empirically demonstrate how differentially private mechanisms influence both statistical accuracy and ML model behavior, offering insight into deploying privacy-preserving analytics in demographic and economic research.

References

[1] Golden Oak Research Group. "U.S. Household Income Statistics with Geo Locations." *Kaggle Dataset*. Available at: https://www.kaggle.com/datasets/goldenoakresearch/us-household-income-stats-geo-locations.

[2] Golden Oak Research Group LLC. "About Golden Oak Research." *Official Website*. Available at: https://www.goldenoakresearch.com/.

[3] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating Noise to Sensitivity in Private Data Analysis." In *Theory of Cryptography Conference (TCC)*, pages 265–284. Springer, 2006. https://link.springer.com/chapter/10.1007/11681878_14.