RUTGERS

THE STATE UNIVERSITY
OF NEW JERSEY

IE 551 Project 1

Median Income US

Deepak Vagish K
dk1156
Nisarg Shah
ns1452
Darshan Senthil
ds1992

March 1, 2024

1 Introduction

Understanding the distribution of income within a population is fundamental for assessing economic well-being and formulating effective policy decisions. Median household income serves as a key metric in this endeavor, offering insights into financial disparities and resource allocation. However, directly analyzing income data for an entire nation like the United States is often impractical.

This project delves into the realm of statistical inference, where the concept of sampling plays a pivotal role. By strategically selecting a representative subset of data, we can gain valuable insights into the broader population. This project focuses on a dataset encompassing median household income for the United States, encompassing national, state, and county levels.

A critical challenge lies in ensuring that the chosen sample accurately reflects the characteristics of the entire population. Sampling error, the inherent discrepancy between sample statistics and population parameters, necessitates the development of robust methods to minimize bias and enhance the generalizability of findings.

2 Objective

The primary objective of this project is to develop techniques for generalizing findings obtained from a sample to a larger population. This involves minimizing the sampling error, which is the discrepancy between the characteristics of the sample and the true population values.

3 Problem Statement

One of the core challenges we encounter is generalizing conclusions from a limited sample to the entire population. Navigating through the complexities of statistical analysis, we strive to ensure that our methodologies effectively capture the underlying characteristics of the dataset. Balancing model complexity with accuracy poses a fundamental dilemma in representing the population parameters faithfully. We endeavor to balance these factors harmoniously to yield robust and reliable insights.

4 Data Collection and Analysis

This research relies on a sample-based approach to analyze income distribution across the United States. To ensure representativeness, we will employ random sampling techniques. Specifically, we will draw 50 random samples from the original dataset of 3200 observations. This approach helps mitigate bias and provides a broader perspective on income distribution compared to analyzing the entire dataset at once.

However, the income distribution often exhibits skewness, with a concentration of values towards the lower and middle income ranges. We will implement a filtering technique based on the Interquartile Range (IQR). We will filter out data points falling outside the range of 1.5 times the IQR below the first quartile (Q1) and 1.5 times the IQR above the third quartile (Q3). This approach effectively removes outliers while preserving the core distribution of the data.

$49K	$100.1K	$63.7K	$54.9K	$54.7K
$52.5K	$78K	$62.2K	$60K	$72.3K
$51K	$42.4K	$68.2K	$43.7K	$83.7K
$75.2K	$73.4K	$73K	$57.6K	$64.6K
$80K	$57.7K	$52.9K	$137.3K	$41.9K
$54.9K	$79.5K	$69.2K	$71.1K	$51.5K
$60.5K	$48.1K	$39.9K	$56.2K	$58.4K
$63.1K	$82.6K	$57.1K	$44.7K	$55.4K
$50.7K	$45.9K	$103.1K	$55.1K	$62.9K
$84.3K	$67.8K	$68.3K	$62.9K	$167.6K

5 Methodology

In addition to fitting probability distributions, we will leverage the CDF (cumulative distribution function) to analyze the income data. The CDF provides the probability that a data point falls below a certain value. By comparing the empirical CDF (derived from the data) with the CDFs of the fitted distributions, we can assess how well each model captures the cumulative distribution of income within each sample.

5.1 Exponential Distribution

A constant rate of occurrence of events characterizes the exponential distribution. This model captures situations where the probability of an event decreases as the time or magnitude of the event increases. It is often used to model the time between events in a Poisson process

5.2 Uniform Distribution

This represents a scenario where all values within a specific range are equally probable. The distribution assigns equal probability to all values within a specified range. The pdf for the uniform distribution is constant within the range and zero elsewhere.

5.3 Lower-Bound Truncated Exponential Distribution

This distribution is an extension of the exponential distribution, introducing a lower bound parameter. The pdf for the 2-parameter exponential distribution is similar to the exponential distribution but includes a shift parameter.

5.4 Normal Distribution

This bell-shaped curve represents the ideal scenario where most incomes cluster around the average, with fewer outliers on either side. It is a common starting point for income analysis, but may not always capture real-world skewness.

5.5 Lognormal Distribution

This model is suitable for data that exhibits positive skewness, where a larger proportion of incomes fall towards the lower end. It can be appropriate for analyzing income distribution, which often shows a concentration of lower and middle-income households.

5.6 Gamma Distribution

This flexible distribution can accommodate a wider range of shapes, from exponential-like decays to bell-shaped curves. It offers more versatility in fitting income data compared to the basic exponential model.

5.7 Beta Distribution

This distribution is useful for modeling data bounded between a minimum and maximum value. It is characterized by two shape parameters, α and β. It could be relevant if there are restrictions on income levels, such as a cap on very high earners.

5.8 Weibull Distribution

This model can capture data with skewed tails, where the probability of extreme values is higher on one side. It might be applicable if the income distribution exhibits a significant number of outliers on either the high or low end. It has two parameters: shape (k) and scale (λ)

5.9 Rayleigh Distribution

This specialized distribution describes situations where data is non-negative and skewed to the right. While less commonly used for income analysis, it could be relevant in specific contexts.

6 Modeling Analysis

6.1 Sum of Squared Error (SSE)

It calculates the squared difference between each predicted value and its corresponding actual value, and then sums these squared differences. A lower SSE indicates a better fit of the model to the data, as it signifies that the predicted values are closer to the actual values.

$$\sum_{i=1}^{D}(x_i - y_i)^2$$

6.2 Mean Squared Error (MSE)

MSE is a measure that quantifies the average squared difference between the estimated values and the actual data points. It provides a comprehensive assessment of the model's accuracy,

where lower MSE values indicate better model performance. Mathematically, MSE is calculated by taking the average of the squared differences between observed and predicted values.

$$\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{n-k}$$

6.3 Akaike Information Criterion (AIC)

AIC is a measure used for model selection, balancing the goodness of fit with the complexity of the model. It penalizes models with higher complexity to avoid overfitting. AIC is calculated based on the likelihood function of the model and the number of parameters. Lower AIC values suggest a better balance between model fit and complexity.

$$-2logL + 2k$$

6.4 Bayesian Information Criterion (BIC)

Similar to AIC, BIC is a criterion for model selection that penalizes complex models to prevent overfitting. However, BIC places a stronger penalty on models with a larger number of parameters. BIC takes into account the sample size and the number of parameters, providing a more stringent measure of model complexity.

$$-2logL + klogn$$

6.5 Pham's Criterion (PC)

Pham's Criterion is another model selection criterion that considers the balance between model complexity and sample size. PC increases the penalty for adding parameters to the model, particularly when dealing with small sample sizes. It aims to prevent overfitting by adjusting the penalty term based on the size of the dataset.

$$SSE + k\frac{n-1}{n-k}$$

6.6 RED

This stands for normalized-Rank Euclidean Distance criteria or RED which selects the best model based on a set of contributing criteria

$$D_i = \sum_{j=1}^{d} \left\{ \left(\sqrt{\left[\sum_{k=1}^{2} \left(\frac{c_{ijk}}{\sum_{i=i}^{s} c_{ijk}} \right)^2 \right] w_j} \right) \right\}$$

7 Results

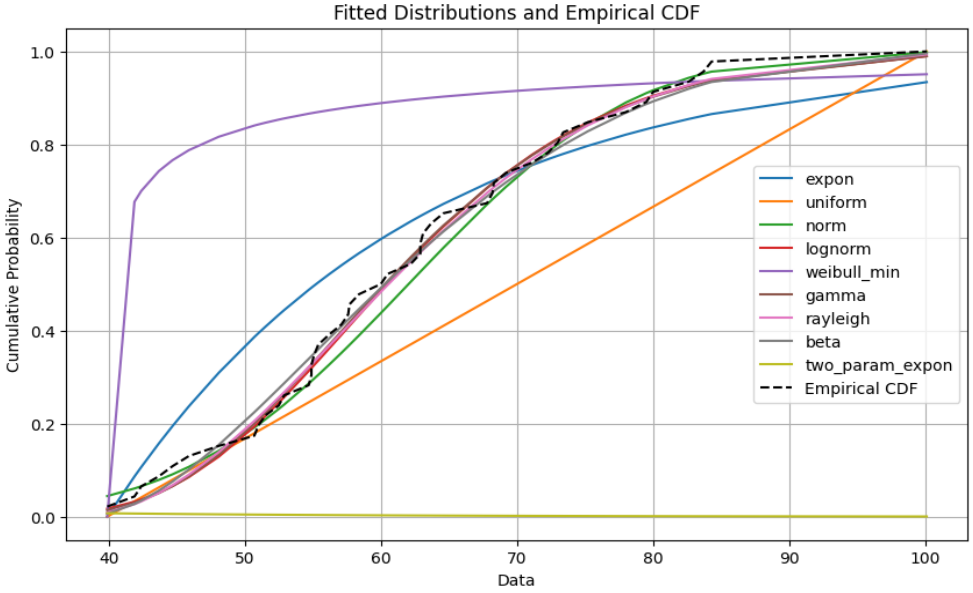

Table 1: Comparison of Model Performance Metrics

Distribution	SSE	MSE	AIC	BIC	PC
expon	0.54	0.01	-218.60	-216.71	-107.05
uniform	1.35	0.03	-171.98	-168.20	-82.36
norm	0.08	0.00	-313.22	-309.43	-150.09
lognorm	0.02	0.00	-368.40	-364.61	-176.56
weibull	8.08	0.17	-84.29	-80.51	-40.30
gamma	0.02	0.00	-372.89	-369.11	-178.71
rayleigh	0.03	0.00	-369.34	-367.45	-180.88
beta	0.03	0.00	-359.89	-356.11	-172.48
two param expon	15.73	0.33	-51.67	-47.89	-24.66

Distribution	SSE Rank	MSE Rank	AIC Rank	BIC Rank	PC Rank
expon	6.00	6.00	6.00	6.00	6.00
uniform	7.00	7.00	7.00	7.00	7.00
norm	5.00	5.00	5.00	5.00	5.00
lognorm	2.00	2.00	3.00	3.00	3.00
weibull	8.00	8.00	8.00	8.00	8.00
gamma	1.00	1.00	1.00	1.00	2.00
rayleigh	3.00	3.00	2.00	2.00	1.00
beta	4.00	4.00	4.00	4.00	4.00
two param expon	9.00	9.00	9.00	9.00	9.00

Table 2: RED metric

Distribution	RED Value	Rank
expon	3.44	6.00
uniform	2.72	7.00
norm	4.88	5.00
lognorm	5.70	3.00
weibull	1.42	8.00
gamma	5.76	1.00
rayleigh	5.73	2.00
beta	5.59	4.00
two param expon	0.98	9.00

8 Conclusion and Findings

Our analysis revealed that the gamma distribution emerged as the most suitable model for capturing the income distribution within our samples. This finding is supported by the highest normalized Euclidean rank score of 5.76 indicating that the gamma distribution achieved the best overall fit compared to other evaluated models. The fitted gamma distribution exhibited a mean of 62.03, a median of 60.29, a variance of 175.55, and a standard deviation of 13.24. There is a 95% confidence level that the true population mean falls within the range of [36.07, 88.01], based on the fitted gamma distribution. These values provide quantitative insights into the central tendency and spread of income within the analyzed samples.