Project Report

STATISTICS FOR MACHINE LEARNING - IE 551

Population vs Sample

Sammed Sunil Admuthe (ssa180)
Shraddha Sanjeev Pattanshetti (sp2304)
Code Available Here

1. Introduction

The project focuses on the critical concepts of population versus samples within the context of statistical analysis for machine learning. Populations encompass the entire set of groups or individuals relevant to a study, while samples are subsets of the population used for analysis. This distinction is pivotal for understanding and applying statistical inferences or hypothesis testing to draw accurate conclusions from sample data about the larger population.

2. Objectives

The primary objective of the research is to critically assess model selection criteria in the context of machine learning and statistical analysis. To choose the best statistical model, it is important to evaluate the usefulness and applicability of several criteria, including the Mean Squared Error (MSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Pham's Criterion (PC). This analysis is essential for reliable data analysis, precise forecasts, and well-informed decision-making in various scientific, engineering, and economic applications.

3. Problem Statements

The problem is to develop statistical techniques that reliably extrapolate these results to the entire population given a sample dataset of median household earnings from different regions of the United States. To provide accurate and dependable findings regarding the median family income across the United States, the work entails minimizing the sampling error, or the difference between the sample statistic and the population parameter. In addition to creative approaches to model and data analysis that are acquired during the course, this calls for the careful application of statistical inference techniques and hypothesis testing.

4. Data Collection

The data for statistical analysis consist of the median household income for the entire United States and each state and county for the recent year. There are approximately 4000 data points in the population. The sample data points comprises 50 random samples drawn from this population.

Sample data consists of following data points :-

69.7	60.3	61.2	56.4	52.8	46.2	44.7	44.0	76.3	73.4
94.7	40.7	81.1	65.0	64.4	67.9	43.9	35.8	87.9	68.5
64.2	46.0	77.0	53.3	67.5	73.0	167.6	76.2	65.2	80.2
47.9	54.0	45.4	96.5	78.8	58.4	56.0	75.5	96.3	54.7
43.7	46.4	56.8	50.0	65.2	77.0	83.5	66.4	55.0	43.7

5. Analysis

The analysis focuses on fitting different probability distributions to a given sample dataset. The objective is to determine which distribution best represents the sample, thereby minimizing the sampling error when generalizing findings to the population. This involves a multi-step process where various statistical distributions are fitted to the data, and their goodness-of-fit is evaluated using criteria such as the Sum of Squared Errors (SSE), Mean Squared Error (MSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Pham's Criterion (PC).

The sample dataset consists of 50 observations believed to represent median household incomes (in thousands of dollars) across different regions. The analysis includes fitting the following distributions: exponential (1-parameter), uniform, normal (Gaussian), log-normal, Weibull, gamma, Rayleigh, beta, and a custom two-parameter exponential distribution.

6. Methodologies

- Data Preparation: We start with a sample dataset of median household incomes, consisting of 50 observations. This dataset represents a small fraction of the population, from which we aim to infer the population's overall characteristics.

- Distribution Fitting: We fit various statistical distributions to our sample data, including exponential, uniform, normal, lognormal, Weibull, gamma, Rayleigh, beta, and a two-parameter exponential distribution. For each distribution, we estimate parameters using both the Method of Moments (MM) and Maximum Likelihood Estimation (MLE) methods, as specified by the calculateExpectedValue(method) function's method argument. Upon applying both Maximum Likelihood Estimation (MLE) and Method of Moments (MM) for parameter estimation across different distributions, the analysis reveals varying degrees of fit. This variance underscores the importance of selecting a fitting method aligned with the nature of the data and the distribution characteristics.

- Criteria Calculation: The goodness-of-fit for each distribution is rigorously assessed using Sum of Squared Errors (SSE), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Pham's Criterion (PC). These criteria serve dual purposes: they quantify the discrepancy between the empirical and theoretical CDFs, and they introduce penalties for models with a higher number of parameters to avoid overfitting. This comprehensive evaluation helps in identifying the distribution that not only fits the sample

data well but also maintains parsimony, striking a balance between complexity and accuracy.

7. Modeling Analysis

The modeling analysis involves evaluating the fitted distributions against the criteria mentioned to determine the best fit. Here's how we approach this:

- Criteria Evaluation: Using the bestFit(criteria_matrix_updated) function, we rank the distributions based on their performance across all criteria. This function calculates the rank for each criterion and sums these ranks to determine an overall score, with lower scores indicating a better fit.

- Model Selection: The culmination of the modeling analysis is the selection of the best-fit model based on the aggregated ranks from the SSE, AIC, BIC, and PC criteria. This selection process is critical for determining which distribution most accurately represents the sample data and, by extension, can be used to infer about the population parameter with minimal sampling error. The model with the lowest total rank across all criteria is deemed the most suitable for generalizing from the sample to the population.

- Expected Value Estimation: We estimate the expected value (mean) of median household income using integrating PDF and Monte Carlo simulation for the best distribution function. This step helps us understand the central tendency of the population based on our sample.

- Plotting Distributions: We visually compare the empirical Cumulative Distribution Function (CDF) of our sample data against the theoretical CDFs of the fitted distributions. This graphical analysis aids in intuitively selecting the distribution that best matches our sample data.

- Implications for Sampling Error Minimization
  The chosen best-fit model has significant implications for sampling error minimization. By accurately representing the sample data and incorporating

penalties for complexity, the model ensures that the extrapolation to the population parameter is both precise and reliable. This not only enhances the validity of the statistical inferences drawn from the sample but also underscores the efficacy of the chosen model in capturing the central tendencies and variability inherent in the population.

8. Results

8.1 Graphs:

In our comprehensive analysis aimed at understanding the distribution characteristics of the sample data, we employed a series of Cumulative Distribution Function (CDF) plots to compare the empirical CDF of the sample against various theoretical distribution functions.

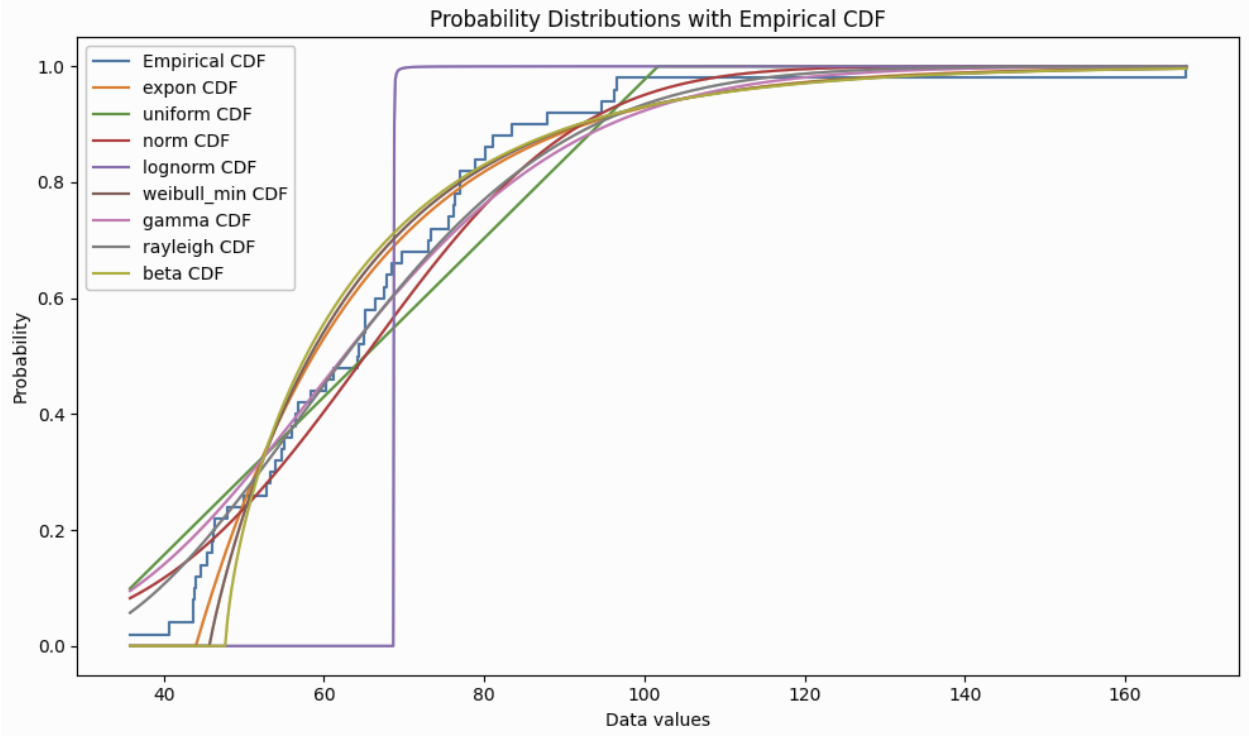

Fig 1: Plot of CDF for distribution functions and empirical CDF (Using MM)

The empirical CDF of the sample data was closely mirrored by the CDFs of most theoretical distributions estimated using MM, with the notable exception of the log-normal distribution. The log-normal distribution's CDF deviated significantly from the empirical CDF, indicating a less satisfactory fit compared to other distributions.

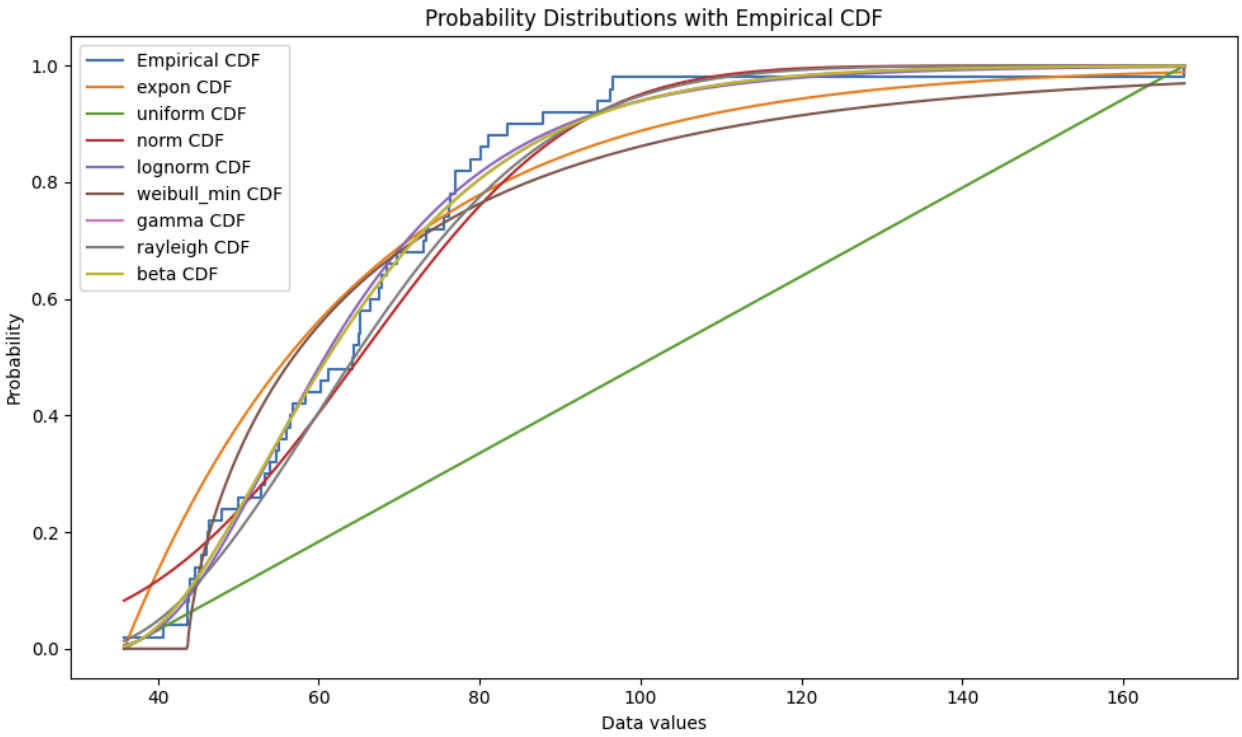

Fig 2: Plot of CDF for distribution functions and empirical CDF (Using MLE)

Similar to the findings with MM, the CDF plots generated using MLE parameters demonstrated a high degree of alignment between the empirical CDF and those of the theoretical distributions, excluding the uniform distribution. The uniform distribution's CDF presented a clear divergence from the empirical data's pattern, indicating a poor fit.

8.2 Ranking Matrix:

A custom ranking table was developed for nine distribution functions, providing a comprehensive comparative analysis.

The MM method's ranking matrix underscored the performance of various distributions but with slight variations in rankings compared to MLE. This difference exemplifies the impact of parameter estimation techniques on the suitability and performance of distribution functions. The gamma distribution outperformed other distributions and managed to secure an overall rank of 5, which is significantly higher than the rest of the distribution fits.

	SSE	MSE	AIC	BIC	PC	SSE_rank	MSE_rank	AIC_rank	BIC_rank	PC_rank	Total_Rank
6	0.113531	0.002271	-300.385070	-296.561024	-144.063167	1.0	1.0	1.0	1.0	1.0	5.0
5	0.171001	0.003420	-277.905562	-272.169493	-130.307955	2.0	2.0	2.0	2.0	4.0	12.0
2	0.193624	0.003872	-273.693110	-269.869064	-131.251026	3.0	3.0	3.0	3.0	2.0	14.0
0	0.200929	0.004019	-271.841252	-268.017206	-130.362134	4.0	4.0	4.0	4.0	3.0	19.0
4	0.308557	0.006171	-248.393542	-242.657473	-116.437305	5.0	5.0	5.0	5.0	5.0	25.0
1	0.386471	0.007729	-239.136058	-235.312012	-114.663641	6.0	6.0	6.0	6.0	6.0	30.0
7	0.413672	0.008273	-231.735213	-224.087121	-106.017329	7.0	7.0	7.0	8.0	8.0	37.0
8	0.468342	0.009367	-229.528984	-225.704938	-110.052246	8.0	8.0	8.0	7.0	7.0	38.0
3	5.606121	0.112122	-103.408196	-97.672127	-48.294192	9.0	9.0	9.0	9.0	9.0	45.0

Table 1: Ranking Matrix for MM

The ranking matrix obtained through MLE revealed distinct preferences for certain distribution functions across the evaluated criteria. However, weibull_min performed significantly better with an overall rank of 5 compared to its counterparts.

	SSE	MSE	AIC	BIC	PC	SSE_rank	MSE_rank	AIC_rank	BIC_rank	PC_rank	Total_Rank
5	0.050707	0.001014	-338.685556	-332.949487	-158.874552	1.0	1.0	1.0	1.0	1.0	5.0
7	0.051025	0.001021	-336.372674	-328.724582	-154.150560	2.0	2.0	2.0	2.0	3.0	11.0
3	0.058714	0.001174	-331.354858	-325.618789	-155.429124	3.0	3.0	3.0	3.0	2.0	14.0
6	0.150933	0.003019	-286.147097	-282.323051	-137.228940	4.0	4.0	4.0	4.0	4.0	20.0
2	0.193624	0.003872	-273.693110	-269.869064	-131.251026	5.0	5.0	5.0	5.0	5.0	25.0
4	0.299427	0.005989	-249.895328	-244.159259	-117.143145	6.0	6.0	6.0	6.0	6.0	30.0
8	0.468342	0.009367	-229.528984	-225.704938	-110.052246	7.0	7.0	7.0	7.0	7.0	35.0
0	0.468342	0.009367	-229.528984	-225.704938	-110.052246	8.0	8.0	8.0	8.0	8.0	40.0
1	5.619457	0.112389	-105.289399	-101.465353	-50.417245	9.0	9.0	9.0	9.0	9.0	45.0

Table 2: Ranking Matrix for MLE

9. Conclusion and Findings

Overall, the study was aimed at identifying the best-fit distribution for our dataset and estimate the expected values and generalize findings from the sample to the entire population.

Using Method of Moments (MM) for parameter estimation

- Through the Method of Moments (MM), the Gamma distribution emerged as the most suitable model for our dataset.
- The expected value calculated using the Gamma distribution's parameters was close to 64.6
- The expected value derived from Monte Carlo simulations for the Gamma model was around 64.806

Using Maximum Likelihood Estimation (MLE) for parameter estimation:

- The Weibull minimum (Weibull_min) distribution was identified as the best-fit model when utilizing the MLE approach.
- The expected value derived directly from the parameters of the Weibull_min distribution was approximately 70.628
- The expected value, as calculated using the Monte Carlo simulation method, was found to be approximately 70.946

Our analysis demonstrates the efficacy of both the Weibull_min and Gamma distributions in modeling our data, as evidenced by the MLE and MM approaches, respectively.