PROJECT REPORT

Analysis of Census Income Dataset:
Imputation, Distributions, and Statistical Testing

Submitted By:
Prabuddha Durge
24BM6JP17
PGDBA Batch 10

Submitted To:
Prof. Amita Pal
Indian Statistical Institute, Kolkata

Course:
Inference

Git Repository:

https://github.com/prabuddhadurge/Inference

Abstract:

This report presents the analysis of a dataset from the UCI Machine Learning Repository, containing both discrete and continuous variables. The exercises include descriptive statistics (frequency distributions, histograms, and box plots) to explore variable distributions, followed by simulation studies on various probability distributions and sample means. Missing data mechanisms and imputation methods are evaluated using RMSE. The report also includes hypothesis testing for median values, confidence intervals, and equality of variances and means. The analysis focuses solely on the observations relevant to the exercises performed.

Date of Submission:
24 Nov 2024

Index

1. Dataset Description and Significance

The **Census Income Dataset**, also known as the "Adult" dataset, was extracted by Barry Becker from the 1994 Census database. It contains **48,842 instances** and **14 attributes**, with a mix of categorical and numerical features. The primary objective is to predict whether an individual's annual income exceeds **$50,000**, based on demographic and employment attributes. This dataset is widely used for classification tasks in machine learning.

The features include personal and professional details, such as:

- **Demographics**: *age*, *sex*, *race*, *native-country*.
- **Education and Work**: *education*, *workclass*, *occupation*, *hours-per-week*.
- **Economic Variables**: *capital-gain*, *capital-loss*.
- **Relationship and Marital Status**: *relationship*, *marital-status*.

Each variable plays a role in understanding socioeconomic factors that influence income levels. For instance, *education-num* quantifies education attainment, while *occupation* and *hours-per-week* highlight employment patterns.

The dataset includes missing values for variables like *workclass*, *occupation*, and *native-country*, which require handling during preprocessing. It is sourced from the UCI Machine Learning Repository and is frequently used to evaluate the performance of algorithms in tasks like income classification.

You can access the dataset at UCI Machine Learning Repository - Adult Dataset. (Note: The dataset is manipulated before being used in the project, which is attached in the mail.)

---

2. Analysis of Occupational and Age Distributions

Objective

This report examines the **distribution of occupations** (*categorical variable*) and **age** (*continuous variable*) from the Census Income Dataset. It includes frequency distribution, visualisations, and statistical insights.

A. Frequency Distribution for Occupation

Occupation	Frequency	Percentage (%)
Prof-specialty	4,038	15.73
Craft-repair	4,030	15.69
Exec-managerial	3,992	15.53
Adm-clerical	3,721	14.50
Sales	3,584	13.96
Other-service	3,212	12.52
Machine-op-inspct	1,966	7.65
Transport-moving	1,572	6.12
Handlers-cleaners	1,350	5.26
Farming-fishing	989	3.85
Tech-support	912	3.55
Protective-serv	644	2.51
Priv-house-serv	143	0.56
Armed-Forces	9	0.04

- **Most Represented Occupations**: *Prof-specialty* (15.73%) and *Craft-repair* (15.69%).
- **Least Represented Occupations**: *Armed-Forces* (0.04%) and *Priv-house-serv* (0.56%).
- **Broad Trends**: Professional and administrative roles dominate the dataset.

B. Visualisations

1. **Histogram for Age**
    - **Method Used**: Bin-width determined using **Sturges' Rule**:
    - **Bin Width = Range/[log2(Number of observations)+1]**
    - Histogram showcases a concentration of individuals aged 20–50 years, with the frequency peaking around the 30s.
2. **Box-and-Whisker Plots**
    - **Occupation**:
      As a categorical variable, a box plot is not meaningful for this variable.
    - **Age**:
      Box plot reveals:
        - Median age around 37 years.
        - Interquartile range (IQR): ~28 to ~45 years.
        - Outliers above 70 years, indicating older individuals.

C. Properties of Distributions

1. **Occupations**:
    - Skewed towards high-frequency categories (*Prof-specialty*, *Craft-repair*, *Exec-managerial*).
    - Low representation in *Armed-Forces* and *Priv-house-serv*.
2. **Age**:
    - Slightly right-skewed, as most individuals are in the younger working-age group.
    - Outliers in the 70+ age group indicate a few older workforce members.

Key Insights

- **Occupational Trends**: High representation in skilled professions suggests a workforce with technical and managerial expertise.
- **Age Dynamics**: The majority in the dataset are in their prime earning years (30s–40s), aligning with income distribution trends.
- **Future Exploration**: Examine how age and occupation interact with income levels to derive socioeconomic insights.

---

3. Report: Analysis of the Distribution of Sample Means

For each dataset:

1.
2. **Histogram of Original Data**: Plotted to visualise the parent distribution.
3. **Sample Means for Varying** $n=10,50,100$
    - 100 independent samples drawn for each n.
    - Means of the samples computed and plotted as histograms.
4. **Visualisation**: Histograms arranged in a 7×4 grid.
    - Row 1–7: Correspond to the datasets.
    - Column 1: Original data distribution.
    - Columns 2–4: Sample mean distributions for $n=10,\ 50,\ 100$

Objective

This report explores the behaviour of sample means for seven datasets under different sample sizes. It analyses how sample means behave as the sample size increases, highlighting the effects of the **Central Limit Theorem (CLT)** across various probability distributions.

Datasets Examined

1. **Occupational Distribution** (*Discrete Variable X from Q2*).
2. **Age Distribution** (*Continuous Variable Y from Q2*).
3. **Simulated Distributions**:
    - **Poisson**: Mean $(\lambda=5)$
    - **Uniform**: Continuous, interval $[3,8]$
    - **Exponential**: $f(x)=7\exp(-7x),\ x>0$.
    - **Beta**: Parameters $\alpha=2,\ \beta=3$
    - **Log-Normal**: Parameters $\mu=0,\ \sigma=1$

Methodology

Observations

1. **General Trends Across All Datasets**:
    - **Variability Reduction**: As n increases ($n=10\rightarrow50\rightarrow100$), the variance of sample means decreases significantly.
    - **Concentration Around the Mean**: Larger samples produce sample means that are more tightly distributed around the population mean.
    - **Normality Approximation**: With larger n, the sample mean distributions become increasingly bell-shaped, validating the **Central Limit Theorem (CLT)**.

2. **Distribution-Specific Observations**:
    - **Occupational Distribution** (Discrete):
        - Original histogram reflects categorical counts.
        - Sample means converge to a stable central value as n grows.
    - **Age Distribution** (Continuous):
        - Right-skewed original data.
        - Sample means become symmetric and normal for $n \geq 50$
    - **Poisson Distribution ($\lambda=5$)**:
        - Original data shows a right-skewed discrete distribution.
        - Sample means transition towards normality for $n=50,100$
    - **Uniform Distribution** $[3,8]$
        - Original histogram is flat.
        - Sample means form a symmetric distribution even for $n=10$
    - **Exponential Distribution ($\lambda=7$)**:
        - Original data is highly skewed.
        - Sample means gradually approximate normality as n increases.
    - **Beta Distribution ($\alpha=2$, $\beta=3$)**:
        - Original histogram is bounded and asymmetric.
        - Sample means show reduced skewness for larger n.
    - **Log-Normal Distribution ($\mu=0$, $\sigma=1$)**:
        - Original histogram is heavily skewed with a long tail.
        - Sample means become nearly normal for $n=100$.

Key Insights

1. **Central Limit Theorem Validation**:
    - Regardless of the parent distribution, the sample mean distributions tend to approximate normality as sample size increases.
    - This convergence is more pronounced for skewed distributions (e.g., Exponential, Log-Normal) and slower for bounded ones (e.g., Beta).
2. **Practical Implications**:
    - For smaller sample sizes ($n=10$), variability in sample means is high, leading to greater uncertainty.
    - Larger samples ($n=50,100$) yield more reliable estimates of population means, reinforcing the importance of sample size in statistical analysis.

---

4. Report on Imputation Methods for Missing Data Mechanisms

Objective

This study evaluates the performance of **Mean Imputation** and **Predictive Mean Matching (PMM)**under three missing data mechanisms:

- **MCAR** (Missing Completely at Random)
- **MAR** (Missing at Random)
- **NMAR** (Not Missing at Random)

The analysis compares the methods based on **Root Mean Square Error (RMSE)** for datasets with **20%** and **30% missingness** in variable X.

Methodology

1. **Data Generation**:
    - **Bivariate Dataset**: $X \sim N(0,1)$, $Y \sim N(6,4)$ 500 observations.
    - Missing data was introduced to X using the **missMethods** package in R, under three mechanisms: MCAR, MAR, and NMAR.
2. **Imputation Methods**:
    - **Mean Imputation**: Missing values in X replaced by the mean of observed values.
    - **Predictive Mean Matching (PMM)**: Missing values imputed by values from observed data that have a predicted mean closest to the predicted mean of the missing value.
3. **Evaluation Criterion**:
    - **RMSE**: Measures the deviation of imputed values from the original complete dataset.

Results

Mechanism	Missing Percentage (%)	RMSE (Mean Imputation)	RMSE (PMM)
MCAR	20	0.8656	1.5166
MCAR	30	0.9815	1.3925
MAR	20	0.9964	1.3285
MAR	30	0.9792	1.4224
NMAR	20	1.7386	2.0684
NMAR	30	1.6549	1.9248

Key Observations

1. **Performance of Imputation Methods**:
    - **Mean Imputation** consistently resulted in **lower RMSE values**, especially under the NMAR mechanism, where data deletion depends on unobserved values of X.
    - **PMM** had **higher RMSE values** overall, particularly for NMAR, though it better preserved relationships between X and Y.
2. **Impact of Missing Data Mechanisms**:
    - **MCAR**: Imputation performed best as missingness was random and unrelated to the data. Both methods had relatively low RMSE values.
    - **MAR**: Imputation became more challenging, as missingness depended on observed Y. This led to slight increases in RMSE for both methods.
    - **NMAR**: The most challenging mechanism, as missingness depended on unobserved values. RMSE values were highest here, especially for PMM.
3. **Impact of Missing Percentage**:
    - Increasing the missing percentage from **20% to 30%** resulted in higher RMSE values for both methods across all mechanisms, reflecting greater imputation difficulty.

Conclusion

- **Mean Imputation** is simple and effective in reducing RMSE, particularly for MCAR and MAR. However, it oversimplifies the data distribution, potentially introducing bias.
- **PMM**, though resulting in higher RMSE, is often preferred in practice as it better preserves the **underlying data structure** and **relationships**.
- The choice of imputation method depends on the **missing data mechanism** and the **analysis objectives**. For NMAR data, advanced methods beyond Mean Imputation and PMM may be needed.

Visualisations

- **Bar Charts**: RMSE comparisons across mechanisms and missing percentages.
- **Scatter Plots**: (X, Y) observations with different colours indicating missing X-values under MCAR, MAR, and NMAR.

5. Report on Binomial Test Analysis

Objective

The goal of this analysis is to evaluate whether the probability (p) of observing values less than or equal to the median (a) in dataset **A** is significantly greater than **0.5** using an exact binomial test.

Analysis Details

1. **Dataset Median (a)**: The median of dataset A is **7**.
2. **Sampling**:
   A simple random sample of size **100** was selected from the dataset.
3. **Sample Observations Below or Equal to Median**:
   - Random Variable (Z): Number of observations in the sample such that $X \leq a$.
   - Observed Value (z): **52** successes.
4. **Hypotheses**:
   - **Null Hypothesis (H_0)**: $p = 0.5$ (true probability of success is 50%).
   - **Alternative Hypothesis (H_1)**: $p > 0.5$ (true probability of success is greater than 50%).
5. **Significance Level (α)**: **0.05**

Results

1. **Probability of Success (Sample Estimate)**:
   - The proportion of successes (observations $X \leq a$) in the sample is:
     $\hat{p} = 52/100 = 0.52$
2. **P-Value**:
   - The p-value for the binomial test is **0.3822**, which is greater than the significance level (0.05).
3. **95% Confidence Interval for p**:
   - Using exact binomial estimation: $[0.4332, 1.0000]$.
   - Alternative estimation method: $[0.4183, 0.6201]$.

Interpretation

1. **Test Results**:
    - Since the p-value (0.3822) is greater than the significance level (0.05), there is **insufficient evidence to reject the null hypothesis** (H_0).
2. **Confidence Interval**:
    - The confidence interval includes 0.5, indicating that the true probability of observing values $X \leq a$ could be equal to or greater than 0.5.
3. **Conclusion**:
    - The analysis does not provide statistically significant evidence to support the claim that the true probability (p) is greater than 0.5.
    - While the observed proportion (0.52) is slightly above 0.5, the deviation is not statistically significant based on this sample.

---

6. Report on Normal Distribution Fitting and Normality Testing

Objective

The objective is to assess whether the variable **education.num (B)** follows a normal distribution by fitting a normal distribution to the data, conducting statistical tests, and visually inspecting the results.

Data and Methodology

1. **Data Description**:
    - The dataset represents numeric values from the variable **education.num (B)**.
2. **Approach**:
    - **Distribution Fitting**: Estimate parameters (mean and standard deviation) using the Maximum Likelihood Estimation (MLE) method.
    - **Normality Testing**: Apply the Shapiro-Wilk test and other goodness-of-fit tests, including the Kolmogorov-Smirnov (KS), Cramer-von Mises, and Anderson-Darling tests.
    - **Visual Analysis**: Overlay the fitted normal distribution on the histogram of the data and inspect deviations.

Results

1. Normal Distribution Fitting

- **Estimated Parameters**:
    - Mean (μ): **12.069**
    - Standard Deviation (σ): **1.944**
- **Parameter Standard Errors**:
    - Mean: **0.087**
    - Standard Deviation: **0.061**
- **Goodness-of-Fit Criteria**:
    - Log-Likelihood: **-1041.738**
    - Akaike Information Criterion (AIC): **2087.476**
    - Bayesian Information Criterion (BIC): **2095.906**

---

2. Shapiro-Wilk Test for Normality

- **Test Statistic (W)**: **0.92561**

- **P-value**: **< 2.2e-16**

Interpretation:

- The p-value is extremely small, leading to the rejection of the null hypothesis (H_0): the data does not follow a normal distribution.

---

3. Goodness-of-Fit Tests

- **Kolmogorov-Smirnov Statistic**: **0.023**
- **Cramer-von Mises Statistic**: **0.048**
- **Anderson-Darling Statistic**: **0.280**

Interpretation:

- The goodness-of-fit statistics suggest minor deviations from normality. However, combined with the Shapiro-Wilk test, these deviations are statistically significant.

Visual Analysis

A histogram of the dataset overlaid with the fitted normal density curve reveals:

- Close alignment between the data and the normal curve for most regions.
- Minor deviations, particularly in the tails, potentially contributing to the rejection of normality in statistical tests.

Conclusion

1. **Normality Assessment**:
    - The Shapiro-Wilk test indicates that the data is not normally distributed, despite the close approximation provided by the fitted normal distribution.
2. **Goodness-of-Fit**:
    - Goodness-of-fit statistics suggest small deviations from normality, which may not substantially impact practical applications unless strict normality assumptions are required.
3. **Recommendations**:
    - If normality is essential for subsequent analyses, consider:
        - Data transformations (e.g., logarithmic or square-root).
        - Alternative methods or models that do not assume normality.
    - Further diagnostics, such as Q-Q plots, can help identify specific deviations.

7. Report on Statistical Analysis of Subsets B1 and B2

Objective

The purpose of this analysis is to evaluate:

- Equality of variances between two randomly divided subsets (**B1** and **B2**) of dataset B.
- Equality of population means between the two subsets.
- A 99% confidence interval for the difference in means.

Data and Methodology

- **Data Description**:
  Dataset B was divided into two subsets (**B1** and **B2**) in a 3:2 ratio. The subsets represent a division of the numeric variable **education.num** for statistical comparison.
- **Methods Used**:
  - **Levene's Test**: To test for equality of variances.
  - **Welch Two-Sample t-Test**: To test for equality of means, accounting for potential unequal variances.
  - **Confidence Interval**: A 99% confidence interval was computed for the difference in population means.

Results

1. Levene’s Test for Equality of Variances

- **Test Statistic (F)**: 3.3522
- **Degrees of Freedom**:
  - Group: 1
  - Total: 498
- **P-value**: 0.06771

Interpretation:

- The p-value is greater than 0.01 (1% significance level), so we fail to reject the null hypothesis of equal variances.
- There is no strong statistical evidence to conclude that the variances of B1 and B2 are unequal.

---

2. Welch Two-Sample t-Test for Equality of Means

- **Test Statistic (t)**: -0.95988
- **Degrees of Freedom (df)**: 395.02
- **P-value**: 0.3377
- **Sample Means**:
    - **B1**: 11.99953
    - **B2**: 12.17366

Interpretation:

- The p-value is greater than 0.01 (1% significance level), so we fail to reject the null hypothesis of equal means.
- There is no significant difference in the means of B1 and B2.

---

3. 99% Confidence Interval for the Difference in Means

- **Confidence Interval**:
- $[-0.531, 0.183]$
- $[-0.531, 0.183]$

Interpretation:

- The interval includes 0, supporting the conclusion that there is no statistically significant difference between the means of B1 and B2.

Conclusion

- **Equality of Variances**:
  - The variances of subsets B1 and B2 are not significantly different at the 1% significance level.
- **Equality of Means**:
  - The means of subsets B1 and B2 are not significantly different based on the Welch t-test.
- **Confidence Interval**:
  - The 99% confidence interval confirms the lack of significant difference between the means of B1 and B2.