Template title

* Statistical Analysis of Real-World Data: A Hypothesis Testing and Clustering Approach

Ciaran Otter
Department of Computer science
Principles of Data Science - CS771
Stellenbosch University, Stellenbosch, South Africa

***Abstract*—**
***Index Terms*—**

I. Introduction

The Adult Income dataset, sourced from the UCI Machine Learning Repository, contains demographic and income-related data for individuals in the United States [1]. The dataset is often used to predict whether a person earns more than $50,000 annually, based on features such as age, education, occupation, hours worked per week, and others. In this project, we aim to apply statistical techniques to analyze the relationship between various demographic factors and income. Specifically, we investigate the following hypotheses:

1) Hypothesis 1: Does education level significantly affect the likelihood of earning more than $50000 per year?
2) Hypothesis 2: Is there a significant difference in hours worked per week between men and women?
3) Hypothesis 3: Does race significantly affect the likelihood of earning more than $50000 per year?

Each hypothesis is tested using appropriate statistical models. Additionally, unsupervised learning techniques, including K-Means, DBSCAN, and Hierarchical Clustering, are applied to cluster individuals based on various features and explore patterns in the dataset.

II. Methodology

A. Hypothesis Testing

1) Hypothesis 1: Does Education Level Affect Income?:

- Null Hypothesis (H_0): Education level does not significantly affect the likelihood of earning more than $50000 per year.
- Alternative Hypothesis (H_1): Education level significantly affects the likelihood of earning more than $50000 per year.

To test this hypothesis, we performed a chi-square test of independence. Since both education level and income are categorical variables, the chi-square test is appropriate to evaluate whether there is a statistically significant association between these variables. Additionally, a logistic regression model was used to assess the predictive power of education level on income.

2) Hypothesis 2: Does Gender Affect Hours Worked per Week?:

- Null Hypothesis (H_0): There is no significant difference in the number of hours worked per week between men and women.
- Alternative Hypothesis (H_1): There is a significant difference in the number of hours worked per week between men and women.

An independent t-test was employed to compare the mean hours worked by men and women. This test is suitable because it compares the means of two independent groups, assuming the hours worked are normally distributed within each group.

III. Results

education	¡=50K	¿50K
10th	761	59
11th	989	59
12th	348	29
1st-4th	145	6
5th-6th	276	12
7th-8th	522	35
9th	430	25
Assoc-acdm	752	256
Assoc-voc	963	344
Bachelors	2918	2126
Doctorate	95	280
HS-grad	8223	1617
Masters	709	918
Preschool	45	0
Prof-school	136	406
Some-college	5342	1336

	chi squared value	P-value
Results	4070.381622	0.000000

IV. Conclusion

References

[1] B. Becker and R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, DOI: https://doi.org/10.24432/C5XW20.