The Impact of Demographics on Employment and Salary

A Hypothesis Testing and Clustering Approach

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

I. Introduction and Project Goal

This project investigates the relationship between demographic factors such as highest education level, age, gender, and ethnicity with employment status and salary in South Africa. Using real-world 2023 household survey data found through stats SA's interactive data portal: 'Nesstar' [1], his investigation aims to conduct statistical hypothesis testing to ascertain whether these factors significantly influence employment outcomes and salary levels in the contemporary South African context. Additionally, clustering techniques will be employed to uncover underlying patterns within the dataset.

The dataset comprises 112 variables, providing a comprehensive overview of household statistics within the South African population. However, for this analysis, we will focus on several key variables that are particularly relevant to our research objectives:

- **Education level**
- **Employment status**
- **Salary**
- **Age**
- **Gender**
- **Ethnicity**

A. Hypotheses

To guide our investigation, we have formulated the following hypotheses regarding the influence of demographic factors on employment status and salary:

B. Hypothesis 1: Education and Employment Status

Research Question: Does an individual's highest level of education influence their employment status?

- **Null Hypothesis (H_0):** There is no significant relationship between education level and the likelihood of employment.
- **Alternative Hypothesis (H_1):** There is a statistically significant relationship between education level and the likelihood of employment.

C. Hypothesis 2: Education and Salary

Research Question: Does an individual's level of education have a significant impact on their salary?

- **Null Hypothesis (H_0):** Education level does not significantly influence salary.
- **Alternative Hypothesis (H_1):** Education level has a statistically significant effect on salary.

D. Hypothesis 3: Age and Salary

Research Question: Does age affect the salary of an employed individual?

- **Null Hypothesis (H_0):** Age has no significant effect on salary compensation.
- **Alternative Hypothesis (H_1):** Age significantly influences salary compensation.

E. Hypothesis 4: Gender and Salary/Education Level

Research Question: Does gender significantly affect salary and education level?

Salary:

- **Null Hypothesis (H_0):** Gender has no significant impact on salary.
- **Alternative Hypothesis (H_1):** Gender has a statistically significant impact on salary.

Education Level:

- **Null Hypothesis (H_0):** Gender does not significantly affect education level.
- **Alternative Hypothesis (H_1):** Gender significantly affects education level.

F. Hypothesis 5: Ethnicity and Salary

Research Question: Does ethnicity influence salary, controlling for education level?

- **Null Hypothesis (H_0):** There are no significant differences in salary based on ethnicity when controlling for education level.
- **Alternative Hypothesis (H_1):** There are statistically significant differences in salary based on ethnicity, even when controlling for education level.

II. Methodology and Implementation

A. Testing Methods

A variety of statistical tests were employed to evaluate the impact of demographic factors on employment and salary outcomes. The methods applied are outlined below:

a) χ^2 Test for Independence: The χ^2 test evaluates the association between two categorical variables. This test was applied to assess relationships between variables such as **education level**, **employment status**, **gender**, and **ethnicity**. It determines whether there is a significant dependence between these demographic factors and employment outcomes.

b) T-Test, ANOVA, and Welch's Test: For continuous variables like **salary** and **age**, t-tests and one-way ANOVA were used to compare means across categorical groups. ANOVA was utilized to identify statistically significant differences in means across three or more independent groups, particularly in analyzing relationships between factors like **education level** and **ethnicity**.

c) Q-Q Plot: A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, typically the normal distribution. By plotting the quantiles of the sample data against the quantiles of the normal distribution, the Q-Q plot visually evaluates how well the data aligns with the theoretical distribution. Deviations from the straight line in the Q-Q plot indicate departures from normality, which can affect the validity of parametric tests.

d) Levene's Test: Levene's test is used to assess the equality of variances across different groups. It evaluates the null hypothesis that the variances are equal between the groups. This test is especially useful in situations where the assumption of homogeneity of variances is critical, such as in ANOVA. If Levene's test shows significant differences in variances, alternative statistical tests like Welch's ANOVA, which do not assume equal variances, may be more appropriate. When comparing the means of two groups, Welch's t-test was used. Unlike the traditional t-test, Welch's test does not assume equal variances, making it more robust in cases where group variances differ. This approach ensures greater accuracy in testing for significant differences between group means, particularly when the assumption of homogeneity of variance is violated.

e) Pearson's Correlation: Pearson's correlation coefficient (denoted as *r*) measures the linear relationship between two continuous variables. It ranges from -1 to 1, where values close to 1 indicate a strong positive linear relationship, values close to -1 indicate a strong negative linear relationship, and values near 0 suggest no linear correlation. Pearson's correlation is commonly used to quantify the strength and direction of relationships between variables, such as age and salary.

f) Tukey's Honest Significant Difference (HSD) Test: Tukey's HSD test is a post-hoc analysis used after an ANOVA to determine which specific groups' means are significantly different from each other. It controls for Type I error when making multiple comparisons, ensuring that the overall significance level is maintained. This test is useful when comparing multiple group means, such as examining whether different education categories have significantly different average salaries.

B. Hypothesis Testing

1) Hypothesis 5: Ethnicity and Salary:

a) Data Preparation: The data is grouped by **Ethnicity**, focusing on the **Salary**.

b) Q-Q Test: A Q-Q test was conducted to assess the normality of salary distributions within each ethnicity group, considering potential non-homogeneity of variance.

c) Kruskal-Wallis Test: The Kruskal-Wallis test is applied to determine the H statistic, which is a measure of the variance between the ranks of the data points in different groups. After computing the H statistic and the corresponding p-value, the results are compared to $\alpha = 0.05$. If the p-value is less than α, the null hypothesis is rejected, indicating that there is a significant difference in the medians of the groups.

d) Dunn's Test: Dunn's test with Bonferroni correction was then applied to the dataset to further explore the p-value and the relationship between **Ethnicity** and **Salary**.

2) Hypothesis 2: Education and Salary:

a) Data Preparation: The dataset was analyzed for outliers, revealing a significant outlier that caused considerable skewness during analysis. To address this issue, the 0.1% and 99.9% quantiles were dropped from the dataset. Additionally, a variable 'lab_amount' was included, indicating whether an individual was willing to disclose their salary. The dataset was filtered to include only the observations where this value was true.

b) Levene's Test: A Levene's test was conducted to assess the homogeneity of variances, ensuring the appropriateness of ANOVA tests for the data.

c) Welch's Test: A Welch's test was performed to evaluate the hypothesis. If the resultant p-value is less than $\alpha = 0.05$, the null hypothesis can be rejected, indicating that education level has a statistically significant impact on salary.

d) Tukey's Multiple Comparison of Means: As a post-hoc analysis, Tukey's Honestly Significant Difference (HSD) test was employed to further investigate the validity of the impact of education on salary at a group-specific level. This combined analysis, utilizing the Welch test with support from Tukey's HSD via `statsmodels`, allowed for a comprehensive examination of education's impact on salary, thus providing robust statistical evidence to support our hypothesis testing.

3) Hypothesis 3: Age and Salary:

a) Data Preparation: The dataset was prepared in a manner consistent with previous hypothesis tests, ensuring data integrity and suitability for analysis.

b) Pearson Correlation: Pearson's correlation coefficient was calculated to initially assess the strength and direction of the linear relationship between **Age** and **Salary**.

c) OLS Regression Test: The data underwent preprocessing using a StandardScaler to standardize the **Age** variable. Subsequently, **Age** was designated as the primary predictor in the Ordinary Least Squares (OLS) regression model aimed at predicting individual salary.

To address potential heteroscedasticity, robust standard errors were implemented. A robust covariance model was em-

ployed to obtain robust covariance estimates, specifically using the HC1 covariance type, which adjusts for small sample sizes.

The coefficients of the model were estimated via the OLS method. The statistical significance of the coefficients was evaluated through p-values, with a threshold of $p < 0.05$ indicating significant relationships.

This methodology facilitated a comprehensive analysis of the relationship between age and salary, yielding insights critical for understanding socio-economic dynamics within the labor market.

4) Hypothesis 4: Gender and Salary/Education Level:

a) Data Preparation: The dataset contained unspecified values in the **Gender** field, which were removed. The data was subsequently divided into two groups based on gender, focusing on salary analysis.

b) Q-Q Test: A Q-Q test was conducted to assess the normality of salary distributions within each gender group, considering potential non-homogeneity of variance.

c) Mann-Whitney U Test: Following the Q-Q test, the Mann-Whitney U test was applied to compare salary distributions between genders. This non-parametric test evaluates whether the ranks of salary values differ significantly between the

5) Hypothesis 5: Ethnicity and Salary:

a) Data Preparation: The data is grouped by **Ethnicity**, focusing on the **Salary**.

b) Q-Q Test: A Q-Q test was conducted to assess the normality of salary distributions within each ethnicity group, considering potential non-homogeneity of variance.

c) Kruskal-Wallis Test: The Kruskal-Wallis test is applied to determine the H statistic, which is a measure of the variance between the ranks of the data points in different groups. After computing the H statistic and the corresponding p-value, the results are compared to $\alpha = 0.05$. If the p-value is less than α, the null hypothesis is rejected, indicating that there is a significant difference in the medians of the groups.

d) Dunn's Test: Dunn's test with Bonferroni correction was then applied to the dataset to further explore the p-value and the relationship between **Ethnicity** and **Salary**.

III. Results

A. Hypothesis Testing

B. Clustering

References

[1] Statistics South Africa, "General household survey 2023 (person file)," identification Number: GHS-2023-PERSON. Accessed via the Stats SA Nesstar interactive data portal. [Online]. Available: http://nesstar.statssa.gov.za:8282/webview/index.jsp?v=2&submode=section&study=http%3A%2F%2F10.131.152.188%3A8282%2Fobj%2FfStudy%2FGHS-2023-PERSON&section=http%3A%2F%2F10.131.152.188%3A8282%2Fobj%2FfStudy%2FGHS-2023-PERSON&mode=documentation&top=yes