Stroke Prediction: Inferential Statistical Analysis of Risk Factors

Course: 21AIC401T Inferential Statistics and Predictive Analytics

Name: Jeevas U

Roll No: RA2211047010010

Abstract:

This study investigates risk factors for stroke using inferential statistics on the publicly available Kaggle Stroke Prediction dataset (n = 4909 after cleaning). The objective was to assess whether critical clinical indicators differ significantly from medical thresholds or across population subgroups. Three types of hypothesis tests were applied: a one-sample t-test, a two-sample independent t-test, and a one-way ANOVA with Tukey HSD post-hoc comparisons.

Findings revealed that average glucose levels were significantly higher than the WHO threshold, indicating elevated metabolic risk. However, no significant BMI difference was observed between smokers and non-smokers. Age differences across work type categories were statistically significant, with Tukey analysis identifying distinct subgroup variations. These results emphasize the role of glucose level and age-related demographic factors in stroke risk, while BMI differences due to smoking status were less conclusive.

Introduction:

Stroke remains one of the leading causes of morbidity and mortality worldwide. Identifying and quantifying risk factors such as blood glucose, BMI, smoking status, and occupational exposure is essential for effective prevention strategies. Inferential statistics provide robust tools to test whether these observed indicators differ significantly from clinical standards.

This study leverages the Kaggle Stroke Prediction dataset to analyze stroke-related risk factors using hypothesis testing methods. The aim is to assess whether glucose levels deviate from WHO guidelines, whether BMI differs between smokers and non-smokers, and whether age varies significantly across different occupational categories.

Dataset Description:

The dataset consists of 4909 anonymized patient records (after cleaning missing BMI values). Key features include:

- Demographics: Age, gender, residence type, work type, marital status
- Clinical Indicators: Hypertension, heart disease, average glucose level, BMI
- Lifestyle: Smoking status
- Outcome Variable: Stroke occurrence (binary: 0 = No stroke, 1 = Stroke)

Hypotheses & Methods:

One-sample t-test

H_0: Mean average glucose level = 110 (WHO threshold)
H_1: Mean average glucose level ≠ 110

Two-sample t-test

H_0: Mean BMI is equal between smokers and non-smokers
H_1: Mean BMI differs between smokers and non-smokers

One-way ANOVA + Tukey HSD

H_0: Mean age is equal across work type groups (Private, Govt_job, Self-employed)
H_1: At least one work type group differs in mean age

Results:

One-sample t-test: Avg. Glucose vs. WHO Threshold

Test Value: 110
t-statistic = -7.40
p-value < 0.001
Interpretation: The mean glucose level is significantly different (and higher) than the WHO guideline.

Two-sample t-test: BMI (Smokers vs. Non-smokers)

t-statistic = 1.77
p-value = 0.077 (> 0.05)
Interpretation: No statistically significant difference in BMI between smokers and non-smokers.

One-way ANOVA: Age across Work Types

F = 214.71, p < 0.001
Interpretation: Age varies significantly across different work categories.

Tukey HSD Post-hoc Test:

Govt_job vs Private: Mean difference = -5.51, $p < 0.001$ → Significant
Govt_job vs Self-employed: Mean difference = 9.20, $p < 0.001$ → Significant
Private vs Self-employed: Mean difference = 14.71, $p < 0.001$ → Significant
Interpretation: All three work types differ significantly in mean age.

Discussion:

The analysis indicates that elevated glucose levels are a strong risk factor for stroke, consistent with known associations between diabetes and cerebrovascular events. Interestingly, BMI did not significantly differ between smokers and non-smokers, suggesting that smoking may impact stroke risk through pathways other than BMI. Significant age differences across work types highlight the role of demographic and occupational variations in stroke vulnerability.

Limitations:

- Cross-sectional dataset; no causal inference
- Missing data on BMI reduced sample size
- Limited variables analyzed; other factors (diet, exercise, stress) not included

Conclusion:

This case study demonstrates the usefulness of inferential statistics in healthcare analytics for stroke prediction. Elevated glucose levels emerged as a key clinical marker, while age varied significantly across work types, offering insights into demographic stroke risks. BMI differences by smoking status were inconclusive. These findings support the integration of statistical testing in predictive modeling and healthcare decision-making to mitigate stroke risk.