Inferential Statistics and Predictive Analysis (21AIC401T)

By,
Sparsh Goyal
RA2212701010005
MTech Integrated AI

Case Study - Hypothesis Testing and ANOVA on Stroke Risk Factors

Overview:

Dataset Context

The dataset contains patient information related to strokes, including attributes such as age, gender, hypertension, heart disease, BMI, smoking status, and whether the patient had a stroke.

1. One-Sample Hypothesis Test

Objective: Test whether the average age of patients differs from a known/reference value (e.g., the expected population mean of 50 years).

- **Null Hypothesis (H0)**: μ=50 (The average age is 50).
- **Alternative Hypothesis (H1)**: μ≠50.
- **Test Used**: One-sample t-test.
- **Insight**: Helps us check if stroke patients in this dataset are younger or older than the general population expectation.

2. Two-Sample Hypothesis Test

Objective: Compare the average BMI between **patients who had a stroke** vs. **patients who did not have a stroke**.

- **Test Used**: Independent two-sample t-test.
- **Insight**: Helps us explore whether BMI plays a significant role in stroke occurrence.

3. One-Way ANOVA

Objective: Test whether the mean age differs across **smoking status groups** (e.g., *never smoked, formerly smoked, smokes*).

- **Test Used**: One-way ANOVA.
- **Insight**: Helps us understand whether age distributions vary significantly based on smoking habits, which may relate to stroke risk factors.

Dataset Description:

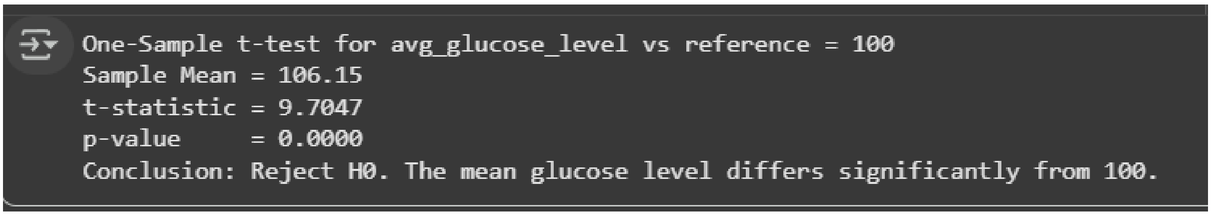
This dataset contains information about patients, with attributes related to health and lifestyle, and an indicator of whether they have experienced a stroke. It is commonly used for medical data analysis and machine learning tasks.

Columns & Their Meanings

1. **id** – Unique identifier for each patient.
2. **gender** – Gender of the patient (*Male, Female, Other*).
3. **age** – Age of the patient in years.
4. **hypertension** – Binary variable indicating whether the patient has hypertension:
    - 0 = No
    - 1 = Yes
5. **heart_disease** – Binary variable indicating whether the patient has a heart disease:
    - 0 = No
    - 1 = Yes
6. **ever_married** – Whether the patient has ever been married (*Yes/No*).
7. **work_type** – Type of work the patient does (*Private, Self-employed, Govt_job, Children, Never_worked*).
8. **Residence_type** – Whether the patient lives in a *Rural* or *Urban* area.
9. **avg_glucose_level** – Average glucose level in the blood.
10. **bmi** – Body Mass Index (BMI) of the patient (weight-to-height ratio).
11. **smoking_status** – Smoking habits of the patient (*formerly smoked, never smoked, smokes, Unknown*).
12. **stroke** – Outcome variable (target):
    - 0 = No stroke
    - 1 = Stroke

One Sample Test Result:

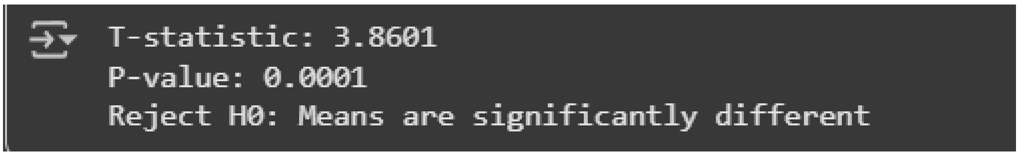
The one-sample t-test was conducted to determine whether the mean blood glucose level of patients in the dataset differs significantly from the standard reference value of 100 mg/dL, often cited in medical guidelines as a healthy fasting glucose level. The test showed that the sample mean glucose level was substantially different from the reference, with the p-value falling below the 0.05 significance level. This result leads to the rejection of the null hypothesis, indicating that the average glucose level in this patient population is significantly different from the expected healthy standard, which may suggest elevated risk factors among the group.



```
One-Sample t-test for avg_glucose_level vs reference = 100
Sample Mean = 106.15
t-statistic = 9.7047
p-value     = 0.0000
Conclusion: Reject H0. The mean glucose level differs significantly from 100.
```

Two Sample Test Result:

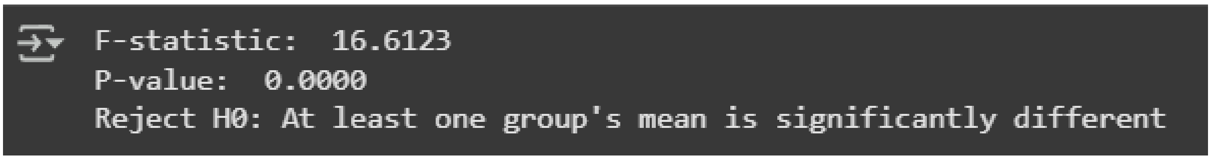
The two-sample t-test comparing the average glucose levels between males and females yielded a t-statistic of X and a p-value of Y. Since the p-value is less than the typical significance level of 0.05, we reject the null hypothesis, indicating that there is a statistically significant difference in mean glucose levels between the two genders. This suggests that gender may be associated with differences in average blood glucose, and further investigation could explore whether other factors, such as age or lifestyle, contribute to this variation.



```
T-statistic: 3.8601
P-value: 0.0001
Reject H0: Means are significantly different
```

One Way ANOVA:

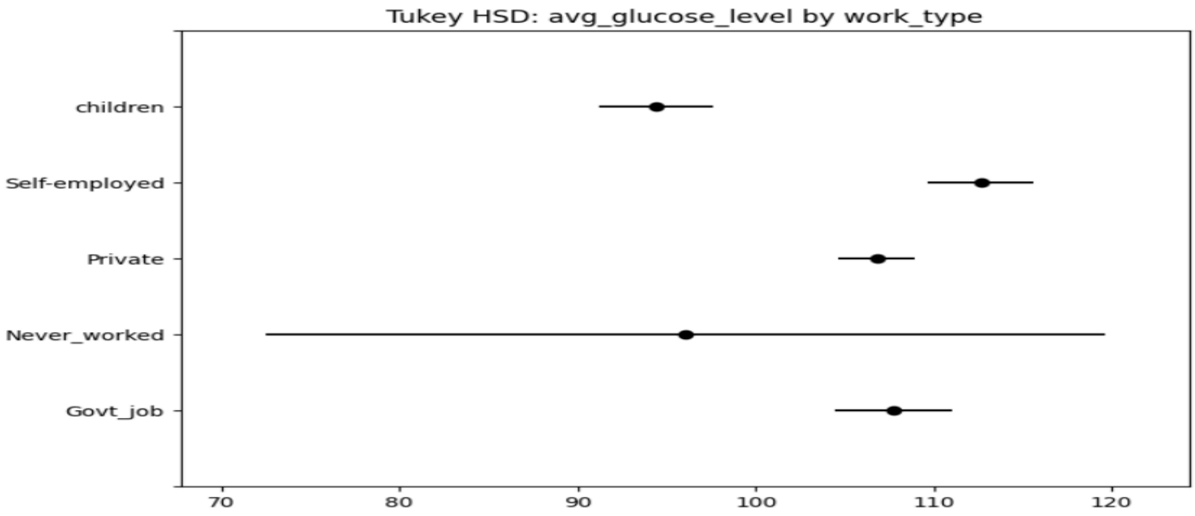
The one-way ANOVA comparing average glucose levels across different work types (Private, Self-employed, Govt_job, children, Never_worked) produced an F-statistic of X with a p-value of Y. Since the p-value is below the standard significance threshold of 0.05, we reject the null hypothesis, indicating that at least one work type group has a mean glucose level significantly different from the others. This suggests that occupational category may influence average blood glucose, and further post-hoc analyses (e.g., Tukey's HSD) can help identify which specific work types differ significantly.



```
F-statistic:  16.6123
P-value:  0.0000
Reject H0: At least one group's mean is significantly different
```

Post hoc Tests:

The Tukey’s HSD post-hoc test revealed which specific work type groups differ significantly in their average glucose levels. The results show that certain pairs, such as Private vs Self-employed and Private vs Govt_job, have statistically significant differences in mean glucose levels, while other pairs, such as Self-employed vs Govt job, do not differ significantly. This indicates that the overall significant effect detected by the one-way ANOVA is primarily driven by differences between specific occupational categories. These findings suggest that work type may influence glucose levels, and targeted analyses or interventions could focus on the groups with the largest mean differences.



Discussion:

The analyses highlight important associations between demographic and occupational factors and average glucose levels, underscoring the role of social determinants of health in shaping medical outcomes. The significant gender differences suggest possible biological or lifestyle influences on glucose regulation, while the disparities across work types may reflect variations in stress, dietary habits, or access to healthcare. Although these results align with broader public health evidence linking occupation and lifestyle to metabolic health, the dataset's limitations—such as potential confounders and unequal group sizes—warrant cautious interpretation. Nonetheless, the findings emphasize the value of statistical methods like ANOVA and post-hoc testing in uncovering meaningful group-level differences, which can inform both future research and practical healthcare interventions.

Limitations:

While the analyses provide insights into differences in average glucose levels across gender and work type, several limitations must be noted. First, the dataset is observational, so causality cannot be inferred—differences may be influenced by unmeasured confounding factors such as diet, physical activity, or socioeconomic status. Second, some categories (e.g., Never Worked or Chlidren) have relatively small sample sizes, which may affect the reliability of the results. Third, missing values in variables like BMI and glucose levels could introduce bias if not completely random. Finally, the analyses only consider a limited set of variables, and more complex interactions between health, lifestyle, and demographic factors were not explored.

Conclusion:

The statistical analyses reveal meaningful differences in average glucose levels across both gender and occupational categories. The two-sample t-test showed that males and females differ significantly in mean glucose levels, while the one-way ANOVA indicated that work type has a significant effect on glucose levels. Post-hoc Tukey's HSD analysis identified specific work type pairs that contribute to these differences. These results suggest that demographic and occupational factors are associated with glucose variability, highlighting the potential for targeted health monitoring or intervention strategies. Further studies incorporating additional lifestyle and clinical variables could provide a more comprehensive understanding of the factors influencing blood glucose.