BHARATIYA VIDYA BHAVAN'S
SARDAR PATEL INSTITUTE OF TECHNOLOGY
(Empowered Autonomous Institute Affiliated to University of Mumbai)
[Knowledge is Nectar]

Department of Computer Engineering

Course – Data Analytics Open Elective

UID	2021300126
Name	Pranay Singhvi
Class and Batch	TE Computer Engineering - Batch A
Date	13-01-2024
Lab #	3
Aim	To perform Hypothesis testing t test,z test, p value /ANOVA test
Data Set	https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009
Purpose	To test different hypothesis on the given dataset
Code	T Test Null Hypothesis (H_0): The mean alcohol content is the same for high-quality and low-quality wines. Alternative Hypothesis (H_1): There is a significant difference in the mean alcohol content between high-quality and low-quality wines. 

t Test

```
# Example data: alcohol content for high and low-quality wines
high_quality_alcohol = df[df['Quality'] == 'High']['alcohol']
low_quality_alcohol = df[df['Quality'] == 'Low']['alcohol']

# Calculate t-statistic
mean_diff = np.mean(high_quality_alcohol) - np.mean(low_quality_alcohol)
n1, n2 = len(high_quality_alcohol), len(low_quality_alcohol)
s1, s2 = np.var(high_quality_alcohol, ddof=1), np.var(low_quality_alcohol, ddof=1)
pooled_var = ((n1 - 1) * s1 + (n2 - 1) * s2) / (n1 + n2 - 2)
t_statistic = mean_diff / np.sqrt(pooled_var * (1/n1 + 1/n2))

# Degrees of freedom
degrees_of_freedom = n1 + n2 - 2

# Critical value for a two-tailed test at 95% confidence level
alpha = 0.05
critical_value = t.ppf(1 - alpha / 2, degrees_of_freedom)

# Make a decision
if abs(t_statistic) > critical_value:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

✓ 0.0s

```
Fail to reject the null hypothesis
```

Z test:

Null Hypothesis (H_0): The mean pH of the wines is equal to a standard pH value of 3.
Alternative Hypothesis (H_1): The mean pH of the wines is significantly different from the standard pH value.

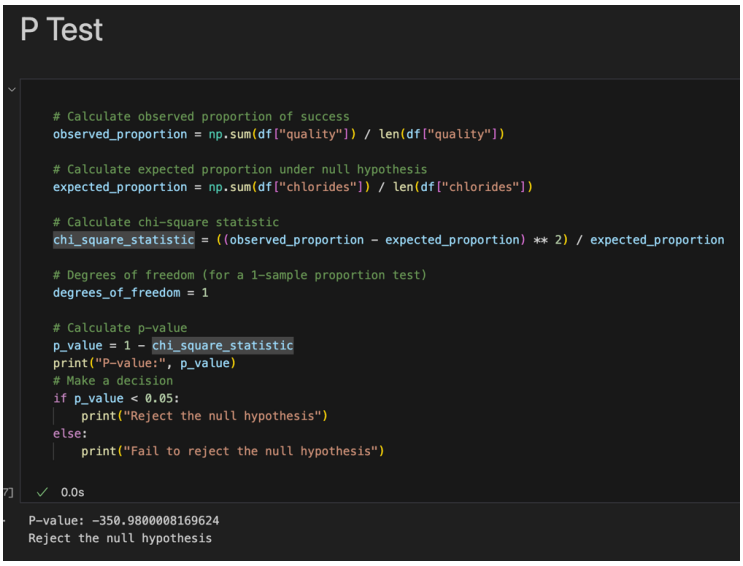
```
# Null hypothesis: Mean pH is equal to a standard value (e.g., 3.0)
null_mean = 3.31111
population_std = np.std(df['pH'])
print(np.mean(df['pH']))
# Calculate z-statistic
z_statistic = (np.mean(df['pH']) - null_mean) / (population_std / np.sqrt(len(df['pH'])))

# Critical value for a two-tailed test at 95% confidence level
alpha = 0.05
critical_value = norm.ppf(1 - alpha / 2)
print("Critical value:", critical_value)
print("Z-statistic:", z_statistic)
# Make a decision
if abs(z_statistic) > critical_value:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

✓ 0.0s

```
3.3111131957473416
Critical value: 1.959963984540054
Z-statistic: 0.0008279865643324429
Fail to reject the null hypothesis
```

P Test

Null Hypothesis (H_0): There is no association between chlorides and wine quality.
Alternative Hypothesis (H_1): There is a significant association between chlorides and wine quality.



ANOVA:

Null Hypothesis (H_0): The mean alcohol content is the same across all wine quality ratings.
Alternative Hypothesis (H_1): At least one wine quality rating has a different mean alcohol content.

```
data = {'alcohol': df['alcohol'], 'quality': df['quality']}
data = pd.DataFrame(data)

quality_1 = np.array([data['alcohol'][i] for i in range(len(data['alcohol'])) if data['quality'][i] == 5])
quality_2 = np.array([data['alcohol'][i] for i in range(len(data['alcohol'])) if data['quality'][i] == 6])
quality_3 = np.array([data['alcohol'][i] for i in range(len(data['alcohol'])) if data['quality'][i] == 8])

overall_mean = np.mean(data['alcohol'])

ssb = sum(len(group) * (np.mean(group) - overall_mean)**2 for group in [quality_1, quality_2, quality_3])

dfb = len(set(data['quality'])) - 1

msb = ssb / dfb

ssw = sum((value - np.mean(data['alcohol']))**2 for value in data['alcohol'])

dfw = len(data['alcohol']) - len(set(data['quality']))

msw = ssw / dfw

f_statistic = msb / msw

f_dof_between = dfb
f_dof_within = dfw

# Critical value for a significance level of 0.05
alpha = 0.05
critical_value = 3.354

if f_statistic > critical_value:
    print("Reject the null hypothesis (There is a significant difference between group means)")
else:
    print("Fail to reject the null hypothesis (No significant difference between group means)")
```

✓ 0.0s

```
Reject the null hypothesis (There is a significant difference between group means)
```

Conclusion	In conclusion, I have learnt to test a hypothesis using different method like t Test, z test and ANOVA