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

Theory

What is Hypothesis Testing?

- Hypothesis testing is a systematic process for evaluating claims or assumptions about a population based on evidence from a sample.
- It involves testing the plausibility of a statement (hypothesis) about a population parameter, such as the population mean or proportion.

Key Components

1. **Null Hypothesis (H_0):** This is the initial assumption you are trying to challenge or examine. It usually represents a statement of "no effect" or "no difference." For example:
    - "The average weight of apples from a new orchard is the same as the national average."
    - "There is no relationship between exercise frequency and stress levels."
2. **Alternative Hypothesis (H_1 or Ha):** This is the statement you will consider supporting if you find enough evidence to reject the null hypothesis. It's often what you aim to demonstrate through your research. For example:
    - "The average weight of apples from the new orchard is greater than the national average."
    - "There is a negative relationship between exercise frequency and stress levels."

3. **Test Statistic:** A numerical value calculated from your sample data, used to compare against a theoretical distribution. Some common test statistics include:
   - **Z-score:** For testing hypotheses about means (when population standard deviation is known)
   - **T-score:** For testing hypotheses about means (when population standard deviation is unknown)
   - **Chi-square statistic:** For testing relationships between categorical variables
4. **Significance Level (α):** A pre-determined threshold of error you are willing to accept in rejecting the null hypothesis when it might be true (called a Type I error). Typical values for α are 0.05 (5%) or 0.01 (1%).
5. **P-value:** The probability of getting a test statistic as extreme or more extreme than the one you observed from your sample data, assuming the null hypothesis is true.

Steps in Hypothesis Testing

1. **State Hypotheses:** Formulate your null and alternative hypotheses clearly.
2. **Set the Significance Level:** Choose your α (usually 0.05).
3. **Select a Test Statistic:** Determine the appropriate test statistic based on your data and question.
4. **Calculate Test Statistic and P-value:** Calculate these values from your sample data.
5. **Decision:** Compare p-value to α.
   - If p-value ≤ α: Reject the null hypothesis (statistically significant result).
   - If p-value > α: Fail to reject the null hypothesis (not enough evidence to conclude the alternative hypothesis is true).

Types of Hypothesis Tests

- **One-tailed vs. Two-tailed Tests:**
  - One-tailed tests specify a direction (greater than or less than) for the potential difference.
  - Two-tailed tests simply look for any difference, regardless of direction.

- **Parametric vs. Non-parametric Tests:**
  - Parametric tests assume the sample data follow a specific distribution (e.g., normal distribution).
  - Non-parametric tests make fewer assumptions about the distribution, useful for non-normal data.

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

```python
# 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.

```python
# 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.

P Test

```python
# Calculate observed proportion of success
observed_proportion = np.sum(df["quality"]) / len(df["quality"])

# Calculate expected proportion under null hypothesis
expected_proportion = np.sum(df["chlorides"]) / len(df["chlorides"])

# Calculate chi-square statistic
chi_square_statistic = ((observed_proportion - expected_proportion) ** 2) / expected_proportion

# Degrees of freedom (for a 1-sample proportion test)
degrees_of_freedom = 1

# Calculate p-value
p_value = 1 - chi_square_statistic
print("P-value:", p_value)
# Make a decision
if p_value < 0.05:
    print("Reject the null hypothesis")
else:
    print("Fail to reject the null hypothesis")
```

7] ✓ 0.0s

```
P-value: -350.9800008169624
Reject the null hypothesis
```

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.

```python
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