2020102005-brsm-1-1

February 14, 2024

1 BRSM - Hypothesis Testing Homework Problemset

1.0.1 Srujana Vanka - 2020102005

[59]:
```
# Importing Libraries
import numpy as np # linear algebra
import pandas as pd # data processing CSV file
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import shapiro
from scipy.stats import ttest_ind
from scipy.stats import mannwhitneyu
from pingouin import compute_effsize
```

1.1 Do students with a GPA<=7 have lower placement TESTSCORES than those with a GPA>7?

To evaluate the provided hypothesis, we initiate the process by formulating a null hypothesis for the question at hand. A two-sample t-test is then conducted on the dataset to assess whether there is sufficient evidence to reject the null hypothesis in favor of our directional hypothesis. The decision to accept our directional hypothesis is contingent upon the p-value being less than the predetermined significance level ( ).

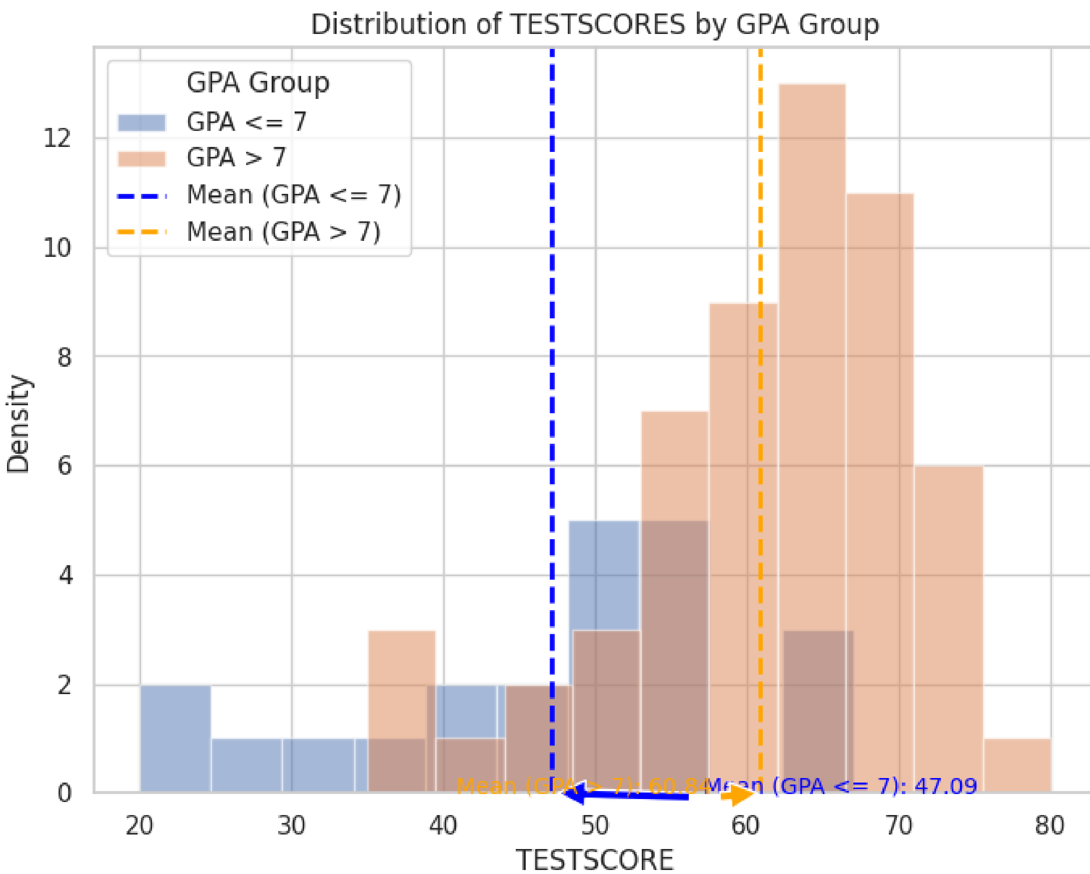
Null Hypothesis (H0): The mean placement TESTSCORES do not differ between students with a GPA 7 and those with a GPA > 7.

Alternative Hypothesis (H1): Students with a GPA 7 exhibit lower mean placement TESTSCORES compared to those with a GPA > 7.

[73]:
```
# Load dataset
df = pd.read_excel('/home/sruj/Downloads/Hypothesis testing data.xlsx')

# Extract data for the two groups
group_low_gpa = df[df['GPA'] <= 7]['Placement \nTESTSCORE']
group_high_gpa = df[df['GPA'] > 7]['Placement \nTESTSCORE']
```

[64]:
```
# Histogram
plt.figure(figsize=(8, 6))
plt.hist(group_low_gpa, bins = 10, alpha = 0.5, label= 'GPA <= 7')
```

```
plt.hist(group_high_gpa, bins = 10, alpha = 0.5,label = 'GPA > 7')
plt.title('Distribution of TESTSCORES by GPA Group')
plt.xlabel('TESTSCORE')
plt.ylabel('Density')
plt.legend(title='GPA Group', labels=['GPA <= 7', 'GPA > 7'])
plt.axvline(x=group_low_gpa.mean(), color='blue', linestyle='dashed',␣
 ↪linewidth=2, label='Mean (GPA <= 7)')
plt.axvline(x=group_high_gpa.mean(), color='orange', linestyle='dashed',␣
 ↪linewidth=2, label='Mean (GPA > 7)')
plt.legend(title='GPA Group')

plt.annotate(f'Mean (GPA <= 7): {group_low_gpa.mean():.2f}',
             xy=(group_low_gpa.mean(), 0),
             xytext=(group_low_gpa.mean() + 10, 0.005),
             arrowprops=dict(facecolor='blue', shrink=0.05),
             color='blue',
             fontsize=10)
plt.annotate(f'Mean (GPA > 7): {group_high_gpa.mean():.2f}',
             xy=(group_high_gpa.mean(), 0),
             xytext=(group_high_gpa.mean() - 20, 0.005),
             arrowprops=dict(facecolor='orange', shrink=0.05),
             color='orange',
             fontsize=10)

plt.show()
```



The histogram shows the distribution of testscores for two GPA groups: GPA <= 7 and GPA > 7. The blue and orange dashed lines represent the means of testscores for each group. The annotations provide the numerical values of these means. The legend distinguishes between the two GPA groups. The plot helps visualize the distribution of testscores and compare the central tendencies between the two GPA groups.

[72]:
```
# Shapiro-Wilk test for normality assumption on GPA <= 7 group
t_stat, p = shapiro(group_low_gpa)
print('GPA<=7')
print('Statistics=%.3f, p=%.3f' % (t_stat, p))

# Check normality assumption and provide interpretation
alpha = 0.05
if p > alpha:
    print('Sample looks Gaussian. Based on the statistical analysis, we fail to␣
 ↪reject the null hypothesis.\n')
else:
    print('Sample does not look Gaussian. Based on the statistical analysis, we␣
 ↪reject the null hypothesis.\n')
```

```
# Shapiro-Wilk test for normality assumption on GPA > 7 group
stat, p = shapiro(group_high_gpa)
print('GPA>7')
print('Statistics=%.3f, p=%.3f' % (stat, p))

# Check normality assumption and provide interpretation
alpha = 0.05
if p > alpha:
    print('Sample looks Gaussian. Based on the statistical analysis, we fail to␣
 ↪reject the null hypothesis.')
else:
    print('Sample does not look Gaussian. Based on the statistical analysis, we␣
 ↪reject the null hypothesis.')
```

```
GPA<=7
Statistics=0.935, p=0.156
Sample looks Gaussian. Based on the statistical analysis, we fail to reject the
null hypothesis.

GPA>7
Statistics=0.942, p=0.010
Sample does not look Gaussian. Based on the statistical analysis, we reject the
null hypothesis.
```

Changing the significance level (alpha) directly impacts the balance between Type I and Type II errors and the level of confidence associated with rejecting the null hypothesis. The significance level is directly related to the confidence level. A significance level of 0.05 corresponds to a 95% confidence level.

[70]:
```
# Mann-Whitney U test for comparing distributions of non-normally distributed␣
 ↪data
stat, p = mannwhitneyu(group_high_gpa, group_low_gpa)
print('Statistics=%.3f, p=%.9f' % (stat, p))

# Check the results of Mann-Whitney U test and provide interpretation
alpha = 0.05
if p > alpha:
    print('Same distribution (fail to reject H0)')
else:
    print('Different distribution (reject H0)')
```

```
Statistics=991.500, p=0.000030844
Different distribution (reject H0)
```

[71]:
```
# Independent Samples T-Test for comparing means between GPA > 7 and GPA <= 7␣
 ↪groups
t_stat, p_value = ttest_ind(group_high_gpa, group_low_gpa)
```

```python
# Compute effect size (Cohen's d)
effect_size = compute_effsize(group_high_gpa, group_low_gpa, eftype='cohen')

# Print T-Test results and effect size
print(f'T-statistic: {t_stat:.3f}')
print(f'P-value: {p_value:.9f}')
print(f'Effect size (Cohen\'s d): {effect_size:.3f}')

# Interpret effect size
if abs(effect_size) < 0.2:
    print('Negligible effect')
elif abs(effect_size) < 0.5:
    print('Small effect')
elif abs(effect_size) < 0.8:
    print('Medium effect')
else:
    print('Large effect')
```

```
T-statistic: 5.056
P-value: 0.000002884
Effect size (Cohen's d): 1.272
Large effect
```

1.1.1 Analysis based on Results:

Normality Assumption (Shapiro-Wilk Test):

- **GPA <= 7:**
  - Statistics: 0.935
  - P-Value: 0.156
  - *Interpretation:* The data for GPA <= 7 group looks Gaussian based on the Shapiro-Wilk test ($p > 0.05$).
- **GPA > 7:**
  - Statistics: 0.942
  - P-Value: 0.010
  - *Interpretation:* The data for GPA > 7 group does not look Gaussian based on the Shapiro-Wilk test ($p < 0.05$).

Mann-Whitney U Test for Different Distributions:

- Statistics: 991.500
- P-Value: 0.000030844
- *Interpretation:* We reject the null hypothesis. There is evidence of a different distribution between GPA <= 7 and GPA > 7 groups.

Independent Samples T-Test and Effect Size:

- T-Statistic: 5.056

- P-Value: 0.000002884
- Effect size (Cohen's d): 1.272 (Large effect)
- *Interpretation:* Students with a GPA <= 7 have significantly lower placement testscores than those with a GPA > 7. The Mann-Whitney U test supports different distributions, and the large effect size indicates substantial practical significance.

1.1.2 Conclusion and Answer to the Question:

- **Yes,** based on the statistical analysis:
  - Students with a GPA <= 7 have significantly lower placement testscores than those with a GPA > 7.
  - The Mann-Whitney U test suggests different distributions between the two GPA groups.
  - The large effect size (Cohen's d) indicates a substantial practical significance in the difference between the two groups.

In conclusion, students with lower GPAs (<=7) achieved significantly lower placement test scores compared to those with higher GPAs (>7), both statistically and practically.