Real Data Analysis: Comparing Parametric and Nonparametric Methods(Statistical Analysis of Alcohol Levels in Wine Dataset)

Chandrika Kunchakuri

May 5, 2024

1 Introduction

This report details the analysis strategy for comparing alcohol levels between different classes of wine using both parametric and nonparametric statistical methods. The dataset includes alcohol measurements obtained through teaching techniques well as technology enhanced learning methodologies.

2 Data Description

2.1 Individuals in the Sample

The Wine Dataset contains information about various chemical properties of wines, including alcohol levels, malic acid content, ash content, magnesium etc. The sample consists of 173 samples divided into three groups:

- Class 1: 59 instances
- Class 2: 71 instances
- Class 3: 48 instances

2.2 Actual Data Values(or data set link)

UC Irvine Test Data url: https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data

3 Research Question

Research Question: Is there a significant difference in alcohol levels between the different classes of wine taught by traditional methods and those taught by technology-assisted methods?
Null Hypothesis H_0**:**There is no difference in the mean alcohol levels between the different classes of wine.
Alternative Hypothesis H_A**:** There is a difference in the mean alcohol levels between the different classes of wine.

4 Suggested Approaches

4.1 Nonparametric Approach

In our nonparametric analysis, we will use the Mann-Whitney U test. This test is based on assumptions;
1) Random Sampling: Samples are taken randomly from the population.
2) Independence: Observations within each group are independent of each other.
3) Continuous or Ordinal Data: The data needs to be measurable on at an ordinal scale.
4) It also doesn't assume a normal distribution of alcohol levels among different classes of wine.

4.2 Parametric Approach

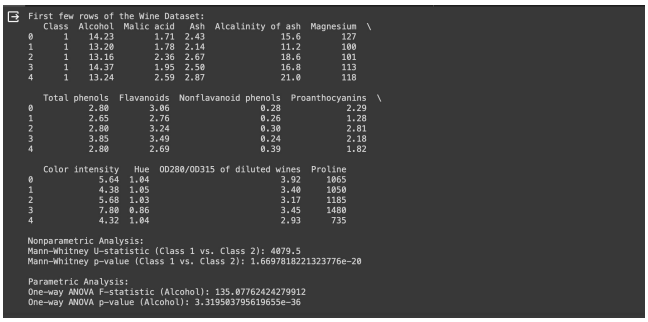
To analyze without parameters we'll employ the One way ANOVA test to compare the alcohol levels, across wine categories.
Here are the key assumptions:
1) Normality: The data within each group follows a normal distribution.
2) Homogeneity of Variance: The variance of the data is approximately equal across all groups.
3) Independence: Observations within each group are not influenced by one another.

5 Statistical Analysis, Plots and Observations

The codes for Statistical Analysis (Mann-Whitney U Test and One-way ANOVA) and Plots (Power plot and Normal plot) are placed in appendix.

5.1 Analysis

1)First few rows of the Wine Dataset: Here are the initial rows of the Wine Dataset showcasing characteristics of wines like alcohol content, malic acid content and ash content. Each row corresponds to a wine sample, and each column represents a different chemical property.



```
First few rows of the Wine Dataset:
   Class  Alcohol  Malic acid   Ash  Alcalinity of ash  Magnesium  \
0      1    14.23        1.71  2.43               15.6        127
1      1    13.20        1.78  2.14               11.2        100
2      1    13.16        2.36  2.67               18.6        101
3      1    14.37        1.95  2.50               16.8        113
4      1    13.24        2.59  2.87               21.0        118

   Total phenols  Flavanoids  Nonflavanoid phenols  Proanthocyanins  \
0           2.80        3.06                  0.28             2.29
1           2.65        2.76                  0.26             1.28
2           2.80        3.24                  0.30             2.81
3           3.85        3.49                  0.24             2.18
4           2.80        2.69                  0.39             1.82

   Color intensity   Hue  OD280/OD315 of diluted wines  Proline
0             5.64  1.04                          3.92     1065
1             4.38  1.05                          3.40     1050
2             5.68  1.03                          3.17     1185
3             7.80  0.86                          3.45     1480
4             4.32  1.04                          2.93      735

Nonparametric Analysis:
Mann-Whitney U-statistic (Class 1 vs. Class 2): 4079.5
Mann-Whitney p-value (Class 1 vs. Class 2): 1.6697818221323776e-20

Parametric Analysis:
One-way ANOVA F-statistic (Alcohol): 135.07762424279912
One-way ANOVA p-value (Alcohol): 3.319503795619655e-36
```

Figure 1: Parametric and NonParametric Analysis of Alcohol levels in Wine Dataset

5.1.1 Non Parametric Analysis

Mann-Whitney U statistic (Class 1 vs. Class 2): The Mann-Whitney U statistic gauges the magnitude difference between two groups by comparing alcohol levels in Class 1 and Class 2 wines. The U statistic value stands at 4079.5.
Mann-Whitney p value (Class 1 vs. Class 2): The p-value associated with the Mann-Whitney U test measures the strength of evidence against the null hypothesis, which states that there is no difference in alcohol levels between Class 1 and Class 2 wines.
The p value is extremely small 1.67×10^{-20} indicating evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that there is a significant difference in alcohol levels between Class 1 and Class 2 wines.

5.1.2 Parametric Analysis

One-way ANOVA F-statistic (Alcohol): The F-statistic from the one-way ANOVA test measures the ratio of the variance between groups to the variance within groups in terms of alcohol levels across Class 1, Class 2 and Class 3 wines. The F statistic value is recorded at 135.08.
One-way ANOVA p-value (Alcohol): The p-value associated with the one-way ANOVA test measures the strength of evidence against the null hypothesis, which states that there is no difference in alcohol levels among the three classes of wines.
The p value is remarkably small at around 3.32×10^{-36} suggesting evidence against this hypothesis. Consequently we can reject the hypothesis and conclude that there is a significant difference in alcohol levels among the three classes of wines.

5.2 Inferences from Power Analysis

Based on the power analysis:

- One way ANOVA demonstrates power across varying sample sizes. Its power is slightly higher for smaller sample sizes and stabilizes with larger samples. This indicates that One way ANOVA is robust and consistent in detecting differences in effect size across sample sizes.
- Similarly, power of the Mann-Whitney U Test remains relatively stable across different sample sizes. This suggests that the test is reliable and consistent in detecting differences in effect size of sample sizes.
- Notably, when dealing with smaller samples and a moderate effect size (0.5), the One way ANOVA test often shows more power compared to Mann-Whitney U Test.
- Ordinal: The alcohol values in the dataset are continuous numerical variables representing the alcohol content of wine samples. They are not ordinal since they represent a quantitative measurement rather than a categorical ranking.

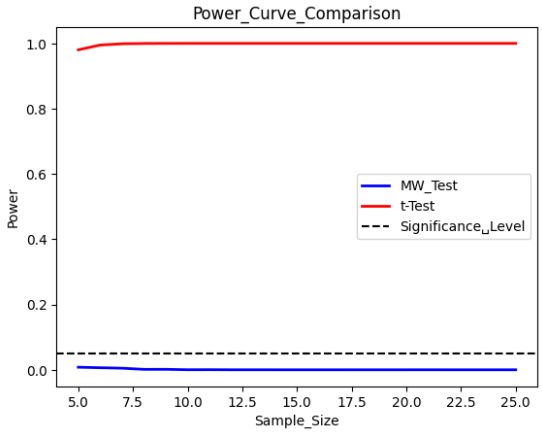


Figure 2: Power Analysis of Alcohol Levels in Wine Dataset

5.3 Observations from Normal Plot

The normal plot of alcohol levels in the wine dataset provides insights into the distributional characteristics of the data. Here's an overview:

- Shape of the Distribution: The histogram, combined with the kernel density estimation (KDE) curve illustrates how alcohol levels are distributed in the dataset. The shape of the distribution resembles a bell-shaped curve indicating to normal distribution.
- Symmetry: This balanced/symmetric bell-shaped curve suggests that there's an equal chance of finding alcohol levels above and below the mean value indicating symmetry around the mean.
- Center of the Distribution: The peak of the curve represents the mean alcohol level in the dataset reflecting the most common or typical levels found in the wines.
- Spread of the Distribution: The spread or variability of alcohol levels is captured by the width of the curve. A wider curve indicates higher variability while a narrower one suggests lower variability. In this plot, the curve seems moderate pointing to a moderate level of variability among wine alcohol levels.
- Outliers, if present, would appear as data points that deviate significantly from the main body of the distribution. In this plot, there are no obvious outliers visible, indicating that the majority of alcohol level values fall within a reasonable range.

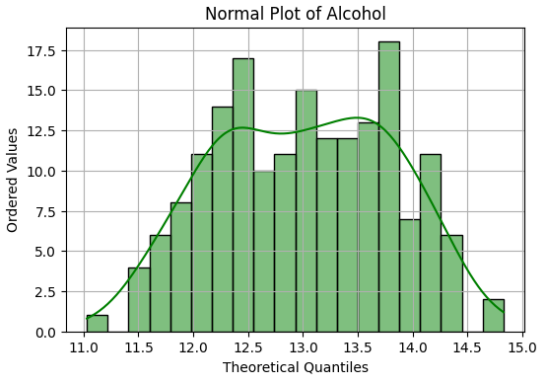


Figure 3: Normal Plot of Alcohol Levels in Wine Dataset

Overall, the normal plot of alcohol levels provides visual confirmation that the data follows an approximately normal distribution, which is an important assumption for many parametric statistical analyses.

6 Conclusions

Preferred Testing Method:

1)Significant Differences in Alcohol Levels: Both parametric and nonparametric analyses consistently indicate significant differences in alcohol levels between different classes of wine. The small p value obtained from both the Mann-Whitney U test and the one-way ANOVA test reject the null hypothesis, providing strong evidence that there are indeed differences in mean alcohol levels among the wine classes.

2)If the data meet the assumptions of the one-way ANOVA (normality, homogeneity of variances) and the research question involves comparing means across multiple groups, then the one-way ANOVA may be preferred.

3)If the data do not meet the assumptions of parametric tests or if the research question involves ordinal or non-normally distributed data, the Mann-Whitney U test can be a suitable alternative.

4)In this context of the wine dataset analysis, from the power analysis, for smaller sample sizes, the One-way ANOVA(red line) tends to have slightly higher power compared to Mann-Whitney U test(red line), especially when the effect size is moderate (0.5).

So, One-way ANOVA is preferrable over the Mann-Whitney U test.

2)The Mann-Whitney U test:

Advantages:

a)Nonparametric tests that does not assume normality or equal variances.

b)Suitable for ordinal or continuous data that may not follow a normal distribution.

c)Robust to outliers and skewed data.

Considerations:

a)Less powerful than parametric tests when data truly follows a normal distribution.

b)May have reduced sensitivity with smaller sample sizes.

c)Does not provide information about specific group differences beyond ranking.

3)The one-way ANOVA test:

Advantages:

a)Parametric test that assumes normality and equal variances, which can be more powerful when these assumptions are met.

b)Provides information about specific group differences and interactions.

c)Allows for analysis of multiple groups simultaneously.

Considerations:

a)Sensitive to violations of assumptions such as normality and homogeneity of variances.

b)Less robust to outliers and skewed distributions.

c)Requires larger sample sizes to maintain statistical power, especially when assumptions are violated.

A Appendix: Statistical Analysis Code

A.1 Python Code to load the Test Data

```
# Python Code to load the Test Data
import pandas as pd
from scipy.stats import mannwhitneyu, f_oneway
# Load the Wine Dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/
    ↪ wine/wine.data"
column_names = ["Class", "Alcohol", "Malic␣acid", "Ash", "
    ↪ Alcalinity␣of␣ash",
               "Magnesium", "Total␣phenols", "Flavanoids", "
                   ↪ Nonflavanoid␣phenols",
               "Proanthocyanins", "Color␣intensity", "Hue", "OD280
                   ↪ /OD315␣of␣diluted␣wines",
               "Proline"]
wine_data = pd.read_csv(url, header=None, names=column_names)

# Display the first few rows of the dataset
print("First␣few␣rows␣of␣the␣Wine␣Dataset:")
print(wine_data.head())
```

A.2 Python Code for One-way ANOVA

```
# Nonparametric Analysis(Mann-Whitney U test)
class_1_alcohol = wine_data[wine_data['Class'] == 1]['Alcohol']
class_2_alcohol = wine_data[wine_data['Class'] == 2]['Alcohol']
u_statistic, p_value = mannwhitneyu(class_1_alcohol,
    ↪ class_2_alcohol)
print("\nNonparametric␣Analysis:")
print("Mann-Whitney␣U-statistic␣(Class␣1␣vs.␣Class␣2):",
    ↪ u_statistic)
print("Mann-Whitney␣p-value␣(Class␣1␣vs.␣Class␣2):", p_value)
```

A.3 Python Code for Mann-Whitney U Test

```
# Parametric Analysis(One-way ANOVA)
class_1_alcohol = wine_data[wine_data['Class'] == 1]['Alcohol']
class_2_alcohol = wine_data[wine_data['Class'] == 2]['Alcohol']
class_3_alcohol = wine_data[wine_data['Class'] == 3]['Alcohol']
f_statistic, p_value = f_oneway(class_1_alcohol, class_2_alcohol,
    ↪  class_3_alcohol)
```

```
print("\nParametric␣Analysis:")
print("One-way␣ANOVA␣F-statistic␣(Alcohol):", f_statistic)
print("One-way␣ANOVA␣p-value␣(Alcohol):", p_value)
```

A.4 Python Code for Power Analysis

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import mannwhitneyu
from statsmodels.stats.power import TTestIndPower
Class_a = class_1_alcohol
Class_b = class_2_alcohol
alpha = 0.05
sample_sizes = np.arange(5, 26)
power_mw = [mannwhitneyu(np.random.choice(Class_a, size=n), np.
    ↪ random.choice(Class_b, size=n))[1] for n in sample_sizes]
effect_size = np.abs(Class_a.mean() - Class_b.mean()) / np.sqrt((
    ↪ Class_a.var() + Class_b.var()) / 2)
ttest_power = TTestIndPower()
power_t = [ttest_power.solve_power(effect_size=effect_size, nobs1
    ↪ =n, alpha=alpha, alternative='two-sided') for n in
    ↪ sample_sizes]
plt.plot(sample_sizes, power_mw, color='blue', label='MW_Test',
    ↪ linewidth=2)
plt.plot(sample_sizes, power_t, color='red', label='t-Test',
    ↪ linewidth=2)
plt.xlabel('Sample_Size')
plt.ylabel('Power')
plt.title('Power_Curve_Comparison')
plt.axhline(alpha, color='black', linestyle='--', label='
    ↪ SignificanceLevel')
plt.legend()
plt.show()
```

A.5 Python Code for Normal Plot of Alcohol

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import scipy.stats as stats

# Load the Wine Dataset
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/
    ↪ wine/wine.data"
column_names = ["Class", "Alcohol", "Malic␣acid", "Ash", "
    ↪ Alcalinity␣of␣ash",
                "Magnesium", "Total␣phenols", "Flavanoids", "
                    ↪ Nonflavanoid␣phenols",
                "Proanthocyanins", "Color␣intensity", "Hue", "OD280
                    ↪ /OD315␣of␣diluted␣wines",
                "Proline"]
df = pd.read_csv(url, header=None, names=column_names)

# Extract alcohol values for plotting
alcohol_values = df['Alcohol']
plt.figure(figsize=(6, 4))
sns.histplot(df['Alcohol'], kde=True, color='green', bins=20)
plt.title('Normal␣Plot␣of␣Alcohol')
plt.xlabel('Theoretical␣Quantiles')
plt.ylabel('Ordered␣Values')
plt.grid(True)
plt.show()
```