Module 2: Data Analytics with Python – Statistics

Assignment 2.1: Advanced Statistics

Objective: By the end of this lab, you will understand and perform advanced statistical methods, including hypothesis testing using various tests such as z-test and t-test. You will also learn how to make decisions based on the results of these tests.

Task 1: Introduction to Hypothesis Testing

Instructions:

- Hypothesis testing is a statistical method for testing a claim or hypothesis about a parameter in a population, using data measured in a sample.
- Follow the steps to perform hypothesis testing:
    1. State the hypothesis.
    2. Set the criteria for a decision.
    3. Compute the test statistic.
    4. Make a decision based on the test results.

Expected Output:

- Understanding of null and alternative hypotheses, significance level, p-value, and how to interpret the results of hypothesis tests.

Task 2: Performing a One-Sample z-Test

Instructions:
- Perform a one-sample z-test to determine if the sample mean is equal to a known population mean.
- Use the systolic blood pressure dataset and test whether the mean of the armsys group is 125.

```python
import pandas as pd
from statsmodels.stats import weightstats as stests

url =
"https://raw.githubusercontent.com/bluedataconsulting/AIMasteryPr
ogram/main/Lab_Exercises/Module2/systolic%20blood%20pressure.csv"
df = pd.read_csv(url)

ztest, pval = stests.ztest(df['armsys'], value=125)
print(pval)

if pval < 0.05:  # 0.05 is the significance level
    print("Mean is not 125")
else:
    print("Mean is 125")
```

Expected Output:
- p-value and conclusion whether to reject or accept the null hypothesis.

Task 3: Performing a Two-Sample z-Test

Instructions:
- Perform a two-sample z-test to compare the means of two independent groups.
- Use the systolic blood pressure dataset and test whether the means of the armsys and fingsys groups are equal.

```python
import pandas as pd
from statsmodels.stats import weightstats as stests

url =
"https://raw.githubusercontent.com/bluedataconsulting/AIMasteryPr
ogram/main/Lab_Exercises/Module2/systolic%20blood%20pressure.csv"
df = pd.read_csv(url)

ztest, pval = stests.ztest(x1=df['armsys'], x2=df['fingsys'],
value=0, alternative='two-sided')
print(pval)

if pval < 0.05:
    print("Mean of two samples is not 0")
else:
    print("Mean of two samples is 0")
```

Expected Output:
- p-value and conclusion whether to reject or accept the null hypothesis.

Task 4: Performing a One-Sample t-Test
Instructions:

- Perform a one-sample t-test to determine if there is a significant difference between the sample mean and a known population mean.
- Use the provided data on the mass of acorns and test whether the average mass of the sample is different from 10.0 g.

```python
import numpy as np
from scipy import stats

x = [8.8, 6.6, 9.5, 11.2, 10.2, 7.4, 8.0, 9.6, 9.9, 9.0,
     7.6, 7.4, 10.4, 11.1, 8.5, 10.0, 11.6, 10.7, 10.3, 7.0]
mu = 10

# Calculate t-statistic
x_bar = np.mean(x)
s = np.std(x, ddof=1)
N = len(x)
SE = s / np.sqrt(N)
t_statistic = (x_bar - mu) / SE
print("t-statistic:", t_statistic)

# One-sample t-test with scipy
t, p = stats.ttest_1samp(x, mu)
print("t = ", t, ", p = ", p)

if p < 0.05:
    print("Reject null hypothesis")
else:
    print("Accept null hypothesis")
```

Expected Output:

- t-statistic, p-value, and conclusion whether to reject or accept the null hypothesis.

Blue Data Consulting
We make you trust data

DATA
ANALYTICS
WITH
PYTHON
Blue Data Consulting

Task 5: Performing a Two-Sample t-Test
Instructions:

- Perform a two-sample t-test to determine if there is a significant difference between the means of two independent groups.
- Use the provided data on the weights of acorns from two different groups.

```python
import numpy as np
from scipy import stats

x1 = [10.8, 10.0, 8.2, 9.9, 11.6, 10.1, 11.3, 10.3, 10.7, 9.7,
      7.8, 9.6, 9.7, 11.6, 10.3, 9.8, 12.3, 11.0, 10.4, 10.4]

x2 = [7.8, 7.5, 9.5, 11.7, 8.1, 8.8, 8.8, 7.7, 9.7, 7.0,
      9.0, 9.7, 11.3, 8.7, 8.8, 10.9, 10.3, 9.6, 8.4, 6.6]

# Two-sample t-test with scipy
t, p_twosided = stats.ttest_ind(x2, x1, equal_var=True)
print("t = ", t, ", p_twosided = ", p_twosided, ", p_onesided =",
p_twosided/2)

if p_twosided < 0.05:
    print("Reject null hypothesis")
else:
    print("Accept null hypothesis")
```

Expected Output:

- t-statistic, p-value, and conclusion whether to reject or accept the null hypothesis.

Task 6: Performing a Paired t-Test
Instructions:
- Perform a paired t-test to compare the means of two related groups.
- Use the provided data on blood pressure before and after treatment.

```
from scipy import stats

x1 = [10.8, 10.0, 8.2, 9.9, 11.6, 10.1, 11.3, 10.3, 10.7, 9.7,
      7.8, 9.6, 9.7, 11.6, 10.3, 9.8, 12.3, 11.0, 10.4, 10.4]

x2 = [7.8, 7.5, 9.5, 11.7, 8.1, 8.8, 8.8, 7.7, 9.7, 7.0,
      9.0, 9.7, 11.3, 8.7, 8.8, 10.9, 10.3, 9.6, 8.4, 6.6]

ttest, pval = stats.ttest_rel(x1, x2)

if pval < 0.05:
    print("Reject null hypothesis")
else:
    print("Accept null hypothesis")
```

Expected Output:
- p-value and conclusion whether to reject or accept the null hypothesis.