Name: Priansh Madan

Section: E Roll Number: 58

Batch: E4

DVA Practical 5

Write a program to perform following statistical test using user defined functions.

1] Z test

2] T-test

3] ANNOVA Test

Use IRIS data set to perform above tests using user defined functions. Verify the results obtained with standard functions.

Z Test

In [ ]:
```
import pandas as pd
from scipy import stats
from sklearn.datasets import load_iris
import numpy as np
```

In [ ]:
```
df=pd.read_csv('iris.csv')
```

In [ ]:
```
df.head()
```

Out[ ]:

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

In [ ]:
```
df.describe()
```

Out[ ]:

	sepal.length	sepal.width	petal.length	petal.width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

In [ ]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal.length  150 non-null    float64
 1   sepal.width   150 non-null    float64
 2   petal.length  150 non-null    float64
 3   petal.width   150 non-null    float64
 4   variety       150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

In [ ]:
```
df.isna().sum()
```

Out[ ]:
```
sepal.length    0
sepal.width     0
petal.length    0
petal.width     0
variety         0
dtype: int64
```

In [ ]:
```
df.columns
```

Out[ ]:
```
Index(['sepal.length', 'sepal.width', 'petal.length', 'petal.width',
       'variety'],
      dtype='object')
```

In [ ]:
```
setosa_sepal_length = df[df['variety'] == 'Iris-setosa']['sepal.length']
```

In [ ]:
```
overall_mean_sepal_length = df['sepal.length'].mean()
overall_std_sepal_length = df['sepal.length'].std()
```

In [ ]:
```
sample_size = len(setosa_sepal_length)
```

In [ ]:
```
z_score = (setosa_sepal_length.mean() - overall_mean_sepal_length) / (overall_std_s
```

```
C:\Users\ACER\AppData\Local\Temp\ipykernel_408\2958648411.py:1: RuntimeWarning: di
vide by zero encountered in double_scalars
  z_score = (setosa_sepal_length.mean() - overall_mean_sepal_length) / (overall_st
d_sepal_length / np.sqrt(sample_size))
```

In [ ]:
```
p_value = 2 * (1 - stats.norm.cdf(np.abs(z_score)))
```

In [ ]:
```
alpha = 0.05

if p_value < alpha:
    print("The difference in sepal length for Iris setosa is statistically signific
else:
    print("The difference in sepal length for Iris setosa is not statistically sign
```

```
The difference in sepal length for Iris setosa is not statistically significant.
```

T-Test

In [ ]:
```
setosa_sepal_length = df[df['variety'] == 'Iris-setosa']['sepal.length']
versicolor_sepal_length = df[df['variety'] == 'Iris-versicolor']['sepal.length']
```

In [ ]:
```
t_statistic, p_value = stats.ttest_ind(setosa_sepal_length, versicolor_sepal_length
```

In [ ]:
```
alpha = 0.05

if p_value < alpha:
    print("The difference in sepal length between Iris setosa and Iris versicolor i
else:
    print("The difference in sepal length between Iris setosa and Iris versicolor i
```

```
The difference in sepal length between Iris setosa and Iris versicolor is not stat
istically significant.
```

ANNOVA Test

In [ ]:
```
setosa_sepal_length = df[df['variety'] == 'Iris-setosa']['sepal.length']
versicolor_sepal_length = df[df['variety'] == 'Iris-versicolor']['sepal.length']
virginica_sepal_length = df[df['variety'] == 'Iris-virginica']['sepal.length']
```

In [ ]:
```
f_statistic, p_value = stats.f_oneway(setosa_sepal_length, versicolor_sepal_length,
```

```
d:\SOFTWARES\PYTHON\lib\site-packages\scipy\stats\_stats_py.py:3869: DegenerateDat
aWarning: at least one input has length 0
  warnings.warn(stats.DegenerateDataWarning('at least one input '
```

In [ ]:
```
alpha = 0.05

if p_value < alpha:
    print("The sepal length differs significantly among the three species.")
else:
    print("There is no significant difference in sepal length among the three speci
```

```
There is no significant difference in sepal length among the three species.
```

In [ ]:
```
setosa_sepal_width = df[df['variety'] == 'Iris-setosa']['sepal.width']
versicolor_sepal_width = df[df['variety'] == 'Iris-versicolor']['sepal.width']
virginica_sepal_width = df[df['variety'] == 'Iris-virginica']['sepal.width']
```

In [ ]:
```
f_statistic, p_value = stats.f_oneway(setosa_sepal_width, versicolor_sepal_width, v
```

```
d:\SOFTWARES\PYTHON\lib\site-packages\scipy\stats\_stats_py.py:3869: DegenerateDat
aWarning: at least one input has length 0
  warnings.warn(stats.DegenerateDataWarning('at least one input '
```

In [ ]:
```
alpha = 0.05

if p_value < alpha:
    print("The sepal width differs significantly among the three species.")
```

```
else:
    print("There is no significant difference in sepal width among the three specie
```

```
There is no significant difference in sepal width among the three species.
```