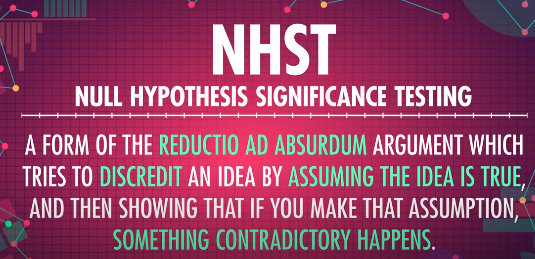
HYPOTHESIS TESTING

In the scope of statistics, there is the branch of Inferential statistics that serves to make logical assumptions that relate between the sample and the population. The former is a small portion of the latter and thus through hypothesis one can make some conclusions from the general attributes. This means Hypothesis are some inferences made of a sample that tend to summarize the general behaviour of the population. Hypothesis testing serves to provide the mathematical reasoning of quantifying such observations.

In general, the Null Hypothesis serves as the Reductio Ad Absurdum which means in any case this is the “formed Truth” that we tend to accept. Any observation or calculation that does not comply which this forms the alternative hypothesis. In other words, the Statistical experiment is assumed to give the Null Hypothesis until proved otherwise.

i

The value that can either give us a Null Hypothesis or an Alternative Hypothesis is the p-value or in full it’s the Probability Value. And to compare this value we define another variable alpha that comes from the statistical experiment. The latter (alpha) can be any value since its from an statistical experiment but the p-value is a fixed number that does not change primarily because it originates from a distribution formula.

# Example: Hypothesis Testing using Python.

1. Required Libraries

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from scipy.stats import f\_oneway

from statsmodels.stats.weightstats import ztest # z-test

1. Loading the Data

df=pd.DataFrame({'id':np.arange(1,51),'val':np.random.randn(50)})

df2=pd.DataFrame({'id':np.arange(1,31),'val':np.random.randint(40,50,size=30)/10})

The two data frames are such that they have different values given the 30 random numbers between 40 and 50 for dataframe 2 & 50 values between 1 and 50 for dataframe 1

Defining the sample Means

# Defining the sample means

sample1=df.val # Values in sample 1

sample2=df2.val # Values in sample 2

sample1\_mean = sample1.mean()

sample2\_mean = sample2.mean()

Defining the Significance Level (Custom Threshold) Alpha as 5%

# Significance Level (alpha)

"If (Probability value i.e., p-value is less than alpha, we reject the null hypothesis"

alpha = 0.05

Performing the Z-Test to obtain the Z-test statist and P-value

# Z Test

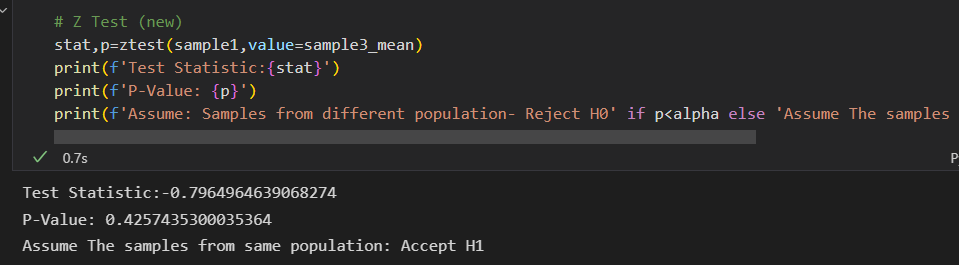
stat,p=ztest(sample1,value=sample2\_mean)

print(f'Test statistic: {stat}')

print(f'p-value: {p}')

print(f'H0 rejected!! Since The samples from different population' if p<alpha else 'H0 not rejected!! Since The samples from same population')

The output from the two samples shows that there are from a different population. Since it rejects the Null Hypothesis (that insinuated they are from same population) as shown with the output below.



This concludes the Hypothesis Testing, But we can spice things up and define a new dataframe with the sample drawn from either sample 1 or Sample 2. For simplicity we choose Sample 3 from Sample 1 using the syntax shown with the code below;

# Taking a sample from the same population to accept the H1 (Alternative Hypothesis)

df3=pd.DataFrame({'id':np.arange(1,31),'val':np.random.choice(df['val'].values,size=30)})

sample3\_mean=df3.val.mean()

To conclude its from the same population we compute the Z-Test statistic using the method below;

# Z Test (new)

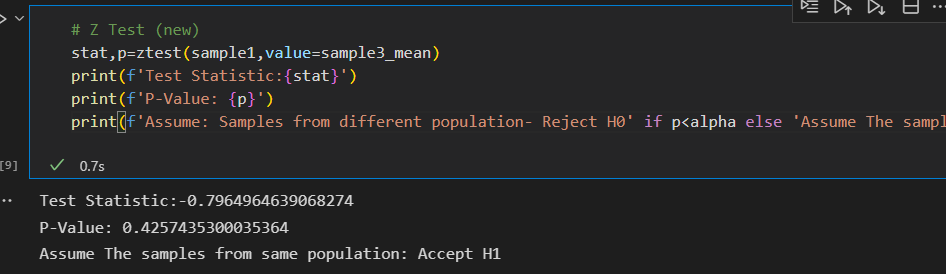
stat,p=ztest(sample1,value=sample3\_mean)

print(f'Test Statistic:{stat}')

print(f'P-Value: {p}')

print(f'Assume: Samples from different population- Reject H0' if p<alpha else 'Assume The samples from same population: Accept H1')

Its output is as shown below;



Plotting the sample distributions for the three samples we use the syntax shown with the code below;

#plt.figure()

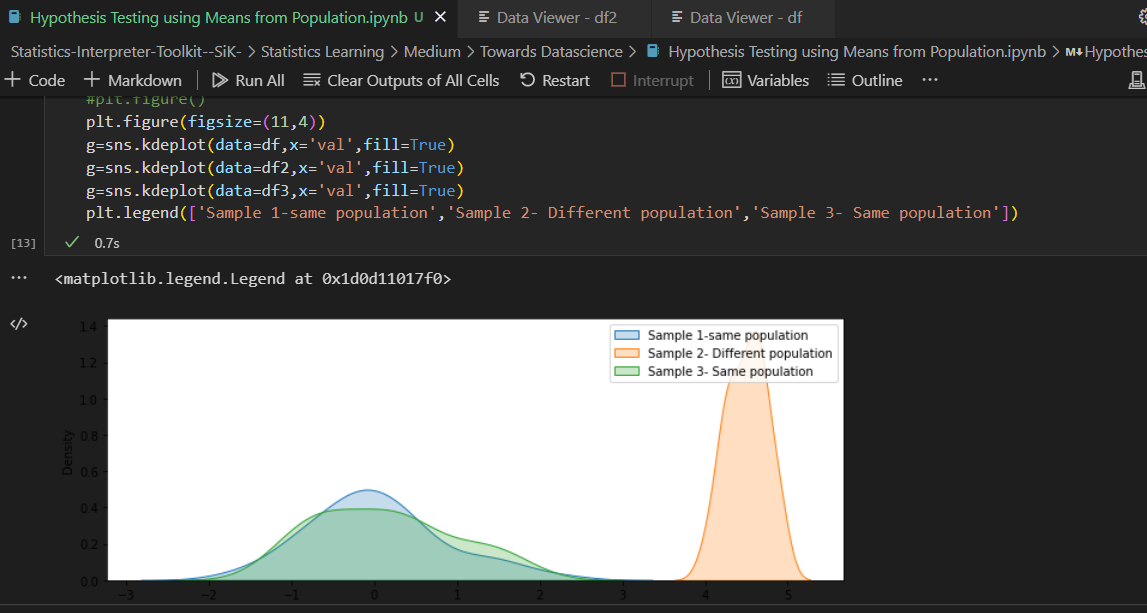
plt.figure(figsize=(11,9))

g=sns.kdeplot(data=df,x='val',fill=True)

g=sns.kdeplot(data=df2,x='val',fill=True)

g=sns.kdeplot(data=df3,x='val',fill=True)

plt.legend(['Sample 1-same population','Sample 2- Different population','Sample 3- Same population'])



*Figure showing the code output for the 3 samples’ normal distribution plots.*