# Hypothesis Testing

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: Vishvash C Batch ID:** 23012024

**Topic: Hypothesis Testing**

**Guidelines:**

**1. An assignment submission is considered complete only when the correct and executable code(s) and documentation explaining the method and results are submitted. Failing to submit either of those will be considered an invalid submission and not a correct submission.**

**2. Ensure that you submit your assignments correctly and in full. Resubmission is not allowed.**

**3. Post the submission you can evaluate your work by referring to the keys provided. (will be available only post the submission).**

**Hints:**

1. Hypothesis Testing Assignments, explanation of the solutions along with Business Objectives & Business Constraints should be documented in black and white along with the codes.
2. All the codes (executable programs) should run without errors
3. Python code, Documentation must be submitted
4. All the tests should be explained in detail (ex: Normality test, Variance test, etc.)

**Problem Statement:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import scipy

from scipy import stats

import statsmodels.stats.descriptivestats as sd

from statsmodels.stats import weightstats as stests

1. A F&B manager wants to determine whether there is any significant difference in the diameter of the cutlet between two units. A randomly selected sample of cutlets was collected from both units and measured. Analyze the data and draw inferences at a 5% significance level. Please state the assumptions and tests that you carried out to check the validity of the assumptions.

Data File: **Cutlets.csv**

# Business Problem:

# Determine whether there is any significant difference in the diameter of the cutlet between two units

# Load the data

cutlet = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/Data Science/Hypothesis Testing/Assignments/Datasets/Cutlets.csv")

cutlet

cutlet.columns = "Unit\_A", "Unit\_B"

# Normality Test - # Shapiro Test

# H0 = Data are Normal

# Ha = Data are not Normal

stats.shapiro(cutlet.Unit\_A) # Shapiro Test

print(stats.shapiro(cutlet.Unit\_B))

# Data are Normal

# help(stats.shapiro)

# Variance test

# H0 = Variances are Equal

# Ha = Variance are not Equal

scipy.stats.levene(cutlet.Unit\_A, cutlet.Unit\_B)

# p-value = 0.417 > 0.05 so p high null fly => Equal variances

# 2 Sample T test

# Ho: Average diameter of both cutlets are equal

# Ha: Average diameter of both cutlets are unequal

scipy.stats.ttest\_ind(cutlet.Unit\_A, cutlet.Unit\_B)

# p-value = 0.47 > 0.05

# P high Ho fly => Fail to reject Ho

**# Conclusion: Average diameter of both cutlets are equal**

1. A hospital wants to determine whether there is any difference in the average Turn Around Time (TAT) of reports of the laboratories on their preferred list. They collected a random sample and recorded TAT for reports from 4 laboratories. TAT is defined as a sample collected to report dispatch. Analyze the data and determine whether there is any difference in average TAT among the different laboratories at a 5% significance level.

Data File: **LabTAT.csv**

# Business Problem:

# Hospital to determine whether there is any difference in average TAT among the different laboratories at a 5% significance level

lab\_tat = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/Data Science/Hypothesis Testing/Assignments/Datasets/lab\_tat\_updated.csv")

lab\_tat

# Normality Test - # Shapiro Test

# H0 = Data are Normal

# Ha = Data are not Normal

stats.shapiro(lab\_tat.Laboratory\_1)

stats.shapiro(lab\_tat.Laboratory\_2)

stats.shapiro(lab\_tat.Laboratory\_3)

stats.shapiro(lab\_tat.Laboratory\_4)

# Variance test

# Ho: All the 4 Laboratories have equal variance TAT

# Ha: All the 4 Laboratories have unequal variance TAT

scipy.stats.levene(lab\_tat.Laboratory\_1, lab\_tat.Laboratory\_2, lab\_tat.Laboratory\_3, lab\_tat.Laboratory\_4)

# Variances are statisticaly equal

# One - Way Anova

# Ho: All the 4 Laboratories have equal mean TAT

# Ha: All the 4 Laboratories have unequal mean TAT

F, p = stats.f\_oneway(lab\_tat.Laboratory\_1, lab\_tat.Laboratory\_2, lab\_tat.Laboratory\_3, lab\_tat.Laboratory\_4)

p

# P low Null go p =2.143740909435053e-58 < 0.05

# All the 4 Laboratories have unequal mean TAT

**#Conclusion: There is any difference in average TAT among the different laboratories at a significance level**

1. Sales of products in four different regions are tabulated for males and females. Find if male-female buyer rations are similar across regions.



* Ho 🡪 All proportions are equal
* Ha 🡪 Not all Proportions are equal

Hint:

Check p-value

If p-Value < alpha, we reject Null Hypothesis

Data file: Buyer Ratio.csv

data = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/Data Science/Hypothesis Testing/Assignments/Datasets/BuyerRatio.csv")

data = data.iloc[:,1:]

# Ho: All regions have equal proportions of male-female buyer %

# Ha: Not all regions have equal proportions of male-female buyer %

Chisquares\_results = scipy.stats.chi2\_contingency(data)

print(Chisquares\_results)

Chi\_square = [['Test Statistic', 'p-value'], [Chisquares\_results[0], Chisquares\_results[1]]]

Chi\_square

# p-value = 0.66 > 0.05 => P hight Null fly

# All regions have equal proportions of male-female buyer %

**# Conclusion: All regions have equal proportions of male-female buyer %**

1. Telecall uses 4 centers around the globe to process customer order forms. They audit a certain % of the customer order forms. Any error in the order form renders it defective and must be reworked before processing. The manager wants to check whether the defective % varies by center. Please analyze the data at a 5% significance level and help the manager draw appropriate inferences.

File: **Customer OrderForm.csv**

# Business Problem:

# Check whether the defective % varies by center.

original\_df = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/Data Science/Hypothesis Testing/Assignments/Datasets/CustomerOrderform.csv")

original\_df

original\_df.replace({'Error Free': 0, 'Defective': 1}, inplace=True)

# Apply value\_counts to each column

result\_df = original\_df.apply(lambda x: x.value\_counts()).fillna(0)

# Print the result

print(result\_df)

# Ho: All centres have equal proportions of defectives %

# Ha: Not all centres have equal proportions of defectives %

Chisquares\_results = scipy.stats.chi2\_contingency(result\_df)

print(Chisquares\_results)

Chi\_square = [['Test Statistic', 'p-value'], [Chisquares\_results[0], Chisquares\_results[1]]]

Chi\_square

# p-value = 0.28 > 0.05 => P high Null fly

# All centres have equal proportions

**# Conclusion: All Proportions are equal**

1. Fantaloons Sales managers commented that % of males versus females walking into the store differs based on the day of the week. Analyze the data and determine whether there is evidence at a 5 % significance level to support this hypothesis.

File: **Fantaloons.csv**

# Business Problem:

# Sales manager has to determine if % of males versus females walking into the store differs based on the day of the week

two\_prop\_test = pd.read\_csv(r"C:/Users/Lenovo/Downloads/Study material/Data Science/Hypothesis Testing/Assignments/Datasets/Fantaloons.csv")

from statsmodels.stats.proportion import proportions\_ztest

tab1 = two\_prop\_test.Weekdays.value\_counts()

print(tab1)

tab2 = two\_prop\_test.Weekend.value\_counts()

print(tab2)

count = np.array([113, 167])

nobs = np.array([400, 400])

# Case1: Two Sided test

# Ho: Males-Female ratio is same on weekdays and on weekend

# Ha: Males-Female ratio is not same on weekdays and on weekend

stats, pval = proportions\_ztest(count, nobs, alternative = 'two-sided')

print("%.2f" % pval)

# P-value = 0.000 < 0.05 => P low Null go

# Ha: Males-Female ratio is not same on weekdays and on weekend

# Case2: One-sided (Greater) test

# Ho: Proportions of Males in weekdays <= Proportions of Males in weekend

# Ha: Proportions of Males in weekdays > Proportions of Males in weekend

stats, pval = proportions\_ztest(count, nobs, alternative = 'larger')

print("%.2f" % pval)

# P-value = 1.0 > 0.05 => P high Null fly

# Ho: Proportions of Males in weekdays <= Proportions of Males in weekend

**# Conclusion: % of males versus females walking into the store differs based on the day of the week**