**HYPOTHESIS TESTING**

**Import packages:**

import pandas as pd

import numpy as npy

import matplotlib as plt

import math

**Import data set:**

dataset=pd.read\_csv('general\_data.csv')

#extract the attrition "Yes"

dataset1=dataset[dataset['Attrition']=='Yes']

# check if null values and duplicated are existing.

dataset3=dataset1.isnull()

dataset1.dropna()

dataset4=dataset1.duplicated()

dataset1.drop\_duplicates()

dataset1.columns

#describe the data

describe=dataset1[['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome','Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender', 'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome', 'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear', 'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager']].describe()



HYPOTHESIS 1:

#sample mean of employee age is 36

# To determine if this is true for a random sample of 35 employee age is taken

# standard derviation is 9,alpha=0.5

#H0=36 and H1!= 36

# this is two tail test

# to get sample of mean

sample=dataset1.Age.sample(35)

samplemeanage=sample.mean()

samplestd=sample.std()

#here we have standard deviation so we have to use Z test

zcalval=(samplemeanage-dataset1.Age.mean())/(dataset1.Age.std()/math.sqrt(35))

print("calculated value:",zcalval)

#table value

df=35-1

print(df)

tablevalue= 1.963423

#confidance intervel

up=dataset1.Age.mean()+(tablevalue\*samplestd)

lo=dataset1.Age.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: 0.4845060121090589

lower confidence level: 12.93151850269723

upper confidence level: 54.2836713707205

# conclusion:Here calculated value is 0.4845060121090589.this is fall into bowndry .so hypothesis is accepted,this statement is true.

HYPOTHESIS 2:

#sample of mean of education

# To determine if this is true for a random sample of 35 employee education is taken

# standard derviation is 1.01,alpha=0.5

#H0=2.87,H1!=2.87

# this is two tail test

# to get sample of mean

sample=dataset1.Education.sample(35)

samplemeaneducation=sample.mean()

samplestd=sample.std()

#here we have standard deviation so we have to use Z test

zcalval=(samplemeaneducation-dataset1.Education.mean())/(dataset1.Education.std()/math.sqrt(35))

print("calculated value:",zcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.Education.mean()+(tablevalue\*samplestd)

lo=dataset1.Education.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

# conclusion:Here calculated value is-2.452767869363454. this is fall out to bowndry. so hypothesis is rejected,so this statement is false.

calculated value: -2.452767869363454

lower confidence level: 0.8850579182879972

upper confidence level: 4.870216343315378

# conclusion:Here calculated value is-2.452767869363454. this is fall out to bowndry. so hypothesis is rejected,so this statement is false.

HYPOTHESIS 3:

#sample of mean of education

# To determine if this is true for a random sample of 35 employee distance home is taken

# standard derviation is 7.77,alpha=0.5

#H0=9.01,H1!=9.01

# this is two tail test

# to get sample of mean

sample=dataset1.DistanceFromHome.sample(35)

samplemeanDistanceFromHome=sample.mean()

samplestd=sample.std()

#here we have standard deviation so we have to use Z test

zcalval=(samplemeanDistanceFromHome-dataset1.DistanceFromHome.mean())/(dataset1.DistanceFromHome.std()/math.sqrt(35))

print("calculated value:",zcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.DistanceFromHome.mean()+(tablevalue\*samplestd)

lo=dataset1.DistanceFromHome.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: 0.055607954629950974

lower confidence level: -9.359525711989308

upper confidence level: 27.384842167685512

# conclusion:Here calculated value is 0.055607954629950974. this is fall into bowndry so hypothesis is accepted.

HYPOTHESIS 4:

#sample of mean of joblevel

# To determine if this is true for a random sample of 35 employee joblevel is taken

# standard derviation is 1.05,alpha=0.5

#H0=2.03,H1!=2.03

# this is two tail test

# to get sample of mean

sample=dataset1.JobLevel.sample(35)

samplemeanJobLevel=sample.mean()

samplestd=sample.std()

#here we have standard deviation so we have to use Z test

zcalval=(samplemeanJobLevel-dataset1.JobLevel.mean())/(dataset1.JobLevel.std()/math.sqrt(35))

print("calculated value:",zcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.JobLevel.mean()+(tablevalue\*samplestd)

lo=dataset1.JobLevel.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: 1.2261323689794104

lower confidence level: -0.18511804253254605

upper confidence level: 4.261067409621154

# conclusion:Here calculated value is 1.2261323689794104.this is fall into bowndry so hypothesis is accepted.

HYPOTHESIS 5:

#sample of mean of MonthlyIncome

# To determine if this is true for a random sample of 35 employee MonthlyIncome is taken

# standard derviation is 44792.1,alpha=0.5

#H0=61682.6,H1!=61682.6

# this is two tail test

# to get sample of mean

sample=dataset1.MonthlyIncome.sample(35)

samplemeanMonthlyIncome=sample.mean()

samplestd=sample.std()

#here we have standard deviation so we have to use Z test

zcalval=(samplemeanMonthlyIncome-dataset1.MonthlyIncome.mean())/(dataset1.MonthlyIncome.std()/math.sqrt(35))

print("calculated value:",zcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.MonthlyIncome.mean()+(tablevalue\*samplestd)

lo=dataset1.MonthlyIncome.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: -0.7493451784579717

lower confidence level: -15477.891385603827

upper confidence level: 138843.12345311436

# conclusion:Here calculated value is -0.7493451784579717. this is fall into the bowndry so hypothesis is accepted,this statement is true.

HYPOTHESIS 6:

#sample of mean of PercentSalaryHike

# To determine if this is true for a random sample of 35 employee PercentSalaryHike is taken

# standard derviation is 3.77,alpha=0.5

#H0=15.48,H1!=15.48

# this is two tail test

# to get sample of mean

sample=dataset1.PercentSalaryHike.sample(35)

samplemeanPercentSalaryHike=sample.mean()

samplestd=sample.std()

#here we have standard deviation so we have to use Z test

zcalval=(samplemeanPercentSalaryHike-dataset1.PercentSalaryHike.mean())/(dataset1.PercentSalaryHike.std()/math.sqrt(35))

print("calculated value:",zcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.PercentSalaryHike.mean()+(tablevalue\*samplestd)

lo=dataset1.PercentSalaryHike.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: -0.17172411933861928

lower confidence level: 8.388066336551159

upper confidence level: 22.57395897990454

# conclusion:Here calculated value is : -0.17172411933861928. this is fall into bowndry. so hypothesis is accepted.

HYPOTHESIS 7:

#sample of mean of TrainingTimesLastYear

# To determine if this is true for a random sample of 35 employee TrainingTimesLastYear is taken

# alpha=0.5

#H0>=2.65,H1!<2.65

# this is one tail test

# to get sample of mean

sample=dataset1.TrainingTimesLastYear.sample(35)

samplemeanTrainingTimesLastYear=sample.mean()

samplestd=sample.std()

samplestd

#this is left tail test

#here we don't have standard deviation so we have to use t test

tcalval=(samplemeanTrainingTimesLastYear-dataset1.TrainingTimesLastYear.mean())/(samplestd/math.sqrt(35))

print("calculated value:",tcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.TrainingTimesLastYear.mean()+(tablevalue\*samplestd)

lo=dataset1.TrainingTimesLastYear.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: 1.7970565432351666

lower confidence level: -0.04309884680462339

upper confidence level: 5.351115724441754

# conclusion:Here calculated value is : 1.7970565432351666. this is fall into the bowndry so hypothesis is accepted. .so the statement is true.

HYPOTHESIS 8:

#sample mean of employee age is 36

# To determine if this is true for a random sample of 35 employee age is taken

# alpha=0.5

#H0>=36 and H1< 36

# this is one tail test

#this is left tail test.

# to get sample of mean

sample=dataset1.Age.sample(35)

samplemeanage=sample.mean()

samplestd=sample.std()

#here we don't have standard deviation so we have to use t test

tcalval=(samplemeanage-dataset1.Age.mean())/(samplestd/math.sqrt(35))

print("calculated value:",tcalval)

#table value

df=35-1

tablevalue= 1.963423

#confidance intervel

up=dataset1.Age.mean()+(tablevalue\*samplestd)

lo=dataset1.Age.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: 1.1024373804205632

lower confidence level: 15.023064789148002

upper confidence level: 52.192125084269726

# conclusion:Here calculated value is : 1.1024373804205632 . this is fall into the bowndry so hypothesis is accepted. .so the statement is true.

HYPOTHESIS 9:

#sample of mean of MonthlyIncome

# To determine if this is true for a random sample of 35 employee MonthlyIncome is taken

# alpha=0.5

#H0>=61682.6,H1<61682.6

# this is one tail test

#this is left side test

# to get sample of mean

sample=dataset1.MonthlyIncome.sample(35)

samplemeanMonthlyIncome=sample.mean()

samplestd=sample.std()

#here we don't have standard deviation so we have to use t test

tcalval=(samplemeanMonthlyIncome-dataset1.MonthlyIncome.mean())/(samplestd/math.sqrt(35))

print("calculated value:",tcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.MonthlyIncome.mean()+(tablevalue\*samplestd)

lo=dataset1.MonthlyIncome.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: 0.6679931682409346

lower confidence level: -27949.47286290725

upper confidence level: 151314.7049304178

# conclusion:Here calculated value is 0.6679931682409346. this is fall into bowndry so hypothesis is accepted,this statement is true.

HYPOTHESIS 10:

#sample of mean of education

# To determine if this is true for a random sample of 35 employee distance home is taken

#alpha=0.5

#H0>=9.01,H1!=9.01

# this is ont tail test

#this is left side test

# to get sample of mean

sample=dataset1.DistanceFromHome.sample(35)

samplemeanDistanceFromHome=sample.mean()

samplestd=sample.std()

#here we don't have standard deviation so we have to use t test

tcalval=(samplemeanDistanceFromHome-dataset1.DistanceFromHome.mean())/(samplestd/math.sqrt(35))

print("calculated value:",tcalval)

#table value

df=35-1

tablevalue= 2.032245

#confidance intervel

up=dataset1.DistanceFromHome.mean()+(tablevalue\*samplestd)

lo=dataset1.DistanceFromHome.mean()-(tablevalue\*samplestd)

print("lower confidence level:",lo)

print("upper confidence level:",up)

calculated value: 1.0145806933888162

lower confidence level: -8.78183815955905

upper confidence level: 26.80715461525525

# conclusion:Here calculated value is 1.0145806933888162.this is fall into bowndry so hypothesis is accepted,this statement is true.