In [1]:
```
import numpy as np
import pandas as pd
import scipy.stats as stats
from scipy.stats import ttest_ind
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [3]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [5]:
```
A=pd.read_csv('/content/drive/My Drive/Statistics Mahesh Anand/Bank.csv',index_col=0)
A.head()
```

Out[5]:

User I.D	Age	Income	Income2	Deposit	Deposit2	Customer_type	Deposit_Scheme
ACX570081	26	32900	20230	14805.0	11935.7	Irregular	Hal-Yearly
ACX570082	43	37390	21410	19442.8	10276.8	Regular	Quaterly
ACX570083	35	11300	22290	5989.0	9361.8	Irregular	Monthly
ACX570084	27	41680	26970	19589.6	15912.3	Irregular	Quaterly
ACX570085	42	27170	27220	14943.5	11160.2	Regular	Quaterly

In [6]:
```
A.shape
```

Out[6]:
```
(264, 7)
```

In [ ]:
```
A.columns
```

Out[6]:
```
Index(['Age', 'Income', 'Income2', 'Deposit', 'Deposit2', 'Customer_type',
       'Deposit_Scheme'],
      dtype='object')
```

In [7]:
```
stats.ttest_rel(A.Income, A.Income2)
```

Out[7]:
```
Ttest_relResult(statistic=7.3845805999351475, pvalue=2.0214409662318323e-12)
```

In [8]:
```
A['Income'].mean(),A['Income2'].mean()
```

Out[8]:
```
(30104.583333333332, 25014.545454545456)
```

In [9]:
```
stats.ttest_rel(A.Deposit, A.Deposit2)
```

Out[9]:
```
Ttest_relResult(statistic=6.9895298325640995, pvalue=2.266359033247305e-11)
```

In [ ]:
```
A['Deposit'].mean(),A['Deposit2'].mean()
```

Out[16]:
```
(15601.345075757581, 12915.172348484848)
```

Two sample Z-Test for difference in proportions

- Verify for the Migraine data, verify the headache proportion is same for male and female patients
- H_0: P1 = P2
- H_a: P1 $\neq$ P2

In [10]:
```
#Holiday preference quiz
p1=209/489
p2=225/473
pp=434/(489+473)
p1,p2
```

Out[10]:
```
(0.4274028629856851, 0.47568710359408034)
```

In [11]:
```
z_data=(p1-p2)/np.sqrt(pp*(1-pp)*(1/489+1/473))
z_data
```

Out[11]:
```
-1.5045828782072506
```

In [ ]:
```
#the inference from the above answer is
#the both proportions are near equal that means less than 1.96
#so pvalue > 5 % , so Ho holds good,that means the proportions are equal
```

In [13]:
```
proportions_ztest([209,225],[489,473])
```

Out[13]:
```
(-1.5045828782072506, 0.13243135094767947)
```

In [14]:
```
#cruise holiday
p1=280/489
p2=248/473
pp=(280+248)/(489+473)
p1,p2
```

Out[14]:
```
(0.5725971370143149, 0.5243128964059197)
```

In [15]:
```
proportions_ztest([280,248],[489,473])
```

Out[15]:
```
(1.5045828782072488, 0.1324313509476798)
```

In [ ]:
```
z_data=(p1-p2)/np.sqrt(pp*(1-pp)*(1/489+1/473))
z_data
```

Out[31]:
```
1.5045828782072488
```

In [16]:
```
M=pd.read_csv('/content/drive/My Drive/Statistics Mahesh Anand/Migraine.csv',index_col=0)
M.head()
```

Out[16]:

	id	time	dos	hatype	age	airq	medication	headache	Gender
1	1	-11	753	Aura	30	9.0	continuing	yes	female
2	1	-10	754	Aura	30	7.0	continuing	yes	female
3	1	-9	755	Aura	30	10.0	continuing	yes	female
4	1	-8	756	Aura	30	13.0	continuing	yes	female
5	1	-7	757	Aura	30	18.0	continuing	yes	female

In [20]:
```
M.shape[0]
```

Out[20]:
```
4152
```

In [17]:
```
#Let P1 be the headache=yes for female
#Let P2 be the headache=yes for male
CT=pd.crosstab(M['headache'],M['Gender'])
print(CT)
```

```
Gender    female  male
headache
no          1266   220
yes         2279   387
```

In [22]:
```
p1=2279/(1266+2279)
p2=387/(607)
p1,p2
```

Out[22]:
```
(0.6428772919605078, 0.6375617792421746)
```

In [ ]:
```
1266+2279
```

Out[35]:
```
3545
```

In [21]:
```
#pooled proportion
pp=(2279+387)/M.shape[0]
pp
```

Out[21]: 0.6421001926782274

In [23]:
```
z_data=(p1-p2)/np.sqrt(pp*(1-pp)*(1/3545+1/607))
z_data
```

Out[23]: 0.2524275906432048

In [12]:
```
from statsmodels.stats.proportion import proportions_ztest
```

In [ ]:
```
proportions_ztest([2279,387],[3545,607])
```

Out[41]: (0.2524275906432048, 0.8007105762350393)

In [ ]:
```
proportions_ztest(1044,1800,.58) #one sample proportion test
```

Out[49]: (0.0, 1.0)

In [25]:
```
proportions_ztest(900,1500,.6) #one sample proportion test
```

Out[25]: (0.0, 1.0)

In [ ]:
```
1044/1800
```

Out[48]: 0.58

Since the p-val >0.05 (5%) it falls in acceptence zone of null hypothesis (Ho) ie., headache proportion is same for male and female patients

Chi-square Test

- Goodness of fit tests are hypothesis tests that are used for comparing the observed distribution of data with expected distribution of the data to decide whether there is any statistically significant difference between the observed distribution and a theoretical

distribution based on the comparison of observed frequencies in the data and the expected frequencies if the data follows a specified theoretical distribution.

Hypothesis	Description
Null hypothesis	There is no statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution
Alternative hypothesis	There is statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution

Chi-square statistic for goodness of fit is given by $\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

Load the Migraine dataset and verify whether the type of migraine is dependent on Gender or not

In [ ]:
```
M['hatype'].value_counts()
```

Out[53]:
```
No Aura    1985
Aura       1710
Mixed       457
Name: hatype, dtype: int64
```

In [ ]:
```
M['Gender'].value_counts()
```

Out[54]:
```
female    3545
male       607
Name: Gender, dtype: int64
```

In [26]:
```
from scipy.stats import chi2_contingency,chisquare
```

In [27]:
```
CT=pd.crosstab(M['Gender'],M['hatype'])
CT
```

Out[27]:

hatype	Aura	Mixed	No Aura
Gender			
female	1593	291	1661
male	117	166	324

In [28]:
```
1593/3545,291/3545,1661/3545
```

Out[28]:
```
(0.44936530324400564, 0.08208744710860366, 0.4685472496473907)
```

In [29]:
```
117/607,166/607,324/607
```

Out[29]:
```
(0.1927512355848435, 0.27347611202635913, 0.5337726523887973)
```

In [30]:
```
chi2_contingency(CT)
```

Out[30]:
```
(259.94962922327386,
 3.569893234435195e-57,
 2,
 array([[1460.00722543,  390.18906551, 1694.80370906],
        [ 249.99277457,   66.81093449,  290.19629094]]))
```

In [ ]:
```
#first value is the chi - square value
#the table at the last is the expected count
```

Post-hoc Analysis

- Female are highly sensitive to aura
- Male are highly sensitive to mixed

In [31]:
```
#Expected Frequencies
(3545*1710)/M.shape[0]
```

Out[31]: 1460.007225433526

In [32]:
```
(3545*457)/M.shape[0]
```

Out[32]: 390.1890655105973

In [ ]:
```
(3545*1985)/M.shape[0]
```

Out[61]: 1694.8037090558767

In [ ]:
```
(607*1710)/M.shape[0]
```

Out[62]: 249.992774566474

In [ ]:
```
(607*457)/M.shape[0]
```

Out[63]: 66.8109344894027

In [ ]:
```
(607*1985)/M.shape[0]
```

Out[64]: 290.1962909441233

In [ ]:
```
#Prop of Aura
1593/3545,117/607
```

Out[12]: (0.44936530324400564, 0.1927512355848435)

In [ ]:
```
#Prop of Mixed type
291/3545,166/607
```

Out[60]: (0.08208744710860366, 0.27347611202635913)

In [ ]:
```
#Prop of No-Aura
1661/3545,324/607
```

Out[61]:
```
(0.4685472496473907, 0.5337726523887973)
```

In [ ]:
```
chi2_contingency(CT)
```

Out[67]:
```
(259.94962922327386,
 3.569893234435195e-57,
 2,
 array([[1460.00722543,  390.18906551, 1694.80370906],
        [ 249.99277457,   66.81093449,  290.19629094]]))
```

In [33]:
```
oc=np.array([1593,117,291,166,1661,324])
ec=np.array([1460.00722543,249.99277457,390.18906551, 66.81093449,1694.80370906,290.19629094])
```

In [34]:
```
np.sum((oc-ec)**2/ec)
```

Out[34]:
```
259.9496292245235
```

- P-val <0.05, which rejects Ho, ie., there is a significant difference in proportions of migraine type among male and female
- We can do the post-hoc analysis and infer the characteristics of migraine types