In [1]:

```
import pandas as pd
import numpy as np
import scipy.stats as st
```

In [2]:

```
df = pd.read_csv("Marketing_Data-1.csv",index_col=0)
my_data = df
lst =['Year_Birth', 'Education', 'Marital_Status', 'Income']
df.columns
```

Out[2]:

```
Index(['Year_Birth', 'Education', 'Marital_Status', 'Income', 'Kidhome',
       'Teenhome', 'Dt_Customer', 'Recency', 'MntWines', 'MntFruits',
       'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts',
       'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases',
       'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth',
       'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1',
       'AcceptedCmp2', 'Response', 'Complain', 'Country', 'Age'],
      dtype='object')
```

In [3]:

```
# Adding columns Purchases which contain total purchases from company
df1 = my_data
df1["Purchases"] = df1["NumDealsPurchases"]+df1['NumWebPurchases']+df1['NumCatalogPurchases
df1["MntTotal"]=df1['MntWines']+df1['MntFruits']+df1['MntMeatProducts']+df1['MntFishProduct
```

Runs test

H0 : The data is produced in a random manner based on Income

H1 : The data is not produced in a random manner based on Income

In [4]:

```
from statsmodels.sandbox.stats.runs import runstest_1samp
# Data = "M_Data.csv"

data = df["Income"]

#Perform Runs test
z,p = runstest_1samp(data,cutoff='median', correction=False)
print("Z- value = ",z)
print("P- value = ",p)
if z<0:
    print("----> A z-score of less than 0 represents an element less than the median.")
alpha = 0.05
z_alpha = st.norm.ppf(1-alpha)
print("Z-table value = ",z_alpha)
if z<0:
    z_alpha = -z_alpha
if z<z_alpha:
    print("----> Z stat value doesnot fall in rejection region Z-value < Z - table value\n-
elif z>z_alpha:
     print("----> Z stat value fall in rejection region Z-value > Z - table value\n----> Fa
```

```
Z- value =  -5.853544528527859
P- value =  4.8120533902742105e-09
----> A z-score of less than 0 represents an element less than the median.
```

```
# Conclusion :
----> As elements n>=30 So we are using Z -test for comparison

----> A z-score of less than 0 represents an element less than the median.

----> Z stat value doesnot fall in rejection region  Z-value < Z - table value

----> Reject H0

The data "M_Data" is not produced randomly based on Income.
```

Sign Test One Sample

Number of median purchases assumed by company is 15

H0 : median = 15

H1 : Median != 15 alpha =0.05

Condition: Sample size< 26 Critical value for two tailed test at n=25 is 6

In [6]:

```
Sign_Sample = df1.sample(n=25,replace=False)
colums = ['Year_Birth','Income','Country', 'Age','Purchases']

def sign_test(lst,median):
    n = len(lst)
    nc=0
    pc=0
    sign = []
    for i in range(0,n):
        if lst[i]-median>0:
            pc=pc+1
            sign.append("+")
        elif lst[i]-median<0:
            nc=nc+1
            sign.append("-")
        elif lst[i]-median == 0:
            nc=nc
            pc=pc
            sign.append("0")
    return nc,pc,sign;

if __name__ == "__main__":
    median = 15
    lst = Sign_Sample["Purchases"].tolist()
    nc,pc,sign = sign_test(lst,median)

    Sign_Sample['sign']=sign
    colums.append('sign')
    print(Sign_Sample[colums])
    count =nc + pc
    x = min(nc,pc)
    print(""+" signs = ",pc,"\t- signs = ",nc)
    print("N value is = ",count)
    print("Find Critical Value for N=",count," in table\n")
    print("Give critical value as Input")
    critical_value = int(input())
    print("Test Value is = ",x)
    if(x<=critical_value):
        print(" Test Value is less than critical Value.")
        print(x,"<=",critical_value,"\n--->Null Hypothesis is Rejected")
        print(" Median is not equal to ",median)
    if(x>critical_value):
        print(" Test Value is greater than critical Value.")
        print(x,">",critical_value,"\n--->Failed to reject Null Hypothesis")
        print(" Median is equal to ",median)
```

```
      Year_Birth   Income Country  Age  Purchases sign
S.No
118         1949  81698.0      SP   72         25    +
1575        1971  74290.0      SP   50         32    +
1352        1971  71969.0     IND   50         19    +
1415        1952  55951.0      SA   69         23    +
358         1962  42769.0      CA   59          8    -
472         1945  70356.0      CA   76         27    +
2047        1959  71232.0      SP   62         25    +
```

```
488         1954  50002.0      SA   67          19    +
1913        1970  44159.0      SA   51          15    0
1815        1973  71128.0      US   48          27    +
359         1982  58582.0      SP   39          22    +
4           1989  21474.0      SP   32           8    -
318         1992  15253.0      CA   29           6    -
574         1944  80589.0     AUS   77          21    +
1855        1981  36038.0      SP   40           8    -
1691        1943  51381.5     AUS   78          22    +
1867        1981  24336.0      SP   40           4    -
850         1958  46692.0      SP   63          15    0
1774        1977  41443.0      SA   44          20    +
599         1975  42160.0      SP   46          11    -
2063        1972  42618.0      SP   49          11    -
1739        1970  58710.0     AUS   51          31    +
682         1959  65031.0     GER   62          30    +
1988        1974  45837.0      SP   47          18    +
740         1958  68281.0      SP   63          25    +
 signs =  16    - signs =  7
N value is =  23
Find Critical Value for N= 23  in table

Give critical value as Input
6
Test Value is =  7
 Test Value is greater than critical Value.
7 > 6
--->Failed to reject Null Hypothesis
 Median is equal to  15
```

Mann Whitney U Test

Checking Whether there is differnece between amount spent on products among singles and married

H0: There is no difference on amount spent on purchasing products between singles and Married

H1: There is difference on amount spent on purchasing products between singles and Married

In [7]:

```
# New data is original data with purchases
new_data = df1
import pandas as pd
import numpy as np
import scipy.stats as st
data1 =df1.loc[df["Marital_Status"]=="Single"]
data2 =df1.loc[df["Marital_Status"]=="Married"]

# data1 = data of singles data2 = data of married persons
singles=data1["MntTotal"].tolist()
Married=data2["MntTotal"].tolist()

# Performing Manwhittney U Test or Wilcoxon Rank Sum Test
from scipy.stats import mannwhitneyu
U1, p = mannwhitneyu(singles, Married)
print("Test Statistic U1 is = ",U1)
nx, ny = len(singles), len(Married)
U2 = nx*ny - U1
print("Test Statistic U2 is = ",U2)
print("manwhitneyu p value is = ",p)

import numpy as np
from scipy.stats import norm
U = min(U1, U2)
N = nx + ny
z = (U - nx*ny/2 + 0.5) / np.sqrt(nx*ny * (N + 1)/ 12)
p = 2 * norm.cdf(z)  # use CDF to get p-value from smaller statistic
print("P-value is = ", p)
print("Z- Value is = ",z)


alpha = 0.05
z_alpha = st.norm.ppf(1-alpha)
print("Z-table value = ",z_alpha)
if z<0:
    z_alpha = -z_alpha
if(z<=z_alpha):
    print(" Test Value is less than critical Value.")
    print(z,"<=",z_alpha,"\n--->Null Hypothesis is Rejected")
    print("There is difference on amount spent on purchasing products between singles and M
elif(z>z_alpha):
    print(" Test Value is greater than critical Value.")
    print(z,">",z_alpha,"\n--->Failed to reject Null Hypothesis")
    print("There is no difference on amount spent on purchasing products between singles an
```

```
Test Statistic U1 is =  563156.5
Test Statistic U2 is =  582585.5
manwhitneyu p value is =  0.2531744435010702
P-value is =  0.5063504962214381
Z- Value is =  -0.6645310290084294
Z-table value =  1.6448536269514722
 Test Value is greater than critical Value.
-0.6645310290084294 > -1.6448536269514722
--->Failed to reject Null Hypothesis
There is no difference on amount spent on purchasing products between single
s and Married
```

Conclusion

There is no difference on amount spent on purchasing products between singles and Married.

Wilcoxon Signed Rank Test

Check whether the customer who purchases meat also purchases fish Non vegeteraian

H0: There is no difference among customers purchasing fish and meat

H1: There is difference among customers purchasing fish and meat

In [8]:

```
wilcox_Sample = df1.sample(n=30,replace=False)
columns =['Year_Birth', 'Education', 'Income','MntMeatProducts', 'MntFishProducts']
print(wilcox_Sample[columns])
meat = wilcox_Sample['MntMeatProducts'].tolist()
fish = wilcox_Sample['MntFishProducts'].tolist()
import numpy as np
from scipy.stats import wilcoxon
w1, p1= wilcoxon(meat,fish,alternative='two-sided')
print("Test statistic value is = ",w1,"\n P-Value is = ", p1)
print("Interfer the result or Give critical value for n=30 at alpha =0.05 from the table")
z=int(input())

if(w1<=z):
    print(" Test Value is less than critical Value.")
    print(w1,"<=",z,"\n--->Null Hypothesis is Rejected")
    print("There is difference among customers purchasing fish and meat")
elif(w1>z):
    print(" Test Value is greater than critical Value.")
    print(w1,">",z,"\n--->Failed to reject Null Hypothesis")
    print("There is no difference among customers purchasing fish and meat")

#length =len(before)
#w2 = (n*(n+1)/2)-w1
#w = min(w1,w2)
#z= (w - ( n*(n+1)/4 )) /np.sqrt(n*(n+1)*(2*n+1)/24)
#print(z)
```

S.No	Year_Birth	Education	Income	MntMeatProducts	MntFishProducts
612	1963	Master	57288.0	21	0
2118	1959	Graduation	24221.0	9	2
1665	1969	Graduation	38361.0	56	20
229	1952	Master	43776.0	71	3
1043	1963	Graduation	80124.0	398	205
2087	1950	Graduation	27203.0	21	4
642	1954	Master	60033.0	57	19
1914	1962	Graduation	76081.0	415	63
1892	1965	Graduation	81168.0	592	147
1091	1954	Graduation	64587.0	16	0
1706	1954	Master	62637.0	48	4
1924	1967	Master	30753.0	25	0
1952	1985	Master	33812.0	19	30
1980	1957	Graduation	78618.0	818	212
1873	1964	Graduation	82224.0	360	138
1559	1973	Graduation	67432.0	341	177
994	1970	Graduation	76467.0	426	210
59	1964	Graduation	60597.0	257	32
1054	1972	PhD	59973.0	168	20
256	1968	Graduation	75693.0	293	72
1367	1986	PhD	82333.0	359	46
1690	1973	Graduation	34853.0	15	2
245	1976	Master	26907.0	7	0
1807	1947	PhD	68117.0	215	0
98	1980	Graduation	80011.0	536	82
1049	1968	Graduation	19514.0	21	2
1228	1967	Master	47821.0	16	6
1780	1963	Graduation	49980.0	54	13
10	1947	Master	81044.0	535	73

```
1932        1959         PhD  38829.0                7                0
Test statistic value is =  5.0
 P-Value is =  2.8716584471854804e-06
Interfer the result or Give critical value for n=30 at alpha =0.05 from t
he table
136
 Test Value is less than critical Value.
5.0 <= 136
--->Null Hypothesis is Rejected
There is difference among customers purchasing fish and meat
```

In [10]:

```
new_data.to_csv("Company_Data.csv")
```