ASSIGNMENT- STATISTICS [MAJOR] BY NISHANT MISHRA

EMAIL- nm9169336@gmail.com

CONTACT NO.- 9873942716

In [1]:

```
# Questions- 1
```

According to a study, the daily average time spent by a user on asocial media website is 50 minutes. To test the claim of this study,Ramesh, a researcher, takes a sample of 25 website users and findsout that the mean time spent by the sample users is 60 minutes andthe sample standard deviation is 30 minutes.Based on this information, the null and the alternative hypotheses will be:

Ho = The average time spent by the users is 50 minutes

H1 = The average time spent by the users is not 50 minutes

Use a 5% significance level to test this hypothesis.

In [2]:

```
import scipy.stats as stats
import numpy as np

sample_mean = 60
sample_std = 30
n = 25
d_mean = 50

t_statistics = (sample_mean - d_mean) / (sample_std / np.sqrt(n))
p_value = 2*(1-stats.t.cdf(abs(t_statistics), df=n-1))

print('t_statistics:', t_statistics)
print('P_value:', p_value)

if p_value < 0.05:
    print('Reject Null Hypothesis')
else:
    print('Failed to Reject null hypothesis')
```

```
t_statistics: 1.6666666666666667
P_value: 0.10858012302472297
Failed to Reject null hypothesis
```

In [3]:

```
# Question-2
```

Height of 7 students (in cm) is given below. What is the median?

(168, 170, 169, 160, 162, 164, 162)

In [4]:

```
# importing libraries
import statistics as stats

heights = [168,170,169,160,162,164,162]
median_heights = stats.median(heights)

print('Median is:', median_heights)
```

```
Median is: 164
```

In [5]:

```
# Question-3
```

Below are the observations of the marks of a student. Find the value of mode.

(84, 85, 89, 92, 93, 89, 87, 89, 92)

In [6]:

```
# importing libraries
import statistics as stats

marks = [84, 85, 89, 92, 93, 89, 87, 89, 92]
mode_marks = stats.mode(marks)
print('Mode is:', mode_marks)
```

```
Mode is: 89
```

In [7]:

```
# Question-4
```

From the table given below, what is the mean of marks obtained by 20 students?

Marks = [3,4,5,6,7,8,9,10]

NO. of students(frequency) = [1,2,2,4,5,3,2,1]

In [8]:

```
import numpy as np

# calculating mean
marks = [3, 4, 5, 6, 7, 8, 9, 10]
freq = [1, 2, 2, 4, 5, 3, 2, 1]

mean = np.average(marks, weights=freq)

print("Mean of marks obtained by students is:", mean)
```

```
Mean of marks obtained by students is: 6.6
```

In [9]:

```
# Question-5
```

For a certain type of computer, the length of time between charges ofthe battery is normally distributed with a mean of 50 hours and astandard deviation of 15 hours. John owns one of these computersand wants to know the probability that the length of time will bebetween 50 and 70 hours.

In [10]:

```
from scipy.stats import norm

mu = 50
sigma = 15

# Probability that length of time will be less than or equal to 70 hours
prob1 = norm.cdf(70, mu, sigma)

# Probability that length of time will be less than or equal to 50 hours
prob2 = norm.cdf(50, mu, sigma)

# Probability that length of time will be between 50 and 70 hours
prob = prob1 - prob2

print("The probability that the length of time will be between 50 and 70 hours is:", prob)
```

```
The probability that the length of time will be between 50 and 70 hours is: 0.4087887802741321
```

So there is a 40.8% chance that the length of time will be between 50 and 70 hours.

In [11]:

```
# Question-6
```

Find the range of the following.

g = [10, 23, 12, 21, 14, 17, 16, 11, 15, 19]

In [12]:

```
g = [10, 23, 12, 21, 14, 17, 16, 11, 15, 19]
range_of_g = max(g) - min(g)
print("Range of g is:", range_of_g)
```

```
Range of g is: 13
```

In [13]:

```
## Question-7
```

It is estimated that 50% of emails are spam emails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?

Solution:

Let us consider events:- A = event that an email is detected as spam, B = event that an email is spam, Bc = event that an email is not spam. Given: P(B) = 0.5, P(A | B) = 0.99, P(A | Bc) = 0.05. By the Bayes’s formula: P(Bc | A) = P(A | Bc)*P(Bc) / (P(A | B)*P(B) + P(A | Bc)*P(Bc)) = 0.05 × 0.5 / (0.05 × 0.5 + 0.99 × 0.5) = 5 / 104 = 0.048

- The probability is 4.8%

In [14]:

```
# Question-8
```

Given the following distribution of returns, determine the lowerquartile:

{10, 25,12, 21, 19, 17, 16, 11, 15, 19}

In [15]:

```
import numpy as np
g =[10,25,12,21,19,17,16,11,15,19]
lower_quartile = np.quantile(g, .25)
print('Lower Quartile-', lower_quartile)
```

```
Lower Quartile- 12.75
```

In [16]:

```
# Question-9
```

For a Binomial distribution, the number of trials(n) is 25, and theprobability of success is 0.3. What’s the variability of the distribution?

In [17]:

```
import numpy as np

n = 25
p = 0.3

q = 1 - p

variance = n * p * q

std_dev = np.sqrt(variance)

print("Standard Deviation : ", std_dev)
```

```
Standard Deviation :  2.29128784747792
```

In [18]:

```
# Question-10
```

Download the Cell Phone Survey Dataset and perform the belowmentioned operations on the dataset:

In [19]:

```
# importing librarires

import pandas as pd
import numpy as np
import statistics as stats
```

In [20]:

```
# uploading dataset
df = pd.read_csv('cell phone survey.csv')
df.head(10)
```

Out[20]:

	Gender	Carrier	Type	Usage	Signal strength	Value for the Dollar	Customer Service
0	M	AT&T	Smart	High	5	4	4
1	M	AT&T	Smart	High	5	4	2
2	M	AT&T	Smart	Average	4	4	4
3	M	AT&T	Smart	Very high	2	3	3
4	M	AT&T	Smart	Very high	5	5	2
5	M	AT&T	Smart	Very high	4	3	5
6	M	AT&T	Smart	Very high	3	4	4
7	F	AT&T	Smart	Very high	3	2	3
8	F	AT&T	Smart	Very high	4	3	4
9	M	AT&T	Smart	Very high	3	3	1

In [21]:

```
# 1. Checking datatypes of each column in the dataset
df.info()
df.dtypes
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52 entries, 0 to 51
Data columns (total 7 columns):
 #   Column                Non-Null Count  Dtype 
---  ------                --------------  ----- 
 0   Gender                52 non-null     object
 1   Carrier               52 non-null     object
 2   Type                  52 non-null     object
 3   Usage                 52 non-null     object
 4   Signal strength       52 non-null     int64 
 5   Value for the Dollar  52 non-null     int64 
 6   Customer Service      52 non-null     int64 
dtypes: int64(3), object(4)
memory usage: 3.0+ KB
```

Out[21]:

```
Gender                  object
Carrier                 object
Type                    object
Usage                   object
Signal strength          int64
Value for the Dollar     int64
Customer Service         int64
dtype: object
```

In [22]:

```
# 2.  Find Mean of Signal strength column using Pandas and Statistics library.

# using pandas
df_mean = df['Signal strength'].mean()
print('Mean of signal strenght is:', df_mean)

# using statistics library
df_mean1 = stats.mean(df['Signal strength'])
```

```
print('Mean of signal strenght is:', df_mean1)
```

```
Mean of signal strenght is: 3.3076923076923075
Mean of signal strenght is: 3.3076923076923075
```

In [23]:

```
# 3. Find the Median of Customer Service column using Pandas and Statistics library.

# using pandas
df_median = df['Customer Service'].median()
print('Median of Customer Service is:', df_median)

# using statistics library
df_median1 = stats.median(df['Customer Service'])
print('Median of Customer Service is:', df_median1)
```

```
Median of Customer Service is: 3.0
Median of Customer Service is: 3.0
```

In [24]:

```
# 4. Find Mode of Signal strength column using Pandas and Statistics library.

# using pandas
df_mode = df['Signal strength'].mode()
print('Mode of Signal Strenth:', df_mode)

# using statistics library
df_mode1 = stats.mode(df['Signal strength'])
print('Mode of Signal Strenth:', df_mode1)
```

```
Mode of Signal Strenth: 0    3
Name: Signal strength, dtype: int64
Mode of Signal Strenth: 3
```

In [25]:

```
# 5. Find Standard deviation of Customer Service column using Pandas and Statistics library.

# using pandas
df_std = stats.stdev(df['Customer Service'])
print("STD. of Customer service is:", df_std)

# using statistics library
df_std1 = df['Customer Service'].std()
print("STD. of Customer Service is:", df_std1)
```

```
STD. of Customer service is: 0.9623375261979595
STD. of Customer Service is: 0.9623375261979594
```

In [26]:

```
# 6. Find Variance of Customer Service column using Pandas and Statistics library.

# using pandas
df_var = stats.variance(df['Customer Service'])
print("Variance Customer Service:", df_var)

# using statistics library
df_var1 = df['Customer Service'].var()
print("Variance of Customer Service:", df_var1)
```

```
Variance Customer Service: 0.9260935143288085
Variance of Customer Service: 0.9260935143288083
```

In [27]:

```
# 7. Calculate Percentiles of Value for the Dollar column using Numpy.

# using numpy library
def_col = df['Value for the Dollar']
df_quartiles = np.quantile(def_col, [0,0.25,0.500,0.75,1])
print("Percentiles for the Dollar:", df_quartiles)
```

```
Percentiles for the Dollar: [1. 3. 3. 4. 5.]
```

In [28]:

```
# 8. Calculate Range of Value for the Dollar column using Pandas.

df_range = df['Value for the Dollar'].max() - df['Value for the Dollar'].min()
print('Range for the dollar column:', df_range)
```

```
Range for the dollar column: 4
```

In [29]:

```
# 9.Calculate IQR of Value for the Dollar column using Pandas.

# using numpy library
qr = df['Value for the Dollar']
qr1 = np.percentile(qr, 25)
qr2 = np.percentile(qr, 75)
IQR = qr2 - qr1
print('Interquartile range of dollar column:', IQR)
```

```
Interquartile range of dollar column: 1.0
```

In [30]:

```
# 10. Hypothesis Testing - Using the data in the Cell Phone Survey dataset,apply ANOVA to determine
#     if the mean response for Value for dollar is the same for different types of cell phones.
```

In [31]:

```
# checking different types of cell phones
df['Type'].unique()
```

Out[31]:

```
array(['Smart', 'Camera', 'Basic'], dtype=object)
```

In [32]:

```
# applying ANOVA Test

import scipy.stats as stats
stats.f_oneway(df['Value for the Dollar'][df['Type'] == 'Basic'],
               df['Value for the Dollar'][df['Type'] == 'Camera'],
               df['Value for the Dollar'][df['Type'] == 'Smart'])
```

Out[32]:

```
F_onewayResult(statistic=3.1111943528010304, pvalue=0.053454200712805613)
```

Since the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that there is no significant difference between the means of different groups.