Student's T-Test

In 1908 **William Sealy Gosset**, an Englishman publishing under the pseudonym Student, developed the t-test and t distribution.

A t-test is a statistical test that is used to compare the means of two groups. It is often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

There are three types of t-tests we can perform based on the data at hand: One sample t-test. Independent two-sample t-test. Paired sample t-test.

In [ ]:
```python
#import Libararies
import pandas as pd
import seaborn as sns
import scipy as sc
import matplotlib.pyplot as plt
```

In [ ]:
```python
kashti = sns.load_dataset("titanic")
kashti.head()
```

Out[ ]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN

In [ ]:
```python
kashti.isna().sum()
```

Out[ ]:
```
survived         0
pclass           0
sex              0
age            177
sibsp            0
parch            0
fare             0
embarked         2
class            0
who              0
adult_male       0
deck           688
embark_town      2
alive            0
alone            0
dtype: int64
```

In [ ]:
```python
kashti.dropna(subset=['age'],axis=0,inplace=True)
kashti.isna().sum()
```

Out[ ]:
```
survived        0
pclass          0
sex             0
age             0
sibsp           0
parch           0
fare            0
embarked        2
class           0
who             0
adult_male      0
deck          530
embark_town     2
alive           0
alone           0
dtype: int64
```

In [ ]:
```python
df = kashti[['sex','age','fare']]
df.head()
```

Out[ ]:

	sex	age	fare
0	male	22.0	7.2500
1	female	38.0	71.2833
2	female	26.0	7.9250
3	female	35.0	53.1000
4	male	35.0	8.0500

One-sample student's t-test

Test a sample with a known standard value. **Assumptions**

- Observations in each sample are independent and identically distributed.
- Observations in each sample are normally distributed.
- **Interpretation**

H0: the means of the samples are equal to the known value.

H1: the means of the samples are unequal to the known value.

In [ ]:
```python
# 1 sample t test to compare the age of male vs female

#1. import libarary
from scipy.stats import ttest_1samp

#2. sub set of male vs female
df_male =  df[df['sex']=='male']
df_female =  df[df['sex']=='female']

#3. t test
stat,p = ttest_1samp(df_male['age'],36)
print('stat=%.3f,p=%.3f'% (stat,p))

#4. make a conditional argument for further case
if p > 0.05:
```

```
    print('There is no significance difference')
else:
    print('There is a significance difference')
```

```
stat=-7.647,p=0.000
There is a significance difference
```

Independent student's t-test

Assumptions

- Observations in each sample are independent and identically distributed.
- Observations in each sample are normally distributed.
- Observations in each sample have the same variance.

Interpretation

H0: the means of the samples are equal.

H1: the means of the samples are unequal

```
In [ ]:   # 2 sample t test to compare the age of male vs female

          #1. import libarary
          from scipy.stats import ttest_ind

          #2. sub set of male vs female
          df_male =  df[df['sex']=='male']
          df_female =  df[df['sex']=='female']

          #3. t test(unpaired/two sample/independent)
          stat,p = ttest_ind(df_male['age'],df_female['age'])
          print('stat=%.3f,p=%.3f'% (stat,p))

          #4. make a conditional argument for further case
          if p > 0.05:
            print('There is no significance difference')
          else:
              print('There is a significance difference')
```

```
stat=2.499,p=0.013
There is a significance difference
```

Paired student's t-test

Tests whether the means of two paired samples are significantly different. **Assumptions**

- Observations in each sample are independent and identically distributed.
- Observations in each sample are normally distributed.
- Observations in each sample have the same variance.
- Observations across each sample are paired.
- **Interpretation**

H0: the means of the samples are equal.

H1: the means of the samples are unequal.

In [ ]:
```python
# 2 sample t test to compare the age of male in first and second class
#1. import libarary
from scipy.stats import ttest_rel

#2. sub set of male vs female
df = kashti[['sex','age','class']]
df.head()
df_male =  df[df['sex']=='male']
df_male.head()
df_male_1st =  df_male[df_male['class']=='First']
df_male_1st.head()
df_male_2nd =  df_male[df_male['class']=='Second']
df_male_2nd.head()

# equaling the rows of the df_male_1st and df_male_2nd
df_male_1st= df_male_1st.sample(n=100,replace=True)
df_male_2nd= df_male_2nd.sample(n=100,replace=True)

#3. t test(paired/two sample/dependent)
stat,p = ttest_rel(df_male_1st['age'],df_male_2nd['age'])
print('stat=%.3f,p=%.3f'% (stat,p))

#4. make a conditional argument for further case
if p > 0.05:
  print('There is no significance difference')
else:
    print('There is a significance difference')
```

```
stat=3.970,p=0.000
There is a significance difference
```

In [ ]:
```python
# 2 sample t test to compare the age of male in first, second class and third class
#1. import libarary
from scipy.stats import ttest_rel

#2. sub set of male vs female
df = kashti[['sex','age','class']]
df.head()
df_male =  df[df['sex']=='male']
df_male.head()
df_male_1st =  df_male[df_male['class']=='First']
df_male_1st.head()
df_male_2nd =  df_male[df_male['class']=='Second']
df_male_2nd.head()
df_male_3rd =  df_male[df_male['class']=='Third']
df_male_3rd.head()

# equaling the rows of the df_male_1st and df_male_2nd
df_male_1st= df_male_1st.sample(n=100,replace=True)
df_male_2nd= df_male_2nd.sample(n=100,replace=True)
df_male_3rd= df_male_3rd.sample(n=100,replace=True)

#3. t test(paired/two sample/dependent)
stat,p = ttest_rel(df_male_1st['age'],df_male_2nd['age'])
print('stat=%.3f,p=%.3f'% (stat,p))

#4. make a conditional argument for further case
if p > 0.05:
  print('There is no significance difference')
else:
    print('There is a significance difference')
```

```
stat=5.996,p=0.000
```

There is a significance difference

In [ ]:
```python
#another way
# 2 sample t test to compare the fare and  age of male
#1. import libarary
from scipy.stats import ttest_rel

#2. sub set of male vs female
df = kashti[['sex','age','fare']]
df.head()
df_male =  df[df['sex']=='male']
df_male.head()

#3. t test(paired/two sample/dependent)
stat,p = ttest_rel(df_male['age'],df_male['fare'])
print('stat=%.3f,p=%.3f'% (stat,p))

#4. make a conditional argument for further case
if p > 0.05:
  print('There is no significance difference')
else:
    print('There is a significance difference')
```