Statistical Test

Data Launching and Data Treatment :

```
import pandas as pd
dataset = pd.read_excel("general_data_Attrition.xlsx",sheet_name=0)
dataset.dropna()
Out[23]:
     Age  Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0     51          0  ...                     0                    0
1     31          1  ...                     1                    4
2     32          0  ...                     0                    3
3     38          0  ...                     7                    5
4     32          0  ...                     0                    4
...  ...        ...  ...                   ...                  ...
4404  29          0  ...                     1                    5
4405  42          0  ...                     0                    2
4406  29          0  ...                     0                    2
4407  25          0  ...                     1                    2
4408  42          0  ...                     7                    8
[4382 rows x 24 columns]

dataset.drop_duplicates()
Out[24]:
     Age  Attrition  ... YearsSinceLastPromotion YearsWithCurrManager
0     51          0  ...                     0                    0
1     31          1  ...                     1                    4
2     32          0  ...                     0                    3
3     38          0  ...                     7                    5
```

Statistical Test

4 32 0 ... 0 4

... ...

4405 42 0 ... 0 2

4406 29 0 ... 0 2

4407 25 0 ... 1 2

4408 42 0 ... 7 8

4409 40 0 ... 3 9

[4410 rows x 24 columns]

n [4]: dataset_AYes = pd.read_excel("general_data_Attrition.xlsx",sheet_name=1)

dataset_ANo = pd.read_excel("general_data_Attrition.xlsx",sheet_name=2)

dataset_AYes.dropna()

Out[17]:

Age Attrition ... YearsSinceLastPromotion YearsWithCurrManager

0 31 1 ... 1 4

1 28 1 ... 0 0

2 47 1 ... 9 9

3 44 1 ... 0 0

4 26 1 ... 0 2

.. ..

706 29 1 ... 0 1

707 33 1 ... 0 4

708 33 1 ... 1 7

709 32 1 ... 1 2

710 37 1 ... 0 0

[705 rows x 24 columns]

dataset_AYes.drop_duplicates()

Out[18]:

Statistical Test

	Age	Attrition	...	YearsSinceLastPromotion	YearsWithCurrManager
0	31	1	...	1	4
1	28	1	...	0	0
2	47	1	...	9	9
3	44	1	...	0	0
4	26	1	...	0	2
..	...	...	...	...	...
706	29	1	...	0	1
707	33	1	...	0	4
708	33	1	...	1	7
709	32	1	...	1	2
710	37	1	...	0	0

[711 rows x 24 columns]

dataset_ANo.dropna()

Out[19]:

	Age	Attrition	...	YearsSinceLastPromotion	YearsWithCurrManager
0	51	0	...	0	0
1	32	0	...	0	3
2	38	0	...	7	5
3	32	0	...	0	4
4	46	0	...	7	7
...	...	...	...	...	...
3693	29	0	...	1	5
3694	42	0	...	0	2
3695	29	0	...	0	2
3696	25	0	...	1	2
3697	42	0	...	7	8

[3677 rows x 24 columns]

Statistical Test

dataset_ANo.drop_duplicates()

Out[20]:

	Age	Attrition	...	YearsSinceLastPromotion	YearsWithCurrManager
0	51	0	...	0	0
1	32	0	...	0	3
2	38	0	...	7	5
3	32	0	...	0	4
4	46	0	...	7	7
...	...	...	...	...	...
3694	42	0	...	0	2
3695	29	0	...	0	2
3696	25	0	...	1	2
3697	42	0	...	7	8
3698	40	0	...	3	9

[3699 rows x 24 columns]

Non Parametric Test :

1. **Mann-whitney :**
   Its is used to compare two independent samples

   H0: There is no significant difference between Attrition yes's DistanceFromHome and Attrition No's DistanceFromHome.
   H1: There is significant difference between Attrition yes's DistanceFromHome and Attrition No's DistanceFromHome.

   ```
   from scipy.stats import mannwhitneyu
   stats,p = mannwhitneyu(dataset_AYes.DistanceFromHome,dataset_ANo.DistanceFromHome)
   print(stats,p)
   1312110.0 0.4629185205822659
   ```

   P < 0.05 => H1 is Accepted.

Statistical Test

2. **Chi-square :**
   It is used to check the dependency between categorical variables.

   H0: There is no significant difference between Attrition and Gender.
   H1: There is significant difference between Attrition and Gender.

   ```
   from scipy.stats import chi2_contingency
   chitable = pd.crosstab(dataset.Attrition,dataset.Gender)
   stats,p,dof,expected = chi2_contingency(chitable)
   print(stats,p)
   1.349904410246582 0.24529482862926827
   ```

   Since p > 0.05 , Ho is accepted.

Parametric Test :

1. **One sample t-test :**
   It is used to compare sample mean with population mean

   H0: There is no significant difference between training time of employee and standard training time as 3.
   H1: There is significant difference between training time of employee and standard training time as 3.

   ```
   from scipy.stats import ttest_1samp
   stats,p = ttest_1samp(dataset.TrainingTimesLastYear,3)
   print(stats,p)
   -10.338997107228291 8.987949368189617e-25
   ```

   Since p < 0.05 => Ho is rejected.

Statistical Test

2. **Two sample Independent :**
   It is used to compare mean of two independent samples

   H0: There is no significant difference between between Attrition yes's MonthlyIncome and Attrition No's MonthlyIncome.

   H1: There is significant difference between Attrition yes's MonthlyIncome and Attrition No's MonthlyIncome.

```
from scipy.stats import ttest_ind
stats,p = ttest_ind(dataset_AYes.MonthlyIncome,dataset_ANo.MonthlyIncome)
print(stats,p)
-2.0708863763619316 0.03842748490605113
```

   Since $p < 0.05$, Ho is rejected.