Student_test

March 25, 2022

Student Name : Sartaj Ahmed Salman

Email: s2140019@edu.cc.uec.ac.jp

Phd Student At UEC Tokyo, Japan

Address: From Skardu, Pakistan

1 Student's t-tests

1. one-sample t-test
2. two-sample t-test
    1. Unpaired or independent t-test
    2. Paired or relational/dependent t-test

1.1 One-sample t-test

Take a sample with a known standard value.

1.2 Assumptions

- Observations in sample is independent and identically distributed.
- Observations in sample is normally distributed.

1.3 Interpretation

H0: the mean of a sample is equal to the known value.

H1: the mean of a sample is unequal to the known value.

[ ]:
```
# Import Libraries
import seaborn as sns
import pandas as pd
from scipy.stats import ttest_1samp
```

[ ]:
```
#load dataset
df = pd.read_csv('sample_30000.csv')
df.head()
```

[ ]:
```
  origin destination airline  refundable  baggage_weight  baggage_pieces  \
0      x           y   gamma           1        0.444444               2
```

```
1      x           y   alpha           1        0.444444               1
2      x           y   gamma           1        0.711111               2
3      x           y   alpha           1        0.333333               1
4      x           y   gamma           1        0.444444               1

  flight_number purchace_date departure_date arival_date departure_time  \
0           c-6    2021-05-17     2021-05-29  2021-05-29       17:00:00
1          a-23    2021-08-13     2021-08-24  2021-08-24       16:00:00
2           c-6    2021-02-12     2021-03-10  2021-03-10       17:00:00
3           a-5    2021-07-13     2021-07-14  2021-07-14       07:00:00
4           c-6    2021-05-16     2021-07-16  2021-07-16       17:00:00

  arival_time purchase_time
0    18:59:00      16:56:59
1    18:00:00      08:37:22
2    18:59:00      17:24:58
3    09:00:00      07:51:11
4    18:59:00      20:45:03
```

[ ]:
```
df1 = df[['airline','baggage_weight','baggage_pieces']]
df1.head()
```

[ ]:
```
  airline  baggage_weight  baggage_pieces
0   gamma        0.444444               2
1   alpha        0.444444               1
2   gamma        0.711111               2
3   alpha        0.333333               1
4   gamma        0.444444               1
```

[ ]:
```
(df1==0).sum()
```

[ ]:
```
airline            0
baggage_weight    807
dtype: int64
```

[ ]:
```
df1 = df1.loc[df1['baggage_weight']*df1['baggage_pieces'] != 0]
```

[ ]:
```
# Data description
df1.describe()
```

[ ]:
```
       baggage_weight  baggage_pieces
count    23710.000000    23710.000000
mean         0.486624        1.204049
std          0.157688        0.403013
min          0.333333        1.000000
25%          0.333333        1.000000
50%          0.444444        1.000000
```

```
75%          0.444444        1.000000
max          0.777778        2.000000
```

[ ]:
```
# Check the Baggage weight and compare with the no value of .8 weight
ttest_1samp(df["baggage_weight"], .8)
```

[ ]:
```
Ttest_1sampResult(statistic=-264.66191503610526, pvalue=0.0)
```

1.4 Two-sample t-test

Independent t-test

1.5 Assumptions

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

1.6 Interpretation

H0: the means of a samples are equal.

H1: the means of a samples are unequal.

[ ]:
```
# comparison of baggage weight of arilines aplha and gamma
df_alpha = df1.loc[df1['airline']=='alpha']
df_gamma = df1.loc[df1['airline']=='gamma']
```

[ ]:
```
df_alpha
```

[ ]:
```
      airline  baggage_weight  baggage_pieces
1       alpha        0.444444               1
3       alpha        0.333333               1
6       alpha        0.777778               1
7       alpha        0.777778               1
10      alpha        0.333333               1
…       …             …                 …
29990   alpha        0.777778               1
29991   alpha        0.333333               1
29992   alpha        0.777778               1
29998   alpha        0.333333               1
29999   alpha        0.777778               1

[12792 rows x 3 columns]
```

[ ]:
```
df_gamma
```

[ ]:
```
      airline  baggage_weight  baggage_pieces
0       gamma        0.444444               2
```

```
2       gamma        0.711111               2
4       gamma        0.444444               1
5       gamma        0.711111               1
8       gamma        0.444444               1
…       …             …              …
29986   gamma        0.444444               2
29987   gamma        0.444444               1
29988   gamma        0.444444               1
29995   gamma        0.444444               1
29997   gamma        0.444444               2

[10063 rows x 3 columns]
```

[ ]:
```
# Independent t-test
from scipy.stats import ttest_ind
stats , p_value =␣
 ↪ttest_ind(df_alpha['baggage_weight'],df_gamma['baggage_weight'])
print('stat = %.3f, p_value =%.3f'%(stats,p_value))

if p_value > 0.5:
    print("Data is probabily normal or Guassian")
else:
    print("Data is probabily not Guassian")
```

```
stat = 25.597, p_value =0.000
Data is probabily not Guassian
```

[ ]:
```
df_alpha.describe()
```

[ ]:
```
       baggage_weight  baggage_pieces
count    12792.000000        12792.0
mean         0.511952            1.0
std          0.204899            0.0
min          0.333333            1.0
25%          0.333333            1.0
50%          0.444444            1.0
75%          0.777778            1.0
max          0.777778            1.0
```

[ ]:
```
df_gamma.describe()
```

[ ]:
```
       baggage_weight  baggage_pieces
count    10063.000000   10063.000000
mean         0.458012       1.480771
std          0.058603       0.499655
min          0.444444       1.000000
25%          0.444444       1.000000
50%          0.444444       1.000000
```

```
75%          0.444444        2.000000
max          0.711111        2.000000
```

[ ]:
```
# Independent t-test
from scipy.stats import ttest_ind
stats , p_value =␣
 ↪ttest_ind(df_alpha['baggage_pieces'],df_gamma['baggage_pieces'])
print('stat = %.3f, p_value =%.3f'%(stats,p_value))

if p_value > 0.5:
    print("Data is probabily normal or Guassian")
else:
    print("Data is probabily not Guassian")
```

```
stat = -108.828, p_value =0.000
Data is probabily not Guassian
```

1.7 Two-sample t-test

Paired t-test Tests whether the means of two paired samples are significantly different.

1.8 Assumptions

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

1.9 Interpretation

H0: the means of a samples are equal.

H1: the means of a samples are unequal.

[ ]:
```
df1.head()
```

[ ]:
```
  airline  baggage_weight  baggage_pieces
0   gamma        0.444444              2
1   alpha        0.444444              1
2   gamma        0.711111              2
3   alpha        0.333333              1
4   gamma        0.444444              1
```

[ ]:
```
df_gamma1 = df1.loc[df1['airline']== 'gamma']
```

[ ]:
```
df_gamma1
```

[ ]:
```
      airline  baggage_weight  baggage_pieces
0       gamma        0.444444              2
```

```
2       gamma        0.711111               2
4       gamma        0.444444               1
5       gamma        0.711111               1
8       gamma        0.444444               1
…       …             …              …
29986   gamma        0.444444               2
29987   gamma        0.444444               1
29988   gamma        0.444444               1
29995   gamma        0.444444               1
29997   gamma        0.444444               2

[10063 rows x 3 columns]
```

[ ]:
```
df_1 = df_gamma1.loc[df_gamma1['baggage_pieces']== 1]
```

[ ]:
```
df_1.head()
```

[ ]:
```
   airline  baggage_weight  baggage_pieces
4    gamma        0.444444               1
5    gamma        0.711111               1
8    gamma        0.444444               1
13   gamma        0.711111               1
19   gamma        0.444444               1
```

[ ]:
```
df_2= df_gamma1.loc[df_gamma1['baggage_pieces']== 2]
```

[ ]:
```
df_2.head()
```

[ ]:
```
   airline  baggage_weight  baggage_pieces
0    gamma        0.444444               2
2    gamma        0.711111               2
11   gamma        0.444444               2
12   gamma        0.444444               2
22   gamma        0.444444               2
```

[ ]:
```
df_1.shape
```

[ ]:
```
(5225, 3)
```

[ ]:
```
df_2.shape
```

[ ]:
```
(4838, 3)
```

[ ]:
```
df_1st = df_1.sample(n=4000)
df_2nd = df_2.sample(n=4000)

print("The number of instances in 1st class are = " , df_1st.shape)
print("The number of instances in 2nd class are = " , df_2nd.shape)
```

```
The number of instances in 1st class are =  (4000, 3)
The number of instances in 2nd class are =  (4000, 3)
```

[ ]:
```python
from scipy.stats import ttest_rel
# Apply test with baggage 1 and 2
stats , p_value = ttest_rel(df_1st['baggage_weight'],df_2nd['baggage_weight'])
print('stat = %.3f, p_value =%.3f'%(stats,p_value))

if p_value > 0.5:
    print("Data is probabily normal or Guassian")
else:
    print("Data is probabily not Guassian")
```

```
stat = 6.765, p_value =0.000
Data is probabily not Guassian
```

[ ]:
```python

```