IAG_test_Oanh_Dang

May 24, 2021

IAG Analyst Screening Exercises

Name: Oanh Dang

Date: 24 May 2021

0.1 Summary

1. The estimate of average effect of the treatment: **-3.97**
2. The 90% confidence interval of effect: **(-5.26, -2.69)**

0.2 Detailed analysis

0.2.1 1. Data processing

Import and exclude data records on weekends for both groups.

[1]:
```
import scipy.stats as st
import numpy as np
import pandas as pd
from math import sqrt
```

[2]:
```
# read data
df_raw = pd.read_csv('Screening_Question.csv')
df_raw.head()
```

[2]:

	A	B	DoW
0	67.829181	64.892863	0
1	80.259588	73.508008	1
2	54.853968	79.869075	1
3	73.039450	75.761761	4
4	74.816165	67.971879	0

[3]:
```
# exclude data record on weekend, i.e. DoW = 0 or DoW = 6
df = df_raw[~((df_raw['DoW'] == 0) | (df_raw['DoW'] == 6))]
df.head()
```

[3]:

	A	B	DoW
1	80.259588	73.508008	1
2	54.853968	79.869075	1
3	73.039450	75.761761	4
5	66.077492	73.170671	2
6	65.367159	82.259513	2

[4]:
```
df.describe()
```

[4]:

	A	B	DoW
count	351.000000	351.000000	351.000000
mean	65.274902	69.246875	2.980057
std	9.825170	10.847636	1.452938
min	39.281765	32.908683	1.000000
25%	58.128391	62.259359	2.000000
50%	65.017882	69.439125	3.000000
75%	71.286383	77.174867	4.000000
max	94.201646	96.307552	5.000000

0.2.2 2. Problem formulation

To calculate the effect of the treatment on the indicator variable, we perform hypothesis testing for two independent samples A and B. The null hypothesis is there are no statistically significant difference between two group means, and the alternative hypothesis is there is statistically significant difference between two group means. ### 2.1. Hypothesis test selection Since both groups are normally distributed and have the same mean and stardard deviation, two-independent-sample t-test is appropriate in this case.

0.2.3 2.2. Performing t-test and calculating 90% CI

0.2.4 T-test

[5]:
```
# Independent T-test
t_test, p_val = st.ttest_ind(df['A'], df['B'])
```

[6]:
```
# p value
p_val
```

[6]: 4.7377430242115836e-07

0.2.5 90% confidence interval

To calculate 90% confidence interval, alpha level is chosen to be 0.1.

The degree of freedom is calculated by `2 * sample size - 2`

[8]:
```
# number of data point
rows = df.shape[0]
# degree of freedom
ddof = rows* 2 -2
```

[9]:
```
# average standard deviations between group A and group B
std_A = df['A'].std()
std_B = df['B'].std()
std_AB = sqrt(((rows - 1)*(std_A)**2 + (rows - 1)*(std_B)**2) / ddof)
```

[10]:
```
# the mean differences between group A and group B
diff_mean = df['A'].mean() - df['B'].mean()
```

[11]:
```
# 90% confidence interval of the effect
interval = st.t.interval(alpha=0.1, df=ddof, loc=diff_mean, scale=std_AB )
interval
```

[11]:
```
(-5.272919358364458, -2.6710275612936574)
```

0.3 3. Interpretation

Since the p-value of **4.74e-07** is less than **0.1** (alpha level), we reject the null hypothesis, i.e. the differences between 2 group means is statistically significant. In other words, we conclude that the effect of the treatment is statistically significant.

The 90% confidence interval is **(-5.26, -2.69)** means that if the experiment is repeated many times, 90% of the times, the mean difference between 2 groups is expected to fall between -5.26 to -2.69. The average effect of the treatment is indicated by the centre of the confidence interval, which is **-3.97**.