Assignment Week 08

May 24, 2021

[12]:
```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import zscore
from scipy.stats import t
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from scipy.stats import f
import scipy.stats as stats
from statsmodels.stats.anova import anova_lm
import statsmodels.api as sm
import statsmodels.stats.multicomp
```

1 Question 1

[ ]:
```
#https://www.statisticssolutions.com/how-to-conduct-the-wilcox-sign-test/
```

[79]:
```
df = pd.DataFrame({"Test":[1,2,3,4,5,6,7,8], "Fan A":[55,52,51,59,60,56,54,54],␣
↪"Fan B":[46,55,59,50,47,62,53,55]})
df
```

[79]:

	Test	Fan A	Fan B
0	1	55	46
1	2	52	55
2	3	51	59
3	4	59	50
4	5	60	47
5	6	56	62
6	7	54	53
7	8	54	55

[80]:
```
df['Difference'] = df['Fan A'] - df['Fan B']
df
```

[80]:

	Test	Fan A	Fan B	Difference
0	1	55	46	9
1	2	52	55	-3
2	3	51	59	-8
3	4	59	50	9
4	5	60	47	13
5	6	56	62	-6
6	7	54	53	1
7	8	54	55	-1

[81]:
```
df['Absolute Difference'] = np.abs(df['Difference'])
df
```

[81]:

	Test	Fan A	Fan B	Difference	Absolute Difference
0	1	55	46	9	9
1	2	52	55	-3	3
2	3	51	59	-8	8
3	4	59	50	9	9
4	5	60	47	13	13
5	6	56	62	-6	6
6	7	54	53	1	1
7	8	54	55	-1	1

[82]:
```
lis1 = list(np.sort(df['Absolute Difference']))
lis1
```

[82]: [1, 1, 3, 6, 8, 9, 9, 13]

[83]:
```
lis2 = list(1+np.arange(9))
lis2
```

[83]: [1, 2, 3, 4, 5, 6, 7, 8, 9]

[84]:
```
df['Rank'] = df['Absolute Difference'].rank()
df
```

[84]:

	Test	Fan A	Fan B	Difference	Absolute Difference	Rank
0	1	55	46	9	9	6.5
1	2	52	55	-3	3	3.0
2	3	51	59	-8	8	5.0
3	4	59	50	9	9	6.5
4	5	60	47	13	13	8.0
5	6	56	62	-6	6	4.0
6	7	54	53	1	1	1.5
7	8	54	55	-1	1	1.5

[89]:
```
df['Signed Rank'] = df.apply(lambda x: (-1*x['Rank'] if x['Difference'] <0 else
↪x['Rank']), axis =1)
df
```

[89]:

	Test	Fan A	Fan B	Difference	Absolute Difference	Rank	Signed Rank
0	1	55	46	9	9	6.5	6.5
1	2	52	55	-3	3	3.0	-3.0
2	3	51	59	-8	8	5.0	-5.0
3	4	59	50	9	9	6.5	6.5
4	5	60	47	13	13	8.0	8.0
5	6	56	62	-6	6	4.0	-4.0
6	7	54	53	1	1	1.5	1.5
7	8	54	55	-1	1	1.5	-1.5

[102]:
```
Wplus = df[df['Signed Rank'] <0]['Rank'].sum()
Wplus
```

[102]: 13.5

[103]:
```
Wminus = df[df['Signed Rank'] >0]['Rank'].sum()
Wminus
```

[103]: 22.5

[101]:
```
#http://users.stat.ufl.edu/~winner/tables/wilcox_signrank.pdf
```

[105]:
```
Wstat = min(Wplus, Wminus)
Wstat
```

[105]: 13.5

[106]:
```
Wcritical = 5 # at n =8 and alpha = 0.05
```

The null hypothesis is that there is no difference in the operating hours of both Fan companies

The alternative hypothesis is that Fan A has lower operating hour potential than Fan B

1.0.1 Since W stat is greater than W critical, we cannot say with 95 percent confidence that Fan B is better than Fan A or that Fan A and B are different in performace because there is not much evidence to suggest that.

2 Question 2

[108]:
```
df =pd.DataFrame({"7 am":[50, 80, 62], "Noon":[45,52,48], "6 pm":[57,74,68]})
df.index= ["Location A", "Location B", "Location C"]
df
```

[108]:

	7 am	Noon	6 pm
Location A	50	45	57
Location B	80	52	74
Location C	62	48	68

3 Part a)

Conduct an ANOVA test to determine whether the mean concentrations of SO2 differ during the three collection periods at = 0.05

[109]:
```
F, p = stats.f_oneway(df["7 am"],df["Noon"],df["6 pm"])
# Seeing if the overall model is significant
print('F-Statistic=%.3f, p=%.3f' % (F, p))
```

```
F-Statistic=2.740, p=0.143
```

[124]:
```
# Second Method
```

[112]:
```
dfb = df.shape[0]-1
dfw = df.shape[0]*df.shape[1] - df.shape[0]
```

[113]:
```
from scipy.stats import f
F_critical = f.ppf(0.95, dfb, dfw)
F_critical
```

[113]:
```
5.143252849784718
```

[116]:
```
df_mean = df.mean()
df_std =df.std()
```

[119]:
```
grand_mean = df.stack().mean()
n_i = df.count()
SSTr = (n_i*(df_mean-grand_mean)**2).sum()
SSTr
```

[119]:
```
574.8888888888886
```

[120]:
```
SSE = ((n_i-1)*df_std**2).sum()
SSE
```

[120]:
```
629.3333333333334
```

[121]:
```
SST = SSTr + SSE
SST
```

[121]:
```
1204.222222222222
```

[122]:
```
MSTr = SSTr/dfb #Mean Square between
MSE = SSE/dfw #Mean Square within
F = MSTr/MSE
F
```

[122]:
```
2.7404661016949134
```

[123]:
```
F > F_critical
```

[123]:
```
False
```

3.1 Since F is less than F critical, the null hypothesis, that there is no difference in the mean pollution levels at each collection period remains standing

[ ]:
```

```

4 Part b)

Use Tukey's multiple comparison procedure to determine which collection periods differ from one another.

[125]:
```
# Since the results are not significant, so it is obvious that there are no␣
↪time periods that are dissimilar in the pollution level
```

[139]:
```
oo = df.stack().to_frame()
oo.reset_index(inplace =True)
```

[140]:
```
oo.columns=['Location', 'Time Period', 'Pollution Level']
oo
```

[140]:

	Location	Time Period	Pollution Level
0	Location A	7 am	50
1	Location A	Noon	45
2	Location A	6 pm	57
3	Location B	7 am	80
4	Location B	Noon	52

```
5  Location B        6 pm              74
6  Location C        7 am              62
7  Location C        Noon              48
8  Location C        6 pm              68
```

[141]:
```
import scipy.stats as stats
import statsmodels.stats.multicomp as mc
turkey = mc.MultiComparison(oo['Pollution Level'],oo['Time Period'])
mc_results = turkey.tukeyhsd(alpha =0.05) #alpha can be changed to 0.1 or other
 ↪values
print(mc_results)
```

```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower    upper  reject
----------------------------------------------------
  6 pm   7 am  -2.3333    0.9 -27.9608 23.2942  False
  6 pm   Noon    -18.0 0.1586 -43.6275  7.6275  False
  7 am   Noon -15.6667 0.2258 -41.2942  9.9608  False
----------------------------------------------------
```

[142]:
```
# Our conclusion that none of the groups are dissimilar are verified
```

[ ]:
```

```