WEEK 13

Recap for final exam

STATISTICAL INFERENCE

- One population sample
  - One sample T-Test
- Two population samples
  - Independence sample T-Test (Two samples T-Test)
  - Paired T-Test
- More than two population samples
  - Analysis of Variance (ANOVA)
- Two categorical data
  - Chi-Square Test

PRACTICE #1

Anorexia Case: The doctor would like to analyze weight changes of anorexia girls who are undergoing a cognitive behavioral therapy. Use anorexia.dat to solve each problem:

a) Compute for the first therapy (cb) the mean and standard deviation of changes (differences between before and after)

b) Refer to last question, compute 95% confidence interval of mean difference

c) Compute 95% confidence interval between the difference of the first therapy and the control group in the experimental study

SOLUTION #1A

```
import numpy as np
import pandas as pd


patient = pd.read_csv(r"C:\Users\iamte\Downloads\Anorexia.dat", sep = '\s+')
#print(patient.head(3))
diff =  patient['after'] - patient['before']
patient['diff'] = diff
print(patient.loc[patient['therapy'] == 'cb']['diff'].describe())
```

SOLUTION #1B

```
diffCB = patient.loc[patient['therapy'] == 'cb']['diff']

import statsmodels.stats.api as sms

# conduct 95% confidence mean change
print(sms.DescrStatsW(diffCB).tconfint_mean())

#conduct 99% confidence mean change
print(sms.DescrStatsW(diffCB).tconfint_mean(alpha = 0.01))
```

PRACTICE #2

Income case: The recruiter company would like to know the difference between income with race and education. Then they decide to randomly collect the data from private employees. Use income.dat to perform the following problem:

a) Generate the ANOVA table and the Tukey comparisons of the difference for three type of therapy

b) Generate the corresponding ANOVA table

SOLUTION #2A

```
import pandas as pd
import statsmodels.formula.api as smf
import statsmodels.api as sm

income = pd.read_csv(r"C:\Users\iamte\Downloads\income.dat", sep = '\s+')

print(income.head(1))
fit = smf.ols(formula = 'income ~ C(race)', data = income).fit()
print(fit.summary())
sm.stats.anova_lm(fit)
```

SOLUTION #2A

```
# Perform Tukey
import statsmodels.stats.multicomp as mc
comp = mc.MultiComparison(income['income'], income['race'])
post_hoc_res = comp.tukeyhsd()
print(post_hoc_res.summary())

post_hoc_res.plot_simultaneous(ylabel = 'race', xlabel = 'mean income difference')

fit2 = smf.ols(formula = 'income ~ C(race) + education', data = income).fit()
```

SOLUTION #2B

```
fit2 = smf.ols(formula = 'income ~ C(race) + education', data =
income).fit()

print(fit2.summary())


sm.stats.anova_lm(fit2, typ=2)
```

PRACTICE #3

Salary case: The recruiter company would like to know the association between the range of salary and education level. Then they decide to randomly collect the data from private employees. Use salary.csv to perform the following problem:

a) Create the two-way contingency table

b) Perform the Chi-Square test with 5% significance level

SOLUTION #3A

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
salary = pd.read_csv(r'C:\Users\iamte\OneDrive\Desktop\ABAC\Stat 1\Salary.csv')


rowlabel = ['Less than 25000 Baht', '25001 - 50000 Baht', '50001 - 75000 Baht', '75001 - 100000 Baht', 'More than 100000 Baht']
collabel = ['Bachelor', 'Master', 'Doctoral']
table = pd.crosstab(salary['Salary'], salary['Education'], margins = False)
table.index = rowlabel
table.columns = collabel
table
```

SOLUTION #3A + 3B

```
## Perform contingency table
import statsmodels.api as sm
table = sm.stats.Table(table)
print(table.fittedvalues)

## Perform Chi-Square Test
Chi_Square = table.test_nominal_association()
print(Chi_Square)
```