L4 Data Analysis Statistics

1. Case 1

A physician is evaluating a new diet for her patients with a family history of heart disease. To test the effectiveness of this diet, 16 patients are placed on the diet for 6 months. Their weights and triglyceride levels are measured before and after the study, and the physician wants to know if either set of measurements has changed. (Data set: dietstudy.csv)

what type of hypothesis test to use?

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import scipy.stats as stats

%matplotlib inline
```

Read the dataset

Mount the google drive

```
from google.colab import drive
drive.mount('/content/drive')
data_dir_path='/content/drive/My Drive/Data/DS4/'
```

```
diet = pd.read_csv(data_dir_path+'dietstudy.csv')
diet.head()
```

```
#Add code
```

```
#Add code
```

```
from google.colab import drive
drive.mount('/content/drive')
data_dir_path='/content/drive/My Drive/Data/DS4/'

    Mounted at /content/drive
```

```
diet = pd.read_csv(data_dir_path+'dietstudy.csv')
```

```
diet.head()
```

	patid	age	gender	tg0	tg1	tg2	tg3	tg4	wgt0	wgt1	wgt2	wgt3	wgt4
0	1	45	Male	180	148	106	113	100	198	196	193	188	192
1	2	56	Male	139	94	119	75	92	237	233	232	228	225
2	3	50	Male	152	185	86	149	118	233	231	229	228	226
3	4	46	Female	112	145	136	149	82	179	181	177	174	172
4	5	64	Male	156	104	157	79	97	219	217	215	213	214

Check dataset size

```
diet.shape
```

```
#Add code
```

```
diet.shape
```

```
    (16, 13)
```

Two Sample T-Test (Paired)

Find the mean for the triglyceride and Weights before and after the diet.

▼ Click here for answer

```
print("The triglyceride levels of patients were {}".format(diet.tg0.mean()))
print("The final triglyceride levels of patients are {}".format(diet.tg4.mean()))
print("The weights of pateints were {}".format(diet.wgt0.mean()))
print("The final weights of pateints are {}".format(diet.wgt4.mean()))
```

▼ Click here for answer

```
#Add code
```

```
print("The triglyceride levels of patients were {}".format(diet.tg0.mean()))
print("The final triglyceride levels of patients are {}".format(diet.tg4.mean()))
print("The weights of pateints were {}".format(diet.wgt0.mean()))
print("The final weights of pateints are {}".format(diet.wgt4.mean()))
```

```
    The triglyceride levels of patients were 138.4375
    The final triglyceride levels of patients are 124.375
    The weights of pateints were 198.375
    The final weights of pateints are 190.3125
```

Triglycerides

H0: levels of Triglycerides of individual before diet == levels of Triglycerides of individual after diet

H1: levels of Triglycerides of individual before diet != levels of Triglycerides of individual after diet

Use:
import scipy.stats as stats
stats.ttest_rel()
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.html

▼ Click here for answer

```
stage1_trig = stats.ttest_rel(a=diet.tg0,b=diet.tg4)

# a =triglyceride levels before diet
# b = triglyceride levels after diet)
```

Get the pvalue

```
stage1_trig.pvalue
```

Check pvalue with confidence level 0.025

```
stage1_trig.pvalue>0.025
```

```
#Add code
#T test
```

```
#Add code
#get p-value
```

```
#Add code
#check significant -check p-value with alpha
```

```
stage1_trig = stats.ttest_rel(a=diet.tg0,b=diet.tg4)

#a =triglyceride levels before diet
#b = triglyceride levels after diet)
```

```
stage1_trig.pvalue
```

```
    0.24874946576903698
```

```
stage1_trig.pvalue>0.025
```

```
    True
```

Weights

H0: weights of individual before diet == weights of individual after diet

H1: weights of individual before diet != weights of individual after diet

▶ Click here for answer

```
#Add code
#T test
```

```
#Add code
#get p-value
```

```
#Add code
#check significant -check p-value with alpha
```

```
stage2_weight = stats.ttest_rel(a=diet.wgt0,b=diet.wgt4)

#a = weights before diet
#b = weights after diet
```

```
stage2_weight.pvalue
```

```
    1.137689414996614e-08
```

```
stage2_weight.pvalue>0.025
```

```
    False
```

Conclusions

▼ Click here for answer

1. However, the p-value greater than 0.025 for change in triglyceride level shows the diet did not significantly reduce their triglyceride levels. Accept Null Hypothesis
2. Since the p-value for change in weight is less than 0.025,we can conclude that the average loss of 8.06 pounds per patient is not due to chance variation, and can be attributed to the diet. Reject Null Hypothesis

2. Case 2

An analyst at a department store wants to evaluate a recent credit card promotion. To this end, 500 cardholders were randomly selected. Half received an ad promoting a reduced interest rate on purchases made over the next three months, and half received a standard seasonal ad. Is the promotion effective to increase sales?

what type of hypothesis test to use?

Read dataset

▼ Click here for answer

```
analyst_cre = pd.read_csv(data_dir_path+'creditpromo.csv')

analyst_cre.head()
```

```
#Add code

analyst_cre = pd.read_csv(data_dir_path+'creditpromo.csv')

analyst_cre.head()
```

	id	insert	dollars
0	148	Standard	2232.771979
1	572	New Promotion	1403.807542

Get mean for Standard and New Promotion

▼ Click here for answer

```
standard = analyst_cre['dollars'].loc[analyst_cre['insert']=="Standard"]

promotion = analyst_cre['dollars'].loc[analyst_cre['insert']=="New Promotion"]

print("The average spent by the normal ad is ${}".format(standard.mean()))

print("The average spent by the promotion ad is ${}".format(promotion.mean()))
```

```
#Add code

standard = analyst_cre['dollars'].loc[analyst_cre['insert']=="Standard"]
promotion = analyst_cre['dollars'].loc[analyst_cre['insert']=="New Promotion"]


print("The average spent by the normal ad is ${}".format(standard.mean()))
print("The average spent by the promotion ad is ${}".format(promotion.mean()))
```

```
    The average spent by the normal ad is $1566.3890309659348
    The average spent by the promotion ad is $1637.4999830647992
```

Observation

▼ Click here for answer

On average, customers who received the interest-rate promotion charged about $70 more than the normal standard season ad, and they vary a little more around their average

T test

H0: Promotion is not effective to increase sales

H1: Promotion is effective to increase sales

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

▼ Click here for answer

```
eq_var = stats.ttest_ind(a= standard,b= promotion)


eq_var.pvalue


p = eq_var.pvalue


print(" For the above test, the p-value is {}".format(p))


if(p<0.025):
    print('We reject null hypothesis')
else:
    print('We accept null hypothesis')
```

```
#Add code


eq_var = stats.ttest_ind(a= standard,
                b= promotion)    # equal variance
eq_var.statistic
eq_var.pvalue
```

```
    0.024225996894147395
```

```
p = eq_var.pvalue

print(" For the above test, the p-value is {}".format(p))

if(p<0.025):
    print('We reject null hypothesis')
else:
    print('We accept null hypothesis')
```

```
     For the above test, the p-value is 0.024225996894147395
    We reject null hypothesis
```

Conclusion

▼ Click here for answer

* Since the significance value of the test is less than 0.025, we can safely conclude that the average of 71.11 dollars more spent by cardholders receiving the reduced interest rate is not due to chance alone. The store will now consider extending the offer to all credit customers.

3. Case 3

An experiment is conducted to study the hybrid seed production of bottle gourd under open field conditions. The main aim of the investigation is to compare natural pollination and hand pollination. The data are collected on 10 randomly selected plants from each of natural pollination and hand pollination. The data are collected on fruit weight (kg), seed yield/plant (g) and seedling length (cm). (Data set: pollination.csv)

a. Is the overall population of Seed yield/plant (g) equals to 200?

what type of hypothesis test to use?

▼ Click here for answer

One Sample t-Test

Load dataset

▼ Click here for answer

```
poll = pd.read_csv(data_dir_path+'pollination.csv')
poll.head()
```

```
#Add code

poll = pd.read_csv(data_dir_path+'pollination.csv')

poll.head()
```

	Group	Fruit_Wt	Seed_Yield_Plant	Seedling_length
0	Natural	1.85	147.70	16.86
1	Natural	1.86	136.86	16.77
2	Natural	1.83	149.97	16.35

Find the mean for Seed yield/plant

▼ Click here for answer

```
poll.Seed_Yield_Plant.mean()
```

```
#Add code
```

```
poll.Seed_Yield_Plant.mean()
```

```
    180.8035
```

One sample T Test

H0: Overall population mean of seed yield plant equal to 200

H1: Overall population mean of seed yield plant not equal to 200

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html

▼ Click here for answer

```
ttest_1 = stats.ttest_1samp(a=poll.Seed_Yield_Plant,popmean = 200)

p_value = ttest_1.pvalue

print("The p value is }".format(p_value))

if(p<0.025):
    print('We reject null hypothesis')
else:
    print('We accept null hypothesis')
```

```
#Add code


ttest_1 = stats.ttest_1samp(a=poll.Seed_Yield_Plant, popmean = 200)

p_value = ttest_1.pvalue
print("The p value is {} ".format(p_value))
```

```
    The p value is 0.032891040921283025
```

```
p_value>0.025
```

```
    True
```

Conclusion

▼ Click here for answer

- Since the significance value of the test is greater than 0.025, we accept the null hypothesis. Therefore overall population of seed yield/plant is equal to 200.

b. Test whether the natural pollination and hand pollination under open field conditions are equally effective or are significantly different.

F-Test/Anova

H0 : mean Fruit_Wt == mean Seed_Yield_Plant == mean Seedling_length

H1 : mean Fruit_Wt != mean Seed_Yield_Plant != mean Seedling_length

Extract data for Natural and Hand pollination

▼ Click here for answer

```
natural_fruit_wt = poll['Fruit_Wt'].loc[poll['Group']=="Natural"]
hand_fruit_wt = poll['Fruit_Wt'].loc[poll['Group']=="Hand"]

natural_seed_yield = poll['Seed_Yield_Plant'].loc[poll['Group']=="Natural"]
hand_seed_yield = poll['Seed_Yield_Plant'].loc[poll['Group']=="Hand"]

natural_seedling_length = poll['Seedling_length'].loc[poll['Group']=="Natural"]
hand_seedling_length = poll['Seedling_length'].loc[poll['Group']=="Hand"]
```

```
#Add code
```

```
#Add code
```

```
#Add code
```

Do F-Test/Anova for Seedling Length, Fruit_W, Seed_Yield_Plant

▼ Click here for answer

```
# Perfrom the Anova
anova = stats.f_oneway(natural_fruit_wt,hand_fruit_wt)
# Statistic :  F Value
f1 = anova.statistic
p1 = anova.pvalue
print("The p value is {}".format(p1))
```

```
# Perfrom the Anova
anova = stats.f_oneway(natural_seed_yield,hand_seed_yield)
# Statistic :  F Value
f2 = anova.statistic
p2 = anova.pvalue
print("The p value is {}".format(p2))
```

```
# Perfrom the Anova
anova = stats.f_oneway(natural_seedling_length,hand_seedling_length)
# Statistic :  F Value

p3 = anova.pvalue
print("The p value is {}".format(p3))
```

```
#Add code
```

```
natural_fruit_wt = poll['Fruit_Wt'].loc[poll['Group']=="Natural"]
hand_fruit_wt = poll['Fruit_Wt'].loc[poll['Group']=="Hand"]

# Perfrom the Anova
anova = stats.f_oneway(natural_fruit_wt,hand_fruit_wt)
# Statistic :  F Value
f1 = anova.statistic
p1 = anova.pvalue
print("The f-value is {} and the p value is {}".format(f1,p1))
```

```
The f-value is 312.228532974426 and the p value is 8.078362076486568e-13
```

```
natural_seed_yield = poll['Seed_Yield_Plant'].loc[poll['Group']=="Natural"]
hand_seed_yield = poll['Seed_Yield_Plant'].loc[poll['Group']=="Hand"]

# Perfrom the Anova
anova = stats.f_oneway(natural_seed_yield,hand_seed_yield)
# Statistic :  F Value
f2 = anova.statistic
p2 = anova.pvalue
print("The f-value is {} and the p value is {}".format(f2,p2))
```

```
The f-value is 194.83303662980398 and the p value is 4.271481585484407e-11
```

```
natural_seedling_length = poll['Seedling_length'].loc[poll['Group']=="Natural"]
hand_seedling_length = poll['Seedling_length'].loc[poll['Group']=="Hand"]

# Perfrom the Anova
anova = stats.f_oneway(natural_seedling_length,hand_seedling_length)
# Statistic :  F Value
f3 = anova.statistic
p3 = anova.pvalue
print("The f-value is {} and the p value is {}".format(f3,p3))
```

```
The f-value is 6.46293337115627 and the p value is 0.020428817064110556
```

Conclusion

▼ Click here for answer

- Since the all the test p-value is less than 0.025, we reject the null hypothesis. Therefore they are significant different .

4. Case 4

An electronics firm is developing a new DVD player in response to customer requests. Using a prototype, the marketing team has collected focus data for different age groups viz. Under 25;

25-34; 35-44; 45-54; 55-64; 65 and above. Do you think that consumers of various ages rated the design differently? (Data set: dvdplayer.csv)

what type of hypothesis test to use?

Load dataset

▼ Click here for answer

```
dvd = pd.read_csv(data_dir_path+'dvdplayer.csv')

dvd.head()
```

```
#Add code

dvd = pd.read_csv(data_dir_path+'dvdplayer.csv')

dvd.head()
```

	agegroup	dvdscore
0	65 and over	38.454803
1	55-64	17.669677
2	65 and over	31.704307
3	65 and over	25.924460
4	Under 25	30.450007

F-Test/Anova

H0 : mean (age_group_1(65 and over)) == mean(age_group_2(55-64)) == mean(age_group3(Under 25)) == mean(age_group4(35-54)) == mean(age_group5(45-54)) == mean(age_group_6(25-34)) (Consumers of various age groups rated the design similarly)

Ha : mean(age_group_1(65 and over)) != mean(age_group_2(55-64)) != mean(age_group3(Under 25))!= mean(age_group4(35-54)) != mean(age_group5(45-54)) != mean(age_group_6(25-34))

(Consumers of various age groups rated the design differently)

Extract data for each group

▼ Click here for answer

```
age_group_1 = dvd['dvdscore'].loc[dvd['agegroup']=="65 and over"]
age_group_2 = dvd['dvdscore'].loc[dvd['agegroup']=="55-64"]
age_group_3 = dvd['dvdscore'].loc[dvd['agegroup']=="Under 25"]
age_group_4 = dvd['dvdscore'].loc[dvd['agegroup']=="35-44"]
age_group_5 = dvd['dvdscore'].loc[dvd['agegroup']=="45-54"]
age_group_6 = dvd['dvdscore'].loc[dvd['agegroup']=="25-34"]
```

```
#Add code

age_group_1 = dvd['dvdscore'].loc[dvd['agegroup']=="65 and over"]
age_group_2 = dvd['dvdscore'].loc[dvd['agegroup']=="55-64"]
age_group_3 = dvd['dvdscore'].loc[dvd['agegroup']=="Under 25"]
age_group_4 = dvd['dvdscore'].loc[dvd['agegroup']=="35-44"]
age_group_5 = dvd['dvdscore'].loc[dvd['agegroup']=="45-54"]
age_group_6 = dvd['dvdscore'].loc[dvd['agegroup']=="25-34"]
```

Do F-Test/Anova
https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html

▼ Click here for answer

```
# Perfrom the Anova
anova = stats.f_oneway(age_group_1,age_group_2,age_group_3,age_group_4,age_group_5,age_grc
# Statistic :  F Value
p = anova.pvalue
print("The p value is {}".format(p))
if(p<0.025):
    print('We reject null hypothesis')
else:
    print('We fail to reject null hypothesis')
```

```
#Add code
```

```
# Perfrom the Anova
anova = stats.f_oneway(age_group_1,age_group_2,age_group_3,age_group_4,age_group_5,
# Statistic :  F Value

p = anova.pvalue
print("The  p value is {}".format(p))
if(p<0.025):
    print('We reject null hypothesis')
else:
    print('We fail to reject null hypothesis')
```

```
The  p value is 3.087324905679639e-05
We reject null hypothesis
```

Conclusion

▼ Click here for answer

* Since the significance value of the test is less than 0.025, we reject the null hypothesis. Therefore,consumers of various ages rated the design differently

5. Case 5

A doctor is evaluating a new medicine for her patients with a high blood pressure. To test the effectiveness of this medicine,the blood pressure before the medicine and after medicine are recorded.
what type of hypothesis test to use?

Load dataset

load the blood_pressure.csv

▼ Click here for answer

```
df = pd.read_csv(data_dir_path+'blood_pressure.csv',',')
df.head()
```

```
#Add code
```

```
df = pd.read_csv(data_dir_path+'blood_pressure.csv',',')
```

```
df.head()
```

	patient	sex	agegrp	bp_before	bp_after
0	1	Male	30-45	143	153
1	2	Male	30-45	163	170
2	3	Male	30-45	153	168
3	4	Male	30-45	153	142
4	5	Male	30-45	146	141

Check the basic statistics for the data

▼ Click here for answer

```
print(df[['bp_before','bp_after']].describe())
```

```
#Add code
```

```
print(df[['bp_before','bp_after']].describe())
```

```
        bp_before    bp_after
count  120.000000  120.000000
mean   156.450000  151.358333
std     11.389845   14.177622
min    138.000000  125.000000
25%    147.000000  140.750000
50%    154.500000  149.500000
75%    164.000000  161.000000
max    185.000000  185.000000
```

Z Test

Since sample size is large use the Z Test

H0 : bp_before == bp_after

H1 : bp_before != bp_after
https://www.statsmodels.org/stable/generated/statsmodels.stats.weightstats.ztest.html

▼ Click here for answer

```
ztest ,pval1 = stests.ztest(df['bp_before'], x2=df['bp_after'], value=0,alternative='two-s

print(float(pval1))

if(pval1<0.025):

    print('We reject null hypothesis')

else:

    print('We fail to reject null hypothesis')
```

```
import pandas as pd
from scipy import stats
from statsmodels.stats import weightstats as stests
#Add code
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: Futur
  import pandas.util.testing as tm
```

```
#Add code
```

```
import pandas as pd
from scipy import stats
from statsmodels.stats import weightstats as stests
ztest ,pval1 = stests.ztest(df['bp_before'], x2=df['bp_after'])
print(float(pval1))

if(pval1<0.025):
    print('We reject null hypothesis')
else:
    print('We fail to reject null hypothesis')
```

```
0.002162306611369422
We reject null hypothesis
```

bp_before mean is not equal bp_after mean

● ✕