Introduction to hypothesis testing

```python
import numpy       as np
import pandas      as pd
import scipy.stats as stats
```

Beware of the problem of testing too many hypotheses; the more you torture the data, the more likely they are to confess, but confessions obtained under duress may not be admissible in the court of scientific opinion - Stephen M Stigler

- Hypothesis is a claim made by a person / organization.
- The claim is usually about the population parameters such as mean or proportion and we seek evidence from a sample for the support of the claim (Example: average salary of Data Scientist with 1 year experience is Rs 5 Lakhs per annum).
- Hypothesis testing is a process used for either rejecting or retaining null hypothesis.

Examples of some claims:

- If you drink Horlicks, you can grow taller, stronger and sharper.
- Two - minute for cooking noodles. (or eating !!)
- Married people are happier than singles (Anon - 2015).
- Smokers are better sales people.

Hypothesis testing is used for checking the validity of the claim using evidence found in sample data.

Type I Error, Type II error and power of the hypothesis test

Type I error:

- It is the conditional probability of rejecting a null hypothesis when it is true, is called **Type I error or False positive.**
- α, the level of significance is the value of Type I error.
- P(Reject null hypothesis | H_0 is true) = α

Type II error:

- It is the conditional probability of retaining a null hypothesis when it is true, is called **Type II error or False Negative.**
- β, is the value of Type II error.
- P(Retain null hypothesis | H_0 is false) = β

Power of the test

- (1 - β) is known as the **power of the test**.
- It is P(Reject null hypothesis | H_0 is false) = 1- β

Steps involved in solving the hypothesis testing

1 Define null and alternative hypotheses

- ### Null hypothesis means no relationship or status quo
- ### Alternative hypothesis is what the researcher wants to prove

Example:

Write the null and alternative hypothesis from the following hypopthesis description: a. Average annual salary of Data Scientists is different for those having Ph.D in Statistics and those who do not.

- Let μ_{PhD} be the average annual salary of a Data scientist with Ph.D in Statistics.
- Let μ_{NoPhD} be the average annual salary of a Data scientist without Ph.D in Statistics.

- Null hypothesis: H_0: μ_{PhD} = μ_{NoPhD}
- Alternative hypothesis: H_A: $\mu_{PhD} \neq \mu_{NoPhD}$

Since the rejection region is on either side of the distribution, it will be a **two-tailed** test.

b. Average annual salary of Data Scientists is more for those having Ph.D in Statistics than those who do not.

- Null hypothesis: H_0: $\mu_{PhD} \leq \mu_{NoPhD}$
- Alternative hypothesis: H_A: μ_{PhD} > μ_{NoPhD}

Since the rejection region is on the right side of the distribution, it will be a one-tailed test.

2 Decide the significance level

- You control the Type I error by determining the risk level, α, the level of significance that you are willing to reject the null hypothesis when it is true. Traditionally, you select a level of 0.01, 0.05 or 0.10. The choice of selection for making Type I error depends on the cost of making a Type I error.
- One way to reduce the probability of making a Type II error is by increasing the sample size. For a given level of α, increasing the sample size decreases β resulting in increasing the power of the statistical test to detect that null hypothesis is false.

3 Identify the test statistic

- ### The test statistic will depend on the probability distribution of the sampling distribution

4 Calculate the p-value or critical values

- ### P-value is the conditional probability of observing the test statistic value or extreme than the sample result when the null hypothesis is true.
- ### Critical value approach
- Critical values for the appropriate test statistic are selected so that the rejection region contains a total area of α when H_0 is true and the non-rejection region contains a total area of 1 - α when H_0 is true.

5 Decide to reject or accept null hypothesis

- ### Reject null hypothesis when test statisic lies in the rejection region; retain null hypothesis otherwise.
- ### OR
- ### Reject null hypothesis when p-value < α; retain null hypothesis otherwise.

Hypothesis testing using the critical value approach

Step 1: Define null and alternative hypotheses

In testing whether the mean volume is 2 litres, the null hypothesis states that mean volume, μ equals 2 litres. The alternative hypthesis states that the mean olume, μ is not equal to 2 litres.

- H_0: μ = 2
- H_A: $\mu \neq 2$

Step 2: Decide the significance level

Choose the α, the level of significance according to the relative importance of the risks of committing Type I and Type II errors in the problem.

In this example, making a Type I error means that you conclude that the population mean is not 2 litres when it is 2 litres. This implies that you will take corrective action on the filling process even though the process is working well (*false alarm*).

On the other hand, when the population mean is 1.98 litres and you conclude that the population mean is 2 litres, you commit a Type II error. Here, you allow the process to continue without adjustment, even though an adjustment is needed (*missed opportunity*).

Here, we select α = 0.05 and n, sample size = 50.

Step 3: Identify the test statistic

We know the population standard deviation and the sample is a large sample, n>30. So you use the normal distribution and the Z_STAT test statistic.

Step 4: Calculate the critical value

We know the α is 0.05. So, the critical values of the Z_STAT test statistic are -1.96 and 1.96.

```python
print(np.abs(round(stats.norm.isf(q = 0.025),2))) # Here we use alpha by 2  for two-tailed test
```

```
1.96
```

- ### Rejection region is Z_{STAT} < -1.96 or Z_{STAT} > 1.96
- ### Acceptance or non-rejection regions is -1.96 $\leq Z_{STAT} \leq$ 1.96

We collect the sample data, calculate the test statistic. In our example,

- $\overline{X}$ = 2.001
- μ = 2
- σ = 15
- n = 50
- $Z_{STAT} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

```python
XAvg  = 2.001
mu    = 2
sigma = 15
n     = 50
Z = (XAvg - mu)/(sigma/np.sqrt(n))
print('Value of Z observed is %2.5f' %Z)
```

```
Value of Z observed is 0.00047
```

5 Decide to reject or accept null hypothesis

In this example, Z = 0.00047 (z observed) lies in the acceptance region because, -1.96 < Z = 0.00047 < 1.96.

Z observed is less than Z critical

So the statistical decision is not to reject the null hypothesis.

So there is no sufficient evidence to prove that the mean fill is different from 2 litres.

One sample test

In one sample test, we compare the population parameter such as mean of a single sample of data collected from a single population.

1) Z test

A one sample Z test is one of the most basic types of hypothesis test.

Example 1: A principal of a prestigious city college claims that the average intelligence of the students of the college is above average.

A random sample of 100 students IQ scores have a mean score of 115. The population mean IQ is 100 with a standard deviation of 15.

Is there sufficient evidence to support the principal's claim?

Solution: Let us work through the several required steps

Step 1: Define null and alternative hypotheses

In testing whether the mean IQ of the students is more than 100, the null hypothesis states that mean IQ, μ equals 100. The alternative hypthesis states that the mean IQ, μ is greater than 100.

- H_0: μ = 100
- H_A: μ > 100

Step 2: Decide the significance level

Here we select α = 0.05 and it is given that n, sample size = 100.

Step 3: Identify the test statistic

We know the population standard deviation and the sample is a large sample, n>30. So you use the normal distribution and the Z_STAT test statistic.

Step 4: Calculate the critical value and test statistic

```python
Zcrit = round(stats.norm.isf(q = 0.025),2)
print('Value of Z critical is %3.6f' %Zcrit)
```

```
Value of Z critical is 1.960000
```

We know the α is 0.05. So, the critical values of the Z_STAT test statistic is 1.96

We collect the sample data, calculate the test statistic. In our example,

- $\overline{X}$ = 115
- μ = 100
- σ = 15
- n = 100
- $Z_{STAT} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

```python
XAvg  = 115
mu    = 100
sigma = 15
n     = 100
Z = (XAvg - mu)/(sigma/np.sqrt(n))
print('Value of Z observed is %2.5f' %Z)
```

```
Value of Z observed is 10.00000
```

Step 5: Decide to reject or accept null hypothesis

In this example, Z = 10 lies in the rejection region because, Z = 10 > 1.96

Z observed is greater than z critical, so we reject Null hypothesis

So there is sufficient evidence to prove that the mean average intelligence of the students of the college is above average.

2) t test

Very rarely we know the variance of the population.

A common strategy to assess hypothesis is to conduct a t test. A t test can tell whether two groups have the same mean. A t test can be estimated for:

- 1) One sample t test
- 2) Two sample t test (including paired t test)

We assume that the samples are randomly selected, independent and come from a normally distributed population with unknown but equal variances.

One sample t test

```python
from scipy.stats              import ttest_1samp,ttest_ind, wilcoxon
from statsmodels.stats.power import ttest_power
import matplotlib.pyplot     as     plt
```

Example 2

Suppose that a doctor claims that 17 year olds have an average body temperature that is higher than the commonly accepted average human temperature of 98.6 degree F. A simple random statistical sample of 25 people, each of age 17 is selected.

ID	Temperature
1	98.56
2	98.66
3	97.54
4	98.71
5	99.22
6	99.49
7	98.14
8	98.84
9	99.28
10	98.48
11	98.88
12	97.29
13	98.88
14	99.07
15	98.81
16	99.49
17	98.57
18	97.98
19	97.75
20	97.69
21	99.28
22	98.52
23	98.82
24	98.81
25	98.22

```python
temperature = np.array([98.56, 98.66, 97.54, 98.71, 99.22, 99.49, 98.14, 98.84,\
                        99.28, 98.48, 98.88, 97.29, 98.88, 99.07, 98.81, 99.49,\
                        98.57, 97.98, 97.75, 97.69, 99.28, 98.52, 98.82, 98.81, 98.22])
```

```python
print('Mean is %2.1f Sd is %2.1f' % (temperature.mean(),np.std(temperature,ddof = 1)))
```

```
Mean is 98.6 Sd is 0.6
```

Step 1: Define null and alternative hypotheses

In testing whether 17 year olds have an average body temperature that is higher than 98.6 deg F,the null hypothesis states that mean body temperature, μ equals 98.6. The alternative hypthesis states that the mean body temprature, μ is greater than 98.6.

- H_0: μ <= 98.6
- H_A: μ > 98.6

Step 2: Decide the significance level

Here we select α = 0.05 and it is given that n, sample size = 25.

Step 3: Identify the test statistic

We do not know the population standard deviation and the sample is not a large sample, n < 30. So you use the t distribution and the t_STAT test statistic.

Step 4: Calculate the p - value and test statistic

scipy.stats.ttest_1samp calculates the t test for the mean of one sample given the sample observations and the expected value in the null hypothesis. This function returns t statistic and two-tailed p value.

```python
t_statistic, p_value = ttest_1samp(temperature, 98.6)
```

```python
print(t_statistic, p_value)
```

```
-0.006668602694974534 0.9947343867528586
```

Step 5 Decide to reject or accept null hypothesis

In this example, p value is 0.9947 and it is greater than 5% level of significance

So the statistical decision is to fail to reject the null hypothesis at 5% level of significance.

So there is no sufficient evidence to prove that 17 year olds have an average body temperature higher than the commonly accepted average human temperature of 98.6 degree F.

Two sample test

Two sample t test (Snedecor and Cochran 1989) is used to determine if two population means are equal. A common application is to test if a new treatment or approach or process is yielding better results than the current treatment or approach or process.

- 1) Data is *paired* - For example, a group of students are given coaching classes and effect of coaching on the marks scored is determined.
- 2) Data is *not paired* - For example, find out whether the miles per gallon of cars of Japanese make is superior to cars of Indian make.

Two sample t test for unpaired data is defined as

- H_0: $\mu1$ = $\mu2$
- H_a: $\mu1 \neq$ = $\mu2$

Test statistic T = $\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{s_1^2}{n1}+\frac{s_2^2}{n2}}}$

- where n1 and n2 are the sample sizes and X1 and X2 are the sample means
- $S_1{}^2$ and $S_2{}^2$ are sample variances

Example 3

Compare two unrelated samples. Data was collected on the weight loss of 16 women and 20 men enrolled in a weight reduction program. At α = 0.05, test whether the weight loss of these two samples is different.

```python
Weight_loss_Male   = [ 3.69, 4.12, 4.65, 3.19,  4.34, 3.68, 4.12, 4.50, 3.70, 3.09,3.65, 4.73, 3.93, 3.46, 3.28, 4.43, 4.13, 3.62, 3.71, 2.92]
Weight_loss_Female = [2.99, 1.80, 3.79, 4.12, 1.76, 3.50, 3.61, 2.32, 3.67, 4.26, 4.57, 3.01, 3.82, 4.33, 3.40, 3.86]
```

```python
from    scipy.stats             import  ttest_1samp,ttest_ind, wilcoxon, ttest_ind_from_stats
import  scipy.stats             as      stats
from    statsmodels.stats.power import  ttest_power
import  matplotlib.pyplot       as      plt
```

Step 1: Define null and alternative hypotheses

In testing whether weight reduction of female and male are same,the null hypothesis states that mean weight reduction, μM equals μF. The alternative hypthesis states that the weight reduction is different for Male and Female, $\mu M \neq \mu F$

- H_0: μM - μF = 0
- H_A: μM - $\mu F \neq 0$

Step 2: Decide the significance level

Here we select α = 0.05 and sample size < 30 and population standard deviation is not known.

Step 3: Identify the test statistic

- We have two samples and we do not know the population standard deviation.
- Sample sizes for both samples are not same.
- The sample is not a large sample, n < 30. So you use the t distribution and the t_STAT test statistic for two sample unpaired test.

Step 4: Calculate the p - value and test statistic

We use the scipy.stats.ttest_ind to calculate the t-test for the means of TWO INDEPENDENT samples of scores given the two sample observations. This function returns t statistic and two-tailed p value.

This is a two-sided test for the null hypothesis that 2 independent samples have identical average (expected) values. This test assumes that the populations have identical variances.

```python
t_statistic, p_value  =  stats.ttest_ind(Weight_loss_Male,Weight_loss_Female)
print('P Value %1.3f' % p_value)
```

```
P Value 0.076
```

Step 5: Decide to reject or accept null hypothesis

In this example, p value is 0.076 and it is more than 5% level of significance

So the statistical decision is to fail to reject the null hypothesis at 5% level of significance.

So there is no sufficient evidence to reject the null hypothesis that the weight loss of these men and women is same.

Two sample t test for paired data

Example 4

Compare two related samples. Data was collected on the marks scored by 25 students in their final practice exam and the marks scored by the students after attending special coaching classes conducted by their college. At 5% level of significance, is there any evidence that the coaching classes has any effect on the marks scored.

```python
Marks_before = [ 52, 56, 61, 47, 58, 52, 56, 60, 52, 46, 51, 62, 54, 50, 48, 59, 56, 51, 52, 44, 52, 45, 57, 60, 45]

Marks_after  = [ 62, 64, 40, 65, 76, 82, 53, 68, 77, 60, 69, 34, 69, 73, 67, 82, 62, 49, 44, 43, 77, 61, 67, 67, 54]
```

Step 1: Define null and alternative hypotheses

In testing whether coaching has any effect on marks scored, the null hypothesis states that difference in marks, $\mu After$ equals $\mu Before$. The alternative hypthesis states that difference in marks is more than 0, $\mu After \neq \mu Before$

- H_0: $\mu After$ - $\mu Before$ = 0
- H_A: $\mu After$ - $\mu Before \neq 0$

Step 2: Decide the significance level

Here we select α = 0.05 and sample size < 30 and population standard deviation is not known.

Step 3: Identify the test statistic

- Sample sizes for both samples are same.
- We have two paired samples and we do not know the population standard deviation.
- The sample is not a large sample, n < 30. So you use the t distribution and the t_STAT test statistic for two sample paired test.

Step 4: Calculate the p - value and test statistic

We use the scipy.stats.ttest_rel to calculate the T-test on TWO RELATED samples of scores. This is a two-sided test for the null hypothesis that 2 related or repeated samples have identical average (expected) values. Here we give the two sample observations as input. This function returns t statistic and two-tailed p value.

```python
import scipy.stats as stats
t_statistic, p_value  =  stats.ttest_rel(Marks_after, Marks_before )
print('P Value %1.3f' % p_value)
```

```
P Value 0.002
```

Step 5: Decide to reject or accept null hypothesis

In this example, p value is 0.002 and it is less than 5% level of significance

So the statistical decision is to reject the null hypothesis at 5% level of significance.

So there is sufficient evidence to reject the null hypothesis that there is an effect of coaching classes on marks scored by students.

Example 5

Alchohol consumption before and after love failure is given in the following table. Conduct a paired t test to check whether the alcholhol consumption is more after the love failure at 5% level of significance.

Step 1: Define null and alternative hypotheses

In testing whether breakup has any effect on alcohol consumption, the null hypothesis states that difference in alcohol consumption, $\mu After$ - $\mu Before$ is zero. The alternative hypthesis states that difference in alcohol consumption is more than 0, $\mu After$ - $\mu Before \neq$ zero.

- H_0: $\mu After$ - $\mu Before$ = 0
- H_A: $\mu After$ - $\mu Before \neq 0$

Step 2: Decide the significance level

Here we select α = 0.05 and sample size < 30 and population standard deviation is not known.

Step 3: Identify the test statistic

- Sample sizes for both samples are same.
- We have two paired samples and we do not know the population standard deviation.
- The sample is not a large sample, n < 30. So you use the t distribution and the t_STAT test statistic for two sample paired test.

Step 4: Calculate the p - value and test statistic

We use the scipy.ttest_1samp to calculate the T-test on the difference between sample scores.

```python
import numpy as np

Alchohol_Consumption_before = np.array([470, 354, 496, 351, 349, 449, 378, 359, 469, 329, 389, 497, 493, 268, 445, 287, 338, 271, 412, 335])
Alchohol_Consumption_after  = np.array([408, 439, 321, 437, 335, 344, 318, 492, 531, 417, 358, 391, 398, 394, 508, 399, 345, 341, 326, 467])

D  = Alchohol_Consumption_after -Alchohol_Consumption_before
print(D)
print('Mean is %3.2f and standard deviation is %3.2f' %(D.mean(),np.std(D,ddof = 1)))
```

```
[ -62   85 -175   86  -14 -105  -60  133   62   88  -31 -106  -95  126
   63  112    7   70  -86  132]
Mean is 11.50 and standard deviation is 95.68
```

```python
import  scipy.stats  as stats
t_statistic, p_value  =  stats.ttest_1samp(D, 0)
print('P Value %1.3f' % p_value)
```

```
P Value 0.597
```

Step 5: Decide to reject or accept null hypothesis

In this example, p value is 0.597 and it is more than 5% level of significance

So the statistical decision is to fail to reject the null hypothesis at 5% level of significance.

There is no sufficient evidence to reject the null hypothesis. So we fail to reject the null hypotheis and conclude that there is no effect of love failure on alcohol consumption

Analysis of variance (ANOVA)

ANOVA is a hypothesis testing technique tests the equality of two or more population means by examining the variances of samples that are taken.

ANOVA tests the general rather than specific differences among means.

Assumptions of ANOVA

- 1) All populations involved follow a normal distribution
- 2) All populations have the same variance
- 3) The samples are randomly selected and independent of one another

One-way ANOVA

Example 1

Consider the monthly income of members from three different gyms - fitness centers given below:

- Gym 1 (n = 22): [60, 66, 65, 55, 62, 70, 51, 72, 58, 61, 71, 41, 70, 57, 55, 63, 64, 76, 74, 54, 58, 73]
- Gym 2 (n = 18): [56, 65, 65, 63, 57, 47, 72, 56, 52, 75, 66, 62, 68, 75, 60, 73, 63, 64]
- Gym 3 (n = 23): [67, 56, 65, 61, 63, 59, 42, 53, 63, 65, 60, 57, 62, 70, 73, 63, 55, 52, 58, 68, 70, 72, 45]
  Using ANOVA, test whether the mean monthly income is equal for each Gym.

```python
import pandas            as pd
import numpy             as np
import matplotlib.pyplot as plt
import seaborn           as sns
```

```python
Gym_1 = np.array([60, 66, 65, 55, 62, 70, 51, 72, 58, 61, 71, 41, 70, 57, 55, 63, 64, 76, 74, 54, 58, 73])
Gym_2 = np.array([56, 65, 65, 63, 57, 47, 72, 56, 52, 75, 66, 62, 68, 75, 60, 73, 63, 64])
Gym_3 = np.array([67, 56, 65, 61, 63, 59, 42, 53, 63, 65, 60, 57, 62, 70, 73, 63, 55, 52, 58, 68, 70, 72, 45])

print('Count, Mean and standard deviation of monthly income of members of Gym 1: %3d, %3.2f and %3.2f' % (len(Gym_1), Gym_1.mean(),np.std(Gym_1,ddof =1)))
print('Count, Mean and standard deviation of monthly income of members of Gym 2: %3d, %3.2f and %3.2f' % (len(Gym_2), Gym_2.mean(),np.std(Gym_2,ddof =1)))
print('Count, Mean and standard deviation of monthly income of members of Gym 3: %3d, %3.2f and %3.2f' % (len(Gym_3), Gym_3.mean(),np.std(Gym_3,ddof =1)))
```

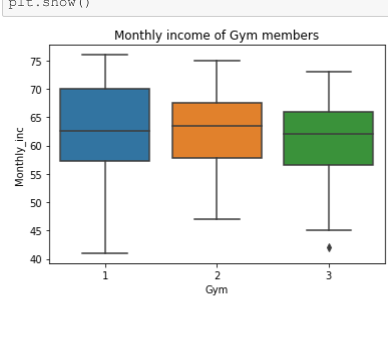
```
Count, Mean and standard deviation of monthly income of members of Gym 1:  22, 62.55 and 8.67
Count, Mean and standard deviation of monthly income of members of Gym 2:  18, 63.28 and 7.79
Count, Mean and standard deviation of monthly income of members of Gym 3:  23, 60.83 and 8.00
```

```python
monthly_inc_df = pd.DataFrame()

df1            = pd.DataFrame({'Gym': '1', 'Monthly_inc':Gym_1})
df2            = pd.DataFrame({'Gym': '2', 'Monthly_inc':Gym_2})
df3            = pd.DataFrame({'Gym': '3', 'Monthly_inc':Gym_3})

monthly_inc_df = monthly_inc_df.append(df1)
monthly_inc_df = monthly_inc_df.append(df2)
monthly_inc_df = monthly_inc_df.append(df3)
```

Let us explore the data graphically

A side by side boxplot is one of the best way to compare group locations, spreads and shapes.

```python
sns.boxplot(x = "Gym", y = "Monthly_inc", data = monthly_inc_df)
plt.title('Monthly income of Gym members')
plt.show()
```



The boxplots show almost similar shapes, location and spreads and group 3 has an low outlier.

Step 1: State the null and alternative hypothesis:

- H_0: $\mu1$ = $\mu2$ = $\mu3$
- H_A: At least one μ differs

Step 2: Decide the significance level

Here we select α = 0.05

Step 3: Identify the test statistic

Here we have three groups. Analysis of variance can determine whether the means of three or more groups are different. ANOVA uses F-tests to statistically test the equality of means.

Step 4: Calculate F, a test statistic

- scipy.stats.f.ppf gives the critical value at a given level of confidence with a pair of degrees of freedom.
- scipy.stats.f.cdf gives the cumulative distribution function for the given random variable - given the calculated F value at a given level of confidence with a pair of degrees of freedom.

```python
import scipy.stats as stats
```

```python
crit = stats.f.ppf(q = 1-0.05, dfn = 2, dfd = 60)
print('F critical value for 2 and 60 df with .95 confidence %3.2f' %crit)
```

```
F critical value for 2 and 60 df with .95 confidence 3.15
```

Directly calculate the fstat and pvalue

```python
stats.f_oneway(Gym_1,Gym_2,Gym_3)[0]
```

```
0.4970745666663714
```

or Calculate p value

```python
p_value = 1 -stats.f.cdf(0.497075, dfn = 2, dfd = 60)
print('P value for 2 and 60 df with .95 confidence for the calculated F value is %3.5f' % p_value)
```

```
P value for 2 and 60 df with .95 confidence for the calculated F value is 0.61079
```

Or formulate an ANOVA table using statsmodels

- statsmodels.formula.api.ols creates a model from a formula and dataframe
- statsmodels.api.sm.stats.anova_lm gives an Anova table for one or more fitted linear models

In the formula. we know that

- 1) ~ separates the left hand side of the model from the right hand side
- 2) + adds new columns to the design matrix
- 3) : adds a new column to the design matrix with the product of the other two columns
- 4) * also adds the individual columns multiplied together along with their product
- 5) C() operator denotes that the variable enclosed in C() will be treated explicitly as categorical variable.

In [30]:
```
import statsmodels.api         as      sm
from   statsmodels.formula.api import ols

mod = ols('Monthly_inc ~ Gym', data = monthly_inc_df).fit()
aov_table = sm.stats.anova_lm(mod, typ=2)
print(aov_table)
```

```
               sum_sq    df         F    PR(>F)
Gym         66.614123   2.0  0.497075  0.61079
Residual  4020.370004  60.0       NaN      NaN
```

Step 5: Decide to reject or accept null hypothesis

In this example, calculated value of F (= 0.497075) is less than Critical value of F(= 3.15)

So the statistical decision is to fail to reject the null hypothesis at 5% level of significance.

So there is no sufficient evidence to reject the null hypothesis that at least one mean monthly income of a gym is different from others .

Two-way ANOVA

The following table shows the quantity of soaps at different discount at locations collected over 20 days.

In [31]:
```
table1  = [['Loc','Dis0','Dis10','Dis20'], [ 1, 20, 28, 32], [ 2, 20, 19, 20],
  [ 1, 16, 23, 29 ],[ 2, 21, 27, 31 ],[ 1, 24, 25, 28 ],[ 2, 23, 23, 35 ],
  [ 1, 20, 31, 27 ],[ 2, 19, 30, 25 ],[ 1, 19, 25, 30 ],[ 2, 25, 25, 31 ],
  [ 1, 10, 24, 26 ],[ 2, 22, 21, 31 ],[ 1, 24, 28, 37 ],[ 2, 25, 33, 31 ],
  [ 1, 16, 23, 33 ],[ 2, 21, 26, 23 ],[ 1, 25, 26, 27 ],[ 2, 26, 22, 22 ],
  [ 1, 16, 25, 31 ],[ 2, 22, 28, 32 ],[ 1, 18, 22, 37 ],[ 2, 25, 24, 22 ],
  [ 1, 20, 24, 28 ],[ 2, 23, 23, 29 ],[ 1, 17, 26, 25 ],[ 2, 23, 26, 25 ],
  [ 1, 26, 28, 23 ],[ 2, 24, 16, 34 ],[ 1, 16, 21, 26 ],[ 2, 20, 30, 30 ],
  [ 1, 21, 27, 33 ],[2, 23, 22, 25 ],[ 1, 24, 25, 28 ],[ 2, 18, 16, 39 ],
  [ 1, 19, 20, 30 ],[ 2, 19, 25, 32 ],[ 1, 19, 26, 30 ],[ 2, 19, 34, 29 ],
  [ 1, 21, 26, 26 ],[ 2, 30, 23, 22 ]]
headers = table1.pop(0) #

df1 = pd.DataFrame(table1, columns=headers)
print(df1)
```

```
    Loc  Dis0  Dis10  Dis20
0     1    20     28     32
1     2    20     19     20
2     1    16     23     29
3     2    21     27     31
4     1    24     25     28
5     2    23     23     35
6     1    20     31     27
7     2    19     30     25
8     1    19     25     30
9     2    25     25     31
10    1    10     24     26
11    2    22     21     31
12    1    24     28     37
13    2    25     33     31
14    1    16     23     33
15    2    21     26     23
16    1    25     26     27
17    2    26     22     22
18    1    16     25     31
19    2    22     28     32
20    1    18     22     37
21    2    25     24     22
22    1    20     24     28
23    2    23     23     29
24    1    17     26     25
25    2    23     26     25
26    1    26     28     23
27    2    24     16     34
28    1    16     21     26
29    2    20     30     30
30    1    21     27     33
31    2    23     22     25
32    1    24     25     28
33    2    18     16     39
34    1    19     20     30
35    2    19     25     32
36    1    19     26     30
37    2    19     34     29
38    1    21     26     26
39    2    30     23     22
```

This is a two-way ANOVA with replication since the data contains values for multiple locations.

Conduct a two-way ANOVA at α = 5% to test the effects of discounts and location on sales.

In [32]:
```
d0_val  = df1['Dis0'].values
d10_val = df1['Dis10'].values
d20_val = df1['Dis20'].values
l_val   = df1['Loc'].values

df1   = pd.DataFrame({'Loc': l_val, 'Discount':'0','Qty': d0_val})
df2   = pd.DataFrame({'Loc': l_val, 'Discount':'10','Qty': d10_val})
df3   = pd.DataFrame({'Loc': l_val, 'Discount':'20','Qty': d20_val})

Sale_qty_df = pd.DataFrame()

Sale_qty_df = Sale_qty_df.append(df1)
Sale_qty_df = Sale_qty_df.append(df2)
Sale_qty_df = Sale_qty_df.append(df3)

pd.DataFrame(Sale_qty_df)
```

Out[32]:

	Loc	Discount	Qty
0	1	0	20
1	2	0	20
2	1	0	16
3	2	0	21
4	1	0	24
...	...	...	...
35	2	20	32
36	1	20	30
37	2	20	29
38	1	20	26
39	2	20	22

120 rows × 3 columns

Step 1: State the null and alternative hypothesis:

The null hypotheses for each of the sets are given below.

- 1) The population means of the first factor (Discount) are equal.
- 2) The population means of the second factor (Location) are equal.
- 3) There is no interaction between the two factors - Discount and Location.

Alternative Hypothesis:

- 1) The population means of the first factor (Discount) are not equal.
- 2) The population means of the second factor (Location) are not equal.
- 3) There is an interaction between the two factors - Discount and Location.

Step 2: Decide the significance level

Here we select α = 0.05

Step 3: Identify the test statistic

Here we have three groups and two factors. There are two independent variables, Discount and Location.

Two-way ANOVA determines how a response (Sale Quantity) is affected by two factors, Discount and Location.

Step 4: Calculate p value using ANOVA table

- statsmodels.formula.api.ols creates a model from a formula and dataframe
- statsmodels.api.sm.stats.anova_lm gives an Anova table for one or more fitted linear models

In [33]:
```
import statsmodels.api         as      sm
from   statsmodels.formula.api import ols
from   statsmodels.stats.anova import anova_lm

formula = 'Qty ~ Discount + C(Loc) + Discount:C(Loc)'
model = ols(formula, Sale_qty_df).fit()
aov_table = anova_lm(model, typ=2)

print(aov_table)
```

```
                      sum_sq     df          F        PR(>F)
Discount         1240.316667    2.0  39.279968  1.055160e-13
C(Loc)              7.008333    1.0   0.443898  5.065930e-01
Discount:C(Loc)    84.816667    2.0   2.686085  7.246036e-02
Residual         1799.850000  114.0        NaN           NaN
```

Step 5: Decide to reject or accept null hypothesis

In this example,

- p value for discount is 1.06e-13 and < 0.05 so we reject the null hypothesis (1) and conclude that the discount rate is having an effect on sales quantity.
- p value for location is 0.5066 and > 0.05 so we retain the null hypothesis (2) and conclude that the location is not having an effect on sales quantity.
- p value for interaction (discount:location) is 0.0725 and > 0.05 so we retain the null hypothesis (3) and conclude that the interaction (discount:location) is not having an effect on sales quantity.

Chi Square

A chi-square distribution with k degrees of freedom is given by sum of squares of standard normal random variables Z_1, Z_2, ... Z_k obtained by transforming normal standard variables X_1, X_2, ... X_k with mean values μ_1, μ_2, ... μ_k and corresponding standard deviation σ_1, σ_2, ... σ_k

${\chi_k}^2$ = ${Z_1}^2$ + ${Z_2}^2$ + ... + ${Z_k}^2$

The probability density function of f(x) =

$\frac{x^{\frac{k}{2}-1} e^{\frac{-x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$ if x > 0 else 0

where Γ(k/2) is a gamma function given by

$\Gamma\frac{k}{2}$ = $\int_0^\infty x^{k-1} e^{-x} dx$

Properties of Chi Square distribution

1. The mean and standard deviation of a chi-square distribution are k and √2k respectively, where k is the degrees of freedom.

2. As the degrees of freedom increases, the probability density function of a chi-square distribution approaches normal distribution.

3. Chi-square goodness of fit is one of the popular tests for checking whether a data follows a specific probability distribution.

4. Chi square test is a right tailed test.

Chi-square Goodness of fit tests

Goodness of fit tests are hypothesis tests that are used for comparing the observed distribution pf data with expected distribution of the data to decide whether there is any statistically significant difference between the observed distribution and a theoretical distribution (for example, normal, exponential, etc.) based on the comparison of observed frequencies in the data and the expected frequencies if the data follows a specified theoretical distribution.

Hypothesis	Description
Null hypothesis	There is no statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution
Alternative hypothesis	There is statistically significant difference between the observed frequencies and the expected frequencies from a hypothesized distribution

Chi-square Goodness of fit tests

Chi-square statistic for goodness of fit is given by

χ^2 = $\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

This test is invalid when the observed or expected frequencies in each category are too small. A typical rule is that all of the observed and expected frequencies should be at least 5.

Chi-square tests of independence

Chi-square test of independence is a hypothesis test in which we test whether two or more groups are statistically independent or not.

Hypothesis	Description
Null Hypothesis	Two or more groups are independent
Alternative Hypothesis	Two or more groups are dependent

The corresponding degrees of freedom is (r - 1) * (c - 1) , where r is the number of rows and c is the number of columns in the contingency table.

scipy.stats.chi2_contingency is the Chi-square test of independence of variables in a contingency table.

This function computes the chi-square statistic and p-value for the hypothesis test of independence of the observed frequencies in the contingency table observed. The expected frequencies are computed based on the marginal sums under the assumption of independence.

Example:

The table below contains the number of perfect, satisfactory and defective products are manufactured by both male and female.

Gender	Perfect	Satisfactory	Defective
Male	138	83	64
Female	64	67	84

Do these data provide sufficient evidence at the 5% significance level to infer that there are differences in quality among genders (Male and Female)?

Step 1: State the null and alternative hypothesis:

Null hypothesis: H_0: There is no difference in quality of the products manufactured by male and female

Alternative hypothesis: H_A: There is a significant difference in quality of the products manufactured by male and female

Step 2: Decide the significance level

Here we select α = 0.05

Step 3: Identify the test statistic

We use the chi-square test of independence to find out the difference of categorical variables

Step 4: Calculate p value or chi-square statistic value

In [34]:
```
import pandas       as pd
import numpy        as np
import scipy.stats as stats

quality_array = np.array([[138, 83, 64],[64, 67, 84]])
chi_sq_Stat, p_value, deg_freedom, exp_freq = stats.chi2_contingency(quality_array)

print('Chi-square statistic %3.5f P value %1.6f Degrees of freedom %d' %(chi_sq_Stat, p_value,deg_freed
om))
```

```
Chi-square statistic 22.15247 P value 0.000015 Degrees of freedom 2
```

Step 5: Decide to reject or accept null hypothesis

In this example, p value is 0.000015 and < 0.05 so we reject the null hypothesis.

So, we conclude that there is a significant difference in quality of the products manufactured by male and female.

End