Inferential Stats:

it is a branch of statistics that involves drawing conclusion about a population based on sample data taken from population.

Hypothesis Testing:

it is a technique used in Inferential stats to compare assumptions.
we assume 2 hypothesis

Null hypothesis
it considered both groups are having same parameter values i.e. there is no difference Alternate hypothesis
it considered both groups are having different parameter values.
we try to reject Null hypothesis on the basis of appropriate test and pvalue.

there are different type of tests used in HT.

Type of tests:

Parametric
Non-parametric

Parametric Tests:

Assumption: Parametric tests assume that the data comes from a specific distribution, usually the normal distribution.

Examples:
t-tests (for two groups),
analysis of variance (ANOVA, for more than two groups),
z-test

Advantages: Parametric tests are often more powerful (sensitive) when the assumptions are met.
Disadvantages: They are sensitive to violations of assumptions, especially the assumption of normality.

Non-parametric Tests:

Assumption: Non-parametric tests do not assume a specific distribution for the data.

Examples:
Mann-Whitney U test (non-parametric equivalent of t-test for two groups),

Kruskal-Wallis test (non-parametric equivalent of ANOVA),
chi-sqr test

Advantages: Non-parametric tests are more robust in the presence of outliers or when the assumption of normality is violated.

Disadvantages: They may be less powerful than parametric tests when the assumptions of parametric tests are met.

Choosing Between Parametric and Non-parametric Tests:

If your data meets the assumptions of parametric tests (e.g., normal distribution, homogeneity of variances), and the sample size is sufficient, parametric tests are often preferred for their higher statistical power.

If your data does not meet the assumptions of parametric tests, or if you are dealing with ordinal or non-normally distributed data, non-parametric tests might be more appropriate.

P-value

it is the probabilty of getting the difference by chance or randomly.

A lower p-value means less chance to happen the difference by chance i.e. stronger evidence against the null hypothesis.

Generally,p-value <=.05 then we reject Null Hypothesis and accept alternate Hypothesis

Parametric Tests

t-test

T-test is used to compare the means of two groups.

It is performed on continuous variables.

The data should be approximately normally distributed.

The data should not contain any outlier

### Types of t test

One Sample t test
Independent(Unpaired) t test
Dependent(Paired) t test

One Sample t test:

The one sample t test compares the mean of your sample data to a known value or population mean.

The **Null hypothesis** is there is no difference between the sample mean and given or population mean.

The **Alternate hypothesis** is there is a difference between the sample mean and given or population mean.

```
**Example**:
- We have a sample of 25 individuals and the question is to test whether the
average weight of sample is equal (approximately equal) to population?
```

```
In [11]: import pandas as pd
          import random as rd
          import numpy as np
          from scipy import stats
         df=pd.read_csv("f:/dataset/analysis/weight-height.csv")
         p=df.Weight
         s=rd.sample(list(p),25)
         svalue,pvalue=stats.ttest_1samp(s,66)
         print(pvalue)
         print(np.mean(p),np.mean(s))
         if(pvalue<=.05):
          print("Rejecting Null Hypothesis and accepting alternate i.e. means are not
         same") else:
          print("Fail to reject Null Hypothesis and accepting Null i.e.means are same")
```

```
0.39026847723619873
66.367559754866 66.75506048640001
Fail to reject Null Hypothesis and accepting Null i.e.means are same
```

Independent(Unpaired) t test:

Independent (or unpaired two sample) t-test is used to compare the means of two unrelated groups of samples.

Example

We have a sample of individuals (25 women and 25 men).
The **Null Hypothesis** is there is no difference between average Height of women & men.
The **Altrnate Hypothesis** is there is a difference between average Height of women & men.

```
In [13]:s1=df[df.Gender=='Male'].sample(25).Height
s2=df[df.Gender=='Female'].sample(25).Height
_,pvalue=stats.ttest_ind(s1,s2)
if(pvalue<=.05):
 print("Rejecting Null Hypothesis and accepting alternate i.e. means are not
same") else:
 print("Fail to reject Null Hypothesis and accepting Null i.e.means are same")
```

```
Rejecting Null Hypothesis and accepting alternate i.e. means are not same
```

Dependent(Paired) t test:

Paired t-test when you want to compare means of the same group at different times.

The **Null Hypothesis** is there is no difference in means.

The **Alternate Hypothesis** is there is a difference in means.

Example:

You could use a paired t-test to understand whether there was a difference in managers' salaries before and after undertaking a PhD.
Reaction times for the same people on a task before or after a training.
Sales of an organization before or after covid.
Effect of a fairness cream before and after on same group.

```
In [16]: s1=[.2,.4,.5,.15,.6,.7,.4,.6,.8,.75]
s2=[.2,.35,.55,.15,.6,.72,.45,.61,.85,.76]

_,pvalue=stats.ttest_rel(s1,s2)
if(pvalue<.05):
 print("Rejecting Null hypo i.e. there is a difference")
else:
 print("Accepting Null hypo i.e. there is no difference")
```

```
Accepting Null hypo i.e. there is no difference
```

ANOVA test:

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.

The samples are independent.

The **Null hypothesis** in ANOVA is valid when all the sample means are equal, or they don’t have any significant difference.

The **Alternate hypothesis** is all the sample means are not same (even if any group mean is not same with other) .

```
In [18]:phy = [34, 36, 30, 35, 37]
chm = [28, 32, 29, 26, 30]
maths = [38, 41, 40, 39, 36]
_,pvalue=stats.f_oneway(phy,chm,maths)
if(pvalue<.05):
 print("Rejecting Null hypo i.e. there is a difference")
else:
 print("Accepting Null hypo i.e. there is no difference")
```

```
Rejecting Null hypo i.e. there is a difference
```

Z-test:

Z-test is used when sample size is large (n>30).
Statsmodels has a ztest function that allows you to compare two means, assuming they are independent and have the same standard deviation.

```
In [28]:from statsmodels.stats import weightstats
s1=np.random.randn(100)
#1 Sample Z-test
svalue,pvalue=weightstats.ztest(s1,value=0)
if(pvalue<.05):
 print("Rejecting Null i.e. means are not equal")
else:
 print("Accepting Null i.e. means are equal")

s1=np.random.randn(100)
s2=np.random.randn(100)
#2 Sample Z-test(Independent)
svalue,pvalue=weightstats.ztest(s1,s2)
if(pvalue<.05):
 print("Rejecting Null i.e. means are not equal")
else:
 print("Accepting Null i.e. means are equal")
```

```
 Accepting Null i.e. means are equal
 Accepting Null i.e. means are equal
```

Non-Paramertic Tests

Chi Square Test:

The Chi Square statistic is commonly used for testing relationships between categorical variables.
The Null hypothesis of the Chi-Square test is that no relationship exists on the categorical variables in the population; they are independent.
The Alternate hypothesis of the Chi-Square test is that relationship exists on the categorical variables in the population; they are dependent.

It is performed on two or more categorical variables

```
In [30]:import pandas as pd
df=pd.read_csv("f:/dataset/analysis/titanic.csv")
frq_df=pd.crosstab(df.Survived,df.gender)
frq_df
```

1 233 109

Out[30]:
gender female

male Survived

0 81 468

```
In [31]: _,pvalue,_,_=stats.chi2_contingency(frq_df)
         if(pvalue<.05):
          print("Rejecting Null hpyo i.e. relationship exists between Survived & Gender ")
         else:
          print("Accepting Null i.e. no relationship exists between Survived & Gender")
```

```
Rejecting Null hpyo i.e. relationship exists between Survived & Gender
```

Mann Whitney U

> it is 2 sample independent test.

```
In [32]:from scipy.stats import mannwhitneyu

         # Example data for two independent groups
         group1 = [23, 45, 67, 12, 89, 34, 78, 56]
         group2 = [11, 55, 76, 23, 90, 44, 67, 32]

         # Perform Mann-Whitney U Test
         statistic, p_value = mannwhitneyu(group1, group2)

         if p_value <= .05:
          print("Reject the null hypothesis. There is a significant difference between the gro
         else:
            print("Fail to reject the null hypothesis. There is no significant difference betwee
```

```
Fail to reject the null hypothesis. There is no significant difference between the group
s.
```

Wilcoxon test

> it is 2 sample paired(dependent test)

```
In [34]:from scipy.stats import wilcoxon

         # Example data for paired samples
         before = [23, 45, 67, 12, 89, 34, 78, 56]
         after = [18, 42, 65, 10, 85, 30, 75, 50]

         # Perform Wilcoxon Signed-Rank Test
         statistic, p_value = wilcoxon(before, after)

         if p_value <=.05:
          print("Reject the null hypothesis. There is a significant difference between the pai
         else:
             print("Fail to reject the null hypothesis. There is no significant difference betwee
```

```
Reject the null hypothesis. There is a significant difference between the paired sample
s.
```

Kruskal test

Anova equivalant of non-parametric

3 sample independent.

```
In [35]:from scipy.stats import kruskal

         # Example data for three independent groups
         group1 = [23, 45, 67, 12, 89, 34, 78, 56]
         group2 = [11, 55, 76, 23, 90, 44, 67, 32]
         group3 = [37, 49, 61, 28, 75, 40, 82, 50]

         # Perform Kruskal-Wallis test
         statistic, p_value = kruskal(group1, group2, group3)

         if p_value <=.05:
          print("Reject the null hypothesis. There is a significant difference between the gro
         else:
             print("Fail to reject the null hypothesis. There is no significant difference betwee
```

```
Fail to reject the null hypothesis. There is no significant difference between the group
s.
```

```
In [ ]:
```