Statistical Inference

An Introduction

Contents

Basic Terms as Prerequisite

- **Variable** (under study) – What you measure (ex. monthly salary of employees)
- **Population**- Set of all units in the study (all employees in the organization)
- **Sample**- Subset of units selected from population (ex. monthly salary of few selected employees in the organization)
- **Distribution**-How values of variable are distributed in the population (ex. normal distribution)
- **Factor**- Defines subgroups in the study.(ex. Gender, where gender wise salary distribution can be studied.)
- **Descriptive Statistics**- mean, median, standard deviation etc of the variable under study.. (ex. Average salary)

What is Statistical Inference ?

- Statistical inference is the process of drawing conclusion about unknown population properties, using a sample drawn from the population.

- These unknown population properties can be:
  - Mean
  - Proportion
  - Variance etc.

- Such unknown population properties are called as ‘Parameters’.

What is Statistical Inference ?

Statistical Inference

Estimation

Point estimation	Interval estimation	Testing of Hypothesis
• Summarize the sample by a single value as an estimate of the population parameter. • Ex. Average salary of junior data scientists is. 55,000 euros.	• A range of values within which, we believe, • the true population parameter lies with high probability. • Ex. Average salary of junior data scientists is in the range of (52,000,55,000) • With 95% confidence level.	• To decide whether a statement regarding population parameter is true or false, based on sample data. • Ex. Claim: Average salary of junior data scientists is greater than.50,000 euros annually.

Parameter, Estimator, Estimate

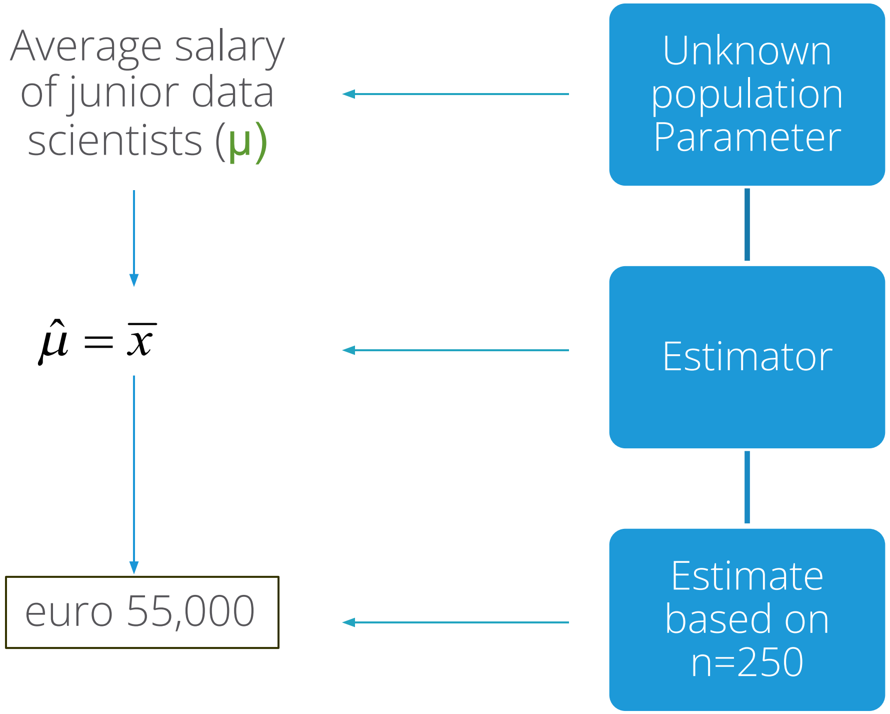
- **Parameter**: Unknown property or characteristic of population
  - (population mean (μ), variance (σ^2), proportion (P))

- **Estimator**: A rule or function based on sample observations which is used to estimate the parameter
  - (sample mean, sample variance, sample proportion)

- **Estimate:** A particular value computed by substituting the sample observations into an Estimator.

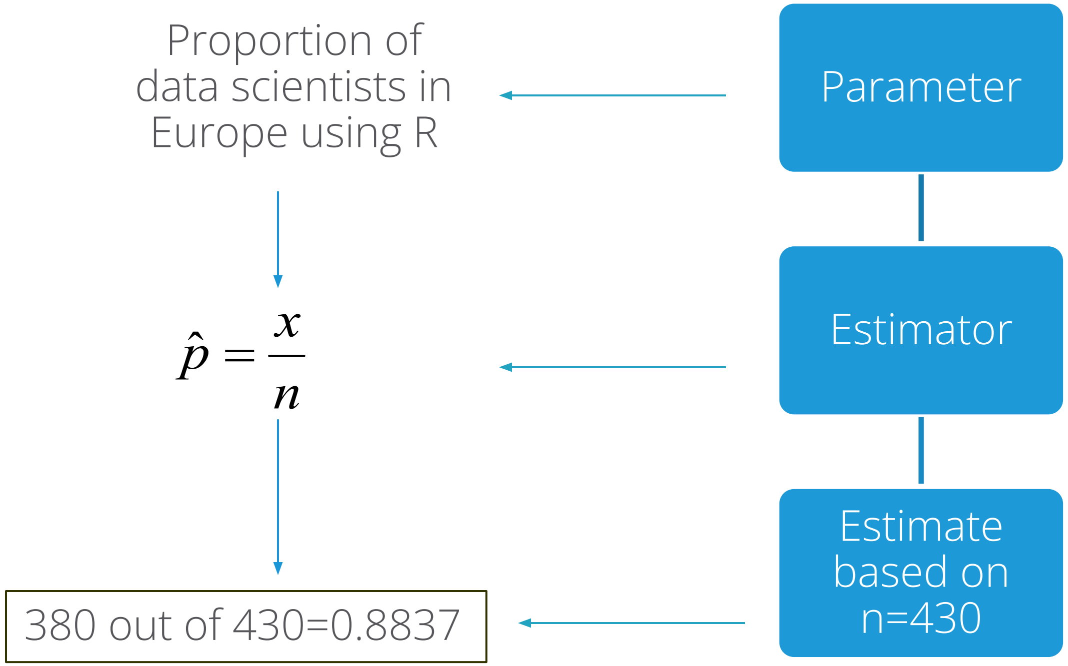
Parameter, Estimator, Estimate

- Research Question: What is the average salary of junior data scientists in Europe?
  - Average salary of junior data scientists in Europe is Population **Parameter.**
  - Sample of 250 junior data scientists is observed and Sample mean is computed.
  - Sample mean is used as **Estimator** of Population Mean.
  - Sample mean “55,000” which is calculated from sample of 250 is the **Estimate.**



Parameter, Estimator, Estimate

- Research Question: What is the proportion of data scientists in Europe who use R for data analysis?
  - Proportion of data scientists in Europe who use R for data analysis is population parameter.
  - Sample of 430 data scientists observed and proportion (or percentage) is calculated.
  - Sample proportion is used as an estimator of population proportion.
  - 380 out of 430 which is calculated from sample is **Estimate.**



Point Estimation vs. Interval Estimation

- In both the previous examples, (estimation of average salary of junior data scientists and proportion of data scientists using R) estimator is a single value estimating unknown population parameter.

- A confidence interval gives an estimated range of values which is likely to include an unknown population parameter with some probability, the estimated range being calculated from a given set of sample data.

- Generally, 95% or 90% Confidence Intervals are used.

- 95% confidence interval is a range estimate within which the true value of the parameter lies with probability 0.95.

Sampling distribution and Sampling error

- Research Question: What is the average salary of junior data scientists in Europe?

  - 50 samples, each of size 250 junior data scientists are observed and sample mean for each of theses 50 samples are computed. Here, sample mean will vary based on sample values.

- A probability distribution of all these means of the sample is called the **sampling distribution** of mean.

- **Standard error** is standard deviation of the these mean values.

Hypothesis Testing

- **Hypothesis**: An assertion about the distribution / parameter of the distribution of one or more random variables.
- **Null Hypothesis (H_0)**: An assertion which is generally believed to be true until researcher rejects it with evidence.
- **Alternative Hypothesis (H_1)**: A researcher's claim which contradicts null hypothesis.
- In simple words, testing of hypothesis is to decide whether a statement regarding population parameter is true or false, based on sample data.
- **Test Statistic**: The statistic on which decision rule of rejection of null hypothesis is defined.
- **Critical region or Rejection region**: the region, in which, if the value of test statistic falls, the null hypothesis is rejected.

Hypothesis Testing : Example

Objective	A consumer protection agency wants to test a Paint Manufacturer's claim, that average drying time of their new paint is less than 20 minutes.

- Sample: n=36 boards were painted from 36 different cans and the drying time was observed.
- Estimator of mean drying time is sample mean $\hat{\mu} = \bar{x}$

Null Hypothesis (H_0): $\mu = 20$

Alternate Hypothesis (H_1): $\mu < 20$

Test Statistic	In this case the test statistic is based on $\bar{x}$
Decision Criteria	Reject null hypothesis if test statistic based on sample mean is less than critical value.

Two types of error

- While testing the hypothesis using any decision rule, one of the following scenario might occur.

Decision	Reality	
	Ho is true	Ho is false
Reject Ho	Type I error	Correct
Do Not Reject Ho	Correct	Type II error

- **For example**, in legal system,
  H_0: person is not guilty H_1: person is guilty

Decision	Reality	
	Not Guilty	Guilty
Guilty	Type I Error -- Innocent person goes to jail	Correct
Not Guilty	Correct	Type II error Guilty person is set free

Two Types of error

- **Level of significance (LOS):** Probability of Type I error is called as 'Level of Significance (α)'
  generally set as 5% (los=0.05) and null hypothesis is rejected if observed risk(p value) is less
  than 0.05

- **p-value:** is the smallest level of significance that would lead to rejection of the null hypothesis (generally if $p < 0.05$, we reject the null hypothesis).

- **α =** Probability [Type I Error] = Probability [Reject H_0 | H_0 is True]

- **β =** Probability [Type II Error] = Probability [Do not reject H_0 | H_0 is not True]

- **Power of the test** is given by: $(1 - \beta)$

One tailed and two tailed tests

- Hypothesis test where the alternative hypothesis is one-tailed (right-tailed or left-tailed), is called a **one-tailed test**.

  $H0: \mu = \mu 0$
  $H1: \mu > \mu 0$ (Right-tailed) or $H1: \mu < \mu 0$ (left-tailed)

- Hypothesis test where the alternative hypothesis is two-tailed is called **two-tailed test**.

  $H0: \mu = \mu 0$
  $H1: \mu \neq \mu 0$

Quick Recap

Statistical Inference	• It is the process of drawing conclusion about unknown population properties, using a sample drawn from the population.
Point Estimation	• Summarize the sample by a single value as an estimate of the population parameter.
Interval Estimation	• A range of values within which, we believe, the true population parameter lies with high probability.
Testing of Hypothesis	• To decide whether a statement regarding population parameter is true or false
Type I error	• α = Probability [Type I Error] = Probability [Reject H0 | H0 is True]
Type II error	• β = Probability [Type II Error] = Probability [Do not reject H0 | H0 is not True] • Power of the test is given by: $(1 - \beta)$

Statistical Inference

Testing Assumption of Normality

Contents

Normality test

- An assessment of the normality of data is a prerequisite for many statistical tests because normal distribution is an underlying assumption in parametric testing.

- Normality can be assessed using two approaches: graphical and numerical.
  - Graphical approach
    - Box-Whisker plot (It is used to asses symmetry rather than normality.)
    - Quantile-Quantile plot (Q-Q plot).

  - Numerical (Statistical) approach
    - Shapiro-Wilk test (Used generally for **small sample**)
    - Kolmogorov-Smirnov test (Used generally for **large sample**)

* Box-Whisker plot is used to asses symmetry rather than normality. Hence, only Q-Q plot method is explained.

Case Study

To assess normality of data in Python, we shall consider the below case as an example.

Background

Data has 2 variables recorded for 80 guests in a large hotel.
Customer Satisfaction Index (csi) & Total Bill Amount in thousand Rs. (billamt)

Objective

To check if variables follow normal distribution

Sample Size

Sample size: 80
Variables: id, csi, billamt

Data Snapshot

Normality Testing
Data

Variables

ervations

id	csi	billamt
1	38.35	34.85
2	47.02	10.99
3	36.96	24.73
4	43.07	7.9
5	38.77	9.38

Column	Description	Type	Measurement	Possible Values
id	Customer ID	Numeric	-	-
csi	Customer Satisfaction Index	Numeric	-	Positive value
billamt	Total Bill Amount in thousand Rs.	Numeric	Rs.	Positive value

Quantile-Quantile plot

- Very powerful graphical method of assessing Normality.
- Quantiles are calculated using sample data and plotted against expected quantiles under Normal distribution.
- If Normality assumption is valid then high correlation is expected between sample quantiles and expected(theoretical quantiles under normal distribution) quantiles.
- The Y axis plots the actual quantiles values based on sample. The X axis plots theoretical values.
- If the data is truly sampled from a Normal distribution, the QQ plot will be linear.

QQ Plot in Python For Variable csi

#Import Data

```
import pandas as pd
data=pd.read_csv('Normality Testing Data.csv')
```

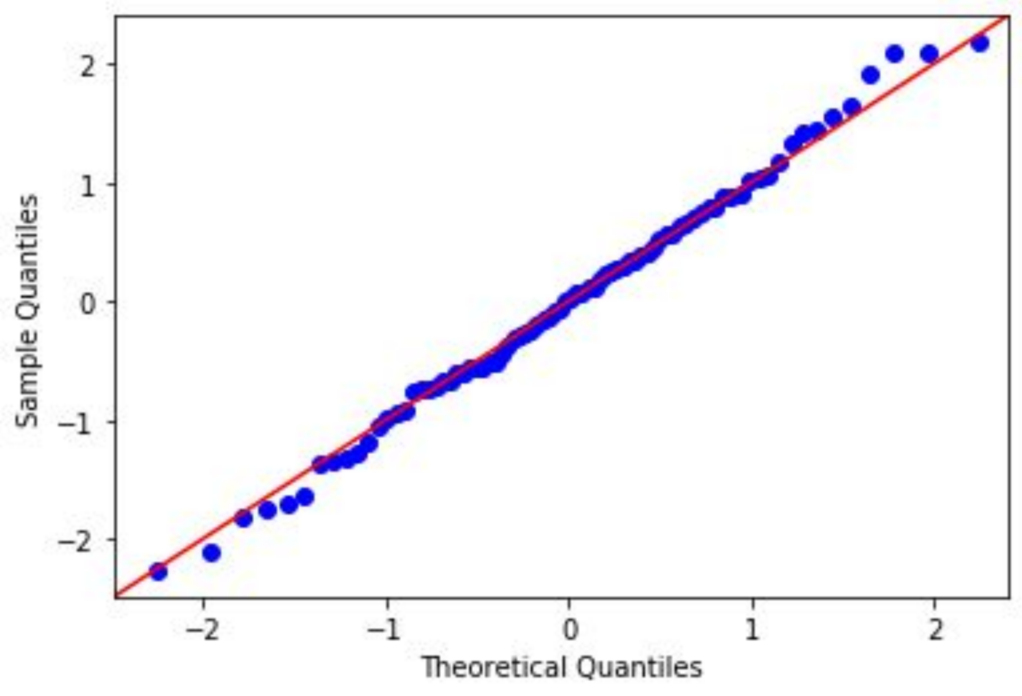
#QQ Plot

```
import statsmodels.api as sm
sm.graphics.qqplot(data.csi, line='45', fit=True)
```

- **qqplot()** produces a plot with theoretical quantiles on x axis against the sample quantiles on y axis. Column for which normality is being tested is specified in the first argument.
- **line=** is an argument that adds reference line to the qqplot. Here it adds a 45-degree line
- **fit=True** indicates, parameters are fit using the distribution's fit() method

QQ Plot in Python For Variable csi

```
# Output:
```



Interpretation :

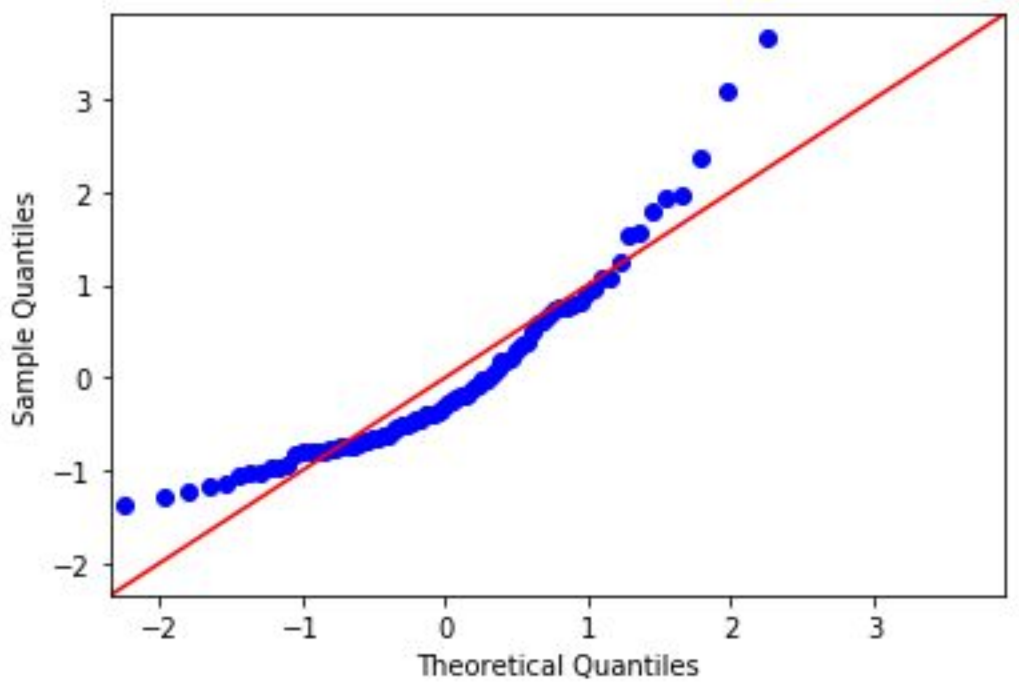
- Q-Q plot is Linear. Distribution of ‘csi’ can be assumed to be normal.

Q-Q plot in Python For Variable billamt

```
# Q-Q plot for the variable billamt
sm.graphics.qqplot(data.billamt, line='45', fit=True)
```

- **data.billamt** is the variable for which normality is to be checked.

```
# Output:
```



Interpretation :

- Q-Q plot is deviated from linearity. Distribution of ‘billamt’ appears to be non-normal.

Shapiro-Wilk test

Shapiro-Wilk test is widely used statistical test for assessing **Normality**.

Objective	To test the **normality** of the data.

Null Hypothesis (H_0): Sample is drawn from Normal Population

Alternate Hypothesis (H_1): Sample is drawn from Non-Normal Population

The test is performed for the variables, 'csi' and 'billamt' separately.

Test Statistic	It correlates sample ordered values with expected Normal scores. (actual calculation is very complex so we will avoid details)
Decision Criteria	Reject the null hypothesis if p-value < 0.05

* Here we continue to use previous data of CSI and Bill Amount for Shapiro-Wilk test.

Shapiro Wilk Test For Variable csi

```
# Shapiro Wilk Test
import scipy as sp
sp.stats.shapiro(data.csi)
```

shapiro() from scipy package, returns correlation coefficient w and p-value.

```
# Output
(0.9919633269309998, 0.9037835597991943)
```

Interpretation :

- Since p-value is >0.05,do not reject H0. Distribution of ‘csi’ can be assumed to be normal.

Shapiro-Wilk test For Variable bilamt

```
# Shapiro Wilk test for the variable billamt
sp.stats.shapiro(data.billamt)
```

- **data.billamt** is the variable for which normality is to be checked.

```
# Output:
(0.890307724758606, 4.858443844568683e-06)
```

Interpretation :

- Since p-value is <0.05, reject H0. Distribution of 'billamt' appears to be non-normal.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov test is another widely used statistical test for assessing Normality.

Objective	To test the **normality** of the data.

Null Hypothesis (H_0): Sample is drawn from Normal Population

Alternate Hypothesis (H_1): Sample is drawn from Non-Normal Population

The test is performed for the variables, 'csi' and 'billamt' separately.

Test Statistic	Kolmogorov-Smirnov Test: It compares empirical (sample) cumulative distribution function (CDF) with Normal distribution CDF. The test statistic is maximum difference between CDF's.
Decision Criteria	Reject the null hypothesis if p-value < 0.05

* Here we continue to use previous data of CSI and Bill Amount for Shapiro-Wilk test.

Kolmogorov-Smirnov test in Python

```
# Kolmogorov Smirnov test
sm.stats.diagnostic.lilliefors(data.csi)
```

- Instead **of lilliefors, kstest_normal()** from **statsmodels** can also be used to perform a Lilliefors (KS) Normality Test.
- Both tests returns Kolmogorov-Smirnov test statistic and p-value.
- **data.csi** is the variable for which normality is to

```
# Output:
(0.04238708824708459, 0.9859314950919987)
```

Interpretation :

- Since p-value is >0.05,do not reject H0. Distribution of ‘csi’ can be assumed to be normal.

Kolmogorov-Smirnov test in Python

```
# Kolmogorov Smirnov test for the variable billamt

sm.stats.diagnostic.lilliefors(data.billamt)
```

- **data.billamt** is the variable for which normality is to be checked.

```
# Output:

(0.1424429511673755, 0.0009999999999998899)
```

Interpretation :

- Since p-value is <0.05, reject H0. Distribution of ‘billamt’ appears to be non-normal.

Statistical Inference

Parametric Tests I

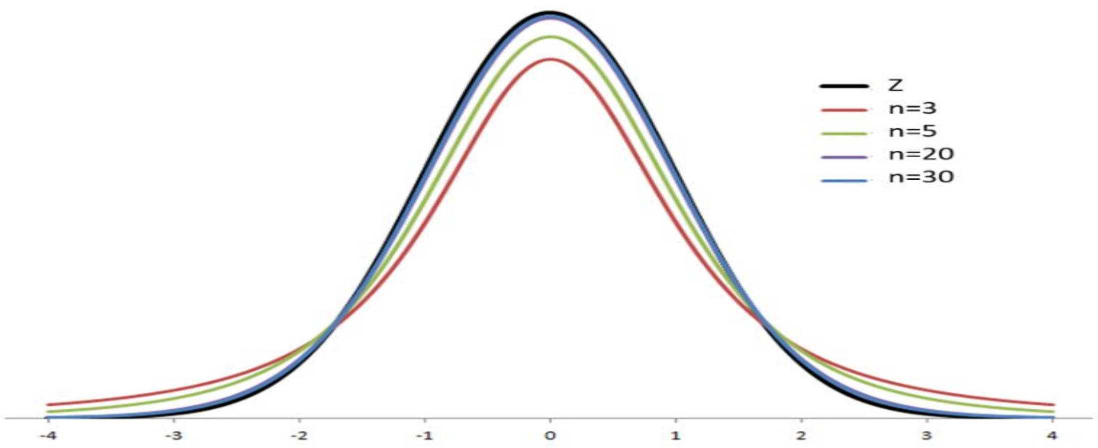
Contents

1. Introduction to t-distribution
2. One Sample t-test
3. Independent samples t-test

t-distribution

- The t distribution is symmetric and its overall shape resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider.

- As the sample size increases so as the number of degrees of freedom grows, the t-distribution approaches the normal distribution with mean 0 and variance 1.



- In the above graph, z is normal distribution with mean 0 and variance 1.

A note on Degrees of Freedom (DF)

- Degrees of freedom (df) is defined as the number of independent terms.
- "Sum of the squared deviations about mean of n values" has n-1 degrees of freedom. Knowing n-1 values, we can find last value since sum of deviations about mean is always zero.
- Sampling distributions like t, F and chi square have shapes based on degrees of freedom.
- Example , Give 5 numbers such that sum is 20. You can use 4 numbers freely but fifth number should be such that sum is 20. Here df =4

One sample t-test

- One sample t test is used to test the hypothesis about a single population mean.
- We use one-sample t-test when we collect data on a single sample drawn from a defined population.
- For this design, we have one group of subjects, collect data on these subjects and compare sample statistic to the hypothesized value of population parameter.
- Subjects in the study can be patients, customers, retail stores etc.

Case Study

To execute Parametric test in Python, we shall consider the below case as an example.

Background

A large company is concerned about time taken by employees to complete weekly MIS report.

Objective

To check if average time taken to complete the MIS report is more than 90 minutes

Sample Size

Sample size: 12
Variables: Time

Data Snapshot

ONE SAMPLE t TEST

Variables

Observations

Time
85
95
105
85
90
97
104
95

Columns	Description	Type	Measurement	Possible values
Time	Time taken to complete MIS	Numeric	Minutes	Positive Values

Assumptions for one sample t-test

- The assumptions of the one-sample t-test are listed below:
  - Random sampling from a defined population
    (employees are selected at random from the company)

  - Population is normally distributed
    (Time taken to complete MIS report should be normally distributed).

  - Variable under study should be continuous.

- Normality test can be performed by any of the methods explained earlier.

- The validity of the test is not seriously affected by moderate deviations from 'Normality' assumption.

One sample t-test

Testing whether mean is equal to a test value.

Objective	To test the average time taken to complete MIS is more than 90 minutes

Null Hypothesis (H_0): $\mu = 90$

Alternate Hypothesis (H_1): $\mu > 90$

Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	12
Mean		93.5833
Standard Deviation	S	6.4731
Standard Error	s/√n	1.8686
Difference		93.5833-90=3.5833
t	$\frac{\bar{x} - \mu_0}{S.E}$	1.9176

One sample t-test in Python

```
# Import data
data2=pd.read_csv('ONE SAMPLE t TEST.csv')
```

```
# t-test for one sample
from scipy.stats import ttest_1samp
ttest_1samp(data2.Time, popmean=90,alternative='greater')
```

- **ttest_1samp()** from scipy package, returns two tailed t and p-value.
- **data.time** is the variable under study.
- **popmean=90** is the value to be tested.

```
# Output:
Ttest_1sampResult(statistic=1.9176218472595046, pvalue=0.04074043079962237)
```

Interpretation :

- **scipy always gives the test statistic as signed**. This means that given p and t values from a two-tailed test, you would reject the null hypothesis of a greater-than test when p/2 < alpha and t > 0, and of a less-than test when p/2 < alpha and t < 0.
- Since p/2 is <0.05, reject H0. Average time taken to complete the MIS report is more than 90 minutes ’

Independent samples t-test

- The independent-samples t-test compares the means of two independent groups on the same continuous variable.

- Following hypotheses are tested in independent samples t test

  - H0: Two population means are equal
  - H1: Two population means are not equal

Case Study

To execute Parametric test in Python, we shall consider the below case as an example.

Background

The company is assessing the difference in time to complete MIS report between two groups of employees :
Group I: Experience(0-1 years)
Group II: Experience(1-2 years)

Objective

To test whether the average time taken to complete MIS by both the groups is same.

Sample Size

Sample size: 14
Variables: time_g1, time_g2

Data Snapshot

INDEPENDENT SAMPLES t TEST

Variables

servations

time_g1	time_g2
85	83
95	85
105	96
85	94

Columns	Description	Type	Measurement	Possible values
time_g1	Time to complete MIS report by group1	Numeric	Hours	Positive Values
time_g2	Time to complete MIS report by group2	Numeric	Hours	Positive Values

Assumptions for independent samples t-test

- The assumptions for independent samples t-test are listed below :

  - The samples drawn are random samples.
    (Employees are selected at random from the company)

  - The populations from which samples are drawn have equal & unknown variances.
    (F-test is used to validate this assumption which will be covered in next presentation)

  - The populations follow normal distribution.
    (**Time taken to complete MIS report should be normally distributed for both groups**)

    Normality assumption can be validated using method explained earlier).

Independent sample t-test

Testing whether means of two groups are equal.

Null Hypothesis (H_0): μ1 = μ2

Alternate Hypothesis (H_1): μ1 ≠ μ2

μ1= average time taken by group1 to complete MIS
μ2=average time taken by group2 to complete MIS .

Objective	To test the average time taken to complete MIS by both the groups is same.

Test Statistic	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Group I	Group II
Sample Size	n1=12	n2=14
Mean		
Variance	S_1^2=41.9015	S_2^2=27.1483
Pooled Variance	Sp^2=33.9102	
Difference		
t	$t = \dfrac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ =0.22345	

Independent samples t-test in Python

```
# Import data
```

```
import pandas as pd
data=pd.read_csv('INDEPENDENT SAMPLES t TEST.csv')
```

```
# t-test for independent samples
```

```
from scipy import stats
stats.ttest_ind(data['time_g1'],data['time_g2'],nan_policy='omit’,
,equal_var=True)
```

- ❑ **ttest_ind()** from scipy, returns t & pvalue
- ❑ **nan_policy=‘omit’** Defines how to handle when input contains nan. 'propagate' returns nan, 'raise' throws an error, 'omit' performs the calculations ignoring nan values. Default is 'propagate'.

* Before performing t test, normality test is done to ensure time variable is normally distributed in both the groups.

Independent samples t-test in Python

```
# Output:
```

```
Ttest_indResult(statistic=0.22345590920212569,pvalue=0.8250717960964372)
```

Interpretation :

- Since p-value is >0.05, do not reject H0. There is no significant difference in average time taken to complete the MIS between both the group of employees.

Independent samples t-test when variances are not equal

- Welch's t test is used to test the equality of two means if variances of two groups can not be assumed equal.
- Welch's t-test defines the statistic t by the following formula:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The denominator is not based on a pooled variance estimate.
- If 2 variance are not equal, t test syntax in Python is given below:

```
data=pd.read_csv('INDEPENDENT SAMPLES t TEST.csv')

stats.ttest_ind(data['time_g1'],data['time_g2'], equal_var=False,
nan_policy='omit')
```

Independent samples t-test when variances are not equal

```
# Output:
```

```
Ttest_indResult(statistic=0.21965992515741178,pvalue=0.8282468548302413)
```

Interpretation :

- Since p-value is >0.05, do not reject H0. There is no significant difference in average time taken to complete the MIS between both the group of employees.

Quick Recap

In this session, we continued to learn various parametric tests . Here is a quick recap :

Independent sample t test	• It compares the means of two independent groups on the same continuous variable. • H0: $\mu1 = \mu2$

Statistical Inference

Parametric Tests II

Contents

1. **Paired sample t-test**
2. **t test for correlation**

Paired samples t-test

- The paired sample t-test is used to determine whether the mean difference between two sets of observations is zero ,where each subject or entity is measured twice resulting in pair of observations.

- Commonly used when observations are recorded ‘before’ and ‘after’ the treatment / training and objective is to test whether the treatment/training is effective.

Case Study

To execute Parametric test in Python, we shall consider the below case as an example.

Background

The company organized a training program to improve efficiency. Time taken to complete MIS report before and after training are recorded for 15 employees.

Objective

To test whether the average time taken to complete MIS before and after training is not different.

Sample Size

Sample size: 15
Variables: time_before, time_after

Data Snapshot

PAIRED t
TEST

Variables

servations

time_before	time_after
85	74
95	91
92	80
102	91

Columns	Description	Type	Measurement	Possible values
time_before	Time to complete MIS report before training	Numeric	Hours	Positive values
time_after	Time to complete MIS report after training	Numeric	Hours	Positive values

Assumptions for paired sample t-test

- The assumptions of the paired-sample t-test are listed below:

  - Random sampling from a defined population
    (employees are selected at random from the company)

  - Population of the testing variable is normally distributed
    (Difference time taken to complete MIS report should be normally distributed).

- Normality test can be performed by any of the methods explained earlier.

- The validity of the test is not seriously affected by moderate deviations from 'Normality' assumption.

Paired sample t-test

Testing whether means of two dependent groups are equal.

Objective	To test the average time taken to complete MIS before and after training is not different.

Null Hypothesis (H_0): There is no difference in average time before and after the training. i.e. D=0

Alternate Hypothesis (H_1): Average time is less after the training. (Training is effective.) D>0

D= μBefore – μAfter

Test Statistic	$t = \frac{\bar{d}}{s_d/\sqrt{n}}$ Where $\bar{d}$ is the sample mean of the difference before-after , s_d is the sample standard deviation of the difference, n is the sample size of difference. The quantity t follows a distribution called as 't distribution' with n-1 degrees of freedom.
Decision	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	12
Mean difference (before-after)	$\bar{d}$	8.3333
Standard Deviation	s_d	3.9219
t	$t = \frac{\bar{d}}{s_d / \sqrt{n}}$	8.2295

Paired sample t-test in Python

```
# Import data
data=pd.read_csv('PAIRED t TEST.csv')
```

```
# t-test for paired samples
stats.ttest_rel(data['time_before'],data['time_after']
,alternative='greater')
```

- **data['time_before']** and **data['time_after']** are the variables under study.
- **ttest_rel()** from scipy, returns t & pvalue

* Before performing t test, normality test is done to ensure difference variable is normally distributed.

Paired sample t-test in Python

```
# Output:
```

```
Ttest_relResult(statistic=8.22948711672449, pvalue=4.918935850301797e-07)
```

Interpretation :

- Since p-value is <0.05, reject H0. Average time taken to


```
Ttest_relResult(statistic=8.22948711672449, pvalue=4.918935850301797e-07)
```


  training is effective.
- 95% C.I does not contain value of D=0 (under H0),reject H0.

t-test for Correlation

- Correlation coefficient summarizes the strength of a linear relationship between two variables.

- t-test is used to check if there is significant correlation between two variables.

- Sample correlation coefficient (r) is calculated using bivariate data.

- Null hypothesis of this test is
  H0: there is no correlation between 2 variables under study ($\rho=0$)

Case Study

To execute Parametric test in Python, we shall consider the below case as an example.

Background

A company with 25 employees has calculated job proficiency score & aptitude test score for its employees

Objective

To test if there is significant correlation between job proficiency and aptitude test score.

Sample Size

Sample size: 25
Variables: Empcode, Aptitude, Job_prof

Data Snapshot

Correlation
test

Variables

bservations

Empcode	aptitude	job_prof
E101	86	88
E102	62	80
E103	110	96
E104	101	76
E105	100	80
E106	78	73

Columns	Description	Type	Measurement	Possible values
Empcode	Employee code	Numeric	-	
Aptitude	Score of aptitude test	Numeric	-	Positive values
Job_prof	Job proficiency score	Numeric	-	Positive values

Correlation t-test

Testing for correlation coefficient value.

Objective	To test whether there exists significant correlation between job proficiency and aptitude score.

Null Hypothesis (H_0): There is no correlation between Job proficiency and Aptitude test

Alternate Hypothesis (H_1): There is correlation between Job proficiency and Aptitude test.

Test Statistic	$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$ **where r is the sample correlation coefficient, the sample size. The quantity t follows a distribution called as 't distribution' with n-2 degrees of freedom.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Notation	Value
Sample Size	n	25
Sample correlation coefficient	r	0.514411
t	$t = \frac{r\sqrt{(n-2)}}{\sqrt{1-r^2}}$	2.8769

Correlation **t-test in Python**

```
# Import data
data=pd.read_csv('Correlation test.csv')

# t-test for correlation
stats.pearsonr(data['aptitude'], data['job_prof'])
```

- **data[‘aptitude’]** and **data[‘job_prof ‘]**are the variables under study.
- **pearsonr()** from scipy, returns t & pvalue

Correlation t-test in Python

```
# Output:
```

```
(0.5144106946654772, 0.008517216152487137)
```

Interpretation :

- Since p-value is <0.05, reject H0. There is correlation between aptitude test and job proficiency.
- 95% C.I does not contain value ρ=0 (under H0), reject H0.

Quick Recap

In this session, we continued to learn various parametric tests . Here is a quick recap

:

Paired sample t test	• Used to determine whether the mean difference between two sets of observations is zero ,where each subject or entity is measured twice resulting in pair of observations. • H0: $\mu1-\mu2=d=0$
t test for correlation	• Used to check if there is significant correlation between two variables. • H0: $\rho=0$

Statistical Inference

Test for equality of variances

F-test for equality of variances

- F test is used to test the equality of two population variances.
- Testing equality of variances is the prerequisite for many statistical test (like Independent sample t-test).
- Under H0 $\sigma_1^2 = \sigma_2^2$

Where σ_1^2 and σ_2^2 are the first and second population variances, respectively.

Assumptions for F-test

- The assumptions of F-test are listed below:
  - Random sampling from a defined population
    (employees are selected at random from the company)

  - Population of the testing variable is normally distributed
    (Time taken to complete MIS report should be normally distributed).

- Note that, generally F test is used to validate assumption of equal variance while performing t test for equality of means. The parent population is assumed to follow normal distribution.

Case Study

To execute Test for Equality of Variance in Python, we shall consider the below case as an example.

Background

The company is analysing time to complete MIS report between two groups of employees.
Group I: Experience (0-1 years)
Group II: Experience(1-2 years)

Objective

To test the equality of the variances in time taken to complete MIS in two groups of employees.

Sample Size

Sample size: 14
Variables: time_g1, time_g2

Data Snapshot

F test for 2 variances

Variables

servations

time_g1	time_g2
85	83
95	85
105	96

Columns	Description	Type	Measurement	Possible values
time_g1	Time to complete MIS report by group1	Numeric	Hours	Positive Values
time_g2	Time to complete MIS report by group2	Numeric	Hours	Positive Values

F-test

Testing equality of variances in two samples.

Objective	To test the **equality** of the variances in time taken to complete MIS in two groups of employees.

Null Hypothesis (H_0): Variances of time are equal in two groups. i.e. $\sigma_1^2 = \sigma_2^2$.
Alternate Hypothesis (H_1): Alternative Hypothesis H1: $\sigma_1^2 \neq \sigma_2^2$

Test Statistic	$F = \frac{s_1^2}{s_2^2} \sim F_{\alpha, n_1-1, n_2-1}$ **Where s_1^2 is the sample variance of first sample and, s_2^2 is the sample variance of the second sample. n_1 and n_2 are sample sizes of first and second sample respectively.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Computation

	Group I	Group II
Sample Size	n_1=12	n_2=14
Mean	$\bar{x}_1$ = 93.5833	
Sample Variance	s_1^2 =41.9015	s_2^2 = 27.1484
F Value	$F = \frac{s_1^2}{s_2^2}$	1.5434

F-test in Python

```
# Import data
data = pd.read_csv('F test for 2 variances.csv')
```

```
# Variance test
import numpy as np
from scipy import stats

x = np.array(data.dropna()['time_g1'])
y = np.array(data['time_g2'])

f = np.var(x, ddof=1)/np.var(y, ddof=1) #calculate F test
statistic
dfn = x.size-1 #define degrees of freedom numerator
dfd = y.size-1 #define degrees of freedom denominator
p = 2*(1-stats.f.cdf(f, dfn, dfd)) #find p-value of F test
statistic
print(f, p)
```

```
# Output :
1.5434275971616587 0.4523632544892888
```

Interpretation :

- Since p-value is >0.05, do not reject H0. There is no significant difference in variances of the two groups

Statistical Inference

Analysis of variance

Contents

Analysis of Variance (ANOVA)

- Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among more than two group means developed by statistician and evolutionary biologist **Ronald Fisher**.

- Example: There are 20 plots of wheat and 5 fertilizers are applied to four different plots. The yield of wheat is recorded for each of 20 plots.
  ANOVA can be used to find out whether effect of these fertilisers on yields is equal or significantly
  different.

ANOVA

- Note that although the name is 'Analysis of Variance', the method is used to analyze the differences among group means.

- Variation in the variable is inherent in nature. In general, the observed variance in a particular variable is partitioned into components attributable to different sources of variation.

- The total variance in any variable is due to a number of causes which may be classified “assignable causes (which can be detected and measured)” and “chance causes (which is beyond control of human and cannot be traced separately)”.

- Hence, ANOVA is the separation of variance ascribable to one group of causes from the variance ascribable to other group.

Assumptions of ANOVA

- The assumptions of ANOVA are listed below:

  - The samples drawn are random samples.

  -The populations from which samples are drawn have equal & unknown variances.

  - The populations follow normal distribution.

Testing Normality assumption

- An assessment of the normality of data is a prerequisite for many statistical tests because normal data is an underlying assumption in parametric testing.

- Normality can be assessed using two approaches: graphical and numerical.

  - **Graphical approach**

    - Box-Whisker plot (It is used to asses symmetry rather than normality.)
    - Quantile-Quantile plot (Q-Q plot).

  - **Statistical approach**

    - Shapiro-Wilks test
    - Kolmogorov-Smirnov test

* **Normality test is already covered Parametric test ppt.**

One Way ANOVA

- One Way Anova can be considered as an extension of the t test for independent samples.
- One Way Anova is used to test the equality of K population means.
  (when K=2, t test can be used.)
- For two levels (K=2),the t test and One Way Anova provide identical results.
- **Mathematical model** is :

  $X_{ij} = \mu_i + \varepsilon_{ij}$

  Where X_{ij} is the jth observation due to ith level of a factor. μ_i is the effect of ith level of a factor. ε_{ij} is the error term. i=1,2,..,k ; j=1,2,...,n_i
- The null hypothesis is

$$H_0 : \mu_1 = \mu_2 = \dots\dots\dots\dots\dots\dots\dots = \mu_K = \mu$$

Partitioning Total Variance

- Total variation is partitioned into two parts:
  Total SS= Between Groups SS + Within Groups SS
  where, SS stands for sum of squares

$$SS_{total} = \sum_i \sum_j \left(X_{ij} - \bar{X}_{..} \right)^2$$

Total variation (Total SS)

$$SS_{between} = \sum_i n_i \left(\bar{X}_{i.} - \bar{X}_{..} \right)^2$$

Variation due to Assignable causes (Between SS)

$$SS_{error} = \sum_i \sum_j \left(X_{ij} - \bar{X}_{i.} \right)^2$$

Variation due to Chance causes (Within SS)

- Total SS is calculated using squared deviations of each value from overall mean.
- Between SS is calculated using squared deviation of each group mean from overall mean.
- Within Group SS can be obtained by subtracting Between SS from Total SS

Case Study

To execute analysis of Variance in Python, we shall consider the below case as an example.

Background

A large company is assessing the difference in 'Satisfaction Index' of employees in Finance, Marketing and Client-Servicing departments.

Objective

To test whether **mean satisfaction index** for employees in three departments (CS, Marketing, Finance) are equal.

Sample Size

Sample size: 37
Variables: satindex, dept

Data Snapshot

One way anova

Variables

Observations

satindex	dept
75	FINANCE
56	FINANCE
72	FINANCE
59	FINANCE
66	FINANCE
58	FINANCE
58	MARKETING
63	MARKETING

Columns	Description	Type	Measurement	Possible values
satindex	Satisfaction Index	Numeric		Positive Values
dept	Department	Character	MARKETING, CS, FINANCE	3

One Way ANOVA

Testing equality of means in one factor with more than two levels.

Objective	**To test whether mean satisfaction index for employees in three departments (CS, Marketing, Finance) are equal.**

Null Hypothesis (H_0): Mean satisfaction index for 3 departments are equal i.e. $\mu1= \mu2= \mu3$
Alternate Hypothesis (H_1): Mean satisfaction index for 3 departments are not equal

Test Statistic	**The test statistic is denoted as F and is based on F distribution.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Calculation

Total SS = (75-65.59)^2+(56-65.59)^2+…………+(65-65.59)^2+(76-65.59)^2
= 1840.92

Between Groups SS = 12*(64.42-65.59)^2+12*(63.25-65.59)^2+13*(68.85-65.59)^2
= 220.0599

Within Groups SS = Total SS – Between SS

Overall Mean	65.59	n=37
Mean for Finance	64.42	n1=12
Mean for Marketing	63.25	n2=12
Mean for CS	68.85	n3=13

One Way ANOVA table

Sources of variation	Degrees of freedom (df)	Sum of Squares (SS)	Mean Sum of Squares (MS=SS/df)	F-Value
Between groups	K-1=3-1 =2	SSA=**220.0599**	MSA=110.03	F=2.3080
Within groups (error)	n-k=37-3 =34	SSE=**1620.86**	MSE=47.6724	
TOTAL	n-1=37-1 =36	TSS=**1840.92**		

One Way ANOVA in Python

```
# Import data
import pandas as pd
data = pd.read_csv('One way anova.csv')
```

```
# ANOVA table
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('satindex ~ C(dept)', data=data).fit()
aov_table = sm.stats.anova_lm(model, typ=2)
aov_table
```

Output:

	sum_sq	df	F	PR(>F)
C(dept)	220.059945	2.0	2.308047	0.114836
Residual	1620.858974	34.0	NaN	NaN

- **ols()** from statsmodels.formula.api is used to fit the model
- Independent variable to be specified as **C()**
- **sm.stats.anova_lm()** from statsmodel.api is used to get ANOVA table
- **typ =** determines how the sum of squares is calculated & **typ = 2** if there is no

Interpretation :

- Since p-value is >0.05, do not reject H0. There is no significant difference in satisfaction index among 3 different departments

Quick Recap

ANOVA

- Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among more than two group means developed by statistician and evolutionary biologist Ronald Fisher.

Partitioning the variance

- The total variance in any variable is due to a number of causes which may be classified “assignable causes (which can be detected and measured)” and “chance causes (which is beyond control of human and cannot be traced separately)”.

One Way ANOVA

- Comparing several means of different levels of one factor.

$$H_0 : \mu_1 = \mu_2 = \dots\dots\dots\dots\dots\dots\dots = \mu_K = \mu$$

Statistical Inference

Two-Way Analysis of Variance

Contents

Two Way ANOVA

- Two Way Anova is used when there are 2 factors under study.
- Each factor can have 2 or more levels . Example: Gender and Age can be 2 factors.
  Gender with 2 levels as Male and Female
  Age with 3 levels as 18-30,31-50 and >50
- Three hypothesis are tested.

  Factor A — H0: All group means are equal
  H1: At least one mean is different from other means

  Factor B — H0: All group means are equal
  H1: At least one mean is different from other means

  Interaction — H0: The interaction is not significant
  H1: The interaction is significant

* For two-way ANOVA with interaction there has to be more than one observation per combination of the levels of factors.

Two Way ANOVA

- **Total variation is partitioned as below :**

Total SS= Between Groups SS due to factor A (SSA)

+ Between Groups SS due to factor B (SSB)

+ Interaction SS due to factor A and B (SSAB)

+ Error SS (SSE)

where, SS stands for sum of squares

* SS formulae for two-way ANOVA with interaction are not specified due to their complexity.

Case Study

We will illustrate Two Way Anova in Python using following case study

Background

A large company is assessing the difference in 'Satisfaction Index' of employees in Finance, Marketing and Client-Servicing departments. Experience level is also considered in the study.(<=5 years and >5 years)

Objective

To test the equality of the satisfaction index among employees of three departments (CS, Marketing, Finance) and among different experience bands.

Sample Size

Sample size: 36
Variables: satindex, dept, exp

Data Snapshot

Two Way Anova

Variables

Observations

satindex	dept	exp
75	FINANCE	lt5
56	FINANCE	lt5
62	FINANCE	gt5
66	FINANCE	gt5
58	FINANCE	gt5
58	MARKETIN	lt5
63	MARKETIN	lt5
53	MARKETIN	lt5
74	MARKETIN	lt5
77	MARKETIN	lt5
69	MARKETIN	lt5
57	MARKETIN	gt5
70	MARKETIN	gt5
68	MARKETIN	gt5
77	CS	lt5

Columns	Description	Type	Measurement	Possible values
Satindex	Satisfaction Index	Numeric	-	Positive Values
Dept	Department	Character	MARKETING, CS, FINANCE	3
Exp	Years of Experience (grouped)	Character	lt5 = less than 5, gt5 = greater than 5	2

Two Way ANOVA

Testing equality of means in two factors.

Objective	To compare employee satisfaction index in three departments (CS, Marketing, Finance) and two experience level based groups.

Null Hypothesis

(H_{01}): Average satisfaction index is equal for 3 departments.

(H_{02}): Average satisfaction index is equal for 2 experience levels.

(H_{03}) Interaction effect(dept*exp) is not significant on satisfaction index.

The test statistic is computed for each of these null hypothesis.

Reject the null hypothesis if p-value < 0.05

Two Way ANOVA in Python

```
# Import data
```

```
import pandas as pd
data = pd.read_csv('Two Way Anova.csv')
```

```
# ANOVA Table
```

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('satindex ~ C(dept) + C(exp) + C(dept) : C(exp)',
data=data).fit()
sm.stats.anova_lm(model, typ=2)
```

- **'sm.stats.anova_lm'** is the Python function for ANOVA .
- **formula** specifies 'satindex' as analysis (dependent) variable and 'dept' and 'exp' as factor (independent) variables.
- **C(dept) : C(exp)** specifies the interaction effect.

Two Way ANOVA in Python

```
# Output:
```

	sum_sq	df	F	PR(>F)
C(dept)	164.222222	2.0	1.678973	0.203624
C(exp)	78.027778	1.0	1.595479	0.216274
C(dept):C(exp)	20.222222	2.0	0.206748	0.814374
Residual	1467.166667	30.0	NaN	NaN

Interpretation :

- Since p-value is >0.05 for all three (dept, exp and dept*exp), do not reject H0 for all three tests. There is no significant difference in satisfaction index among 3 different departments and 2 experience levels.
- Also interaction effect is not significant.

Knowledge check question

- A large retailer is testing a marketing campaign on 24 stores. 8 stores are selected randomly from each of 3 zones.
- The variable of interest is ' sales increment(%) during campaign month’. Objective is to test whether the campaign is equally effective in 3 regions. Data is given below.

NORTH	WEST	SOUTH
8	10.2	5.3
12.5	9.3	5.8
9.2	9.9	6
6.7	8.7	7.1
9.4	9.1	7
5.9	10.2	6.1
7.7	9.5	6.3
6.9	10	7.3

- Is this One-way ANOVA problem or Two-way ANOVA problem?

ANSWER : One-way ANOVA

EXPLANATION : There is only one factor (zone) with 3 levels (North, West, South).

Quick Recap

Two Way Anova

- The two way anova is extension of one way anova when we have 2 factors in the study instead of one.

Null Hypothesis Drawing Inference

- Equality of means for levels in factor A
- Equality of means for levels in factor B
- No Interaction effect between 2 factors
- Total sum of squares is split into 4 parts and each hypothesis is tested.

Statistical Inference

Non-Parametric Tests 1

Contents

Non-parametric statistical test

- Tests based on t and F distribution assume that populations are normally distributed.
- A large body of statistical methods is available which do not make assumptions about the nature of the distribution(e.g. normality)
- These testing procedures are termed as Nonparametric tests or distribution-free tests.
- If the underlying assumptions of the parametric test are met, then the parametric test will be more powerful than nonparametric test.

* Note : Always check for the normality assumptions using test explained earlier and then decide which hypothesis test is more accurate depending upon the problem statement.

Mann-Whitney test

- The Mann-Whitney test is considered as nonparametric alternative to t test for independent samples.
- The Mann-Whitney U test is used to compare differences between two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.
- The test is equivalent to Wilcoxon rank-sum test (WRS).
- The null hypothesis is that the distributions of both groups are identical, so that there is a 50% probability that an observation randomly selected from one population exceeds an observation randomly selected from the other population.

Mann-Whitney test

- **Steps to follow :**
  - Combine the two samples.
  - Rank all the observations from smallest to largest.
  - Keep track of the group to which each observation belongs.

- Tied observations(observations with same value) are assigned a rank equal to the mean of the rank positions for which they are tied.

- The test statistic is

  $$U = T - \frac{m(m+1)}{2}$$

  Where T is sum of the ranks of first sample in combined ordered sample, m and n are sample sizes.

  $$E(U) = \frac{mn}{2} \qquad V(U) = \frac{mn(m+n+1)}{12}$$

- Standardized U is assumed to follow normal distribution.

- Compare p-value with level of significance & conclude.

Case Study - 1

To execute Non-Parametric test in Python, we shall consider the below case as an example.

Background

Data consist of aptitude score of 2 groups of employees.

Objective

To compare Aptitude scores of two groups and test if they come from the same population.

Sample Size

Sample size: 13
Variables: aptscore, Group

Data Snapshot

Mann Whitney test

Variables

Observations

aptscore	Group
11.5	G1
23.6	G1
12.1	G1
14.5	G1
16.1	G1
19.8	G1
11	G2
25	G2
26.3	G2

Combine the two sample from smallest to largest

Ordered Sample	Group	Rank
11	GroupII	1
11.5	GroupI	2
12.1	GroupI	3
12.2	GroupII	4
14.5	GroupI	5
15	GroupII	6
15.7	GroupII	7
16.1	GroupI	8
18.5	GroupII	9
19.8	GroupI	10

Rank ordered samples

Columns	Description	Type	Measurement	Possible values
aptscore	Aptitude score	Numeric	-	Positive values
Group	group of employees	character	G1,G2	2

- T is sum of the ranks of first sample in combined ordered sample. m and n are sample sizes.
  T=39 , m=6, n= 7
  U=18 , E(U)=21 ,V(U)= 49

Mann-Whitney test

Testing distribution of two samples

Objective	To test the **null hypothesis** that **median** of both the samples is same

Null Hypothesis (H_0): The two samples come from the same population

Alternate Hypothesis (H_1): The two samples do not come from the same population

Test Statistic	$$U = T - \frac{m(m+1)}{2}$$	**Where T is sum of the ranks of first sample in combined ordered sample, m and n are sample sizes**
Decision Criteria	Reject the null hypothesis if p-value < 0.05	

Mann-Whitney test in Python

```
# Import the CSV file
```

```
import pandas as pd
data = pd.read_csv('Mann Whitney test.csv')
```

```
# Mann-Whitney test
```

```
from scipy.stats import mannwhitneyu

# similar to aptscore ~ Group in R
# create objects with aptscore for G1 & G2 separately
group1 = data[data['Group'] == 'G1']['aptscore']
group2 = data[data['Group'] == 'G2']['aptscore']
mannwhitneyu(group1, group2, alternative="two-sided")
```

- **mannwhitneyu from scipy.stats** gives the value of U(as statistics) and p-value.
- **alternative =** Defines the alternative hypothesis. The following options are available None(default), less, greater, two-sided

Mann-Whitney test in Python

```
# Output:
```

```
MannwhitneyuResult(statistic=18.0, pvalue=0.7307692307692307)
```

Interpretation :

- Since p-value is >0.05, do not reject H0. aptitude score is same for both the groups i.e. samples come from the same population.

Wilcoxon Signed Rank Test for paired data

- The Wilcoxon Signed Rank test is considered as nonparametric alternative to paired t test .

- The Wilcoxon Signed Rank test is used to compare differences between two related or paired groups when the variable is either ordinal or continuous, but not normally distributed.

- H0: The median of difference in the population is zero
  H1: Not H0.

Wilcoxon Signed Rank Test for paired data

- **Steps to follow :**
  - Define Di = Xi- Yi which are the differences between two values for each pair.
  - Obtain |Di| which are absolute values of differences.
  - Rank all |Di| from smallest to largest.
  - Define Ri = rank of |Di|.
  - Obtain 'W' which is sum of the ranks associated with positive Di.

- The test statistic is W: which is sum of the ranks associated with positive Di. n is the sample size.

$$E(W) = \frac{n(n+1)}{4}$$

$$V(W) = \frac{n(n+1)(2n+1)}{24}$$

- Standardized W is assumed to follow normal distribution.

- Compare p-value with level of significance & conclude.

Case Study - 2

To execute Non-Parametric test in Python, we shall consider the below case as an example.

Background

A company organized a training program and the scores before and after training were recorded.

Objective

To test whether the median of paired samples is same.

Sample Size

Sample size: 12
Variables: Before, After

Data Snapshot

- A company organized a training program and the scores before and after training were recorded.

Variables

Observations

Before	After
58	74
52	65
61	60
48	45
50	58
39	53
51	61

Combine the two sample to get Di

Di=Before-After	Abs(Di)	Rank (Di)
-16	16	12
-13	13	10
1	**1**	**1**
3	**3**	**3**
-8	8	5
-14	14	11
-10	10	7

Rank associated with positive Di

Columns	Description	Type	Measurement	Possible values
Before	Score before training	Numeric	-	Positive values
After	Score after training	Numeric	-	Positive values

- W is sum of the ranks associated with positive Di. n is sample size.
  W=4 , n= 12
  E(W)=39 , V(W)= 162.5

Wilcoxon Signed Rank Test for paired data

Testing distribution of paired smaples

Objective	To test the **null hypothesis** that **median** of paired samples is same.

Null Hypothesis (H_0): The median of the difference in the population is zero
Alternate Hypothesis (H_1): The median of the difference in the population is less than zero.

Test Statistic	**w=sum of the ranks associated with positive Di.** **Di = Xi- Yi which are the differences between data and specified median value.**
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Wilcoxon Signed Rank Test for paired data **in** Python

```
# Import the CSV file
data = pd.read_csv('Wilcoxon Signed Rank test for paired data.csv')
```

```
# Wilcoxon Signed Rank test
from scipy.stats import wilcoxon
wilcoxon(data['Before'], data['After'], alternative = "less")
```

- **wilcoxon from scipy.stats** gives the value of W (as statistics) and p-value.
- **wilcoxon** function performs Wilcox signed rank test for paired data
- **alternative=less** specifies one tail test .since, score will be more if training program is effective.

Wilcoxon Signed Rank Test for paired data in Python

```
# Output:
WilcoxonResult(statistic=4.0, pvalue=0.001708984375)
```

Interpretation :

- Since p-value is <0.05, reject H0. Training program is effective as score after training is more than before training.

Quick Recap

In this session, we learnt various non parametric tests . Here is a quick recap :

Non Parametric Test	• Non parametric tests are performed if normality assumption is not satisfied.
Mann-Whitney test	• Nonparametric alternative to t test for independent samples.
Wilcoxon Signed Rank test	• Nonparametric alternative to t test for paired samples.

Statistical Inference

Non-Parametric Tests II

Contents

1. **Kruskal Wallis test**
2. **Chi-square test of association**

Kruskal Wallis test

- The Kruskal Wallis test is considered as nonparametric alternative to one way analysis of variance (ANOVA).
- The Kruskal Wallis test is used to compare differences between more than two independent groups when the dependent variable is either ordinal or continuous, but not normally distributed.
- H0: K samples come from the same population
  H1: Not H0.

Kruskal Wallis test procedure

- Combine all the observations from k samples into a single sample of size n and arrange them in ascending order .
- Assign ranks to them from smallest to largest as 1 to n. if there is a tie at two or more places, each observation is given the mean of the ranks for which it is tied.
- The ranks assigned to observations in each of the k groups are added separately to give k rank sums.
- The test statistic is

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$

$$n_j = number\ of\ observations\ in\ j^{th}\ sample$$

$$n = number\ of\ observations\ in\ the\ combined\ sample$$

$$R_j = sum\ of\ the\ ranks\ in\ the\ j^{th}\ sample.$$

- H follows Chi Square Distribution with k-1 df

Case Study - 1

To execute Non-Parametric test in Python, we shall consider the below case as an example.

Background

Data consist of aptitude score of 3 groups of employees.

Objective

To check whether there is difference in the score among three groups.

Sample Size

Sample size: 20
Variables: aptscore, Group

Data Snapshot

Kruskal Wallis Test

Variables

Observations

aptscore	Group
11.5	GroupI
12.1	GroupI
14.5	GroupI
16.1	GroupI
19.8	GroupI
23.6	GroupI
11	GroupII
12.2	GroupII
15	GroupII
15.7	GroupII
18.5	GroupII
25	GroupII
26.3	GroupII
11.1	GroupIII

Combined sample arranged from smallest to largest

Ordered sample	Group	Rank
11	GroupII	1
11.1	GroupIII	2
11.5	GroupI	3
12.1	GroupI	4
12.2	GroupII	5
14.5	GroupI	6
15	GroupII	7
15.7	GroupII	8
16.1	GroupI	9
18.5	GroupII	10
19.1	GroupIII	11
19.8	GroupI	12
21	GroupIII	13
22.2	GroupIII	14

Rank ordered sample

Columns	Description	Type	Measurement	Possible values
aptscore	Aptitude score	Numeric	-	Positive values
Group	Group of employees	character	GroupI, GroupII, GroupIII	3

Kruskal Wallis test

Testing distribution of more than two samples

Objective	To test the **null hypothesis** that all the samples came from same population

Null Hypothesis (H_0): The three samples are from the same population
Alternate Hypothesis (H_1): The three samples do not come from the same population

Test Statistic	$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1)$$ n_j = number of observations in j^{th} sample; n = number of observations in the combined sample; R_j = sum of the ranks in the j^{th} sample.
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Kruskal Wallis test example

Calculations :

	Value
Sample size	$n_1 = 6$ $n_2 = 7$ $n_3 = 7$
R_1	50
R_2	68
R_3	92
H	2.2309
p-value	0.3278

Kruskal Wallis test in Python

```
# Import the CSV file
import pandas as pd
data = pd.read_csv('Kruskal Wallis Test.csv')
```

```
# Kruskal wallis test
from scipy.stats import kruskal

group1 = data[data['Group'] == 'GroupI']['aptscore']
group2 = data[data['Group'] == 'GroupII']['aptscore']
group3 = data[data['Group'] == 'GroupIII']['aptscore’]

kruskal(group1, group2, group3)
```

- **kruskal from scipy.stats** performs the Kruskal wallis test on the data.
- **aptscore** is the analysis variable.
- **Group** is the factor variable.

Kruskal Wallis test in Python

```
# Output:
```

```
KruskalResult(statistic=2.230929090974231, pvalue=0.3277629827136111)
```

Interpretation :

- Since p-value is >0.05, do not reject H0. Aptitude score is same for all three groups of employees.

Chi-square test of Association

- The chi-square test for independence, also called as Pearson's chi-square test or the chi-square test of association, is used to test if there is a relationship between two categorical variables.

- The two categorical variables can be nominal or ordinal.

- H0: Two attributes are independent (not associated)
  H1: Not H0.

Chi-square test procedure

- Assume that there are ‘r’ categories of attribute A and ‘c’ categories of attribute B. Therefore, we have a cross table of r*c (r rows and c columns).
- Let Ri be the total of ith row and Cj be the total of jth column.
- Observed frequencies are calculated from the data.
  Oij: Observed frequency in ith row and jth column.
- Expected frequencies are given by Eij = (Ri * Cj)/ n where n is total sample size.
  Expected frequencies are computed under null hypothesis.
- Test statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where Oij are the observed frequencies in the ith row and jth column.
Eij are the expected frequencies in the ith row and jth column.

- χ^2 follows a Chi-Square Distribution with **(r-1)(c-1)** degrees of freedom.

Case Study - 2

To execute Non-Parametric test in Python, we shall consider the below case as an example.

Background

Data consist of information regarding the Performance & Recruitment Source of employees.

Objective

To check whether Performance & Source of Recruitment are associated.

Sample Size

Sample size: 870
Variables: sn, performance, source

Data Snapshot

chi square test of association

Variables

sn	performan	source
1	Excellent	Internal
2	Excellent	Internal
3	Excellent	Internal
4	Excellent	Internal
101	Excellent	Campus
102	Excellent	Campus
251	Excellent	Jobportal
252	Excellent	Jobportal
253	Excellent	Jobportal
254	Excellent	Jobportal

vations

Columns	Description	Type	Measurement	Possible values
sn	Serial number	Numeric	-	-
performance	Employee performance	Character	Excellent, Good,Poor	3
source	Source of recruitment	Character	Campus, Internal, Jobportal	3

- Get the observed frequency (count) table from this data.

Chi-square test of Association

Testing association between two categorical variables

Objective	To test the **null hypothesis** that two categorical variables are **independent**

Null Hypothesis (H_0): performance and source are not associated

Alternate Hypothesis (H_1): performance and source are associated

Test Statistic	$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$ Oij = observed frequencies in the ith row and jth column. Eij = expected frequencies in the ith row and jth column.
Decision Criteria	Reject the null hypothesis if p-value < 0.05

Chi-square test example

- Observed Frequency table

	Recruitment Source			
Performance	Campus	Internal	Jobportal	Total
Excellent	150	100	40	290
Good	100	100	100	300
Poor	80	50	150	280
Total	330	250	290	870

- Expected Frequency table

	Recruitment Source			
Performance	Campus	Internal	Jobportal	Total
Excellent	=(330*290)/870	83	97	290
Good	114	=(250*300)/870	100	300
Poor	106	80	=(290*280)/870	280
Total	330	250	290	870

	Value
r	3
c	3
χ^2	107.3786

Chi-Square test in Python

```
# Import the CSV file
data = pd.read_csv('chi square test of association.csv')

# create cross table of 2 categorical
cont_table = pd.crosstab(data.performance, data.source)

# Chi-square test of association
from scipy.stats import chi2_contingency
chi2_contingency(cont_table)
```

- **chi2_contingency from scipy.stats** function performs Chi-square test of association.
- It returns chi2(test statistic), p valve, dof, expected frequencies.

Chi-Square test in Python

```
# Output:
```

```
(107.37856396477088,
 2.6359873347121296e-22,
 4,
 array([[110.        ,  83.33333333,  96.66666667],
        [113.79310345,  86.20689655, 100.        ],
        [106.20689655,  80.45977011,  93.33333333]]))
```

Interpretation :

- Since p-value is <0.05, reject H0. Recruitment source and employee performance are associated.

Quick Recap

In this session, we continued learning non parametric tests . Here is a quick recap :

Kruskal Wallis test	• Nonparametric alternative to one way ANOVA.
Chi-Square test	• Also called as Pearson's chi-square test or the chi-square test of association, is used to test if there is a relationship between two categorical variables (nominal or ordinal).