Basics of Mechanical Engineering-1

Prof. J. Ramkumar

Dr. Amandeep Singh

Department of Mechanical Engineering

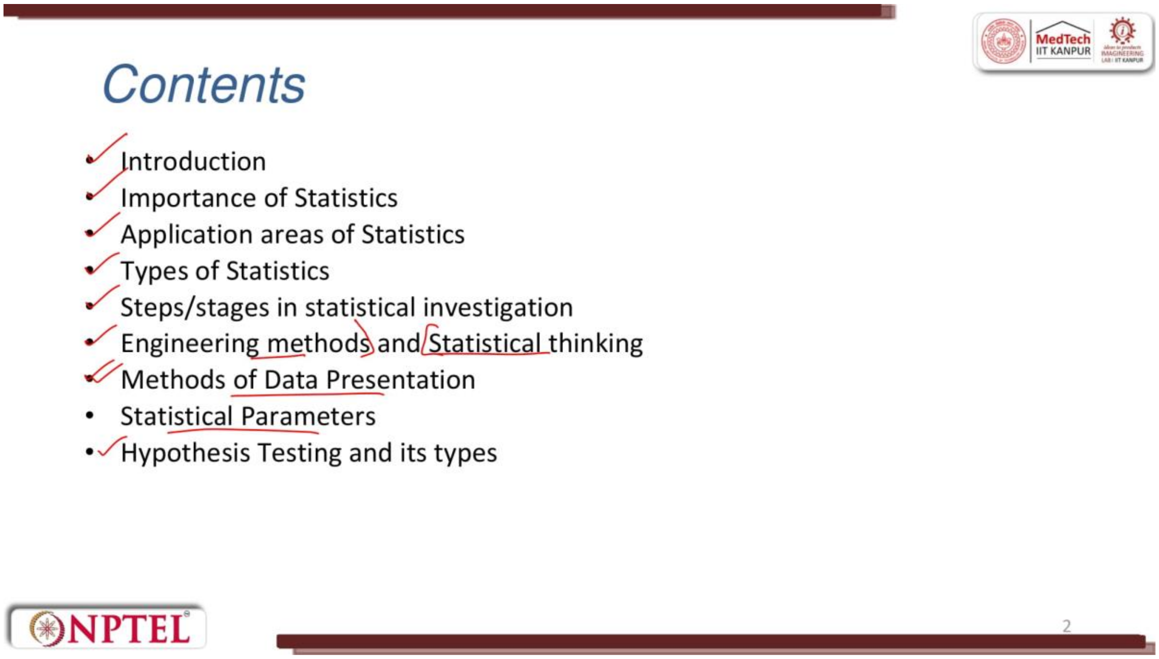
Indian Institute of Technology, Kanpur

Week 12

Lecture 50

Introduction to Engineering Statistics

Welcome to the last topic of this course, Basics of Mechanical Engineering. In this lecture, I will discuss Introduction to Engineering Statistics.



Contents

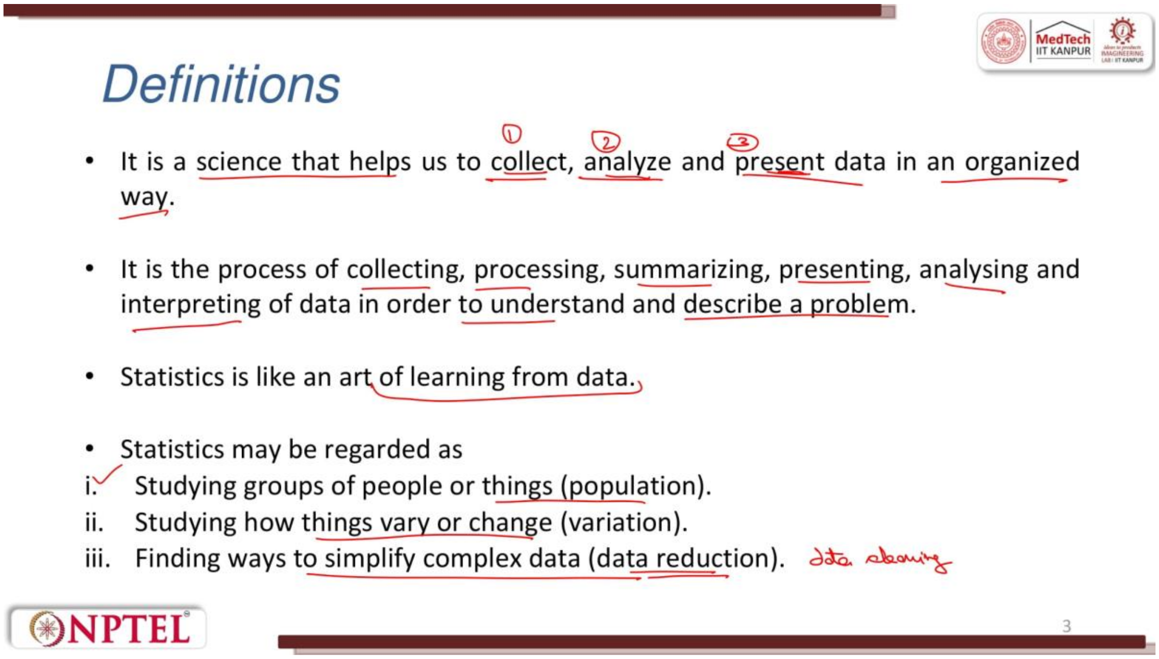
- Introduction
- Importance of Statistics
- Application areas of Statistics
- Types of Statistics
- Steps/stages in statistical investigation
- Engineering methods and Statistical thinking
- Methods of Data Presentation
- Statistical Parameters
- Hypothesis Testing and its types

The flow of the lecture would go in this way. I will introduce what is statistics, what is the Importance of Statistics we will talk about. Then we will see Application areas of Statistics like it can have applications in engineering, business, sociology, multiple topics.

Types of Statistics we will talk about whether earlier we are describing something or we are trying to bring some output of it. Steps and stages in Statistical Investigation we will discuss. Engineering Methods and Statistical Thinking. How does an engineer thinks when the statistics is a tool in his hand. Methods of Data Presentation we will see, statistical parameters we will see and also we will put some light on the Hypothesis Testing and its types.

Data is available everywhere, data could be structured or unstructured. Whatever you are getting readings while taking your tests, whatever you are taking readings while you are producing something you are sending it to, if you keep on noting that, that becomes your data. Weekly production, daily throughput, monthly production, that all is data. If you keep on noting for this material testing, this is hardness number for test 1, test 2, test 3, that all is data. Data has to be put in an organized form.

When you put it in organized form, it has to be analyzed, so that you calculate the required output. For instance, we use certain relationships to calculate force, we use certain relationships to calculate the hardness number or so. So, that is all we are working with data and without statistics we cannot work with it.



Definitions

- It is a science that helps us to ① collect, ② analyze and ③ present data in an organized way.
- It is the process of collecting, processing, summarizing, presenting, analysing and interpreting of data in order to understand and describe a problem.
- Statistics is like an art of learning from data.
- Statistics may be regarded as

i. ✓ Studying groups of people or things (population).
ii. Studying how things vary or change (variation).
iii. Finding ways to simplify complex data (data reduction). data cleaning

3

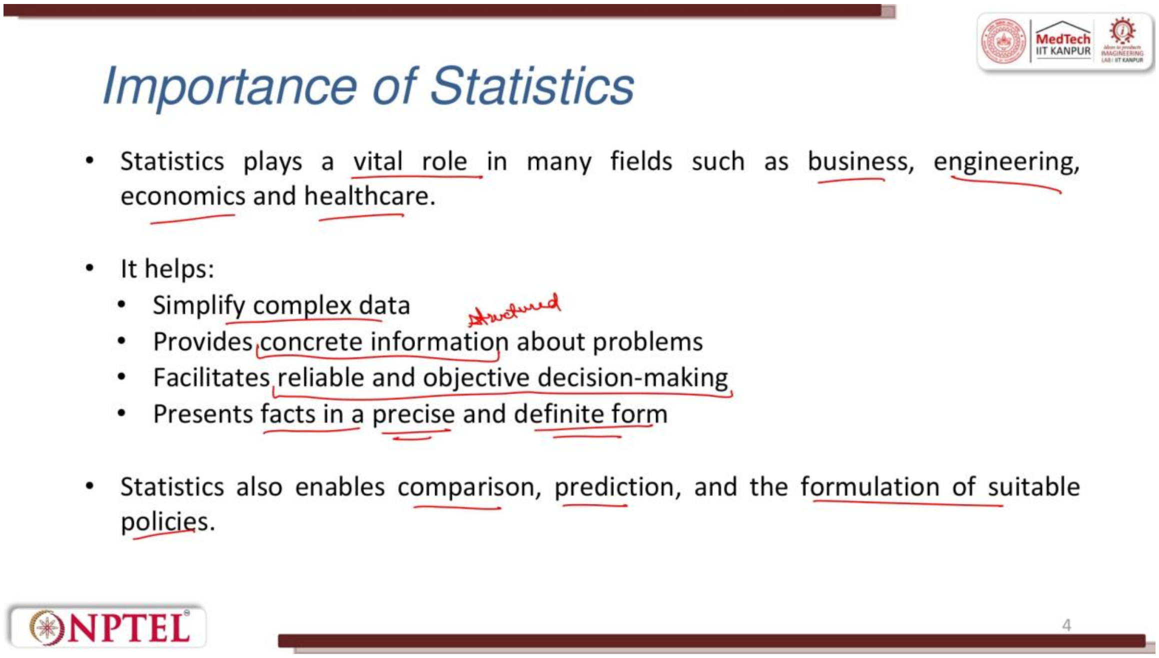
Definitions. Statistics as a definition is a science that helps us to collect, analyze and present data in an organized way.

Collect, analyze, present. Three parts we will cover. Collecting the data that is observation, experimentation or you collect the data. Secondary data, analyzing is whether we work with descriptive analytics or descriptive statistics or we work with inferential statistics that I will discuss. Present the data means how do you put the data in your reports.

In those reports, you can put that in a paragraph form. The hardness is increased after the heat treatment. It is increased by 2 units or it is increased by 5 percent. This you can give paragraphs or statements to us or you can put that in a tabular form or you can put it in a graphical form that is presentation of data. So, in an organized way we will see what kind of data where do you put.

It is a process of collecting, processing, summarizing, presenting, analyzing and interpreting of data in order to understand and describe a problem. Statistics is like an art of learning from data. Testetics may be regarded as studying groups of people or things that is population. Studying how things vary or change like what is contributing where. If we change the load, what is the time of the failure?

If we change the position of a specific work piece, so what is the variation in the output of the strength? And third point is finding ways to simplify complex data. That is data reduction. A lot of reading, lot of observation could be taken. Whatever are useful are taken, whatever are not useful are not taken at that point, but those are kept as a record to be taken separately. So, those could be taken in future if required. So, this is known as data cleaning.



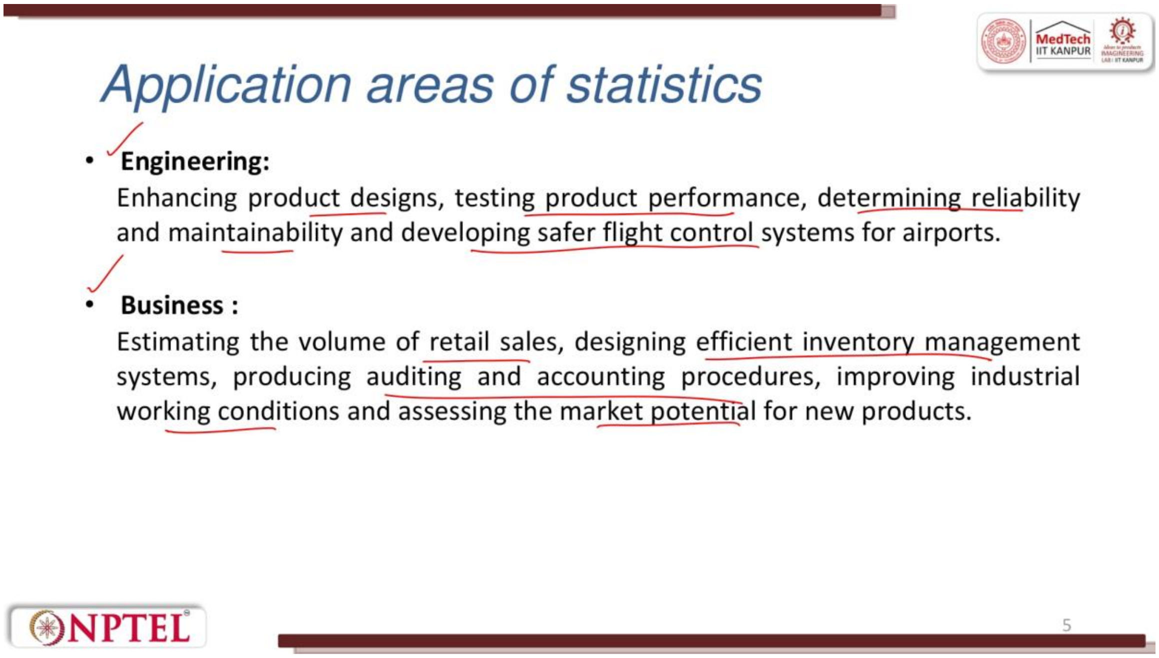
Importance of Statistics

- Statistics plays a vital role in many fields such as business, engineering, economics and healthcare.
- It helps:
  - Simplify complex data
  - Provides concrete Structured information about problems
  - Facilitates reliable and objective decision-making
  - Presents facts in a precise and definite form
- Statistics also enables comparison, prediction, and the formulation of suitable policies.

Statistics play a vital role in many fields such as business, engineering, economics and healthcare. It helps simplify complex data. It provides concrete information about the problems.

As I say, it can provide structured information. It facilitates reliable and objective decision making. There are certain analytics tools which helps to take the decision like decision tree is itself a technique that helps you to take decision based upon the data analysis. It presents facts in a precise and definite form. I will talk about the presentation of data in a graphical form.

One illustration or one graphics is equivalent to thousand words it is said. So, whatever you could see how the graph is flowing or the different entities in the graph different variables are behaving whatever you could present in thousand or more words could be presented in a single figure. So, that is it provides the facts and figures in a precise and definite form. Statistics also enables comparison, prediction and formulation of suitable policies based upon the data that you keep collecting over time.



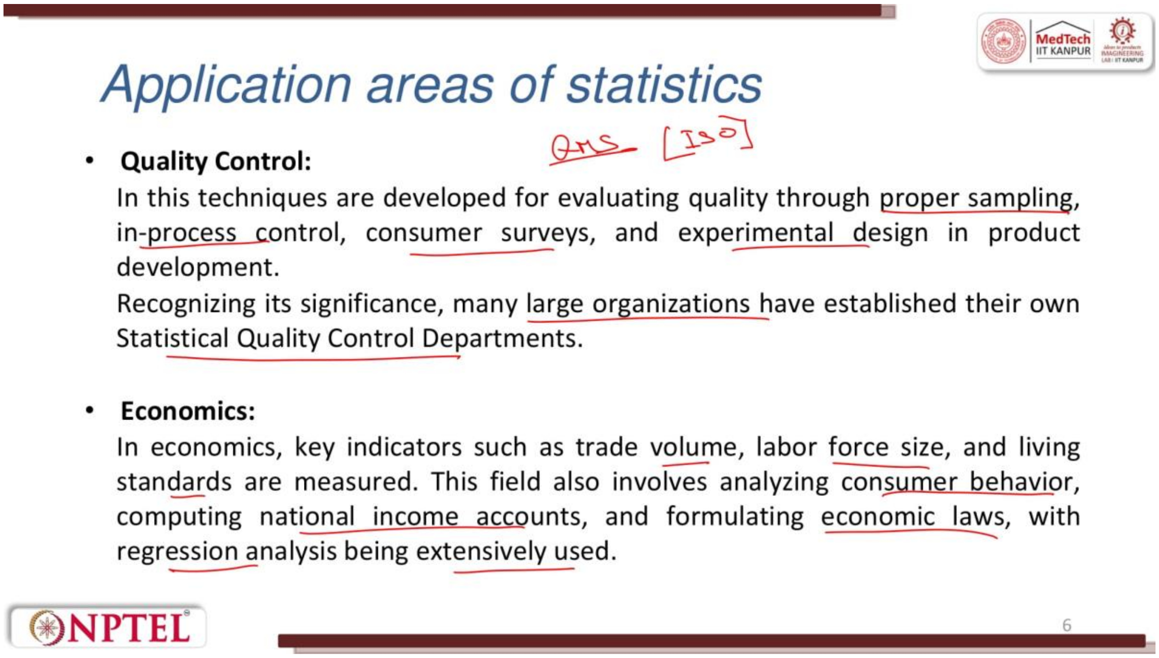
Application areas of statistics

- **Engineering:**
  Enhancing product designs, testing product performance, determining reliability and maintainability and developing safer flight control systems for airports.

- **Business :**
  Estimating the volume of retail sales, designing efficient inventory management systems, producing auditing and accounting procedures, improving industrial working conditions and assessing the market potential for new products.

Application areas of statistics, engineering, business, quality control, economics, health, medicine, biology, psychology wherever you call statistics is applied because data is there all the time.

Enhancing product designs, testing product performance, determining reliability and maintainability and developing safer flight control system, etcetera for airports in engineering could be applications. In business, estimating the volume of retail sales, designing efficient inventory management systems, producing auditing and accounting procedures, improving industrial working conditions and assessing the market potentials are a few applications which could be there in business.



Application areas of statistics

QMS [ISO]

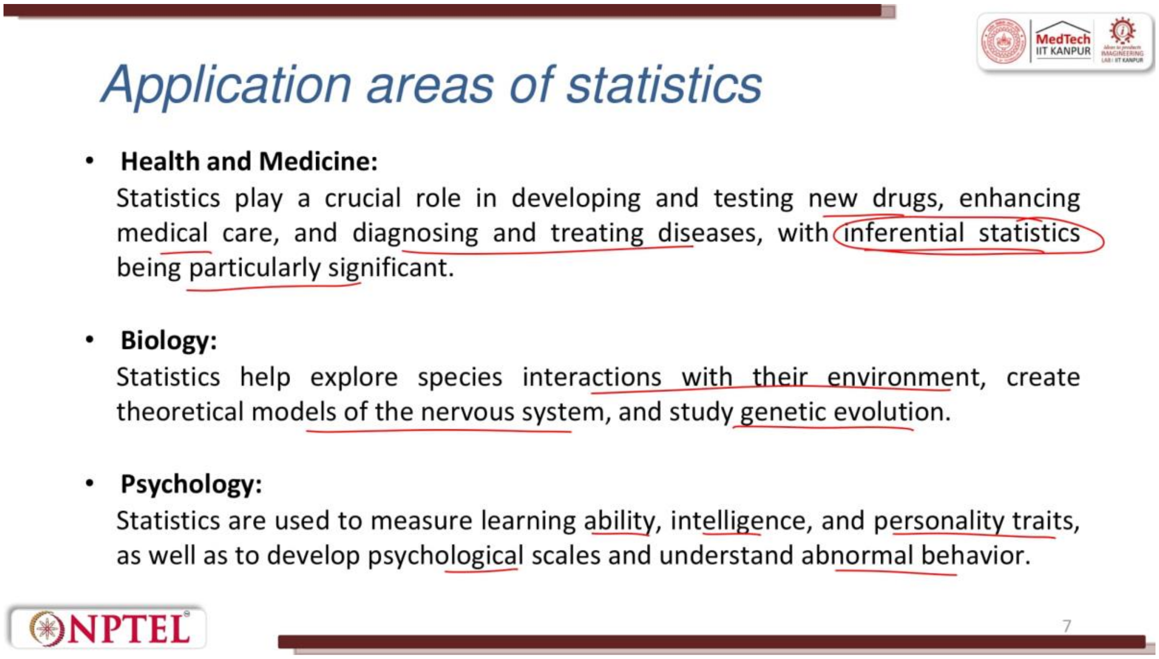
- **Quality Control:**
  In this techniques are developed for evaluating quality through proper sampling, in-process control, consumer surveys, and experimental design in product development.
  Recognizing its significance, many large organizations have established their own Statistical Quality Control Departments.

- **Economics:**
  In economics, key indicators such as trade volume, labor force size, and living standards are measured. This field also involves analyzing consumer behavior, computing national income accounts, and formulating economic laws, with regression analysis being extensively used.

In Quality Control, the techniques are developed for evaluating quality through proper sampling. We will see definition of what is sample and what is population, then in process control that is in situ inspection, etcetera.

Then consumer service, experimental design, quality control itself is a department nowadays a big departments are there. There are QMS (Quality Management System) the ISO system that you call which are based upon the data recording whatever you do you record it and also data analysis.

So, there are many large organizations have established with own big statistical quality control departments. In Economics, key indicators such as volume, labor force size, living standards are measured. This field also involves analyzing consumer behavior, computing national income accounts, formulating economic laws with regression analysis and multiple other analytics techniques which are extensively used.



Application areas of statistics

- **Health and Medicine:**
  Statistics play a crucial role in developing and testing new drugs, enhancing medical care, and diagnosing and treating diseases, with inferential statistics being particularly significant.

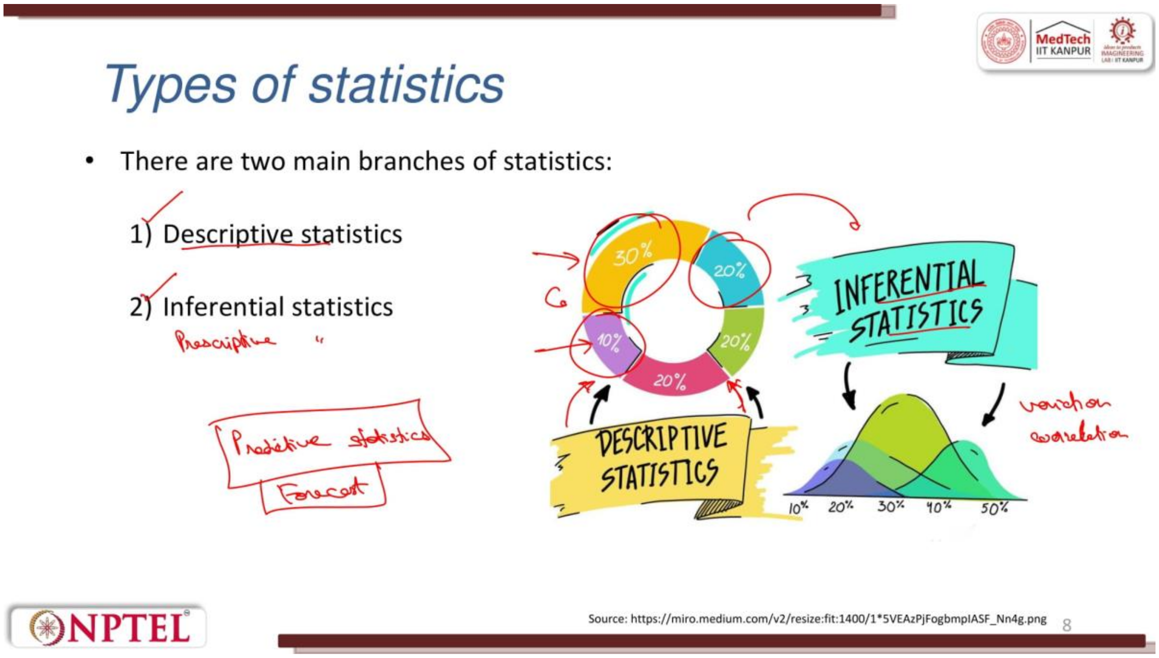
- **Biology:**
  Statistics help explore species interactions with their environment, create theoretical models of the nervous system, and study genetic evolution.

- **Psychology:**
  Statistics are used to measure learning ability, intelligence, and personality traits, as well as to develop psychological scales and understand abnormal behavior.

7

In Health and Medicine, statistics play a crucial role in developing and testing new drugs, enhancing medical care and diagnosing and treating diseases with influential statistics being particularly significant. Influential statistics is a key word here that we will talk about. Biology statistics help explore species interactions with their environment. Create critical models of nervous system. Study genetic evolution.

In Psychology, statistics are used to measure learning ability, intelligence and personality traits as well as to develop psychological skills and understand abnormal behavior. But there could be n number of applications.



Types of statistics

- There are two main branches of statistics:

1) Descriptive statistics

2) Inferential statistics

Prescriptive "

Predictive statistics

Forecast

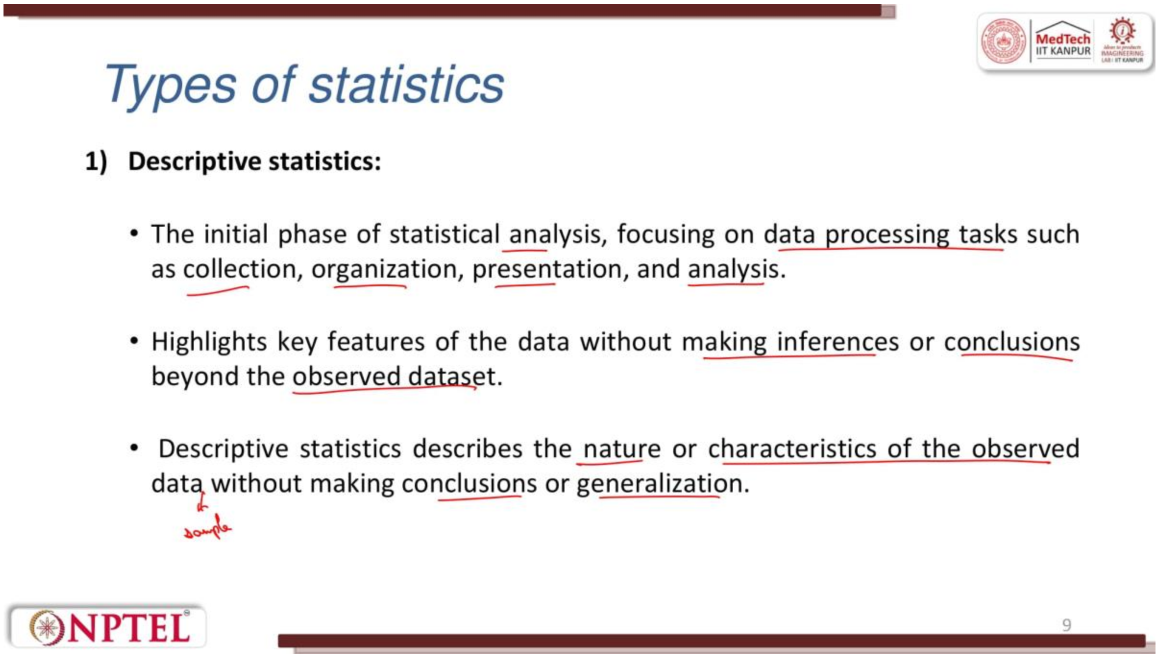
Source: https://miro.medium.com/v2/resize:fit:1400/1*5VEAzPjFogbmpIASF_Nn4g.png

What are the types of statistics? Statistics is majorly of two kinds descriptive statistics or inferential statistics. Inferential is also known as prescriptive.

Descriptive is where you only describe the data. You have a data that you keep on recording and you keep on put it on a paper or keep it recording on a computer. If you try to take mean, if you try to take median, if you try to calculate something else out of it and try to describe the data, yes 10 percent of the population is in this sector, 30 percent is in the other sector, 20 percent in this sector, this is how you are describing, the data only you are presenting the data, yes data has this kind of the structure or this kind of the inferential statistics more focuses upon the behavior of the data that what is variation in the data, what is the correlation with the other factors.

Just putting 10 percent of one section, 30 percent of other section that is one part. Correlation between them if we put that becomes the influential statistics. Also there is statistics field that is known as predictive statistics that we are not going to cover.

Predictive means when you try to forecast something based upon the past data, if you try to forecast this has happened in past, similar things could happen in future based upon the past data, past results that you have taken in your bending movement tests at certain intervals or in the certain season, similar would be behavior in the future in the same kind of the environment.



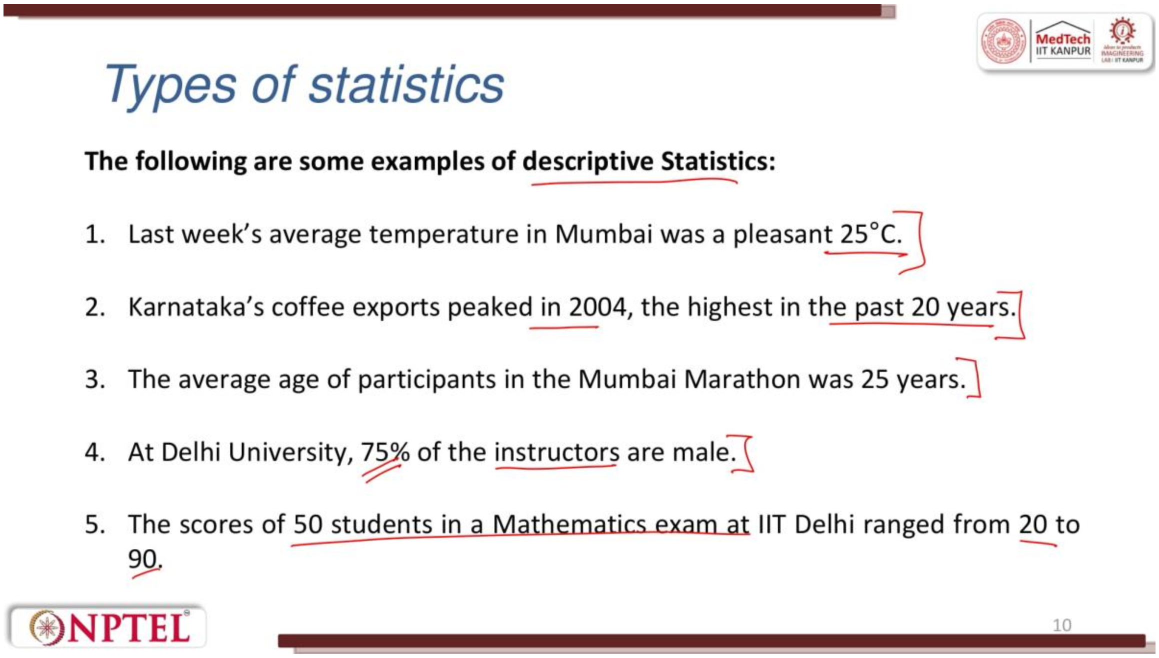
Types of statistics

1) **Descriptive statistics:**

- The initial phase of statistical analysis, focusing on data processing tasks such as collection, organization, presentation, and analysis.
- Highlights key features of the data without making inferences or conclusions beyond the observed dataset.
- Descriptive statistics describes the nature or characteristics of the observed data without making conclusions or generalization.

sample

Descriptive statistics, the initial phase of statistical analysis focuses on data processing tasks such as collection, organization, presentation and some trivial analysis. This highlights key features of data without making inferences, without making any suggestions or conclusions. Beyond the observed data set, it only highlights key features that this data has this kind of a structure. Descriptive statistics describes the nature or characteristics of the observed data without making conclusions or generalizations. So, this data is generally a sample.



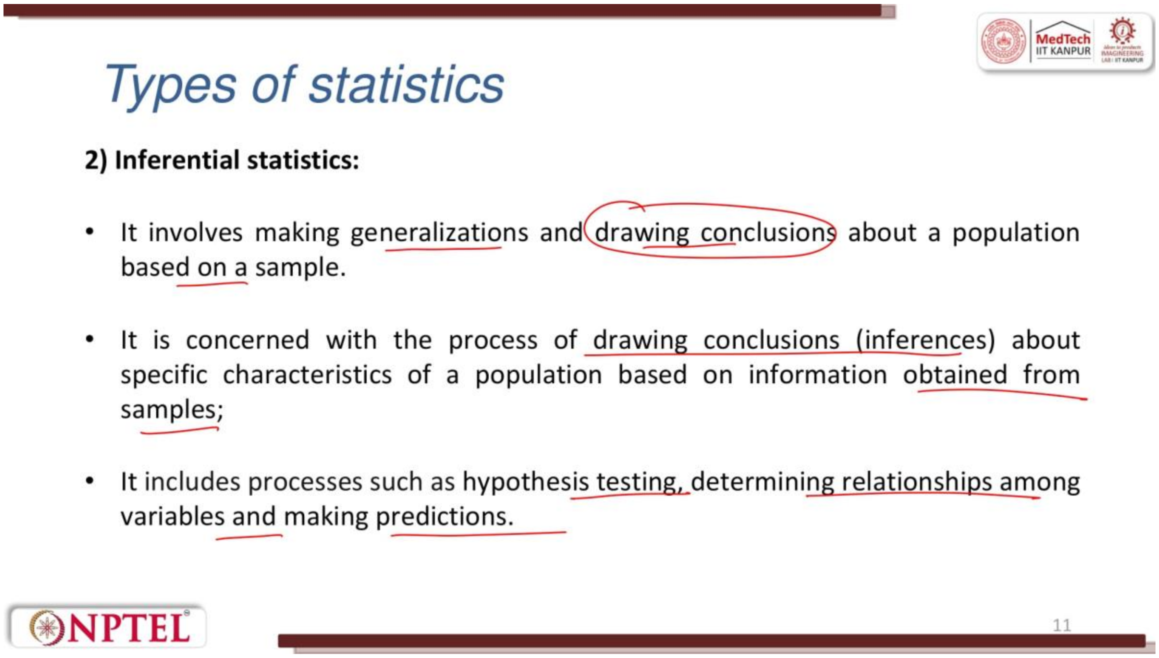
Types of statistics

The following are some examples of descriptive Statistics:

1. Last week’s average temperature in Mumbai was a pleasant 25°C.
2. Karnataka’s coffee exports peaked in 2004, the highest in the past 20 years.
3. The average age of participants in the Mumbai Marathon was 25 years.
4. At Delhi University, 75% of the instructors are male.
5. The scores of 50 students in a Mathematics exam at IIT Delhi ranged from 20 to 90.

These statements are some examples of descriptive statistics. Last week's average temperature in Mumbai was a pleasant 25 degree centigrade. Karnataka's coffee exports peaked in 2024, the highest in the past 20 years. So, this is presentation of data in a text form. The average age of participants in Mumbai marathon was 25 years.

At Delhi University, 75 percent of the instructors are male. So, this data is collected and we are only trying to present it that 75 percent of the instructors are male. So, this is describing the data. The scores of 50 students in mathematics exam at IIT Delhi range from 20 to 90. These are all statements which are presenting my descriptive statistics.



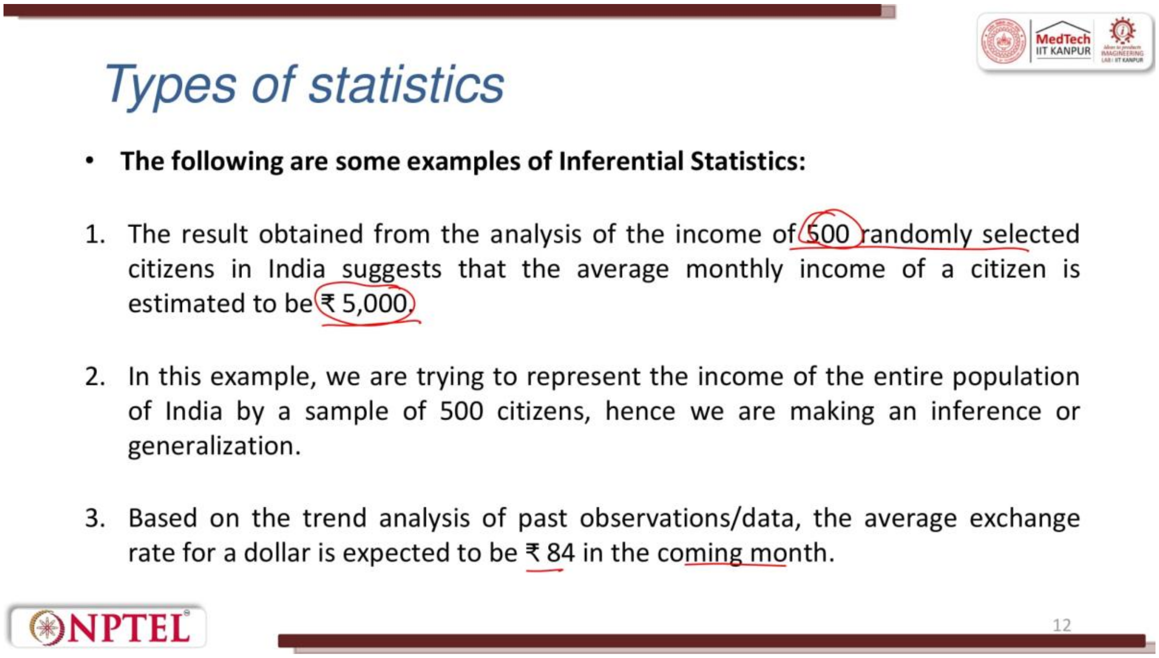
Types of statistics

2) Inferential statistics:

- It involves making generalizations and drawing conclusions about a population based on a sample.
- It is concerned with the process of drawing conclusions (inferences) about specific characteristics of a population based on information obtained from samples;
- It includes processes such as hypothesis testing, determining relationships among variables and making predictions.

Now, let us now talk about the Inferential Statistics. It involves making generalizations and drawing conclusions about a population based upon a sample. It is drawing conclusions. That means a lot more calculations would be now taken. It is concerned with process of drawing conclusions or inferences about specific characteristics of a population based on information obtained from samples.

It includes processes such as hypothesis testing, determining relationship among variables and making predictions. It aims to provide reasonable estimates of unknown population parameters through statistical analysis based upon the sample that we have selected.

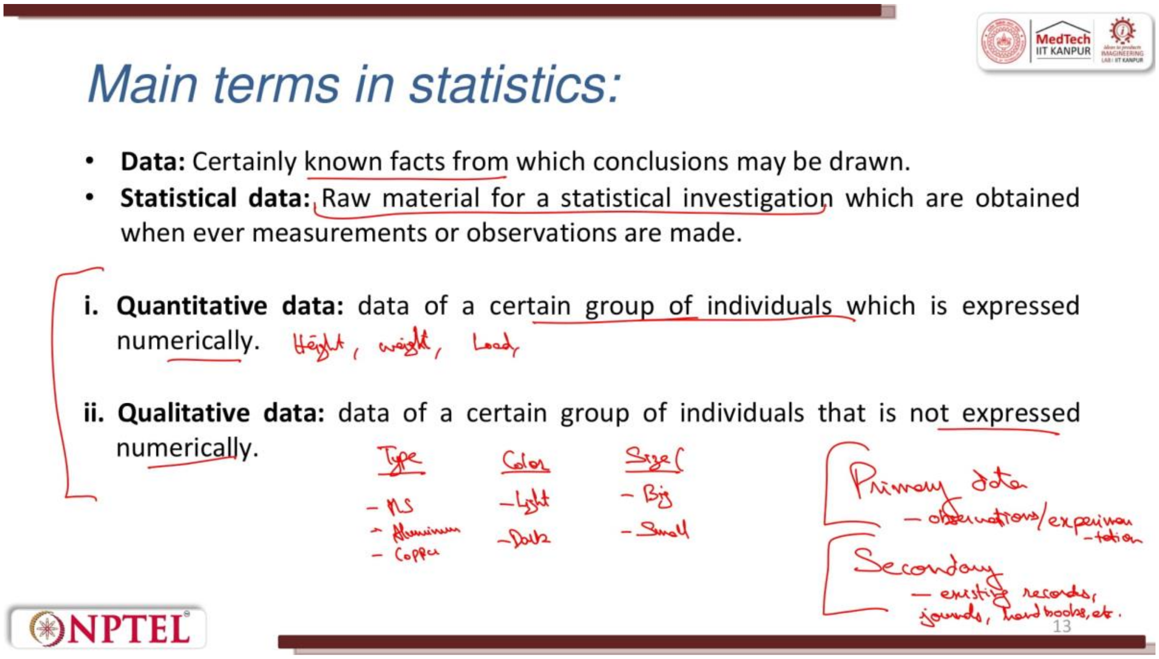


Types of statistics

- **The following are some examples of Inferential Statistics:**

1. The result obtained from the analysis of the income of 500 randomly selected citizens in India suggests that the average monthly income of a citizen is estimated to be ₹ 5,000.
2. In this example, we are trying to represent the income of the entire population of India by a sample of 500 citizens, hence we are making an inference or generalization.
3. Based on the trend analysis of past observations/data, the average exchange rate for a dollar is expected to be ₹ 84 in the coming month.

Some example statements could be the result obtained from the analysis of income of 500 randomly selected citizens in India suggests that average monthly income of a citizen is estimated to be rupees 5000. So, 500 is my sample size and influence is average is 5000. In this example, we are trying to represent the income of the entire population of India by taking a sample of 500 citizens. Hence, we are making an inference or generalization based upon the trend analysis of the past observations or data, the average exchange rate of dollar is expected to be rupees 84 in the coming month. These are all inferential statistics statements.



Main terms in statistics:

- **Data:** Certainly known facts from which conclusions may be drawn.
- **Statistical data:** Raw material for a statistical investigation which are obtained when ever measurements or observations are made.

i. **Quantitative data:** data of a certain group of individuals which is expressed numerically. Height, weight, Load,

ii. **Qualitative data:** data of a certain group of individuals that is not expressed numerically.

Type	Color	Size
- MS	-Light	- Big
- Aluminium	-Dark	- Small
- Copper		

Primary data
— observations/experimen-tation

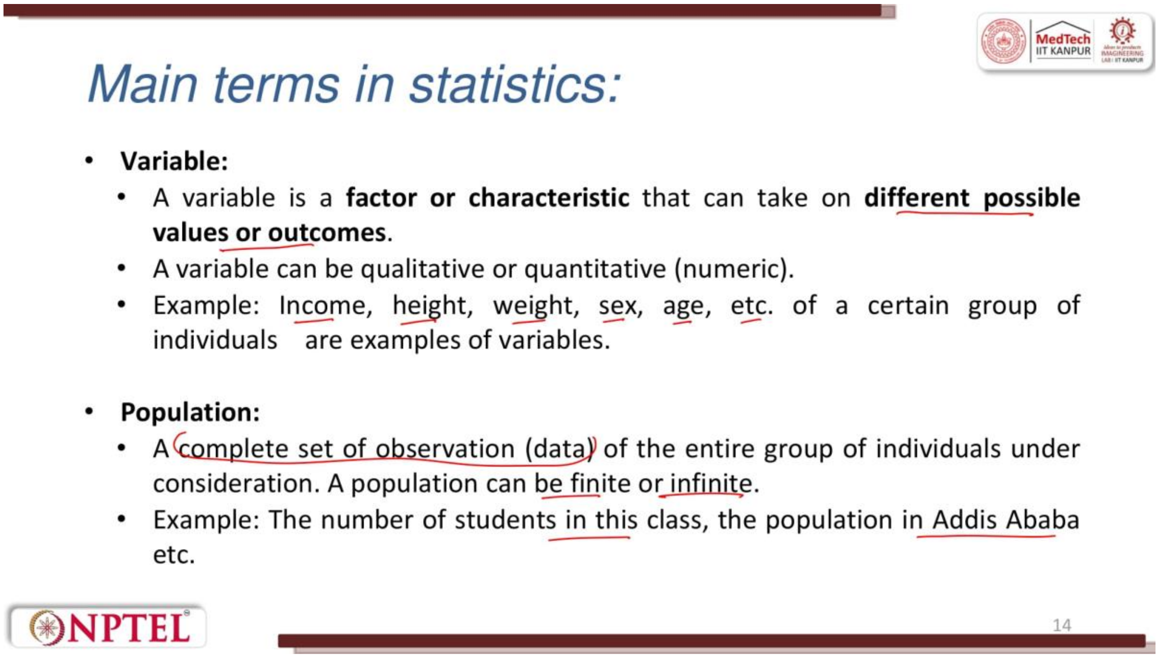
Secondary
— existing records, journals, hand books, etc.

Some terms in statistics, ‘Data’, certainly known facts from which conclusions may be drawn, statistical data that acts as a raw material for statistical investigation which are obtained whenever measurements or observations are made. Data could be quantitative, data could be qualitative.

Quantitative data refers to a certain group of individuals which is expressed numerically. Numerical data, height of the sample, weight of the sample, the hardness number, anything that you calculate using the formula that is variable. So, that is quantitative data. We will put here it is height, weight or I would say load, these are all quantitative data pointers. Qualitative data, data of a certain group of individuals that is not expressed in numerically that is known also a categorical data.

For example, if I am talking about the material type, material type could be it is mild steel, it is aluminum, it is copper. So, this is material type is my qualitative data or it could be colour or it could be the size. Size could also put in the quantitative data if size is specific. For example, if the size is 10 millimetre, 12 millimetre that becomes a quantitative data. If the size is only big or small that becomes my qualitative data.

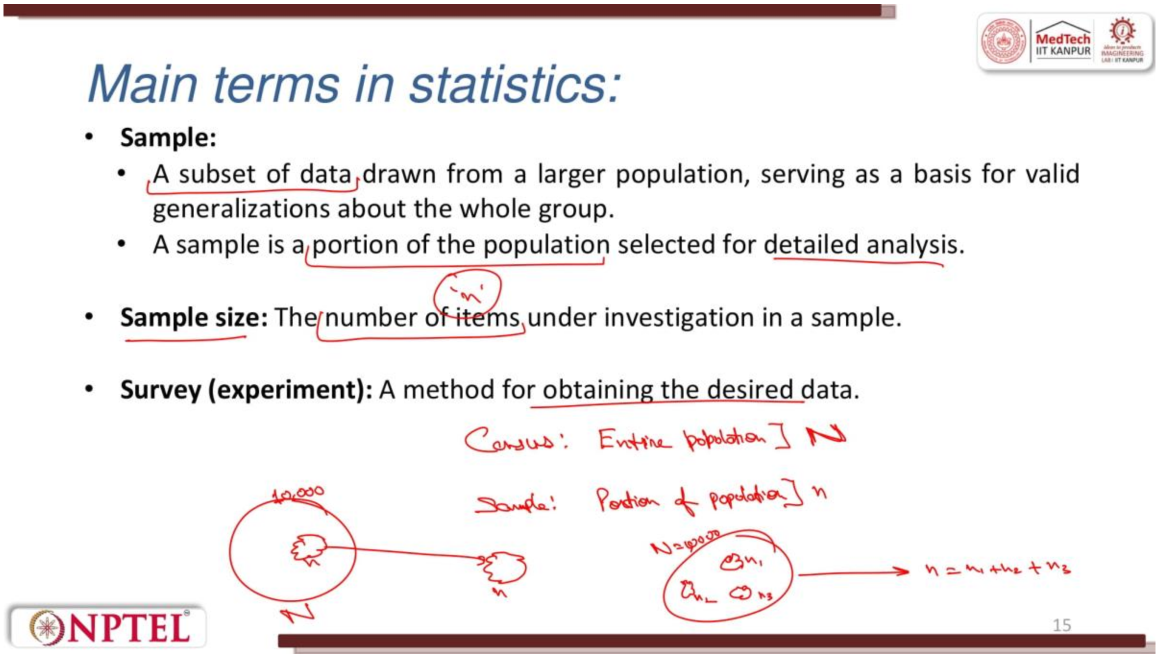
If the colour is light or dark that becomes my qualitative data. So, this data could even be collected or could be taken from the already existing records that is collected is known as primary data and that is taken from the existing records is known as secondary data. Data that is taken from observations or experimentation that is primary data. Secondary data is taken from existing records that could be your journals, handbooks, etcetera.



Main terms in statistics:

- **Variable:**
  - A variable is a **factor or characteristic** that can take on **different possible values or outcomes**.
  - A variable can be qualitative or quantitative (numeric).
  - Example: Income, height, weight, sex, age, etc. of a certain group of individuals are examples of variables.
- **Population:**
  - A complete set of observation (data) of the entire group of individuals under consideration. A population can be finite or infinite.
  - Example: The number of students in this class, the population in Addis Ababa etc.

Variable is a factor or characteristic that can take on different possible values or outcomes. Variable can be quantitative or qualitative. Example could be income, height, weight, sex, age, etcetera of certain group of individuals could be variables. Population, a complete set of observations of the entire group of individuals under consideration is population. A population can be finite or infinite. Example, the number of students in class, the population in a Addis Ababa, etc.,



Main terms in statistics:

- **Sample:**
  - A subset of data drawn from a larger population, serving as a basis for valid generalizations about the whole group.
  - A sample is a portion of the population selected for detailed analysis.
- **Sample size:** The number of items under investigation in a sample.
- **Survey (experiment):** A method for obtaining the desired data.

Census: Entire population] N

Sample: Portion of population] n

When we talk about population, we have to talk about Sample. Sample is a subset of data drawn from larger population serving as a basis for valid generalizations about the whole group. Sample is a portion of population selected for detailed analysis, selection of sample is also very important factor, sampling techniques are there, the way you select the sample, random sampling, stratified sampling, systematic sampling, multistage sampling.

So, that your sample that is selected that is only you are trying to test or trying to measure to present the behavior of the whole population. The sampling techniques are also important. When you select a sample, you also develop an hypothesis. So, whatever this sample behavior is, this will represent my population behavior as well. There could be error in that.

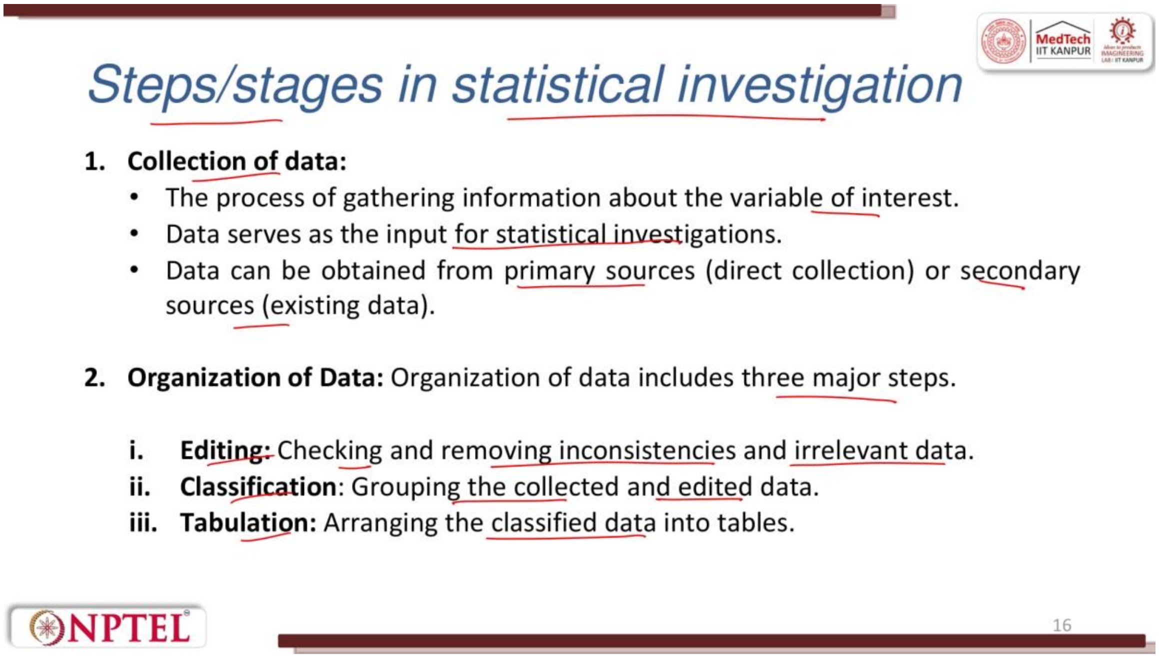
This we will discuss. We will try to come to the hypothesis testing slide. Sample size. The number of items under investigation in a sample. It is sample size.

It is generally written by small n. Survey or experiment. It is a method for obtaining the desired data. It could be census survey or it could be sample survey. As you know, each year census survey is taken in the country. It is obtaining, when I say census survey, it is taking into account the entire population.

Each and every piece is tested for its size, for its weight, for its color, that is census. Sample is when you select a specific set this only the sample size that is selected small n is only inspected. So, census survey is generally for capital N that is population size. Sample survey is only the portion of entire population that is taken that is small n where we conduct the hypothesis. So, we say generally, this is my large population.

This is a too big data set. For instance, there are maybe 10,000 number of pieces those are there. I just select a small sample from here. I bring this sample here. So, this is my capital N. This is small n. Small n is only tested.

Whatever is the behavior of this small sample, whether it is acceptable, not acceptable, whether it is oversized, undersized, whether it is to be heat treated, whether it is to be further processed or whether it is to be finally passed, this all represents the whole population. It depends upon the way we select the sample. It could be just selecting one sample. It could be selecting samples from certain places and then we test for example, this is n1, n2, n3 out of the total n that is 10,000. So, small n size that would be n1 + n2 + n3 is a kind of a stratified sampling.



Steps/stages in statistical investigation

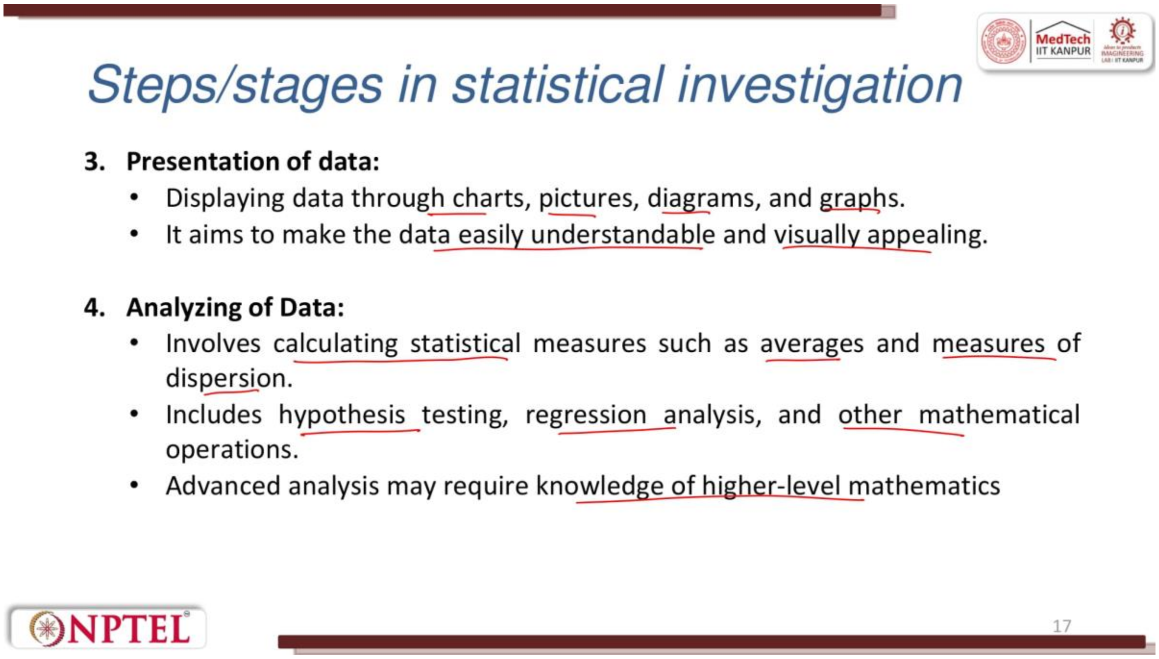
1. **Collection of data:**
   - The process of gathering information about the variable of interest.
   - Data serves as the input for statistical investigations.
   - Data can be obtained from primary sources (direct collection) or secondary sources (existing data).

2. **Organization of Data:** Organization of data includes three major steps.

   i. **Editing:** Checking and removing inconsistencies and irrelevant data.
   ii. **Classification**: Grouping the collected and edited data.
   iii. **Tabulation:** Arranging the classified data into tables.

Next is Steps or Stages in Statistical Investigation. Collection of data, the process of gathering information about the variable of interest. Data serve as input for statistical investigations. Data can be obtained from primary sources or from secondary sources that is existing data.

Organization of data includes three major steps. Editing, Classification, Tabulation. Editing is checking or removing inconsistencies that are called as data cleaning and irrelevant data is also taken off. Classification is grouping the collected data and the edited data. Tabulation is arranging the classified data into tables.



Steps/stages in statistical investigation

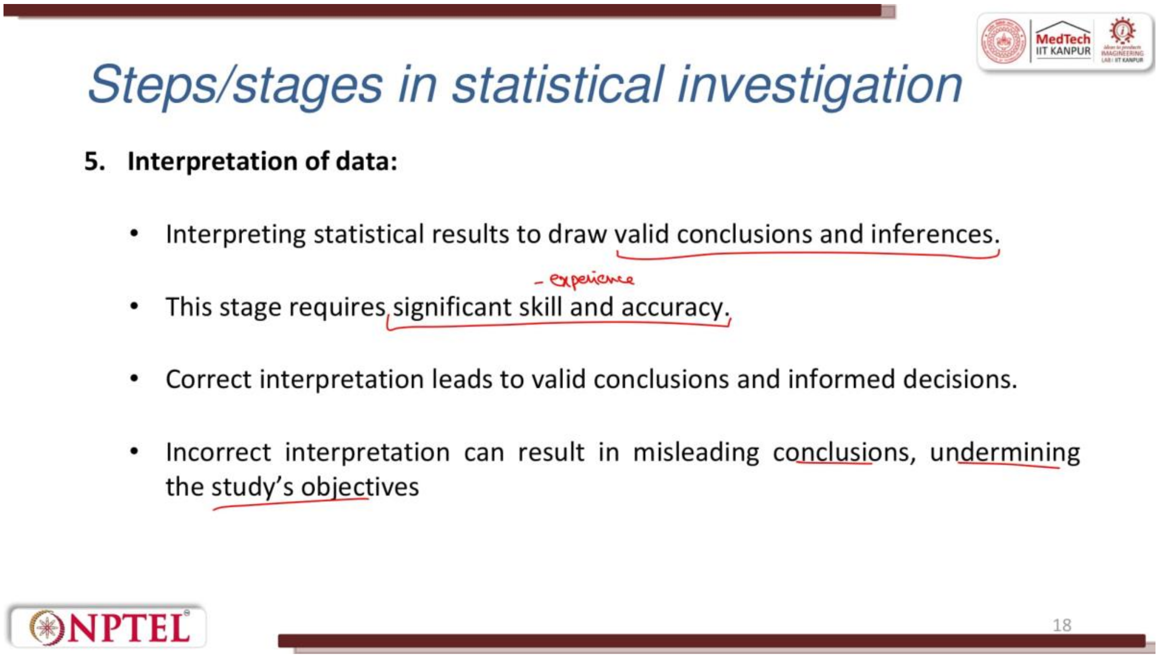
3. **Presentation of data:**
   - Displaying data through charts, pictures, diagrams, and graphs.
   - It aims to make the data easily understandable and visually appealing.

4. **Analyzing of Data:**
   - Involves calculating statistical measures such as averages and measures of dispersion.
   - Includes hypothesis testing, regression analysis, and other mathematical operations.
   - Advanced analysis may require knowledge of higher-level mathematics

17

Then next step comes that is Presentation of Data. Displaying the data through charts, pictures, diagrams, graphs, etc. It aims to make data easily understandable and visually appealing. Analysing of data is the next step in which we try to involve certain statistical calculations and certain measures such as averages or measures of dispersions are taken that in measures of central tendency or measure of dispersion those are taken.

This includes hypothesis testing, regression analysis and other mathematical operations. Advanced analysis may require knowledge of higher level mathematics.



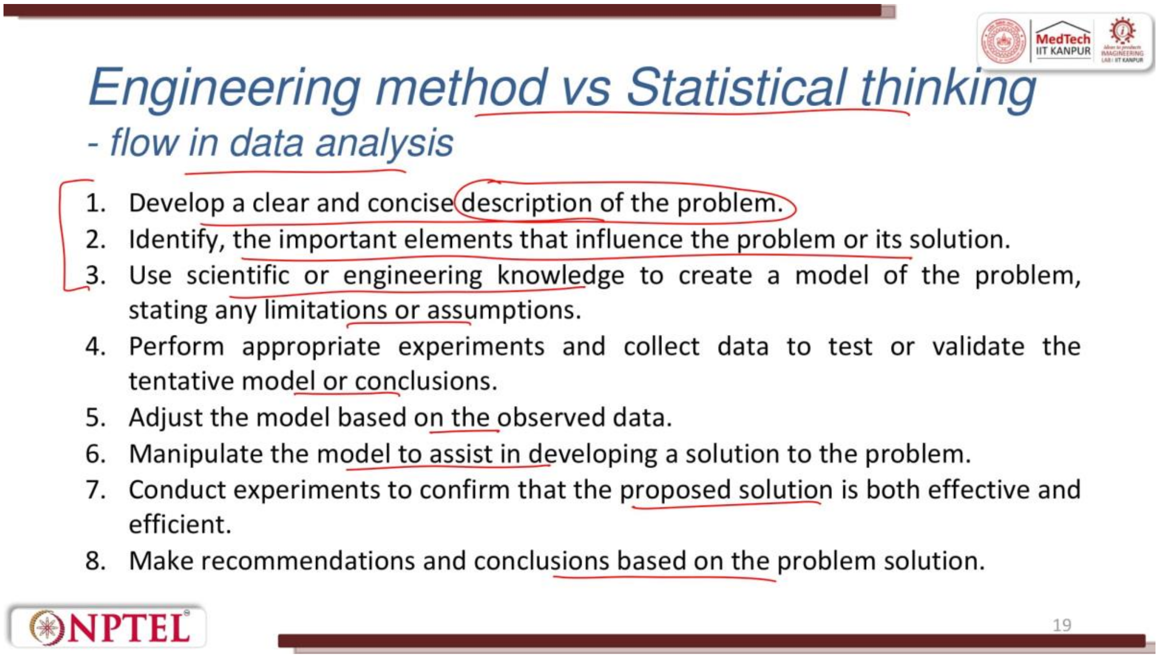
Steps/stages in statistical investigation

5. Interpretation of data:

- Interpreting statistical results to draw valid conclusions and inferences.
- This stage requires significant skill and accuracy. - experience
- Correct interpretation leads to valid conclusions and informed decisions.
- Incorrect interpretation can result in misleading conclusions, undermining the study's objectives

Interpretation of data. Interpreting statistical results to draw valid conclusions and inferences. This stage requires significant skill and accuracy. Skill means an experience has to be there. Correct interpretation leads to valid conclusions and informed decisions. Incorrect interpretation can result in misleading conclusions undermining the study's objectives. An engineer is someone who solves problems of interest to society by efficient application of scientific principles.

So, engineers accomplish this by either refining an existing product or process or by designing a new product or process that meets the customer's expectation. For all that, these five stages to finally come to an inference that what are you going to design, what is you going to present before the society as a larger picture, these steps are to be followed.



Engineering method vs Statistical thinking
- flow in data analysis

1. Develop a clear and concise description of the problem.
2. Identify, the important elements that influence the problem or its solution.
3. Use scientific or engineering knowledge to create a model of the problem, stating any limitations or assumptions.
4. Perform appropriate experiments and collect data to test or validate the tentative model or conclusions.
5. Adjust the model based on the observed data.
6. Manipulate the model to assist in developing a solution to the problem.
7. Conduct experiments to confirm that the proposed solution is both effective and efficient.
8. Make recommendations and conclusions based on the problem solution.

19

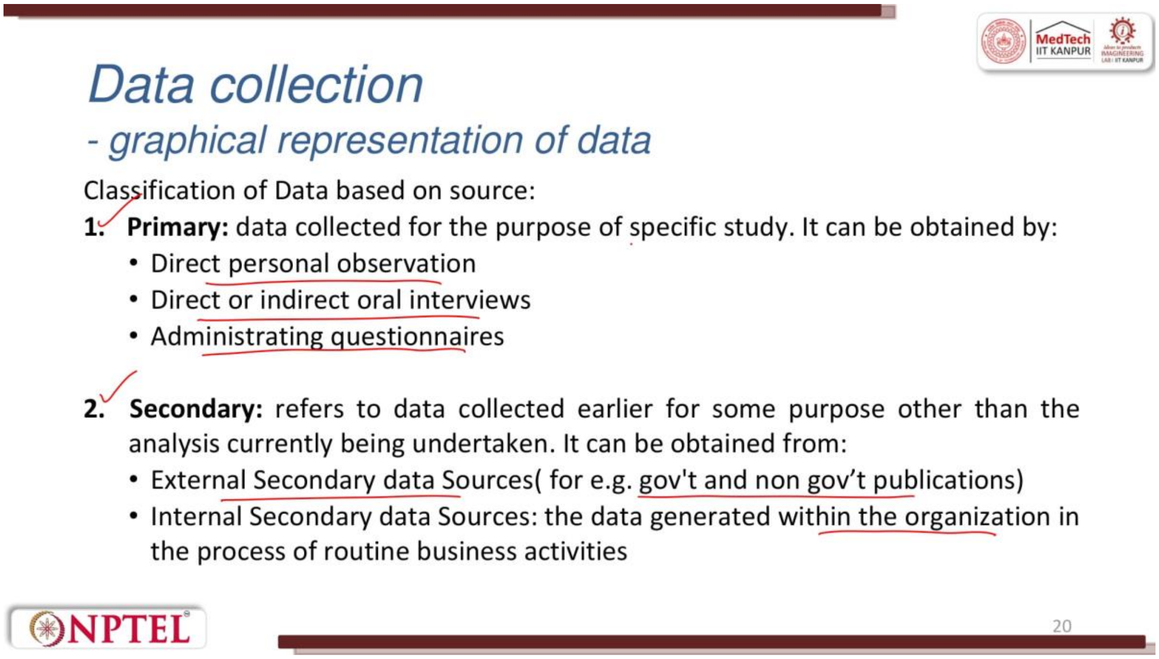
So, let us now talk about Engineering method versus Statistical thinking flow in data analysis. Develop a clear and concise description of the data problem. Identify the important elements that influence the problem or its solution.

Use scientific or engineering knowledge to create a model of the problem stating any limitations or assumptions. Perform appropriate experiments and collect data to test or validate the tentative model and conclusions. Adjust the model based on observed data. Manipulate the model to assist in developing a solution to the problem. Conduct experiments to confirm that the proposed solution is both effective and efficient.

Make recommendations and conclusions based upon the problem solution. These are the five steps that we have discussed in the previous slides. Those are more detailed when you try to talk about engineering problem. Only the first three steps you are talking about is clear and concise description of the problem statement itself. What are you going to do?

What is your problem statement? Are you only going to change the dimensions? Are you going to change the product completely? Are you going to bring the new product? Are you going to change or update the existing product in the terms of the shape only?

Are you going to change only packaging? It depends upon what is the problem statement. So, this is how the flow in data analysis comes, so that you come to a conclusion that is set as an objective in the beginning itself.



Data collection

- graphical representation of data

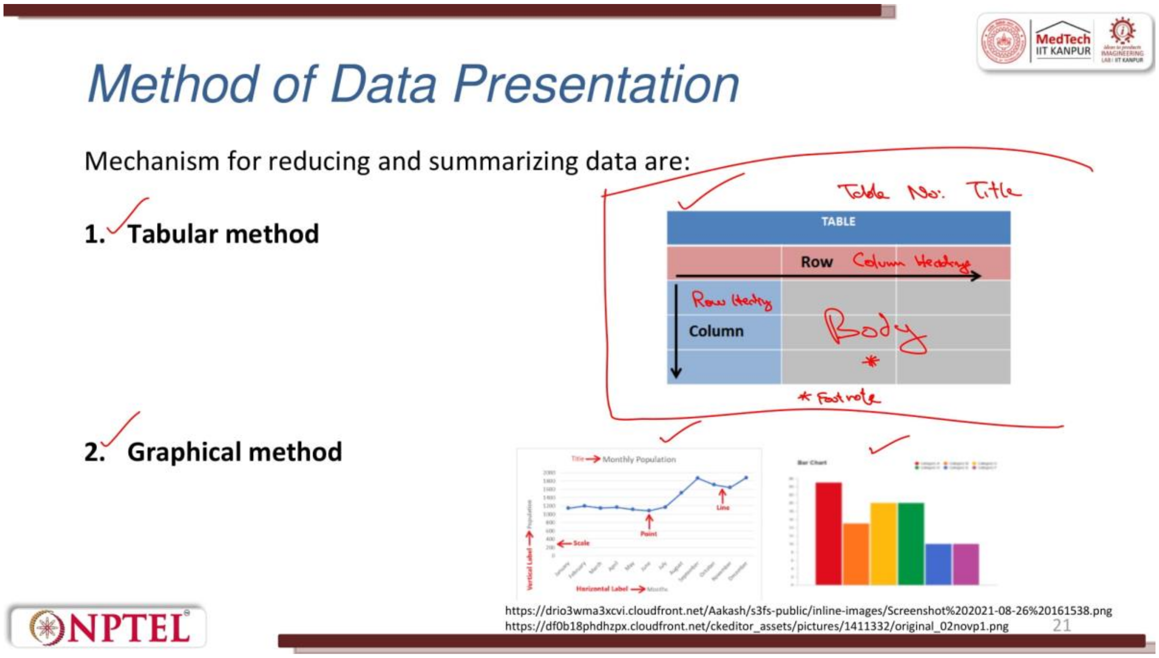
Classification of Data based on source:

1. **Primary:** data collected for the purpose of specific study. It can be obtained by:
   - Direct personal observation
   - Direct or indirect oral interviews
   - Administrating questionnaires

2. **Secondary:** refers to data collected earlier for some purpose other than the analysis currently being undertaken. It can be obtained from:
   - External Secondary data Sources(for e.g. gov't and non gov't publications)
   - Internal Secondary data Sources: the data generated within the organization in the process of routine business activities

Then comes Data Collection, primary and secondary I have already discussed data collected for purpose for specific study it can be obtained by direct personal observation, direct or indirect oral interviews, administrating questionnaires. Secondary data refers to data collected already for some purpose other than analysis that is currently being undertaken.

It can be obtained from external sources of data for example, governments or non-government Internal, secondary data sources, data generated within organization in the process of routine business activities, that could be secondary data.



Method of Data Presentation

Mechanism for reducing and summarizing data are:

1. ✓ **Tabular method**

Table No: Title

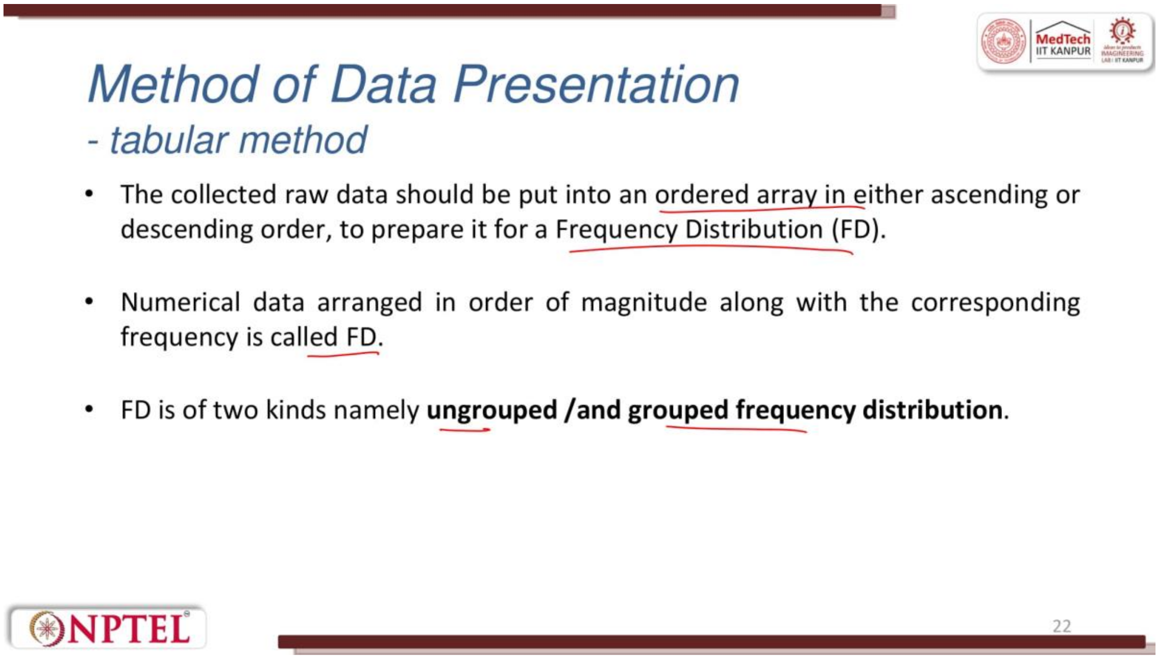
TABLE		
	Row Column Heading	
Row Heading		
Column	Body	
	*	

* Footnote

2. ✓ **Graphical method**

https://drio3wma3xcvi.cloudfront.net/Aakash/s3fs-public/inline-images/Screenshot%202021-08-26%20161538.png
https://df0b18phdhzpx.cloudfront.net/ckeditor_assets/pictures/1411332/original_02novp1.png

Data presentation is majorly in two forms, Tabular and Graphical. This is a table, these are graphs, this is a frequency polygon, this is a bar chart. Let me talk about the tabular data first.

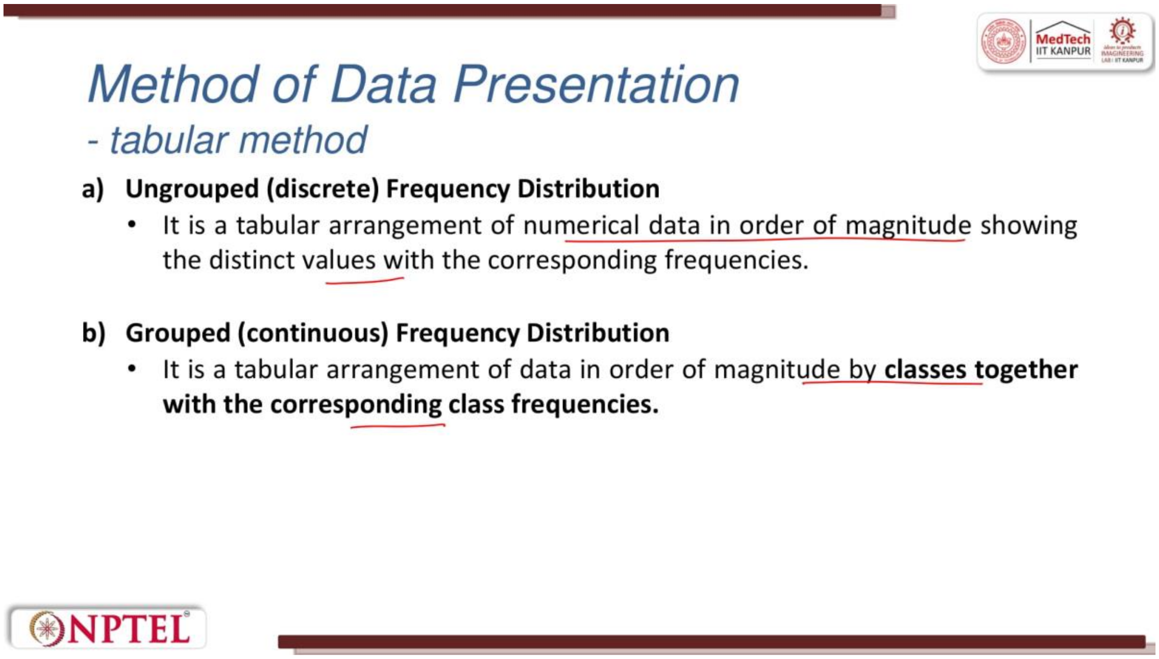


Method of Data Presentation
- tabular method

- The collected raw data should be put into an ordered array in either ascending or descending order, to prepare it for a Frequency Distribution (FD).
- Numerical data arranged in order of magnitude along with the corresponding frequency is called FD.
- FD is of two kinds namely **ungrouped /and grouped frequency distribution**.

The collected raw data should be put into an ordered array in either ascending or descending order to prepare for the frequency distribution. The numerical data range in order of magnitude along with the corresponding frequencies called frequency distribution of two kinds namely ungrouped and grouped frequency distribution.

If I try to talk about a table, in table there are certain parts are there, there is a table number. We have table title, you could have a caption for a table, you could have the row heading for each row, you could have column headings and then is a body of a table. Also, we can have some pointer for example, this is specific data point that is of more interest for you to present you can have a footnote here. So, this is table.



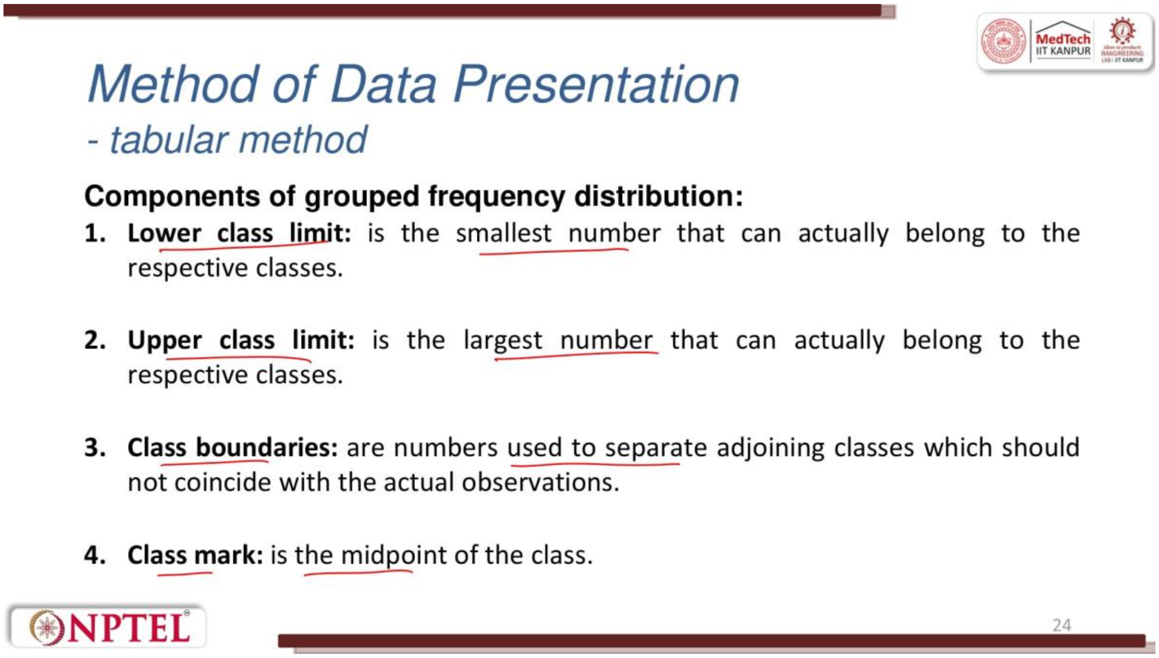
Method of Data Presentation

- tabular method

a) **Ungrouped (discrete) Frequency Distribution**
- It is a tabular arrangement of numerical data in order of magnitude showing the distinct values with the corresponding frequencies.

b) **Grouped (continuous) Frequency Distribution**
- It is a tabular arrangement of data in order of magnitude by **classes together with the corresponding class frequencies.**

Regarding the Grouped and Ungrouped Frequency distribution. Ungrouped or discrete frequency distribution is a tabular arrangement of numerical data in order of magnitude showing the distinct values with corresponding frequencies. Grouped frequency distribution is a tabular arrangement of data in order of magnitude of classes together that is it could be in an ascending order or in a descending order and it is a continuous data that is a group data.



Method of Data Presentation

- tabular method

Components of grouped frequency distribution:

1. **Lower class limit:** is the smallest number that can actually belong to the respective classes.

2. **Upper class limit:** is the largest number that can actually belong to the respective classes.

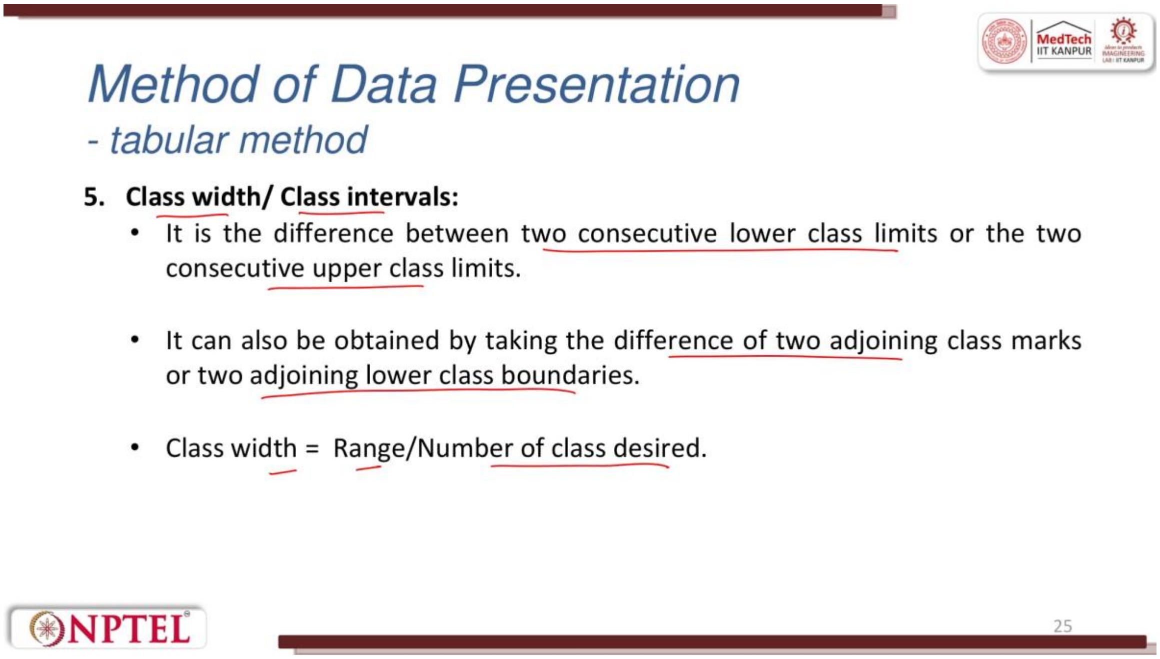
3. **Class boundaries:** are numbers used to separate adjoining classes which should not coincide with the actual observations.

4. **Class mark:** is the midpoint of the class.

24

In a table itself, we can present these Lower class limit, Upper class limit, Class boundaries, Class mark or Class width, etcetera.

Lower class limit is the smallest number that can actually belong to the respective classes. Upper class is the largest number that can actually belong to the respective classes. Class boundaries are numbers used to separate adjoining classes which should be not coincide with actual observation. Class mark is the midpoint of the class. When I am trying to talk about class, we can also draw out of it.

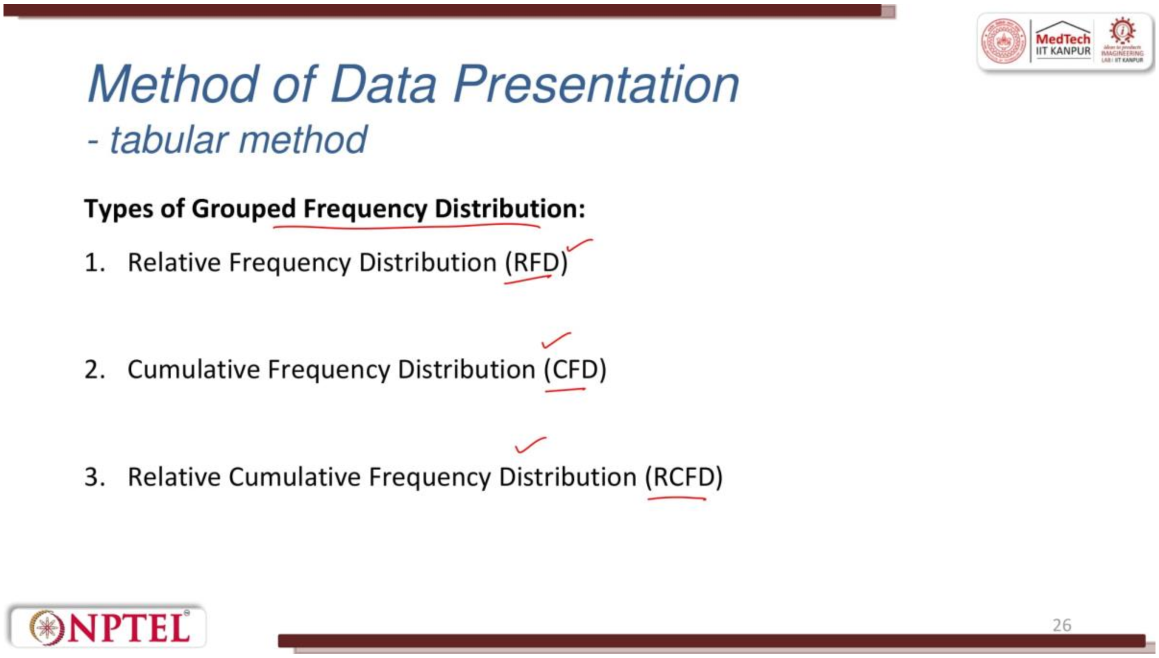


Method of Data Presentation

- tabular method

5. **Class width/ Class intervals:**
   - It is the difference between two consecutive lower class limits or the two consecutive upper class limits.
   - It can also be obtained by taking the difference of two adjoining class marks or two adjoining lower class boundaries.
   - Class width = Range/Number of class desired.

Class width or Class intervals it is a difference between two conjunctive lower class limits and two conjunctive upper class limits. It can also be obtained by taking the difference of two adjoining class marks or two adjoining lower class boundaries. Class width is range by number of class desired. So, when we I will just try to draw a histogram and that would be more clear.



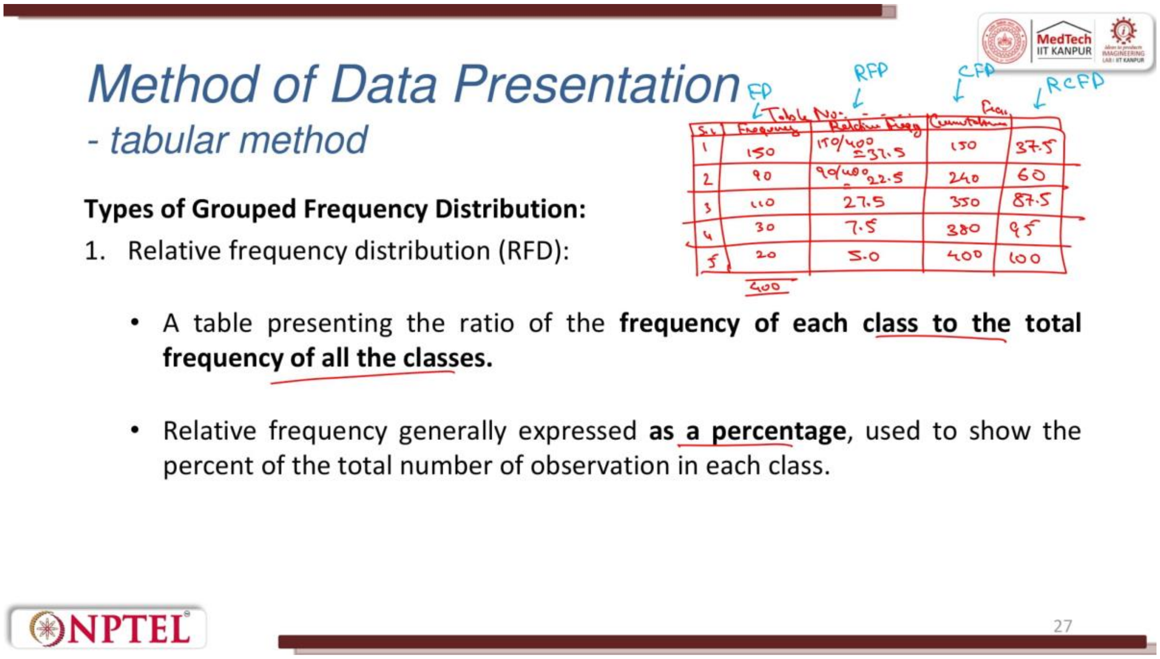
Method of Data Presentation

- tabular method

Types of Grouped Frequency Distribution:

1. Relative Frequency Distribution (RFD) ✓
2. Cumulative Frequency Distribution (CFD) ✓
3. Relative Cumulative Frequency Distribution (RCFD) ✓

It is important to understand the Types of Grouped Frequency distribution. Relative Frequency distribution, cumulative frequency distribution, relative cumulative frequency distribution.



Method of Data Presentation

- tabular method

FD RFD CFD RCFD

Table No. -

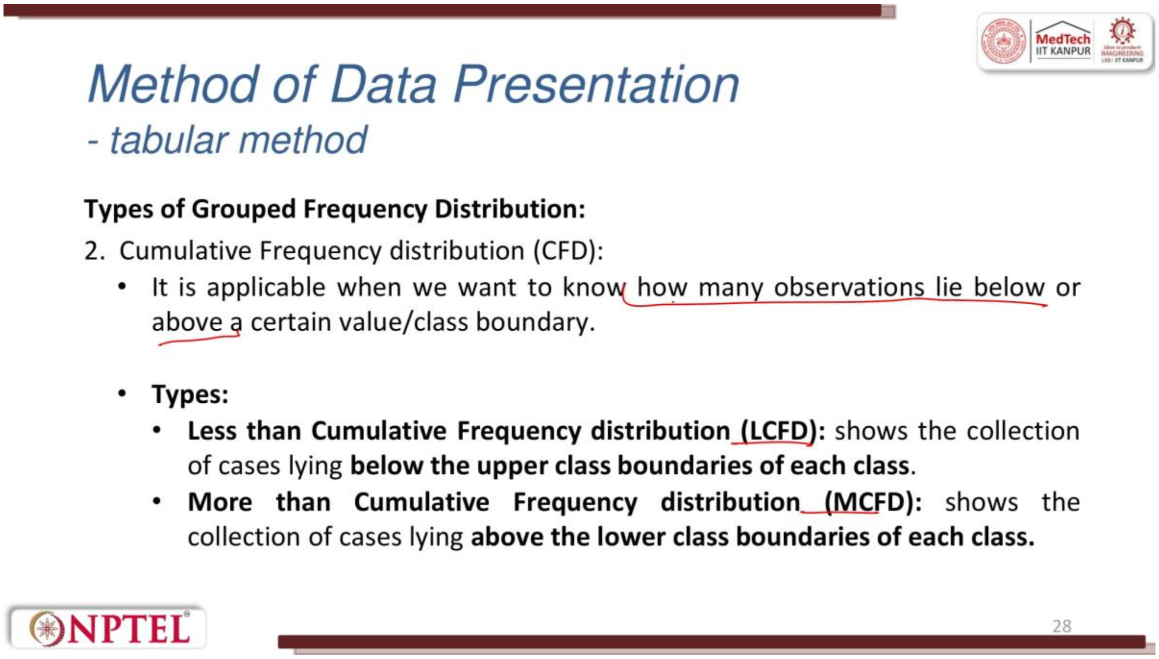
S.l	Frequency	Relative Freq	Cumulative Freq.	
1	150	150/400 = 37.5	150	37.5
2	90	90/400 = 22.5	240	60
3	110	27.5	350	87.5
4	30	7.5	380	95
5	20	5.0	400	100
	400			

Types of Grouped Frequency Distribution:

1. Relative frequency distribution (RFD):
   - A table presenting the ratio of the **frequency of each class to the total frequency of all the classes.**
   - Relative frequency generally expressed **as a percentage**, used to show the percent of the total number of observation in each class.

Relative frequency distribution is table representing the frequency of each class to the total frequency of all class. It could be presented as a percentage. For instance, if it is a table where specific number of sets are presented 1, 2, 3, 4, 5, So, this is table number and title.

So, if I say frequency this is frequency this is serial number frequency of occurring of first observation you know in the first observation where suppose if I have taken some readings this number 1 is coming. 150 times number 2 is coming 90 times this number 3 is coming 110 times this is coming 30 times this is coming 20 times. So, this becomes my frequency distribution to have a relative frequency distribution I have to take the total of the value out of the total I take the relative frequency this total is 150 + 90 + 110 + 30 + 20 is 400. So, this relative frequency Relative frequency becomes 150 by 400 into percentage I can even put it as 37.5 percent this becomes 90 by 400 this is = 22.5 percent. I will just put all the percentages here 27.5, 7.5, 5.0 this is relative frequency.



Method of Data Presentation

- tabular method

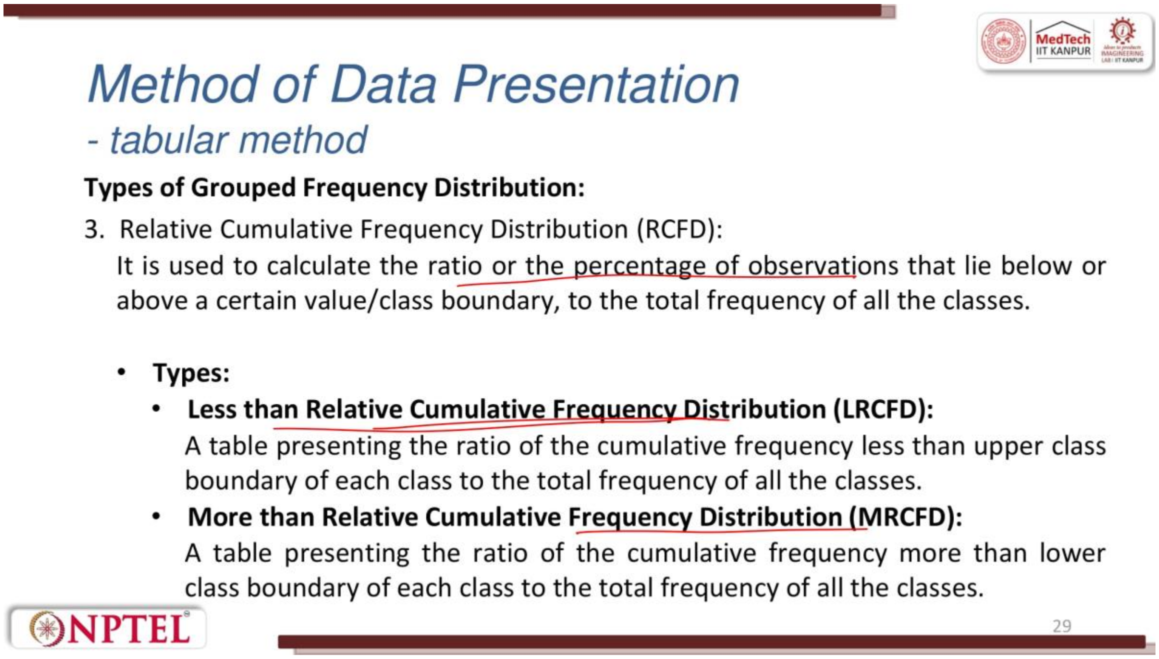
Types of Grouped Frequency Distribution:

2. Cumulative Frequency distribution (CFD):
   - It is applicable when we want to know how many observations lie below or above a certain value/class boundary.
   - **Types:**
     - **Less than Cumulative Frequency distribution (LCFD):** shows the collection of cases lying **below the upper class boundaries of each class.**
     - **More than Cumulative Frequency distribution (MCFD):** shows the collection of cases lying **above the lower class boundaries of each class.**

Then I can have cumulative frequency and cumulative relative frequency, cumulative frequency here the first observation 120 second is 120 + 90 that is 240 + 110 is 350 + 30 is 380 + 20 is 400, this is cumulative frequency. Then I have relative cumulative frequency that is 37.5 + 22.5 would be 60 + 27.5 would be 87.5 + 7.5 would be 95 + 5 would be 100. So, I will put these here RFT (Relative Frequency Distribution), cumulative frequency distribution, relative cumulative frequency distribution.

So, this is my frequency distribution, this is a relative frequency distribution, this is cumulative frequency distribution and this is a relative cumulative frequency distribution in a tabular form. So, cumulative frequency distribution, it is applicable when we want to know many observations lie below or above certain value or boundary. It is applicable when we want to know how many observations lie below or above a certain value of class or boundary.

Types could be less than cumulative frequency distribution, what are the observations those more than cumulative frequency distribution what are the observations which are lying.



Method of Data Presentation

- tabular method

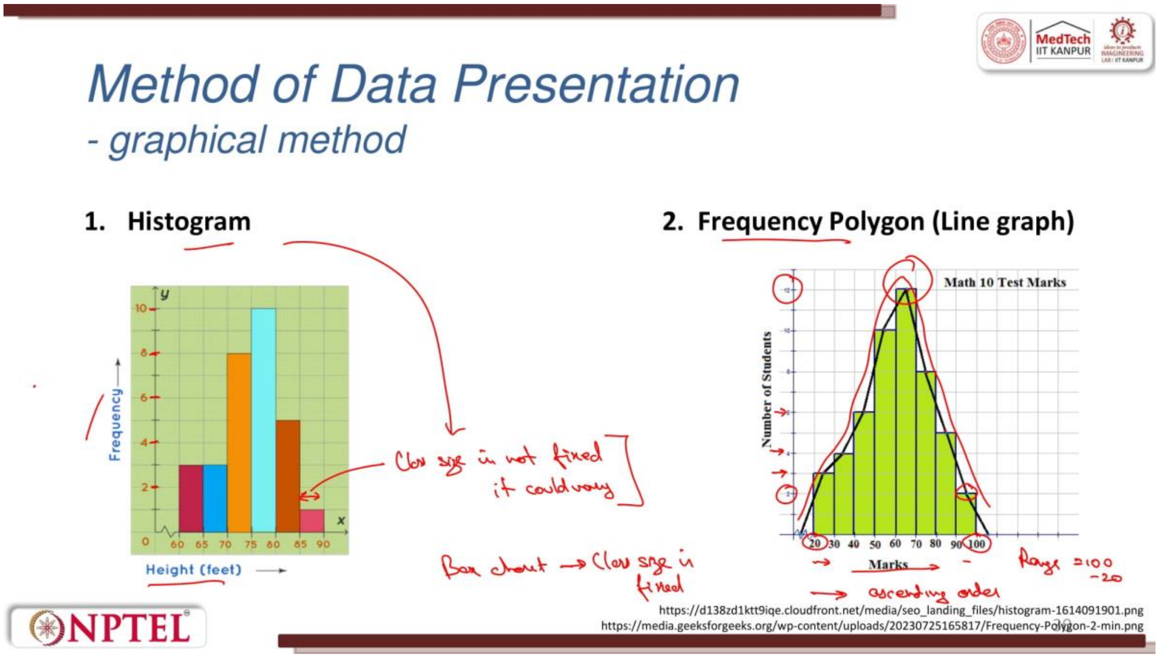
Types of Grouped Frequency Distribution:

3. Relative Cumulative Frequency Distribution (RCFD):
   It is used to calculate the ratio or the percentage of observations that lie below or above a certain value/class boundary, to the total frequency of all the classes.

- **Types:**
  - **Less than Relative Cumulative Frequency Distribution (LRCFD):**
    A table presenting the ratio of the cumulative frequency less than upper class boundary of each class to the total frequency of all the classes.
  - **More than Relative Cumulative Frequency Distribution (MRCFD):**
    A table presenting the ratio of the cumulative frequency more than lower class boundary of each class to the total frequency of all the classes.

Types of group frequency distribution Relative Cumulative Frequency Distribution it is used to calculate ratio or percentage of the observations that lie below or above a certain value or class of a boundary.

In this as well it could be less than cumulative frequency distribution what are the number of observations which are lying lower than the specific observation of interest more than relative frequency distribution is which are higher than the observation of interest.



Method of Data Presentation

- graphical method

1. **Histogram**
2. **Frequency Polygon (Line graph)**

Class size is not fixed
it could vary

Bar chart → Class size is fixed

→ ascending order

Range = 100
−20

https://d138zd1ktt9iqe.cloudfront.net/media/seo_landing_files/histogram-1614091901.png
https://media.geeksforgeeks.org/wp-content/uploads/20230725165817/Frequency-Polygon-2-min.png

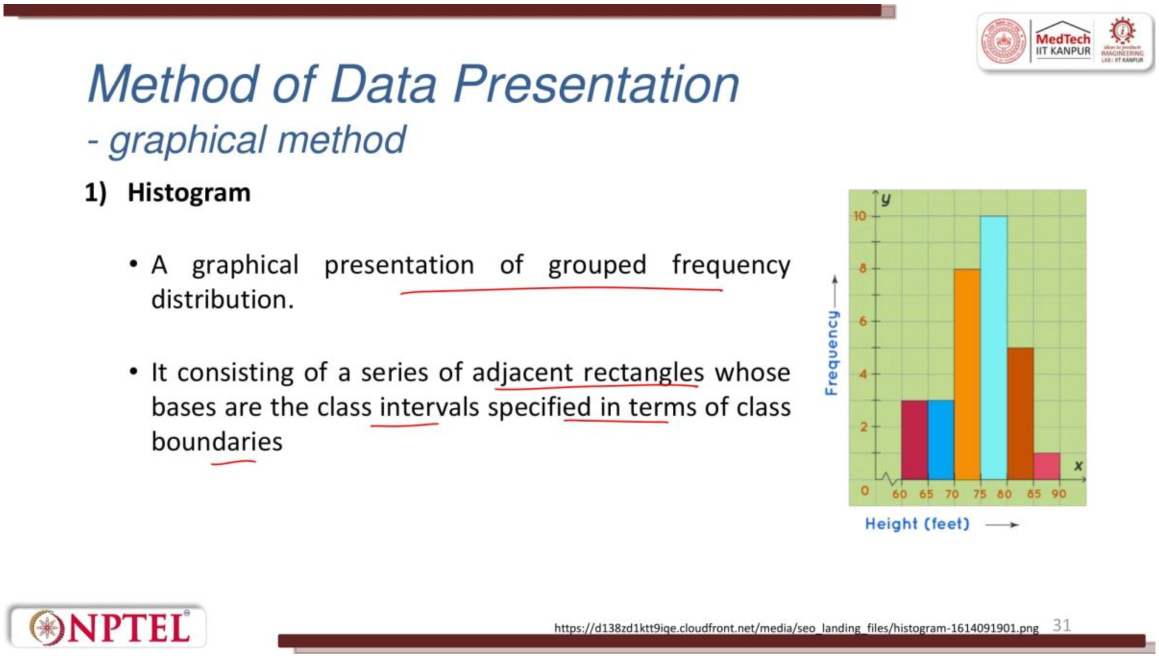
Then comes graphical presentation methods that I will discuss. There are multiple other graphs or charts available, but I will focus upon how do you construct an Histogram. For instance, this is a table that you have constructed. So, this table itself 150, 90, 110, 30, 20 this could be put into a histogram.

The difference between histogram and a chart is in histogram this class size is not fixed or it could vary. This is for histogram. In bar chart, class size is fixed. This is the specific size. It is the difference between histogram and in bar chart.

This is the histogram in which the observations 2, 4, 6, 8, 10. Both the frequency distribution and relative frequency distribution can put it in this way. Let me take it here. How do we draw the frequency polygon? What I am doing is from 20 to 30, 40 to 50, what are the number of observations that are coming between 20 to 30?

For example, number of students those have got marks between 20 to 30 is 3. Number of students those have scored marks between 30 to 40 is 4. Those scored marks between 40 to 50 is 6. You can see maximum number of students would be lying almost at the center that is between 60 to 70 out of 100. There are 12 students who have secured these marks and there are few students who have secured between 90 to 100 like there are two students those have secured between 90 to 100.

This is how we draw histogram or small graphs here and when we try to join the midpoint of each of the class this is that I have discussed before class midpoint. If I try to join them I obtain my polygon, this is the polygon frequency, polygon that we have obtained and remember in this thing this is in an order and there is a range that is maximum value 100 - minimum value that is range. Range = 100 - 20.



Method of Data Presentation

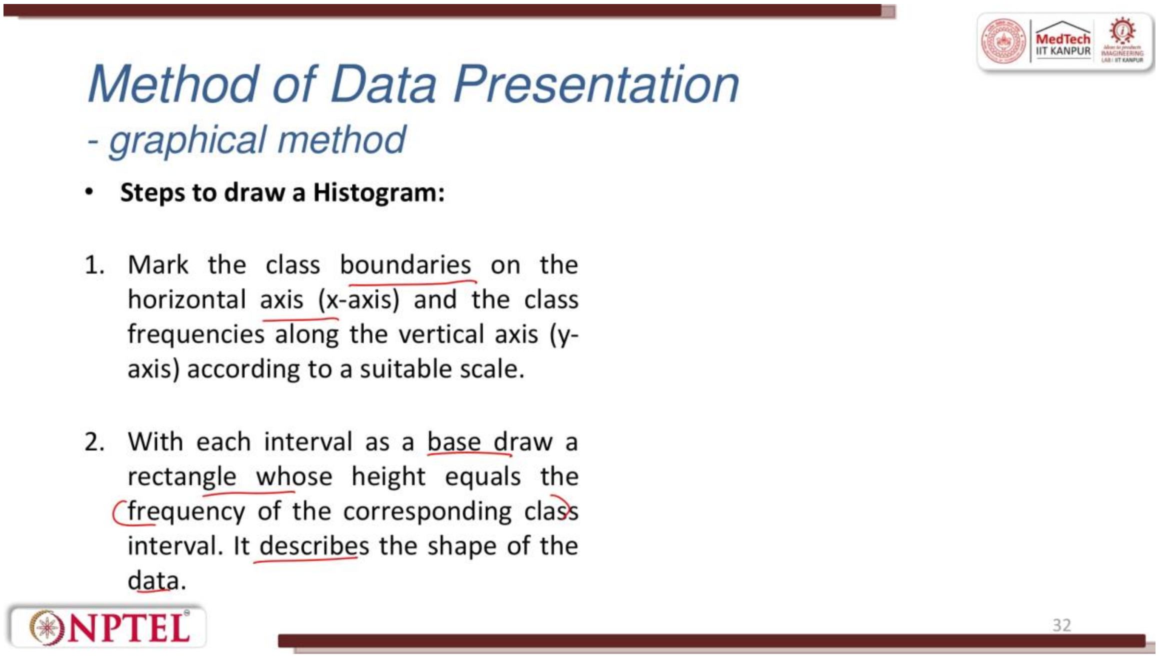
- graphical method

1) **Histogram**

- A graphical presentation of grouped frequency distribution.
- It consisting of a series of adjacent rectangles whose bases are the class intervals specified in terms of class boundaries

https://d138zd1ktt9iqe.cloudfront.net/media/seo_landing_files/histogram-1614091901.png

Histogram is a graphical presentation of group frequency free distribution. It consists of series of adjacent rectangles whose bases are the class intervals specified in terms of class boundaries.



Method of Data Presentation

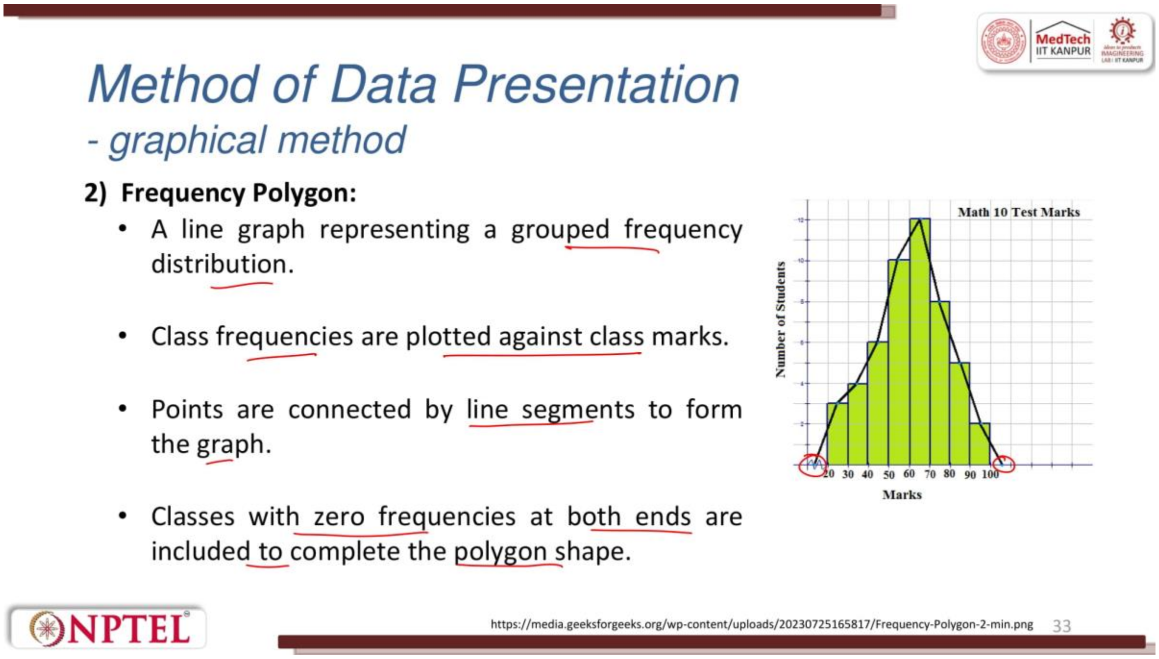
- graphical method

- **Steps to draw a Histogram:**

1. Mark the class boundaries on the horizontal axis (x-axis) and the class frequencies along the vertical axis (y-axis) according to a suitable scale.

2. With each interval as a base draw a rectangle whose height equals the frequency of the corresponding class interval. It describes the shape of the data.

The certain steps to draw histogram mark the class boundaries on the horizontal axis. It is x-axis and classes frequencies along the vertical axis that you have seen in the figure here. This is frequency, this is horizontal axis where height and feet is marked. With each interval as base, draw a rectangle whose height equals the frequency of the corresponding class interval. It describes the shape of the data.



Method of Data Presentation

- graphical method

2) Frequency Polygon:

- A line graph representing a grouped frequency distribution.

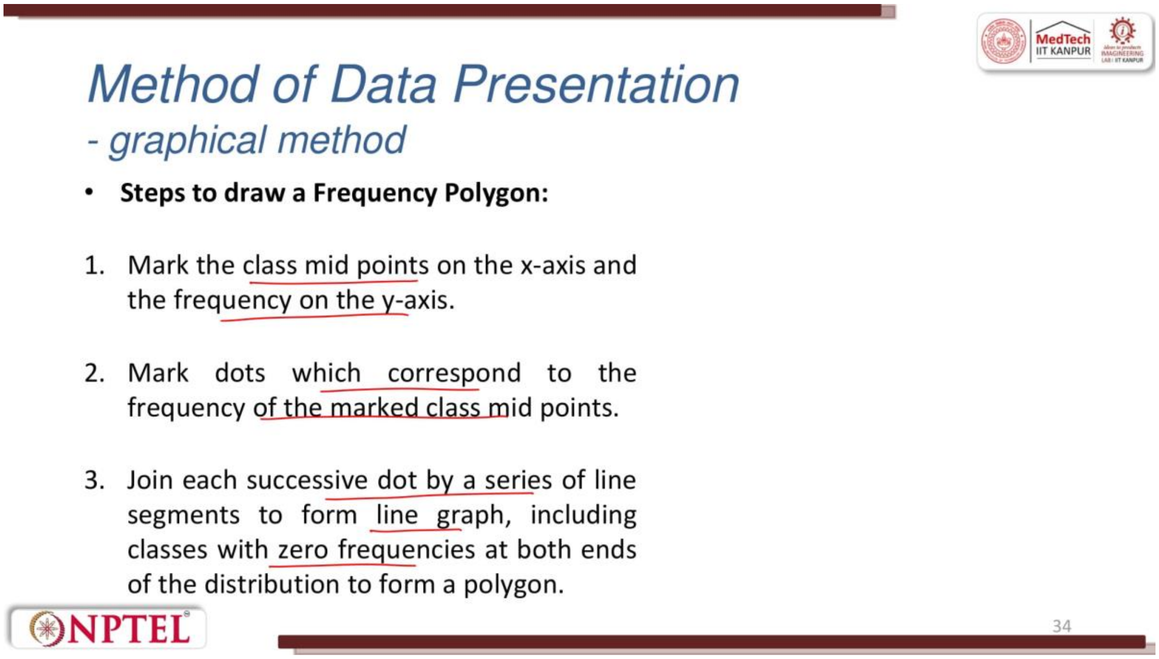
- Class frequencies are plotted against class marks.

- Points are connected by line segments to form the graph.

- Classes with zero frequencies at both ends are included to complete the polygon shape.

https://media.geeksforgeeks.org/wp-content/uploads/20230725165817/Frequency-Polygon-2-min.png

Frequency Polygon, it is a line graph presenting a group frequency distribution. Class frequencies are plotted against class marks. Points are connected by line segments to form the graph. This is how it is connected I have shown you in the previous slides. Classes with zero frequencies at both ends are included to complete the polygon shape. So, this is with zero frequency these two are included. So, that we complete the polygon.



Method of Data Presentation

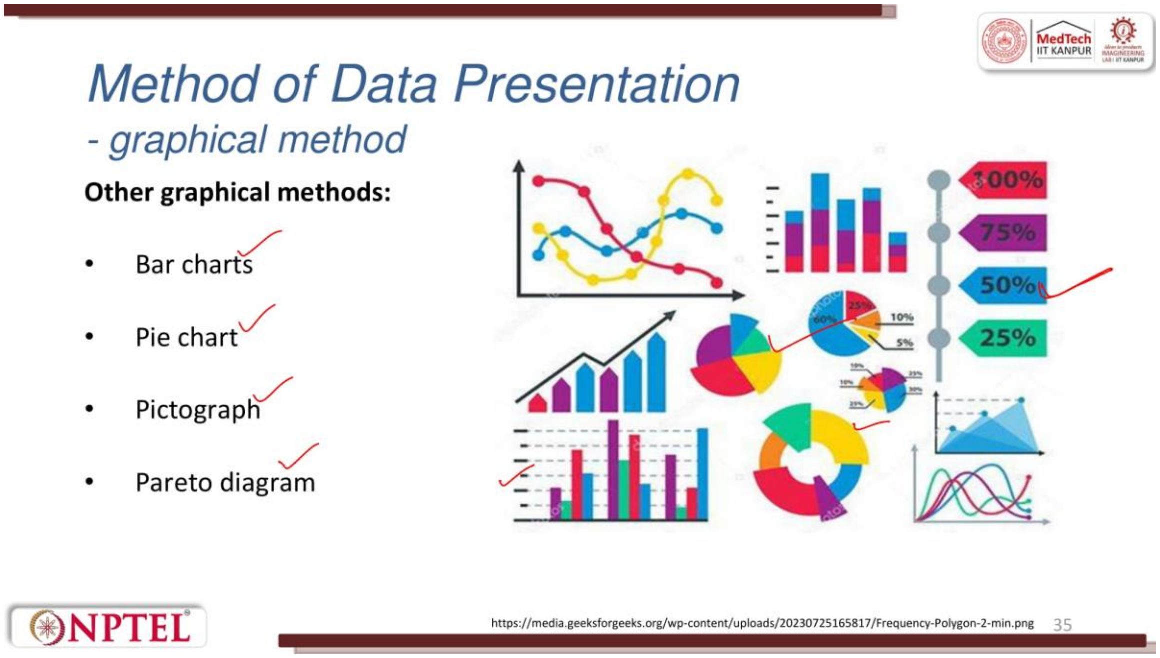
- graphical method

- **Steps to draw a Frequency Polygon:**

1. Mark the class mid points on the x-axis and the frequency on the y-axis.
2. Mark dots which correspond to the frequency of the marked class mid points.
3. Join each successive dot by a series of line segments to form line graph, including classes with zero frequencies at both ends of the distribution to form a polygon.

34

Steps are we mark the midpoints on x axis and the frequency on the y-axis mark dots which correspond to the frequency of the mark class in the midpoints join each successive dot by a series of line segments, to form a line graph including classes with zero frequencies at both ends of the distribution. This is what I did to develop a polygon from a histogram.



Method of Data Presentation

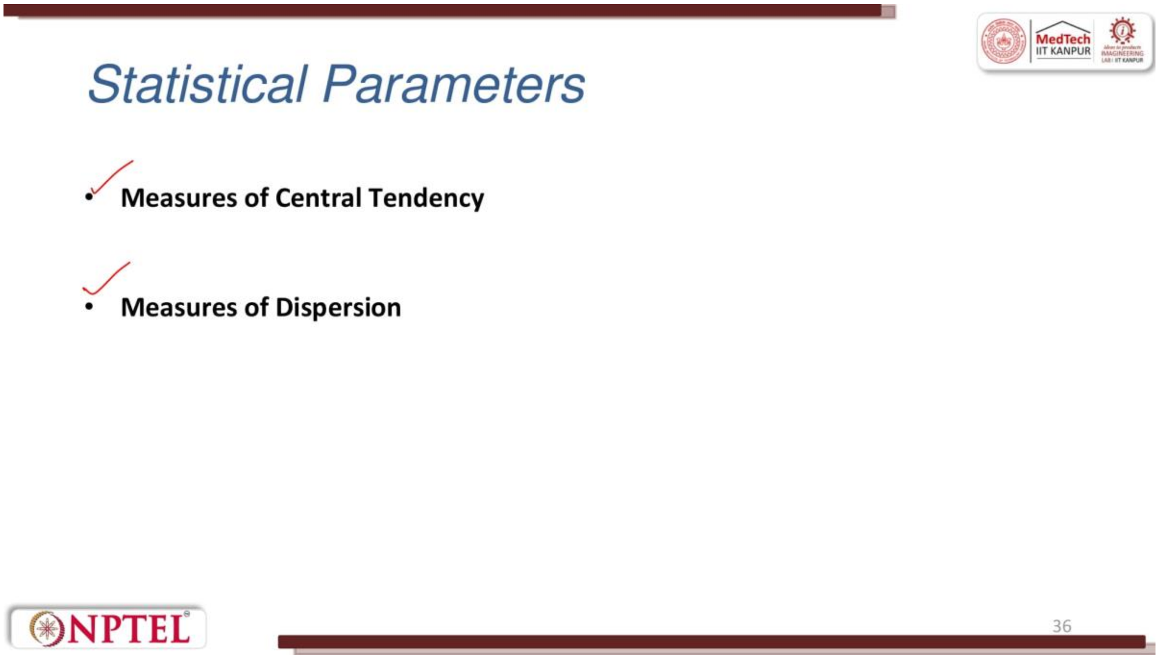
- graphical method

Other graphical methods:

- Bar charts
- Pie chart
- Pictograph
- Pareto diagram

https://media.geeksforgeeks.org/wp-content/uploads/20230725165817/Frequency-Polygon-2-min.png

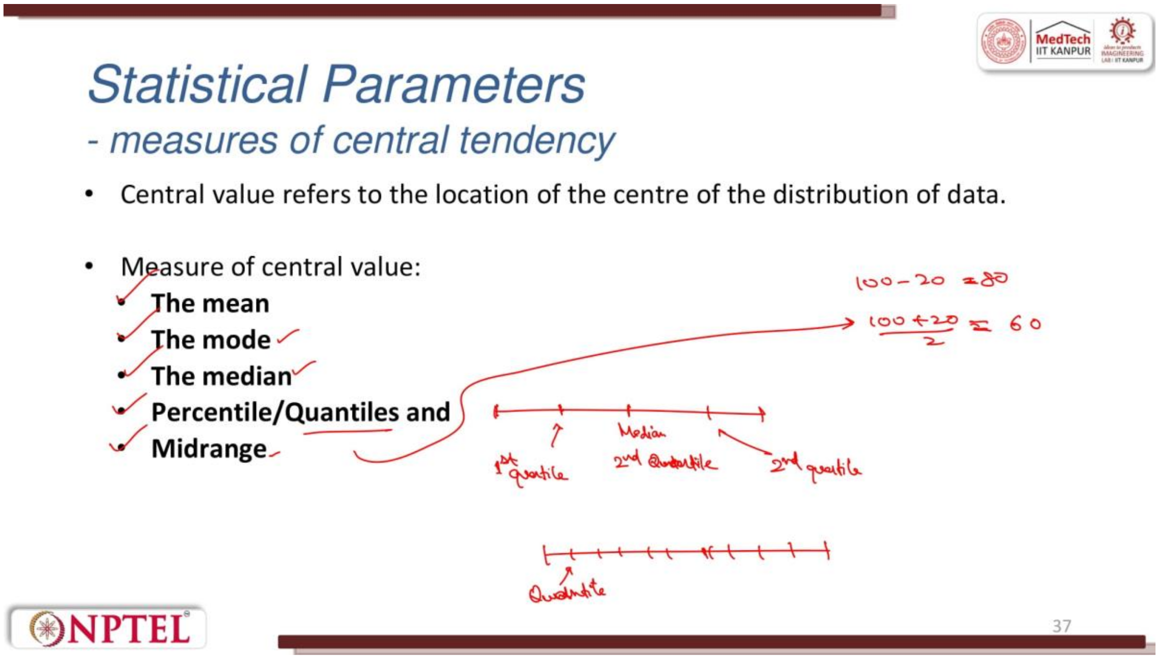
There could be many other charts. For example, bar charts are there, pie charts are there. Pictogram is there, based upon the pictogram we say this is how the humidity is varying, it could be Pareto diagram that represent how the data is behaving in different forms. So, I am not covering all of them.



Statistical Parameters

- **Measures of Central Tendency**
- **Measures of Dispersion**

Statistical parameters are important Measures of Central Tendency and Measures of Dispersion.



Statistical Parameters

- measures of central tendency

- Central value refers to the location of the centre of the distribution of data.
- Measure of central value:
  - **The mean**
  - **The mode**
  - **The median**
  - **Percentile/Quantiles and**
  - **Midrange**

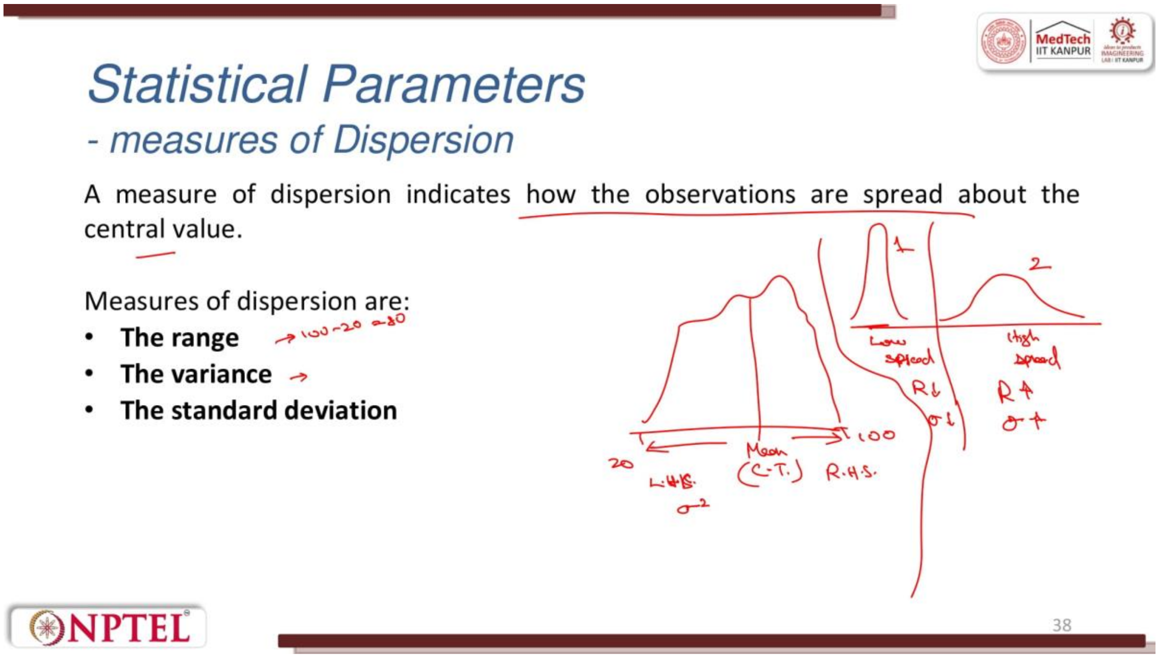
$100 - 20 = 80$

$\frac{100+20}{2} = 60$

What are Measures of Central Tendency? Measures of central tendency is measuring the centre value, the mean, the mode, the median, percentile, mid range these are all measures of central tendency. Mean you know is the central value, mode is the highest frequency that is there, medium is the middle value, percentile and quantiles could be the out of the whole data this is a median that is also known as second. When I say quartile, it is divided into 4 quarters. This is first quartile, this is second quartile.

Other than quartiles, we can also have quantiles. Quantiles could be suppose if I divide this into 10 equal parts. So, they can say this is a quantile. Quantile could be any division, 10 parts, 20 parts or so. Then is a mid-range.

Range is, for example, it was 100 minus 20, 80 was the range given in the previous problem statement. Highest value + lowest value by 2 is mid-range, this is = 120 by 260. These are all measures of this central tendency, where is the center of the data.



Statistical Parameters

- measures of Dispersion

A measure of dispersion indicates how the observations are spread about the central value.

Measures of dispersion are:

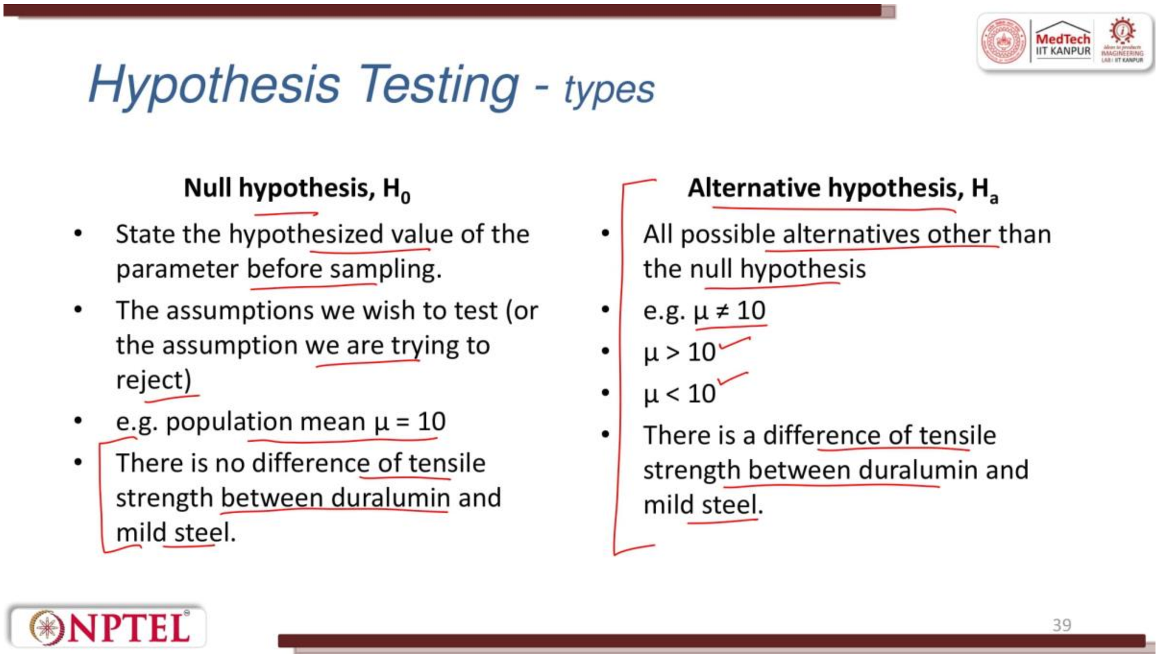
- **The range** → 100-20 = 80
- **The variance** →
- **The standard deviation**

Then comes measures of dispersion. Measure of dispersion indicates how the observation are spread about the central value.

Central tendency is for instance this was the data that we presented in the polygon Where is the center? This is mean, this is central tendency and this is dispersion. On the left hand side and on the right hand side, how dispersed is the theta? The value 20 and 100 represented my range, this was 100 minus 20.

80 gives me the spread of the data. The variance, if I put sigma value here, there is a formula to calculate sigma. If I calculate sigma scale value here, that is my variance, standard deviation. That gives me the spread. Spread could be for the same number of observations.

It could be something like this. You can see in the first case and in the second case, if I compare, here the spread is 80. Low, here the spread is high that means, in the case 1 and 2, in the case, the range value would be low, the sigma value would be low, here the range value would be high and sigma value would be high in case of the high spread.



Hypothesis Testing - *types*

Null hypothesis, H_0

- State the hypothesized value of the parameter before sampling.
- The assumptions we wish to test (or the assumption we are trying to reject)
- e.g. population mean $\mu = 10$
- There is no difference of tensile strength between duralumin and mild steel.

Alternative hypothesis, H_a

- All possible alternatives other than the null hypothesis
- e.g. $\mu \neq 10$
- $\mu > 10$
- $\mu < 10$
- There is a difference of tensile strength between duralumin and mild steel.

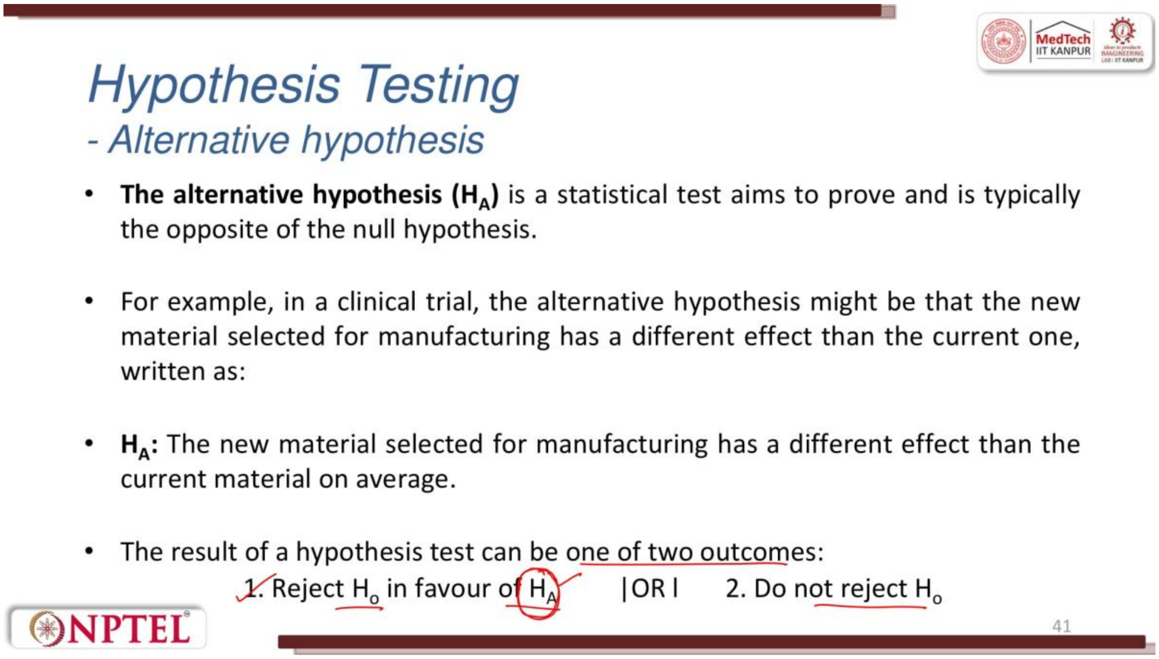
Let me try to talk about hypothesis testing. Null hypothesis and Alternative hypothesis.

Null hypothesis present the hypothesized value of a parameter before sampling assumptions that we wish to test or the assumptions that we are trying to reject that is population mean nu is = 10. Hypothesis testing is a statistical method used to take decisions about a data set without having to examine every element in the data set. That is we try to pick a sample and we will try to measure the sample behavior and we will try to reflect that upon the overall population. Here it is saying population means 10. There is no difference of tensile strength between duralumin and mild steel.

This is how null hypothesis is stated. Duralumin is an alloy of aluminium whose strength is equivalent to mild steel or I would say the alternative hypothesis would be whatever you have set in the null hypothesis, if it is not true, that means there is a difference. All possible alternatives other than null hypothesis is alternative hypothesis. For example, we say mu is not = 10, mu could be greater than 10, mu could be less than 10.

This is or the statement could be there is a difference of tensile strength between duralumin and mild steel.

Null hypothesis (Ho) represents a theory or assumption that has not been proven and is often used as a basis for testing. For example, in a critical trial, null hypothesis might state that there is no significant difference between new drug and the existing one and can be written as Ho. The new drug is no better than the current drug on average. This is null hypothesis.



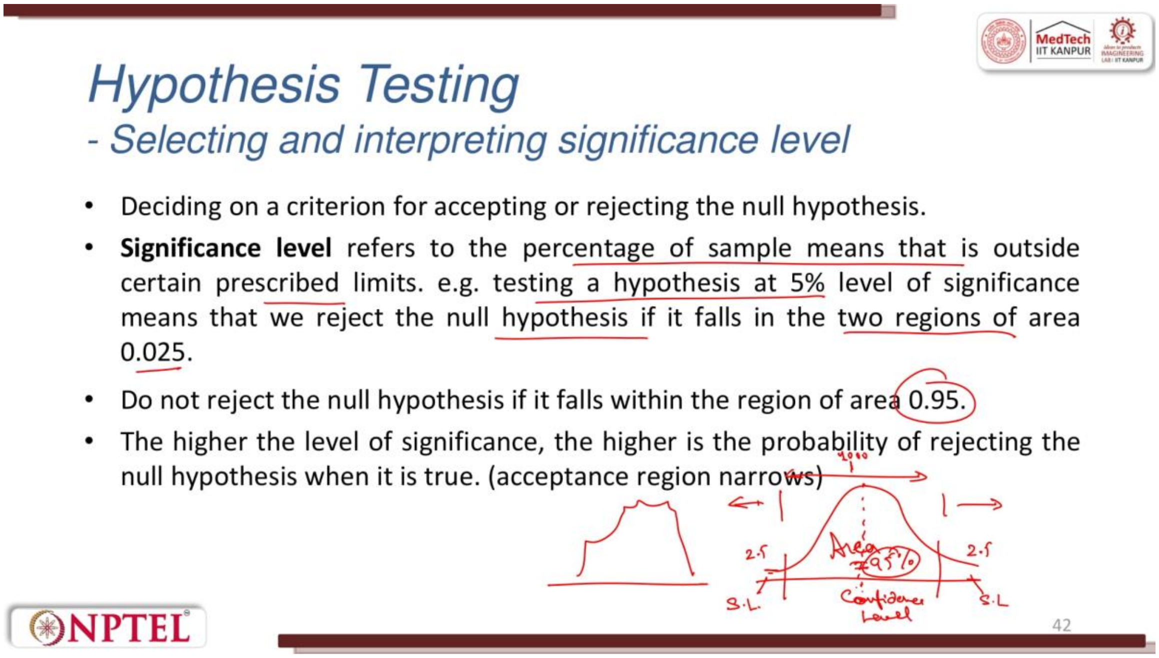
Hypothesis Testing
- Alternative hypothesis

- **The alternative hypothesis (H_A)** is a statistical test aims to prove and is typically the opposite of the null hypothesis.
- For example, in a clinical trial, the alternative hypothesis might be that the new material selected for manufacturing has a different effect than the current one, written as:
- **H_A:** The new material selected for manufacturing has a different effect than the current material on average.
- The result of a hypothesis test can be one of two outcomes:

1. Reject H_o in favour of H_A |OR| 2. Do not reject H_o

41

Alternative hypothesis (HA) is a statistical test that aims to prove and it is typically opposite the null hypothesis. The example could be in clinical trial the alternative hypothesis might be that the new drug has a different effect than the current one written as alternative hypothesis (HA) the new drug has different effect than current drug on average. The result of a hypothesis test can be one of the two outcomes either we reject Ho that is null hypothesis in favor of HA that is alternate hypothesis or we do not reject Ho that is alternate hypothesis is designed in such a way that when null hypothesis is rejected this is the option that is left with us.



Hypothesis Testing

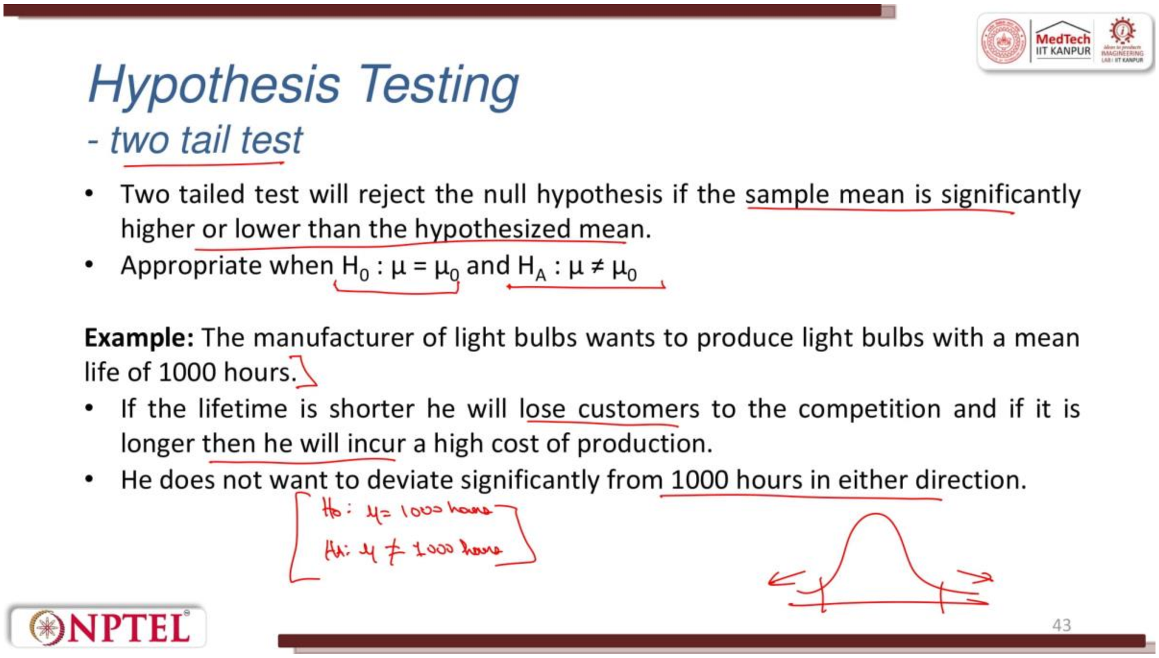
- Selecting and interpreting significance level

- Deciding on a criterion for accepting or rejecting the null hypothesis.
- **Significance level** refers to the percentage of sample means that is outside certain prescribed limits. e.g. testing a hypothesis at 5% level of significance means that we reject the null hypothesis if it falls in the two regions of area 0.025.
- Do not reject the null hypothesis if it falls within the region of area 0.95.
- The higher the level of significance, the higher is the probability of rejecting the null hypothesis when it is true. (acceptance region narrows)

Deciding on criteria for accepting or rejecting the null hypothesis, we have to select a significance level.

Significance level refers to the percentage of sample means that is outside certain prescribed limit. For example, testing a hypothesis at 5 percent of significance means we reject the null hypothesis if it falls in two regions of the area of 0.25. So, this is something your frequency distribution was something like this, right. If I try to smoothen it down it will become something like this that is known as normal distribution. Here in the normal distribution if I say I am 95 percent confident that my sample will represent my population, 95 percent of confidence limits I am in setting which means 5 percent of confidence is still not there.

So, therefore, it means the area here. 95 percent that I am confident, this is my confidence level and this becomes my significance level. If 95 percent of area is inside, 5 percent are outside that is 2.5 percent on this side, 2.5 percent on this side.



Hypothesis Testing

- two tail test

- Two tailed test will reject the null hypothesis if the sample mean is significantly higher or lower than the hypothesized mean.
- Appropriate when $H_0 : \mu = \mu_0$ and $H_A : \mu \neq \mu_0$

Example: The manufacturer of light bulbs wants to produce light bulbs with a mean life of 1000 hours.

- If the lifetime is shorter he will lose customers to the competition and if it is longer then he will incur a high cost of production.
- He does not want to deviate significantly from 1000 hours in either direction.

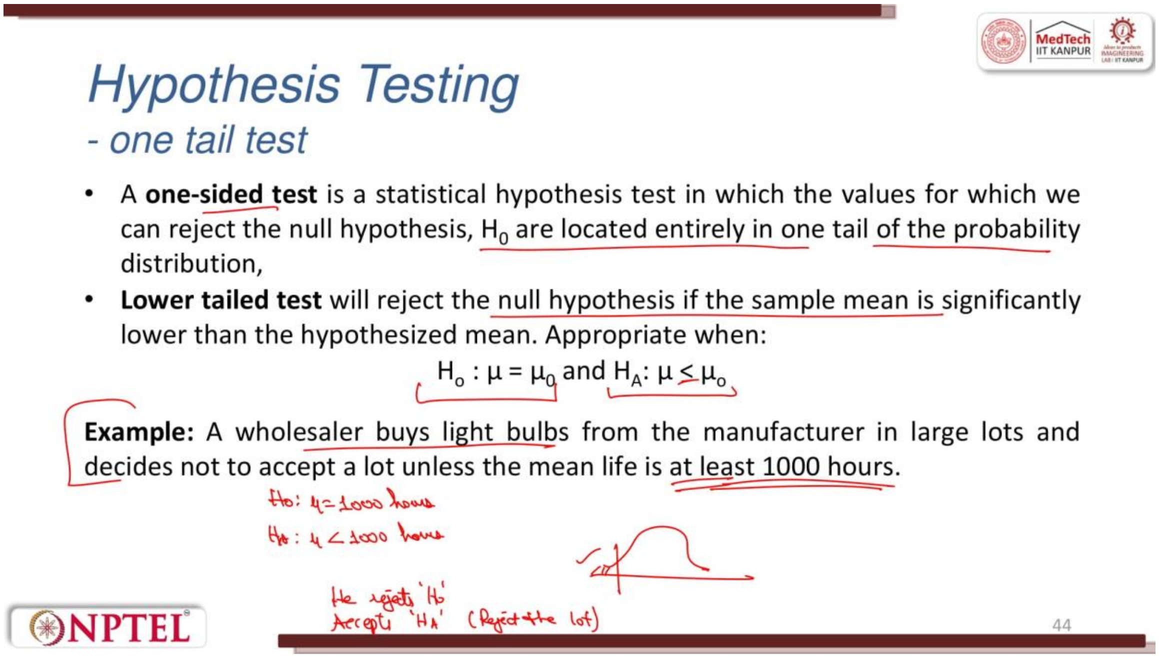
$H_0 : \mu = 1000$ hours

$H_A : \mu \neq 1000$ hours

Here it is two tailed test. What is two-tailed test? Two-tailed test will reject the null hypothesis if the sample mean is significantly higher or lower than the hypothesized mean. Appropriate Ho that is null hypothesis is telling that mu is = µo and alternate hypothesis H is telling that µ is not = µo. Example is a manufacturer of light bulbs wants to produce light bulbs with a mean of 1000 of the life, if the lifetime is shorter, he will lose customers to the competition and if it is longer, he will incur higher cost, this is very important.

So, he cannot lose he is telling 1000 hours is there he does not want to deviate significantly from 1000 hours in either direction that means, in the graph that I have shown here he does not want to go this side or does not want to come this side, he wish to stay here where the central value is 1000; I have only given mean value, it could be standard deviation and range as well here. So, here it is two tail test. So, this is a two way hypothesis this is two tail test. So, it could be written in this way Ho is µ that the manufacturer want is 1000 hours and HA that is alternative hypothesis could be that µ is not = 1000 hours.

That is in both the directions it could be rejected or not rejected. Either this side or this side we have a rejection region. I will show you details of this frequency polygon in the end. This is a two tailed cistern.



Hypothesis Testing

- one tail test

- A **one-sided test** is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0 are located entirely in one tail of the probability distribution,
- **Lower tailed test** will reject the null hypothesis if the sample mean is significantly lower than the hypothesized mean. Appropriate when:

$$H_o : \mu = \mu_0 \text{ and } H_A : \mu \leq \mu_o$$

Example: A wholesaler buys light bulbs from the manufacturer in large lots and decides not to accept a lot unless the mean life is at least 1000 hours.

H_o: $\mu = 1000$ hours

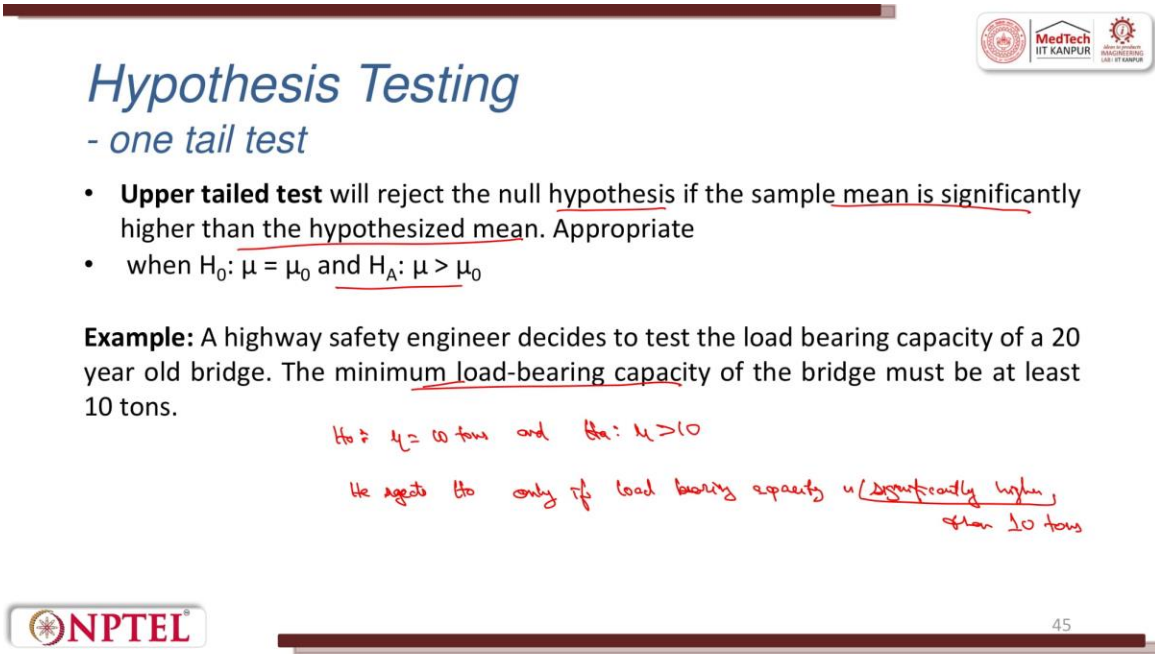
H_A: $\mu < 1000$ hours

He rejects 'H_o'
Accepts 'H_A' (Reject the lot)

One tail test could also be there that is one sided test that is lower tailed or upper tailed. One sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis Ho are located entirely in one tail of the probability distribution. Lower tail test will reject the null hypothesis if the sample mean is significantly lower than hypothesized mean here; null hypothesis is μ is = μo alternative hypothesis is μ is less than μo. Example is a wholesaler buys light bulbs from the manufacturer in large lots and decides not to accept a lot unless mean life is at least 1000, at least you see it is less than, at least 1000 hours. This is one tail test and the lower tailed one.

So, let me put it in the terms of Ho and HA when we are talking about a one tail test. It is wholesaler who is accepting or not accepting the bulbs manufactured by the manufacturer or provided by the manufacturer in large lots. So, Ho is that μ is = 1000 hours and H alternative because it is one tail test and one sided lower tail test it is μ is less than 1000 hours. So, that means if μ is less than 1000 hours, he rejects Ho here. He rejects Ho and accepts H alternative.

What is HA here? That he will reject the lot or does not accept the lot. This is lower tail test.



Hypothesis Testing

- one tail test

- **Upper tailed test** will reject the null hypothesis if the sample mean is significantly higher than the hypothesized mean. Appropriate
- when H_0: $\mu = \mu_0$ and H_A: $\mu > \mu_0$

Example: A highway safety engineer decides to test the load bearing capacity of a 20 year old bridge. The minimum load-bearing capacity of the bridge must be at least 10 tons.

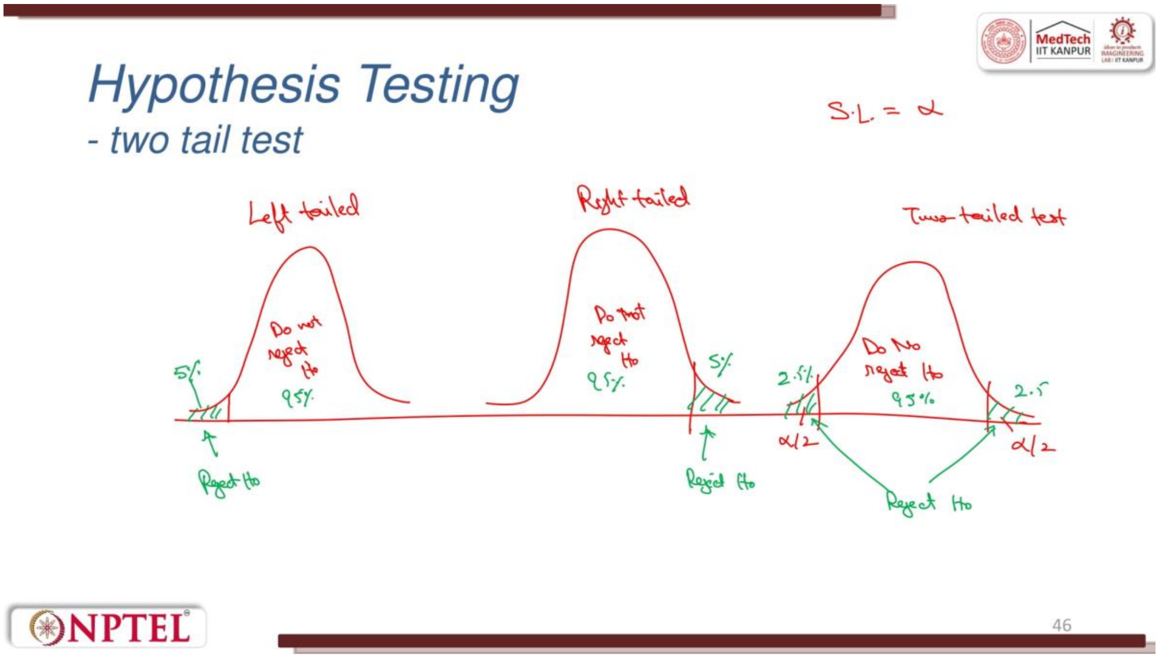
H_o: $\mu = 10$ tons and H_a: $\mu > 10$

He rejects Ho only if load bearing capacity is (significantly higher than 10 tons

Upper tail test will reject the null hypothesis if the sample mean is significantly higher than hypothesized mean that is µ is = µo that is the Ho and H alternative is µ is greater than µo. Example is a highway safety engineer decides to test load bearing capacity of 20 year old bridge.

The minimum load bearing capacity for the bridge must be at least 10 tons. He says the minimum value has to be there. So, this becomes an upper tail test that is Ho here is µ is = 10 tons per and H alternative here is µ is greater than 10 and we use upper tail test.

Now, here he rejects Ho that is null hypothesis only if the load bearing capacity of the bridge is significantly higher than 10 tons is significantly higher than 10 tons. Here the key word is significantly based upon the significance that you said.



Hypothesis Testing

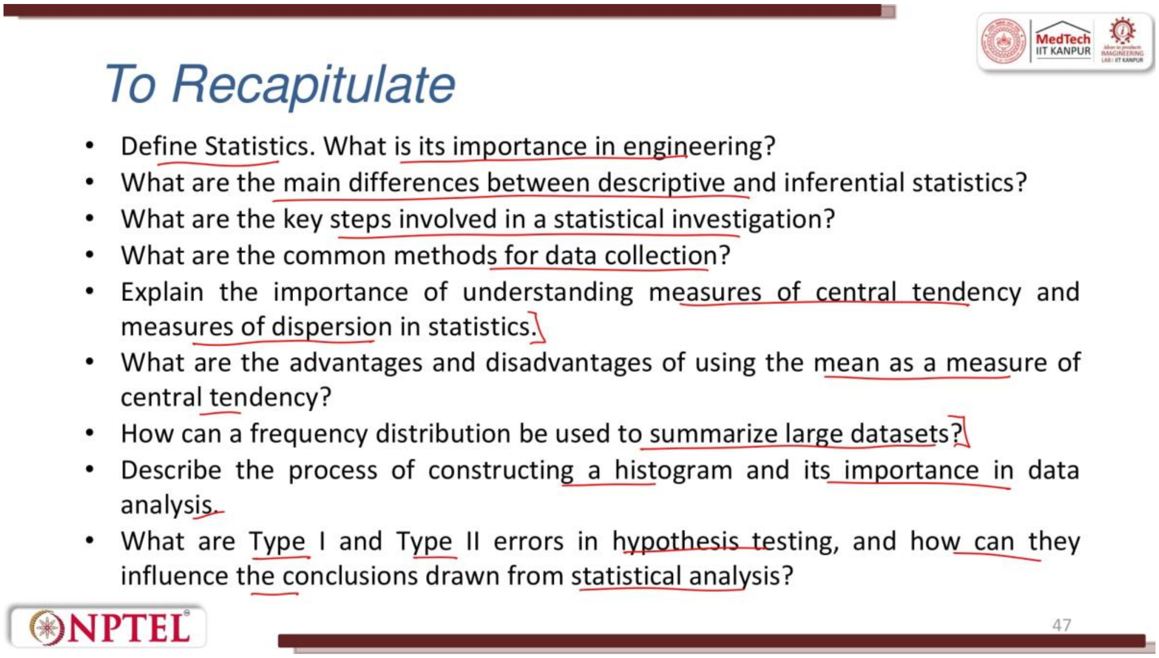
- two tail test

For that let me just draw the different scenario. There could be two tail test for a frequency distribution which is the normal curve where on the left hand side and right hand side you have a two tail test in between if your observation falls, we do not reject Ho and if it is here, the significance level is termed as alpha.

So, this becomes alpha by 2, this also becomes alpha by 2. For instance, it was 95 percent inside, it could be 2.5 percent on both the sides. So, we do not reject Ho here and these two regions become where we reject Ho. Reject Ho here. Let me say this is 95 percent, this is 2.5 percent, this is 2.5.

One-tailed and two-tailed test could be put in this way, one-tailed test that is lower-tailed test and upper-tailed test I will try to draw. In the lower-tailed test in this side, in the both side, this area is do not reject Ho. So, this is left tailed and on the left hand side I have a significance value this is right tailed and which I showed on the right hand side here is two tailed. So, what is here? Here I have 95 percent value and this becomes my 5 percent value on the left hand side.

Here I have 95 percent value that is do not reject Ho and I have a 5 percent value on the right hand side and these two suggests reject Ho.



To Recapitulate

- Define Statistics. What is its importance in engineering?
- What are the main differences between descriptive and inferential statistics?
- What are the key steps involved in a statistical investigation?
- What are the common methods for data collection?
- Explain the importance of understanding measures of central tendency and measures of dispersion in statistics.
- What are the advantages and disadvantages of using the mean as a measure of central tendency?
- How can a frequency distribution be used to summarize large datasets?
- Describe the process of constructing a histogram and its importance in data analysis.
- What are Type I and Type II errors in hypothesis testing, and how can they influence the conclusions drawn from statistical analysis?

To recapitulate this lecture, the question that you can ask is define statistics, what is its importance in engineering? What are main differences between descriptive and information statistics? What are key steps involved in statistical investigation? What are common methods of data collection?

Primary method, secondary method, in primary what are the ways? Do you take observations? Do you take interviews or so? Explain importance of understanding measures of central tendency and measures of dispersion in statistics. What are advantages and disadvantages of using mean as a measure of central tendency?

How can a frequency distribution be used to summarize large data sets. You can draw a frequency distribution, you can take your own examples and try to work upon them. Describe the process of constructing a histogram and its importance in data analysis. And what are type 1 and type 2 errors in hypothesis testing? This you can study and try to answer that and how can they influence the conclusions drawn from statistical analysis.

So, with this I am closing this lecture and there could be many things that could be discussed in statistics like I talked about sampling techniques, like I talked about the test which are conducted for the hypothesis. Then there are types of errors in hypothesis as well type 1 and type 2 error, I am not covering that in this course to keep it limited to the basic understandings of what is hypothesis.

To design a hypothesis, we could either have one direction that is left or right or both the directions could be there. A generic understanding of the presentation of the data, collection of the data, primary and secondary data, qualitative and quantitative data, these are all there. There will not be any numericals from this topic in the exam, only you need to understand what is the role of statistics.

And I hope you have learnt through this course and you will secure good marks in the coming exam. Now, we will meet in the next course that is Basics of Mechanical Engineering Part 2.

Thank you.