

**INDIAN INSTITUTE  
OF  
TECHNOLOGY  
KHARAGPUR**

**NPTEL  
National Programme  
on  
Technology Enhanced Learning**

**Applied Multivariate Statistical Modeling**

**Prof. J. Maiti  
Department of Industrial Engineering and Management  
IIT Kharagpur**

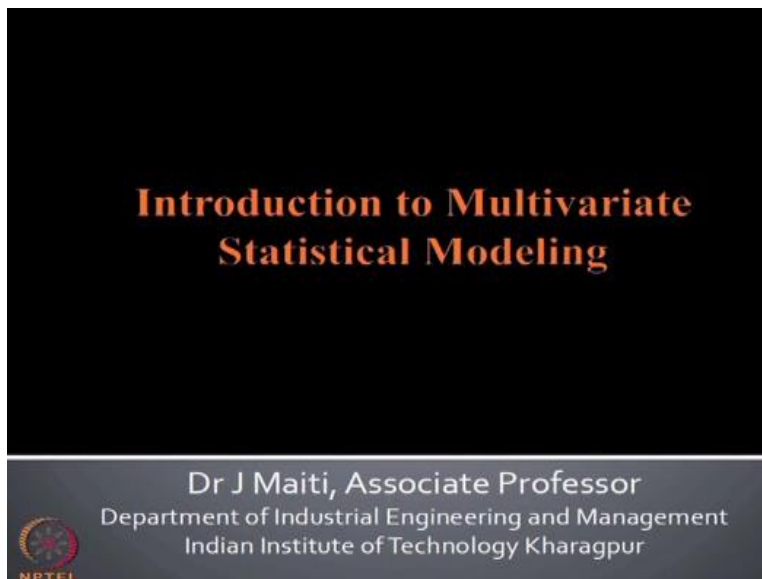
**Lecture – 01**

**Topic**

**Introduction to Multivariate  
Statistical Modeling**


Good morning welcome to the first lecture of applied multivariate statistical modeling.

(Refer Slide Time: 00:28)



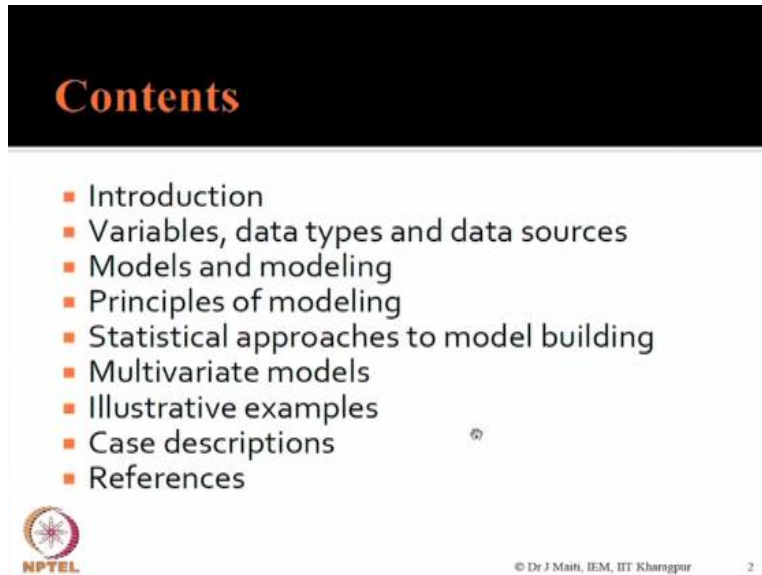
**Introduction to Multivariate  
Statistical Modeling**

Dr J Maiti, Associate Professor  
Department of Industrial Engineering and Management  
Indian Institute of Technology Kharagpur




Let me tell you the content of these today's presentation.

(Refer Slide Time: 00:33)



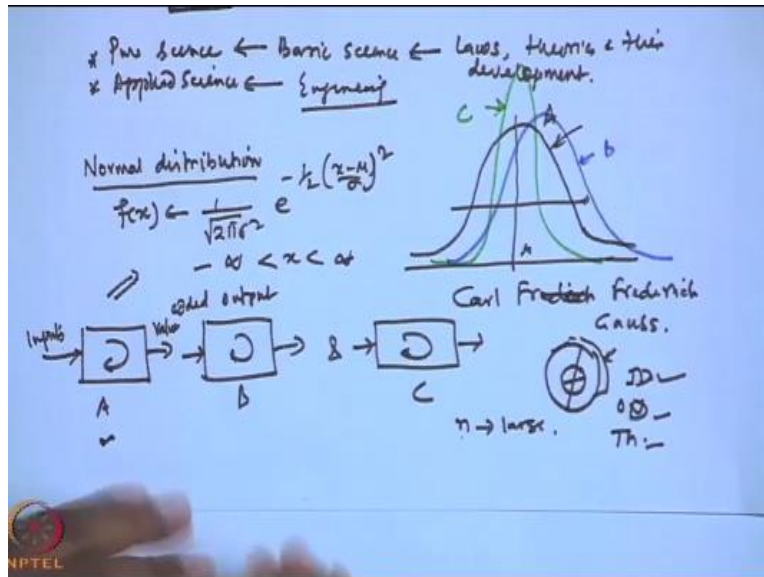
**Contents**

- Introduction
- Variables, data types and data sources
- Models and modeling
- Principles of modeling
- Statistical approaches to model building
- Multivariate models
- Illustrative examples
- Case descriptions
- References

 © Dr J Maiti, IEM, IIT Kharagpur 2

So we will start with introduction the variables data types, data sources, models and modeling followed by principles of modeling statistical approaches to model building multivariate models some I illustrative examples three cases followed by references the entire content will be covered in two hours today I will try to finish up to principles of modeling let us start with defining what is applied multivariate statistical modeling.

(Refer Slide Time: 01:21)



Applied multivariate statistical modeling let us defined whatever you want first is applied what do you mean by applied in science there is pure science and applied science pure science we generally understand which is basic science which is basically talks about laws theories and their development definitely it links with the phenomena which we usually observe in different aspects of our life now applied science which uses the knowledge of the pure science and developed something for the benefit of the mankind.

So applied science one of the benefit we can say then when you talk about engineering it is basically applied now when I talk about applied statistics what do you mean I am assuming that you have knowledge on preliminary basic statistics for example normal distribution if you know normal distribution then also you know the probability density function  $F(x)$  which is  $1 / \sqrt{2 \pi \sigma^2} e^{-1/2 (x - \mu / \sigma^2)}$  where  $x$  varies from  $-\infty$  to  $+\infty$ .

This is the so called this bell shaped curve which is developed by Carl Friedrich Gauss the theoretical devolvement so that development of these type of distributions this is it is coming under basics now if I want to apply these knowledge to real life situation I can find out a satiation like this for example, let us there are three process. Process A, B and C take certain

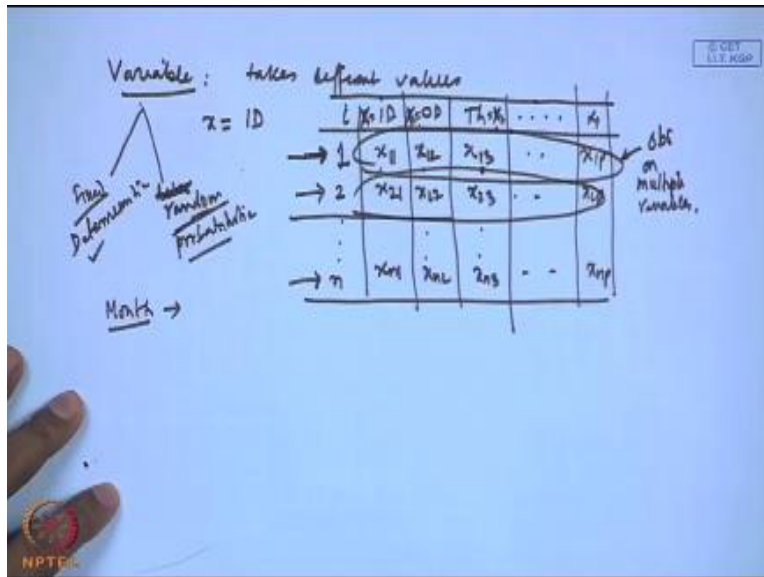
inputs convert into value added outputs all cases great there are basically three identical machines which is producing steel washers will be shape like this where there is inner diameter ID there is outer diameter OD.

As well as there will be certain thinness of this washers so I can say Th now if you produce a large amount of steel washers that means the number of items produce is large  $n$  is large then the suppose the quality characteristics of the steel voters which is of important to the people the customer the ID if you plot you may get this type of distribution which is normally distributed and where you will be getting mean here and there will be definitely standard deviation for ID similarly for OD and similarly of techniques.

Now then what you were doing by what is applied here a production process A for example in this case which is producing steel washers is converted into a statistical process in the sense in terms of a distribution like normal distribution where we are seeing that the production process can be interpreted the behavior of this process when we interrupted like this now here to get it further clarified if we do like this suppose this one is for A production process A and if I say this is for production process B and third one this one for production process C.

Then using these things you will be able compare A, B and C is their performance in terms of mean and standard deviation there is possibility also so to see that whether the mean ID produced by C is equal to that of B or A this type of complacence and other things possible so when we actually when we develop something which will be useful to the society for the man kind then we say it is applied so.

(Refer Slide Time: 07:42)



Now come to the second what which is basically multivariate now in order to understand multivariate you have to understand what is variable I think it is known to you that variable is something which takes different values that since I can say takes different values for example if I say ID  $x$  is ID in a diameter then if I produced one item  $i$  stands for the items sample first items and the ID value it may take value  $x_1$  when you go for second washer then it may take  $x_2$  so if I such way if I go for  $n$  washers produced that  $x_n$  come to configuration so these are the values so ID takes different values as a result ID is a variable now in statistics we basically talk about two types of variable one is fixed variable and other one is deterministic random variable.

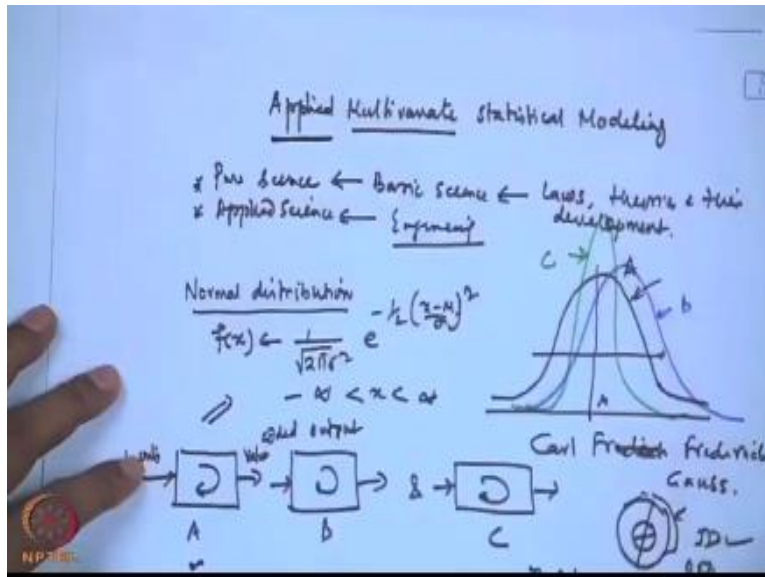
So fixed other way we can say deterministic and random we can say probabilistic for example if I create another variable which is month it varies over here so but we know all the month suppose what will be the next month if this month is your December next month will be January it is known with certainty so it should deterministic one but in this case when you are going to produce the second lot suppose been the second lot even in one lot what is the value of ID for the second item or second version it is not know with certainty it is governed through probabilistic distribution.

So that sense let us say it is a random one with one or the other will exactly and this value is coming based on certain random experiment in this case the process which is produced in this item so if I go on saying like this then other variable here is OD similarly other one is our thickness now in order to accumulate more than one variable we will write this  $x_1$  is ID  $x_2$  is OD and  $x_3$  is  $x_1$ ,  $x_2$ , and  $x_3$  is thickness. Then for the first item was is produced then this will be  $x_{11}$ ,  $x_{21}$  and  $x_{n1}$  similarly for OD  $x_{12}$ ,  $x_{22}$  like  $x_{n2}$  and if I go for the  $x_3$  variable that is also for first observation it is  $x_{13}$  second one  $x_{23}$  so like  $x_{n3}$

So what we are trying to say it here that we are considering three variables  $x_1$ ,  $x_2$ ,  $x_3$  which are nothing but the characteristics of this steel washes in this example which as inner diameter which is outer diameter which is thickness values. Now if you produced n number of washers then what will happen every washers have will be having different values for ID, OD and thickness okay.

So this is my observation first one is observation number one second one observation number two like there is observation number n and you see observation number one if I consider only ID that value is  $x_{11}$  if I consider all three together the value observation one  $x_{11}$   $x_{12}$   $x_{13}$  okay so similarly we should go on increasing the number variables up to  $x_p$  then here it will be  $x_{1p}$ ,  $x_{2p}$  like this  $x_{mp}$  now each of these as well as these are observations on multiple variables observation on multiple variables what we want to define here.

(Refer Slide Time: 13:03)



We want to define here multi variate okay so where do so we know variable deterministic variable probabilistic data random variable and this is that this is one example why here every observation is measured on several variables then when multiple variables coming to picture then each observation is a variable work vector for example if I take the  $i$ th observation here then  $x_i$  will be  $x_{i1}, x_{i2}$  like this  $x_{ip}$  so it is a variable vector that is  $i$ th observation on  $p$  variables so we deal with this type of situation.

Where our observations or each of the observations had multiple values which of the variance is multiple values in the sense values are multiple number of variables more than one when the situation is multi variant situation now with different variable with different multi variant situation let us understand what is variant I think so institute out saying that  $x_i$  is like this if I create something different based on all those observations that then each linear combination, combination of variables.

For example here in this you know for example there are three variable  $x_1, x_2$  and  $x_3$  if I create a combination linear combination LC which is  $\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$  so these combining line will give a quantity or a value or other way we can also see a variable which is we are saying linear

combination variable which talk which is valid and this variate and then what is the definition of variate linear combination of variables with imperial determinates so that means  $\beta_1$   $\beta_2$  and  $\beta_3$  will be determined this tone observations there are in observations so will be able to be telling all those 1 flow linear combination of our weighted linear combination of variable square the weights are determined empirically that is variate.

Now in this case we can go for one variables into one variable that means if I say there are  $p$  variables we are going for one variable  $p = 1$  then that will be in variate when you go for  $p = >$  than equal to 2 that is multi variate that is which multi variate okay usually the statistics which we will be finding uni variate statistics for example in terms of normal uni variate normal distribution by variate normal distribution multi variate normal distribution so although by vary part of multi variate.

We basically talk about when uni variate  $p = 1$  by variate  $p = 2$  multi variate is  $p >$  than equal to so this what is multi variate by what multi variate we definitely talked about something about linear combination of variables so here more than one variable is there and there are multiple observations notice single observations in number of observations and which will be deleted for empirically based on the  $x$  observations in observation that will be collected from the population for which we want to in for something al those in parents are things will be discussed later well okay so third one the third.



(Refer Slide Time: 18:03)

Applied Multivariate Statistical Modeling

\* Pure science ← Basic science ← Laws, theories & their development.  
\* Applied science ← Experiment

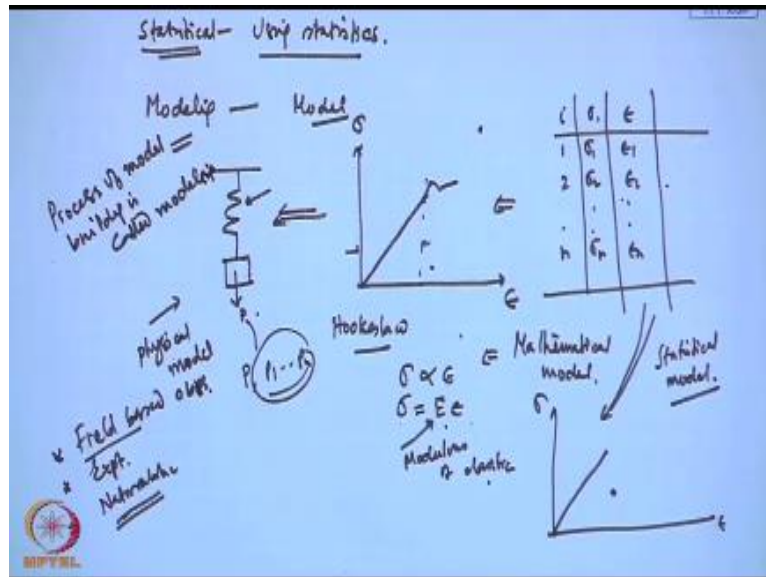
Normal distribution  $-k \cdot \frac{1}{\sigma} e^{-k \left(\frac{x-\mu}{\sigma}\right)^2}$   
 $f(x) \propto \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$   
 $-\infty < x < \infty$

Carl Friedrich Gauss.

The whiteboard contains handwritten notes on a light blue background. At the top, the title 'Applied Multivariate Statistical Modeling' is written in blue. Below it, a flowchart shows 'Pure science' and 'Applied science' branching from 'Basic science'. To the right, a normal distribution curve is drawn with a mean line and standard deviation markers labeled 'a' and 'b'. The name 'Carl Friedrich Gauss' is written below the curve. On the left, the normal distribution formula is written in two forms. At the bottom, there are small diagrams, including a box with a circle inside, and logos for 'NPT' and 'IIT KGP'.

Is statistical now.

(Refer Slide Time: 18:07)



What is statistical, statistical by statistical we want to say that it is basically using statistics using statistics that is what I want to whatever you are developing something using the statistical tools and techniques then these development of statistical is statistical development okay now what is statistics if I say that statistics is nothing but collecting then organizing then analyzing then representing and inter relating what I mean to say collecting data organizing data analyzing data representing the ranges and inter relating the ranges for the population.

For each the statistical model are the statistics which used is purpose full some purpose full work will be shaped so we when you talk about statistical so that means with talk about the a population then a sample consist of data from the population and we have some purpose in our mind objective in our mind we want to in for something from of the population and we collect data accordingly we organize the data we will analyze the data then we find the ragels that ragels we summarize and they told this summarization in that we findings we in for about the population so that is what if though.

That is why what the body statistical is used now last word but the important either modelip, modelip if you want to understand the first one is tell with model now model there are many

types of model actually very simple one is you know in school days I can remember we talk about this spring balance like this so what happen this is a spring the elastic one will only attached with this, this  $p$  and it beat in some way that we if you increase the load the elimination here will be more.

It will reduce it will be less so when this is a this the spring balloons module so in your toll show that be aware of the spring this type of physical model a table so this is one model which is basically with physical model which is a physical model okay now same thing when I came to by engineering studies I founded there is one, one important concept call or development of theory called hooks law where that  $\sigma$  the stage develop on the spring and the  $E$  longer than stream develop on it.

They are our modeled in such may not that there is a relationship like this and this is the range of electricity the another concept of electricity so what we I have seen there are well have seen their that  $\sigma$  varies so  $\sigma$  is  $E$  it is all on  $E$  is a modulus or modulus of electricity so this module theory we are here of elastic body when the load is so develop that which will not go to the  $n$  point or dual  $n$  point that in gymnastic so, so long the body is just praised within the within the elastic limit what is the perpendicular if you remove the load that hr go for back at to the original position.

So this type this development is possible because the feelings of this particular issue that spring is known and item wean say see if I know the illus models of electricity I be able to tell the relationship between  $\sigma$  and  $x$  alone and that time in any mechanic system the metrical subject we are all know this things this is mathematical model so in reality you will get different types of mathematical model so that mean what I am into say here that physical model in a mathematical model okay now what will be the statistical model in this case for example you take a case I think the.

In the beginning of the this particular study for example the hooks how the data probabilistic so bar as to experiment I have no idea but suppose you do you do not know the model of electricity but you know that say the steel body then you want to find out the relationship in that case you

and do experiment we seen  $p$  variant  $p$  from  $p_1$  to  $p_a$  so that means you will create in a different combination then you will be getting 123 in observations and  $\sigma$  and  $\epsilon$  values you will be getting  $\sigma_1, \sigma_2, \sigma_n$  the  $\epsilon_1, \epsilon_2$ , and then  $\epsilon_n$  okay, so now if you plot these what will happen you may get a plot like this, here it is  $\sigma$  and  $\epsilon$ . Essentially what is the difference between this and this here what I am saying I have stated without when you show, I have shown you this spring balance then I immediately say that is in it is elastic body, if it is elastic body this is the diagram because I know that is, this would flossing to me.

So, mathematics is known to me I am showing but in case it is not known I have done several experiment here and based on this I am trying to, I will do plot like this did not be the perfect state line you will get when you go for empirical model. So this is what is the empirical one model, so this empirical model when you talk about empirical model or like this experiment based or data based models like this, these are basically using statistical these are all statistical, this is, so for me this is your or for all of us this is our statistical model.

Okay, now what is modeling then modeling is basically you want to get this type of brases it is not that immediately will get all those things, there is a process, steps I have to understand what is my purpose, I have to understand in one or two fulfill the purpose what are the different variables they are effecting there, and I have to identify all important variables, then I have to see that how that data on the variable will be collected.

For example, here I have shown you the experiment but it may so happen that you cannot do the experiment. So in that case is there any other way of collecting data, for example naturalistic observation, so one is interested to see the prior of a particular animal. So we cannot do experiment maybe that is a wild animal but that there are large number of wild animal, animals of that particular spaces so we can observe that we are just going and observing field based, so okay so field based observation what this one is your experiment sometimes what happen we will go for some naturalistic observations which I talk about the wild animal case.

Field based observation mean the products go suppose the steel worst of case it both do the product some so you can see that what is happening there collect data, and accordingly you do

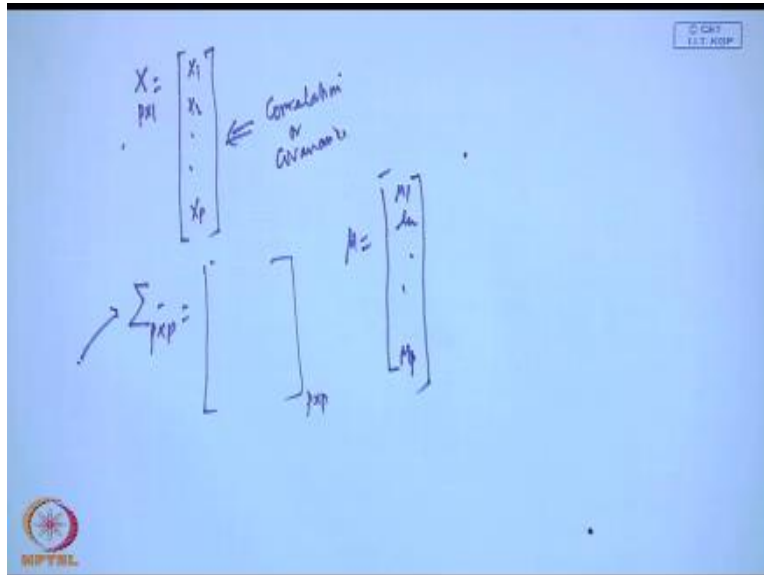
some modeling naturalistic observations, so all those type of data collection mechanism comes under empirical modeling, and you have to understand all those things, so there is a process the process a modeling, then the process model building is called modeling. The process of model building is called of model building is called model, okay. So let us see some of the slides now.

(Refer Slide Time: 27:51)

The slide features a black header with the word "Introduction" in orange. Below the header, a vertical line of four colored circles (dark red, red, orange, light orange) is connected to four horizontal orange-bordered boxes containing the following text: "Multivariate!!!! What's that?", "Why should I use it ?", "When to use it ?", and "How to do it? Model!!". In the bottom left corner is the NPTEL logo, and in the bottom right corner is the text "© Dr J Maiti, IIM, IIT Kharagpur" followed by a small number "3".

That I told what is multivariate and what is this it is discussed why should I use it and it is obvious question and with that why should I go for multivariate things, if I can do by some other way why multivariate. So there are some keys who is which basically will be known to you later on that when you talk about multivariate, we talk about multiple variables that is.

(Refer Slide Time: 28:20)



$p \times 1$ , if  $p$  is the number of variables  $x_1, x_2$  like your  $x_p$  now there is possibility that these variables are interrelated there is correlation one of the which correlation in between the variables, so that means you may get a correlation matrix or other way it is basically the covariance between the variables or covariance by covariance what I mean to say that in one variable vary there is a possibility that in particular way or some other way variable will also vary, then there will be covariance and standardized covariance is correlation this is will be discussed and in the subsequent lectures.

So covariance that will be  $p \times p$  that matrix will come, okay so all those things especially the mean values for all those variables  $\mu_1, \mu_2$ , like  $\mu_p$  this things will be there. Now so my answer to your question is that why should I use it because no physical process or a search any other systems also, which is characterized by multiple variables they should be analyst other way I can say the behavior should be analyst taking into consideration of all the variables characterized in it, when this variables consider is very, very important for that design development or improvement of the system for which it is developed.

And as it is obvious that there will be covariance or correlation between the variables, so if I go for univariate key analysis we will lose substantially information about the behavior because of known inclusion of the covariance structure. So we require to control this covariance structure and it is multivariate statistics covariance is very big issue and which will be one in multivariate distribution discrete I will be discussed in all those covariance things. So it is required of bracelet, it is required.

For example, for this case like our this one, steel versus this case, the steel versus three variables are basically controlling its quantity inner diameter, outer diameter and thickness there is chance that inner and outer diameter will be related and also the thickness, so in that a customer will not will be able to apply it or fit it to its own situation if there is huge mismatch. Now if I control individually inner outer or diameter or thickness then what will happen the correlation structure will not be considered and ultimately you will not be able to satisfy the customer.

(Refer Slide Time: 31:37)

**Introduction**

- Multivariate!!!! What's that?
- Why should I use it ?
- When to use it ?
- How to do it? Model!!

NPTEL

© Dr J Maiti, IIM, IIT Kharagpur 3

Okay, so you will be using multivariate and statistics or multivariate modeling when your system is complex in terms of number of variables it may be in conditions like this the correlation covariance structure is in depth, okay and you know where to extract those correlation

information you want to extract the pattern from this data that is why you will be using, okay. So how do I do it, it is two different models, so this models will be describe little late.

(Refer Slide Time: 32:33)

## An example

Sl. No.	Months	Profit in Rs million	Sales volume in 1000	Absenteeism in %	Machine breakdown in hours	M-Ratio
1	April	10	100	9	62	1
2	May	12	110	8	58	1.3
3	June	11	105	7	64	1.2
4	July	9	94	14	60	0.8
5	Aug	9	95	12	63	0.8
6	Sep	10	99	10	57	0.9
7	Oct	11	104	7	55	1
8	Nov	12	108	4	56	1.2
9	Dec	11	105	6	59	1.1
10	Jan	10	98	5	61	1.0
11	Feb	11	105	7	57	1.2
12	March	12	110	6	60	1.2



© Dr. J. Mani, IEM, IIT Kharagpur

Now what is next, next one example, here we are saying that a particular company operating may be in a city market, and we want to see the organizational health of this company with respect to profit in rupees million, with respect to sales volume in rupees 100 absenteeism, machine breakdown and M-ratio actually this is equate intentionally first one is profit and sales volume these are the organizational issue and that health.

If you sell more the profit may be more and if your profit is more your healthy in financially and another this we is absenteeism if you are paying substantially and if you are taking care the well being of the employees absenteeism will be less and if you are maintaining the health of the process here we are saying machine your machine breakdown will be less and if you are able to coordinate with customer as well as your supplier and your M-ratio that marketing ratio particularly if I say marketing ratio is related to the customer and that will be high.

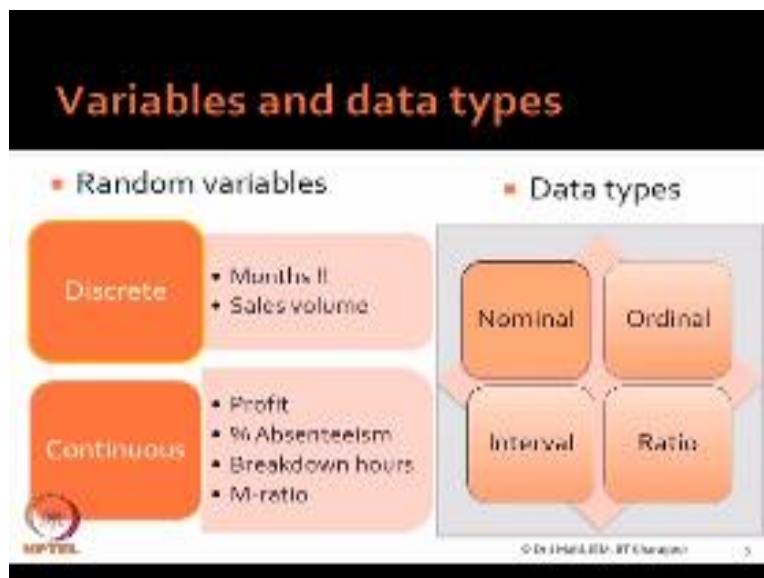


So if this is the case and then we are basically observing from April, May, June, July that 12 months data and in some units majored and this is nothing but a case of multi variant situation where are each of the row like starting from the one the first row these values are talking about multivariate observation for the month April for similarly for second year multi variable observation so there are two all multi variant observations.

Oaky now you may be could may be interested to know how profit variance over the months then it will be invariant one if you want to say that how sales volume variable the month it will also be invariant to one, if you want to know absentees in variance over the year a months that is also invariant like this.

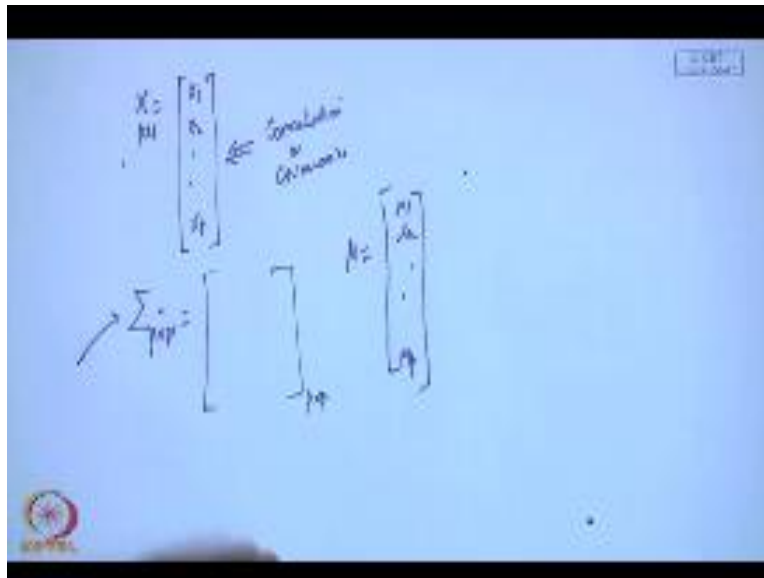
But if you are interstate to see that how the profit and sales volume go vary and they are own variation then you have to have to consider two variable and then this will be a multi variation situation sometimes you may interested to know that how sales volume will be dependent on absentees in breakdown and marketing ratio and then this will dependent model and that is also a multi variant issue. So this is inertial what we talking about multivariate observations.

(Refer Slide Time: 35:26)



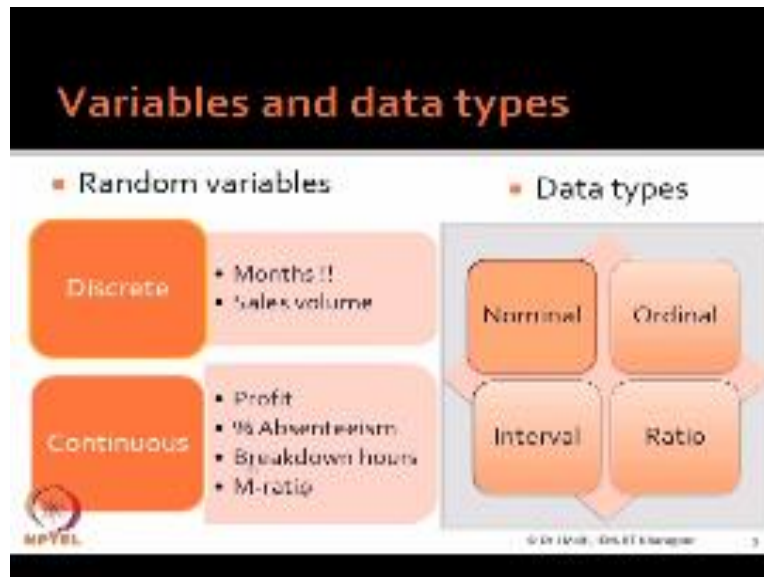
So now we have discussed some of the things and some of the variables interconsideration and we have seen that we have assigned them some values but how where those values are coming? For example if I say

(Refer Slide Time: 35:45)



The steel was a thickness that mean be the inner or the outer thickness so how it is known so you have used some measurements scale to measure these if I got to say that may be you have used for near caliph to major the outer diameter may be used calliper to the inner diameter so you have used some instrument and as well as there is a scale of measurement in this case this scale it is basically length which may be in terms of millimetre in terms of millimetre so you have to use some scale of measurement. And based on that scale used whatever data you gain those data will be of different types ok so.

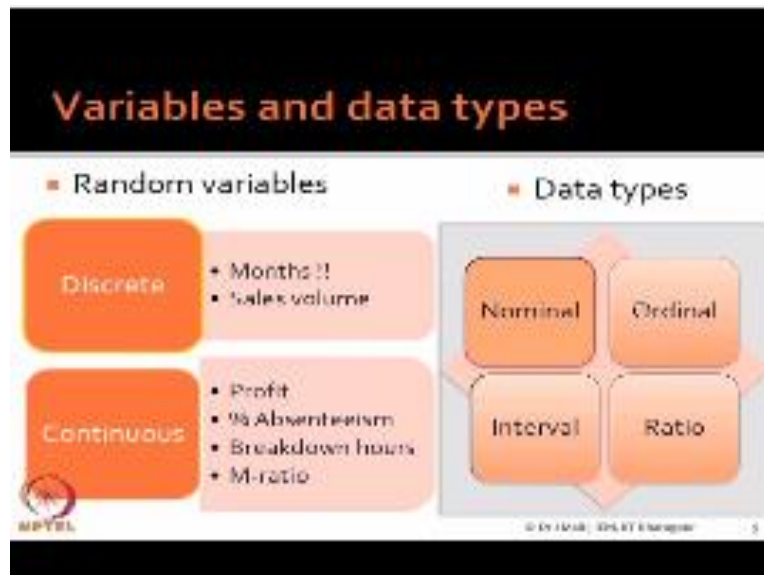
(Refer Slide Time: 36:36)



You see this slide here the left side is we are talking about random variables and right hand side we are talking about data types I have explained you these random variables earlier so I will not spend much time here but rather that you must understand one thing only that the random variable there will be discrete and contentious random variable by discrete random variable we mean to say they will take the some counted a count values like 0,1,2 or something like these or January, February march something like this ok so and our continuous space that profit absentees in breakdown hours in ratio here what is that any values possible.

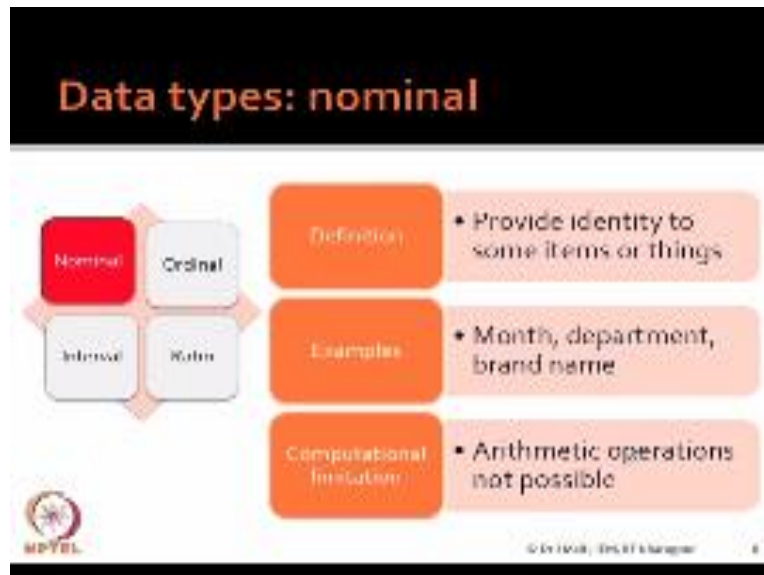
So please understand one thing here sales volume and coming under your discrete before it is a countable one but many countable such count values can also be considered contentious in many situations, but any how so there are two types now your data types I told you that what measurement scale your are using based on this data types it will be known means the data will be having certain properties because data is nothing but information ,how much information is available in a data getting me ?so it all depends on what scale you used to measure this data so based on that

(Refer Slide Time: 38:08)



There are four types of data one is dominal data ordinal data interval data and ratio data.

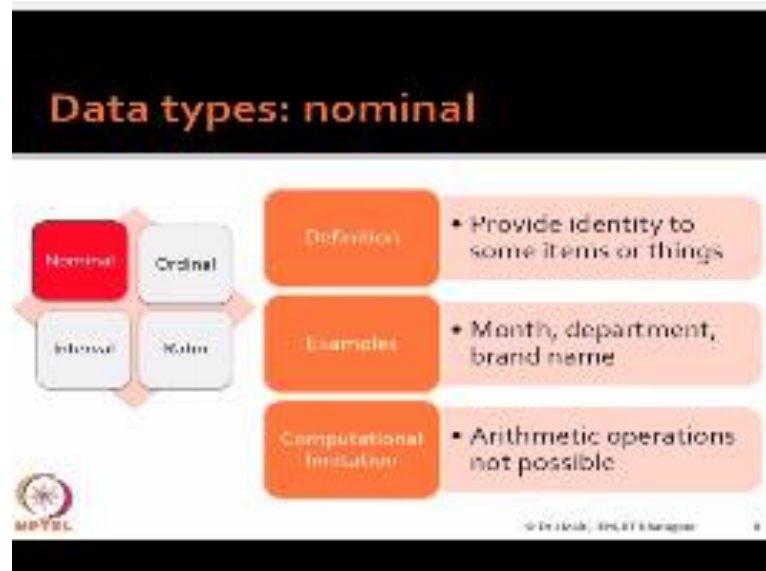
(Refer Slide Time: 38:16)



Let us discuss something about nominal data my definition it is provide identity to some items or things if I see the month the company small component that that business I shown you that they want to know how hard the different month what is the status so if the month is variable starting from January to December because it changes.

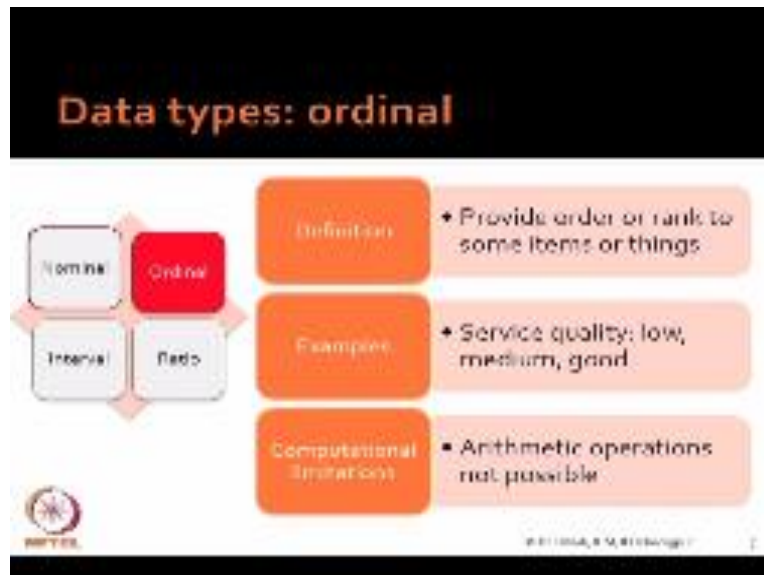
So then it is January and February all those things nothing but they are the identity of the period of time identity of particular period suppose you just think of you are trying to know the some performance or status of the different dependence of for example IIT so then if I say IIT is the department of chemistry department of physics department of mathematics department of computer science department of industrials and systems industrial engineering management so all those things they are basically providing identity which sometimes require this type of data to include in our analysis.

(Refer Slide Time: 39:28)



So this is nothing but nominal data now what is the problem with nominal data problem in nominal data is that there is use computational limitations you cannot do any arithmetic operations we cannot add the department of chemistry plus department of physics like this we cannot say department of chemistry is one and department of physics is two and accordingly we will add we cannot subtract we cannot multiply we cannot make divisions also so this is the problem.

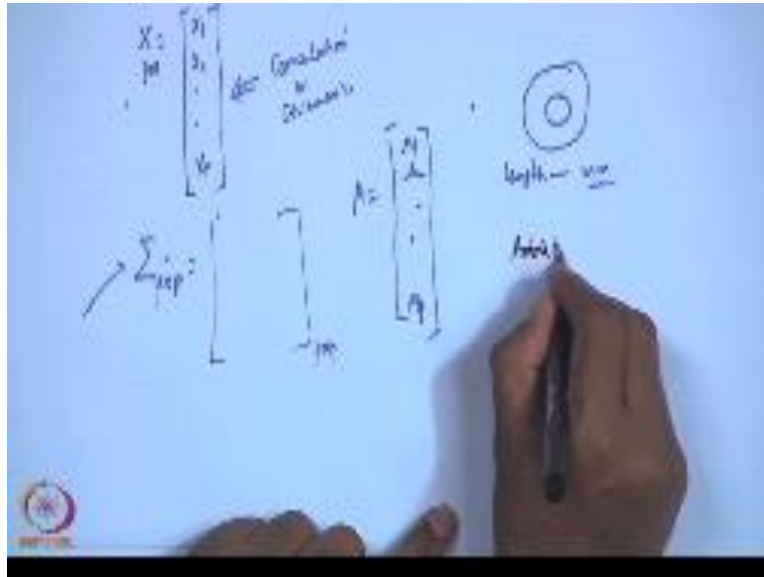
(Refer Slide Time: 40:05)



Next data type is your ordinal data type what is ordinal data type? suppose you have seen that you have travelled in flights several times maybe on train or some other place if you go to restaurants and where you have taken food and you might have seen that you are giving feedback from there asking that please read the in case suppose the hotel food qualities are dish quality, room quality all those things in terms of no satisfactory of totally unsatisfactory find to extremely satisfying in this type of scale we have used.

For example, for the food case your case okay taste wise very good, good or something like this so this type of ordering when ordering is there this is called ordinal data ok what it does it basically provided some order rank to some items on things examples are dis-quality low medium or good and computational limitations here also we cannot do any arithmetic operations like your addition.

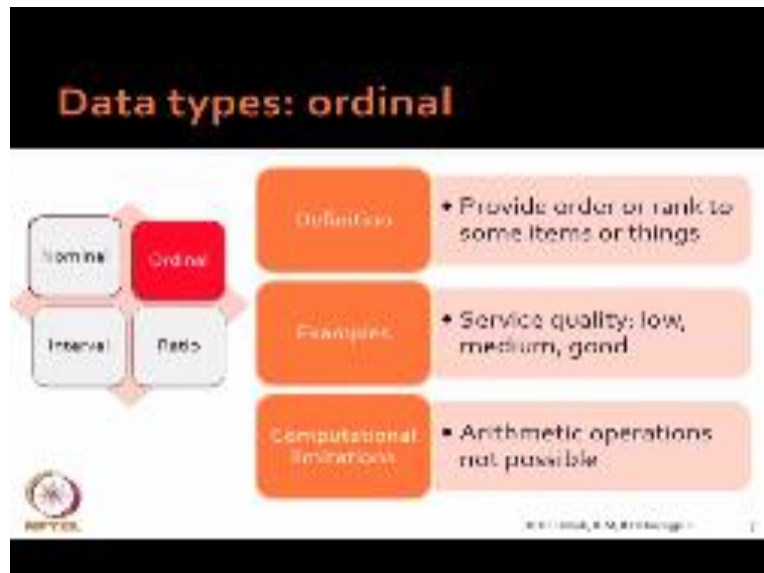
(Refer Slide Time: 41:17)



Subtraction multiplication, and division and you cannot do then what way it is better then.

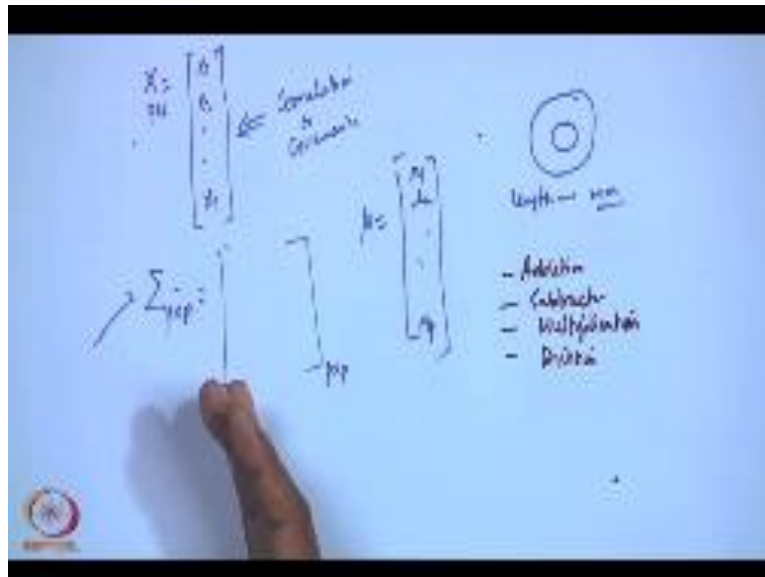


(Refer Slide Time: 41:34)



Nominal data because here you are getting order a rank you are getting if I say the performance my student performance is low average very good excellent like this the person who is getting excellent is definitely better then the person or the student who got very good.

(Refer Slide Time: 42:02)



So I have a ranking scheme here ranking ability with this data ok the ordinal data is rich compare to nominal data.

(Refer Slide Time: 42:18)

The slide is titled "Data types: interval" in orange text on a black background. Below the title, there is a diagram showing four data types in a 2x2 grid: Nominal, Ordinal, Interval, and Ratio. The "Interval" box is highlighted in red. To the right of the grid, there are three orange boxes with definitions and examples:

- Definition:** Provide continuous data in storage
- Examples:** Temperature
- Comparison and operations:** Use same unit, provide

At the bottom right, there are two thermometers. The left one is labeled "Fahrenheit" and has markings at 32°F, 70°F, and 100°F. The right one is labeled "Celsius" and has markings at 0°C, 20°C, and 100°C. Horizontal lines connect the 0°C mark to the 32°F mark, the 20°C mark to the 70°F mark, and the 100°C mark to the 212°F mark, illustrating the mapping between the two scales.

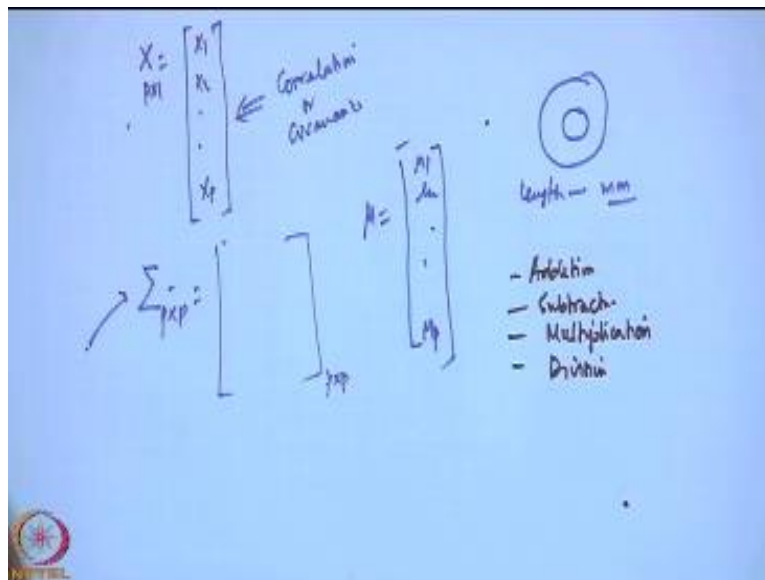
Next data type I said that interval data what is interval data it is basically a well understood if he take this example here temperature you are measuring two scales one is Celsius and another one is Fahrenheit in developing these two scales for example Fahrenheit scale as well as your Celsius scale the reference point is taken in two different places means locations so it is not the same you getting me.

So and if you see the horizontal lines here you see that  $0^{\circ}\text{C}$  to  $100^{\circ}\text{C}$  then the corresponding Fahrenheit range 32, 70 and 212 Fahrenheit fantastic. So and there is a range when if I say the difference from 100 to  $0^{\circ}$  you are getting this range here also that one tend to 12 to 32 the corresponding range is this so that we weather we measure using Celsius scale or Fahrenheit scale so we will be getting the equal range.

Now what will happen suppose I measured a temperature today's temperature day temperature is  $20^{\circ}\text{C}$  today about  $30^{\circ}$  and may be day after tomorrow again  $21^{\circ}$  then I find 1 to average in I can add them and then divided by 3 that average I will get if I do the same thing in parameter so it is possible I can do the similar thing I can do but what will happen suppose I want to say that what is the how many times temperature of today is compare to the tomorrows each standard

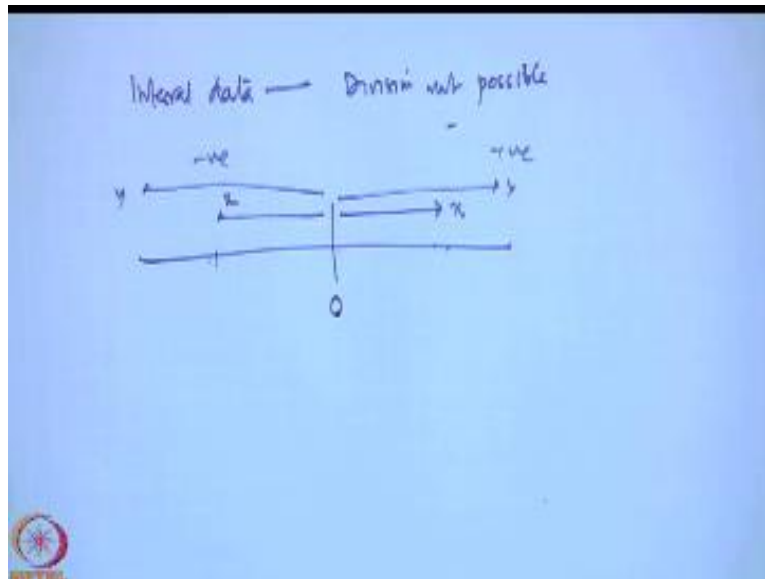
temperature then if I use Elsevier scale at in 5 degabged suppose if 22/20 and then here it will be maybe 70 and some other things then you will find out that they are not natural. So that means internal scale is some scale were you will get interval data range data.

(Refer Slide Time: 44:39)



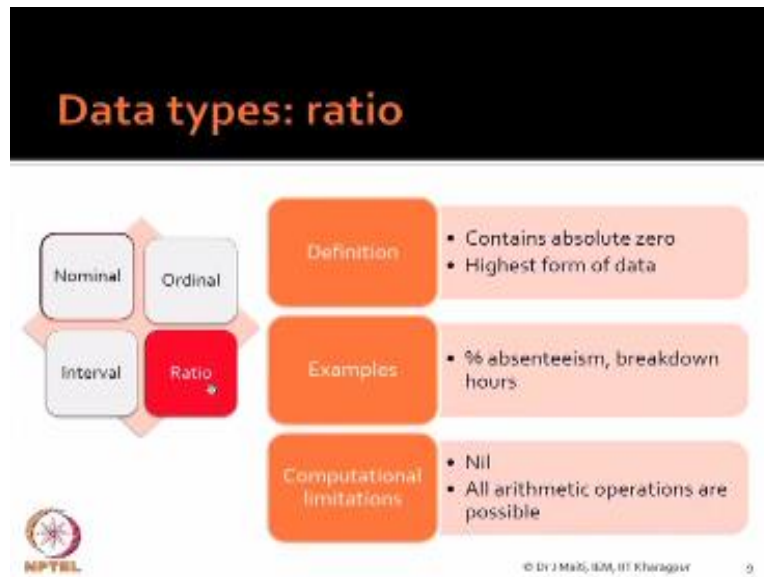
And they are all they having all types of continuous properties except at they can do three arrange mating operations very easily addition subtraction and multiplication but when do division you will find out the those you use changes the scale ultimately what will happen you will find that they are not match.

(Refer Slide Time: 45:17)



So in interval data you cannot go for division all other automatic operations possible let us go to the next slide.

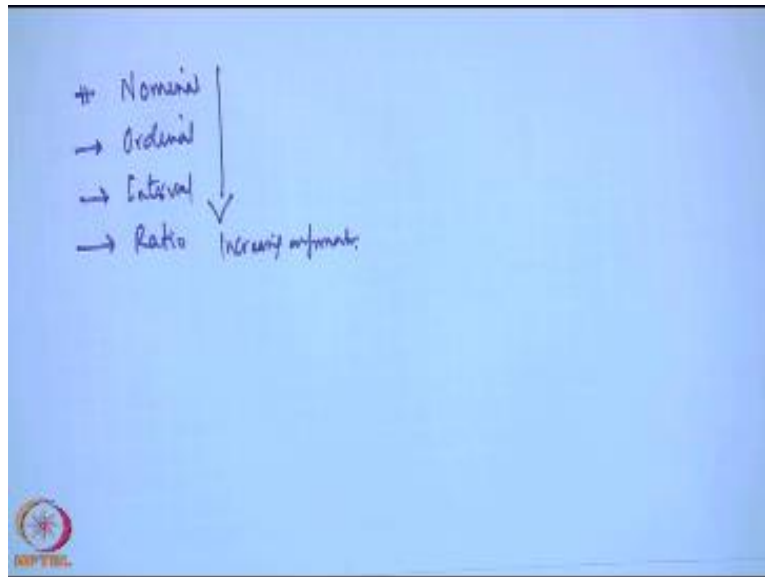
(Refer Slide Time: 45:37)



That is we are talking about ratio data ratio data is something where there is absolute 0 in the scale of measurement this is 0 if I move towards right it is x amount and towards left also x amount then the difference this difference is same if I go for y also this side also y also that is same so that mean if you go in the to the left it is negative this side it is positive but there is absolute 0 in between okay so this 0 is the reference point not in terms of that the permanent centigrade scale that were is the two different in corresponding.

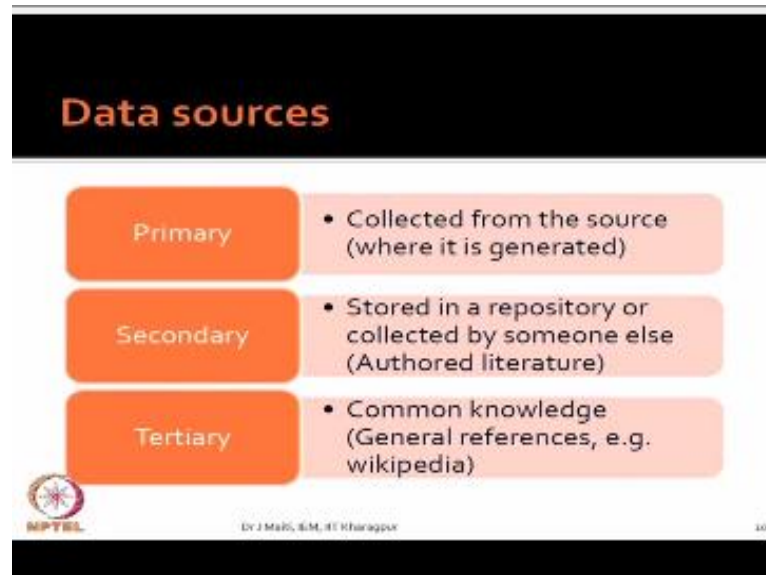
So definition is constant absolute 0 highest form of data sorry so the ratio data is constant absolute 0 highest form of data example absenteeism breakdown hours for drive zone and earlier and computation limitation in all arithmetic operations are possible here, okay so no if I go by the order of information available then definitely.

(Refer Slide Time: 47:07)



Your first one is if it is nominal then followed by ordinal then your interval then you ratio then definitely in order of increasing in population this is the case. My base data is this next base is this next base is this and the is the lowest form of information data okay.

(Refer Slide Time: 47:43)



So you know that different data types now you know that as you will apply multi variant mode statistical modeling you must require collect data so you require to know that data sources so primary data collected from the sources where it is generated for example in the case of steel vases example if you collect data where it from the production shop just going they are collecting data or that is what is known at primary data.

Okay suppose you want to see the behavior of the animals in the jungle when observe and then accordingly no down and that will be your primary data okay so for the production that city can that example profit may have come and sales volume case that is also primary data so long you are collected from the source. What is secondary data is stored in repository or collected by someone else getting me so you are not collecting it is already there.

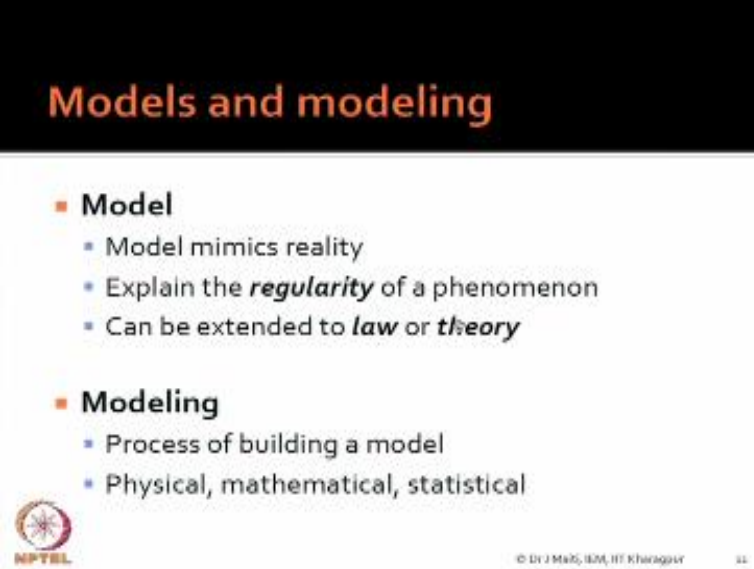
We have different sources for example you will make it the financial data from some sources and suppose company is maintaining records of their production and suppose the maintenance or the health oblations or many things okay so you have not collected so company is to would have gone there and collected this things or it is better that in literature you studying in something in your own area and found that a paper is there were some data is are giving.



So this type of data is secondary but secondary data must have must be arithmetic in the sales that reference of the data available authored references are they are this are authored literature data but this is definitely as it is done by somebody else it is not primary that you have to rely on the arithmetic city on the data collected by somebody else. Then Tertiary data which is basically a common knowledge type of things okay suppose you will keep it here you will find many things are there.


Actually when in terms of modeling when you start with the subject area you start with this that when your knowledge is not that much clear and you will start with Tertiary sources and then slowly go to secondary source and then finally when you do actual work you may will go for the primary data sources.

(Refer Slide Time: 50:27)



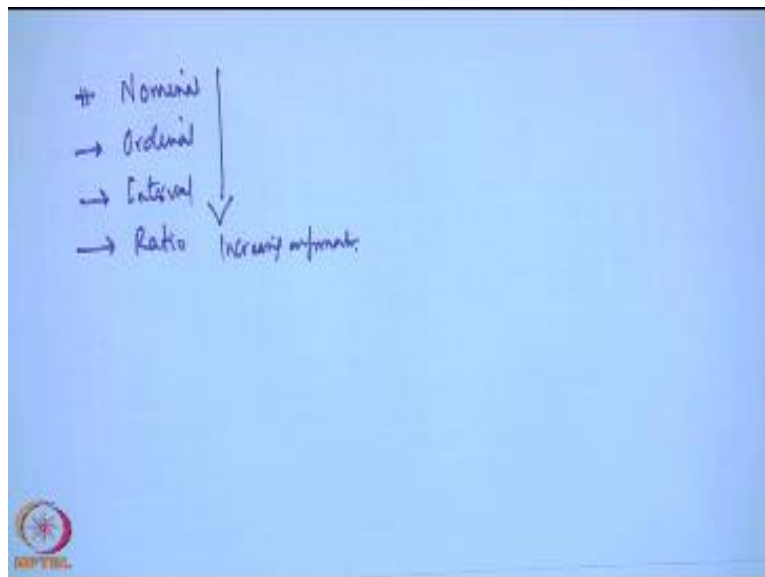
**Models and modeling**

- **Model**
  - Model mimics reality
  - Explain the *regularity* of a phenomenon
  - Can be extended to *law* or *theory*
- **Modeling**
  - Process of building a model
  - Physical, mathematical, statistical

 © D17 MIT, IIT Kharagpur 22

I told you earlier what is model let me repeat this again that model mimics reality when if you develop a model that without considering the reality real thing you are not doing the justice. So model mimic reality so it should be it should have real applications that is what is the meaning for example suppose you think of a car.

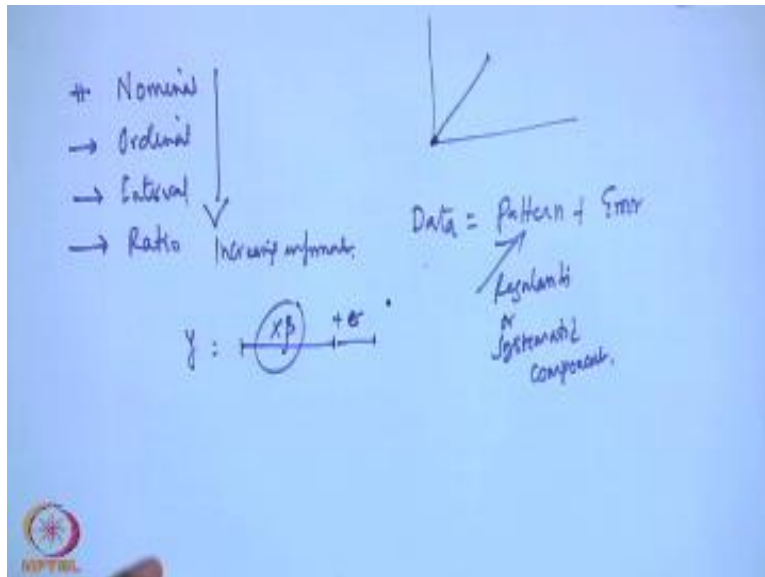
(Refer Slide Time: 51:04)



Which is what is by suppose any automobile company and they develop this things what I mean to said the develop a model simulation model in computer first before going for developing the car one after another manufacturing the car in the manufacturing shop what that there must be some simulation model okay and means how the car will work so that type of thing are known as that means it is in terms of the reality the car, car is the real thing.

So you are modeling some things so let us explain balance example also that the mathematics is related to the elastic behavior of the that is the reality, okay in statistical terms when we talk about the how sells volume is depended on other things like this you have seen digamous ratio all those things that also going to the talk about the reality so actually if statistical sense a model talks about explain the regularity of if endow manner, okay.

(Refer Slide Time: 52:25)




In Hooks law the regularity so long it is within the elastically meet when the load is relived it will come back to the original shape that is the regularity, in case of our statistical model building we talk about data and data is nothing but equal to this paten plus error this pattern is the regularity pattern is the regularity or systematic component okay so you, we must know what is our problem and accordingly what data you will collected and you want to extra pattern from this data okay.

In case of prednisone model suppose you want to predict some  $y$  value then with respect to some  $x$  values and then you will find out there is some linear combination of variable that is  $x\beta + \text{error}$  will be there this is my regularity or pattern this one is the error, okay now when you repeat that similar development under different situations then what will happen if it performs well under the different situations for which it is develop then one day we may say it is a law or the theory like Hooks law or the Hooks law is this one the elastic that elastic theory okay, so will all know that neutrons large emotions and we all know that.

(Refer Slide Time: 53:59)

## Models and modeling

- **Model**
  - Model mimics reality
  - Explain the *regularity* of a phenomenon
  - Can be extended to *law* or *theory*
  
- **Modeling**
  - Process of building a model
  - Physical, mathematical, statistical



© Dr J.Maiti, IEM, IIT Kharagpur 33

Dalton's atom theory and many other things, not that only everything is developed and if you full accepted, it basically developed, tested and verified, validate after several years and then other side, that is the research has accepted the fact and then it will applied to the different situations and founded it is working. Okay, I told you modeling also, process of building a model physical, mathematical and statistical which is already I have explained to you.

(Refer Slide Time: 54:53)

## Principles of modeling

- Do not build a complicated model when a simple one will suffice
- Beware of modeling the problem to fit the technique
- The design phase of modeling must be conducted rigorously
- Model should be verified prior to validation
- A model should never be taken too literally



Source: Ravindran et al. (2006)

I hope that you got the glimpse of, actually the purpose of applied multi varied statistical modeling, and actually we want to develop empirical modeling. Okay those empirical models, these are all data based, data based in the sense is given, you have the data and you are going for model, build your building model and to find out the regularity of the data of the patent of the data, to show that you will be able to describe the relationship of the behavior of the population or system in consideration.

You will be able to establish the state of the relationship, you will able to predict something, you may be interested to prescribe something also, but when you talk about statistical modeling, usually this is the description and prediction first important things. Description, explanation and prediction. Ok these three things come into consideration. So, you will be knowing different types of statistical model together and you will be tempted to develop different models also, based on the data whatever available to you.

But before going for modeling or applying any statistical techniques, what is happening? What we want to say that, you have to have some principles in mind, before going for this. Here I am just jotting down some of the principles, which I have taken from the book by Ravindran et al.

You just see, what he said 'Do not build a complicated model when a simple one will suffice'. For example, if I know the meaning of different, lots of mean value of a particular constraint, for example the inner diameter of different lots for this too okay. And if that suffice by purpose, go for mean or utmost we require the standard deviation mean of the inner diameter that differ by the different suffice I told you.

So there you may, you do not go by instruction, many other things it may not need mean data. So you do not go, If it is needed you go. Beware of modeling the problem to fit the technique, many times, I have seen in my case. It is there one model which will be discussed later on known as structural equation modeling.

People are using structures, where simple regression model can be ok. But people are interested to fit the structural equation model. So please be little bit of cautious all those things, that given model is for problem solving and model comes for the problem not fit a statistical technique. The design phase of modeling must be conducted rigorously and it will be discussed later, what you mean by design phase, coming under study design.

Model should be verified where to validation, verification means, suppose you where you collect data, you would split the data into two halves what for trading or there for test or other way you can say, one side for model building and another side for verification. And validation basically talks about when you take some new data again and you find it is working that is validation.

A model should never be taken too literally; many times what I found that, if is the model ok there are more variables, statistics is taken very loose. Say there are many variables. Let us find the relationship is there or not. These type of or whatever variable is there let us find that relationship without considering the purpose.

(Refer Slide Time: 59:21)

## Principles of modeling

- A model should neither be pressed to do, nor criticized for failing to do that for which it was never intended
- Beware of overselling a model
- Some of primary benefits of modeling are associated with the process of developing the model
- A model cannot be any better than the information that goes into it
- Model cannot replace decision makers



Dr. J. Maiti, IEM, IIT Kharagpur

Source: Ravindran et al. (2006)

13

A model should neither be pressed to do, nor criticized for failing to do that for which it was never intended. For example you are interested to see the relationship between several variables of a particular population. Now, later on you want to see that, how I want to predict something, See if developed model to see the patterns and state the relationship and not to predict. Show how can your model will predict, which was not intended for, so that is a another issue, so if it fails to do prediction when it was just to understand the coherent structure.

Then you should not criticize for this or you should not praise the model to be it. Beware of overselling a model. Many times we basically mix shore of; I can say recommendation based on the model and many of the things are basically from common sense, and so that type of selling is prohibited. Some of primary benefits of modeling are associated with the process of developing the model.

So see as all of you are learning multi varied statistics, so do not think that always you will be doing something great in type of modeling, you are learning, so the learning process while you develop something, you know the physics of the problem, may be you know the process to which

data generated, you how the data to be captured, how the data to be analyzed, what basis is applicable with, so this the entire gamet, the gamet of the process is very, very important.

So very, very fits have come out of it. A model cannot be any better than the information that goes into it. So you cannot say that you are using nominal data and you will be basically talking about the model of regression, where y variable is nominal, so you have to go for the some other type of model that may be that logistic regression. So the information, what you are the quality of information what is fit into the model, that is important.

Because if the input is not good, output also you should not expect good, so model cannot replace decision maker. So you cannot think that your model is superior to you. The decision maker the analyst who has been the system knowledge, that do be next part, so they are more important paper. So whatever you develop whatever you do, for what purpose you are developing all these, so that is your brain, if the group who are working they do better than any model.

So in this case, what I want to say that, you please, take all those issues what I have discussed now, the principles particularly, in this stage seriously and accordingly develop the model and today it is up to these, and next class we will be starting the statistical approaches to problems solving. Thank you for your patient hearing.

**NPTEL Video Recording Team**

**NPTEL Web Editing Team**

**Technical Superintendents**

**Computer Technicians**

**A IIT Kharagpur Production**

**[www.nptel.iitm.ac.in](http://www.nptel.iitm.ac.in)**

**Copyrights Reserved**