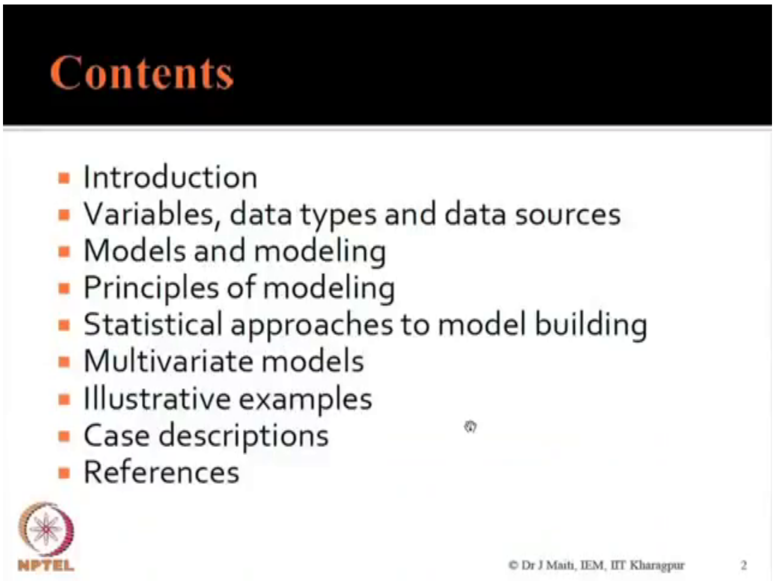
Applied Multivariate Statistical Modeling

Prof. J. Maiti
Department of Industrial Engineering and Management
Indian Institute of Technology, Kharagpur

Lecture - 1
Introduction to Multivariate Statistical Modeling

Good morning, welcome to the first lecture of applied multivariate statistical modeling. Let me tell you the content of this today's presentation.

(Refer Slide Time: 00:34)



Contents

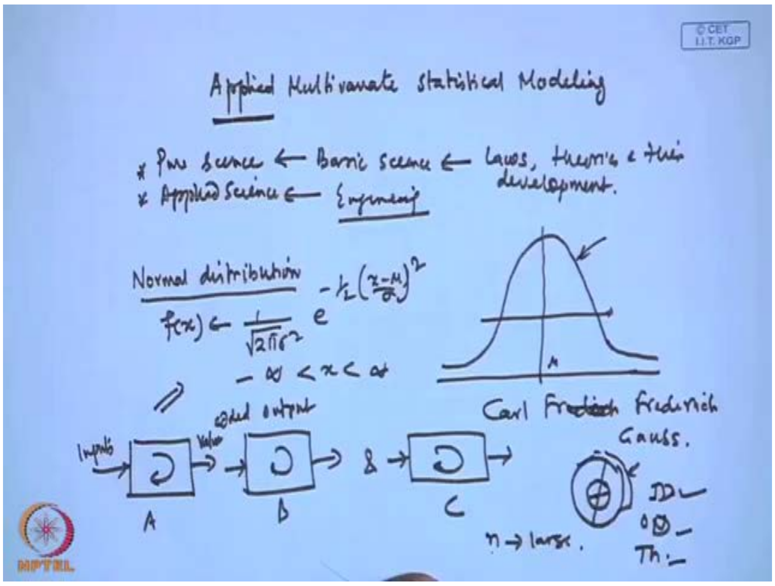
- Introduction
- Variables, data types and data sources
- Models and modeling
- Principles of modeling
- Statistical approaches to model building
- Multivariate models
- Illustrative examples
- Case descriptions
- References

NPTEL

© Dr J Maiti, IEM, IIT Kharagpur 2

So, we will start with introduction, then variables, data types, data sources, models and modeling followed by principles of modeling, statistical approaches to model building, multivariate models, some illustrative examples, three cases followed by references. The entire content will be covered in two hours.

(Refer Slide Time: 01:20)



Applied Multivariate Statistical Modeling

* Pure Science ← Basic science ← Laws, theories & their development.
* Applied Science ← Engineering

Normal distribution

$f(x) \leftarrow \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

$-\infty < x < \infty$

Carl ~~Friedrich~~ Friedrich Gauss.

Today, I will try to finish up to principles of modeling, let us start with defining what is applied multivariate statistical modeling? Let us define whatever you want, first is applied. Now, 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 it basically talks about laws theories and their development, and their development, definitely it links with the phenomena, which we usually observe in different aspects of our life. Now, applied science which will use the knowledge of the pure science and develops 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.

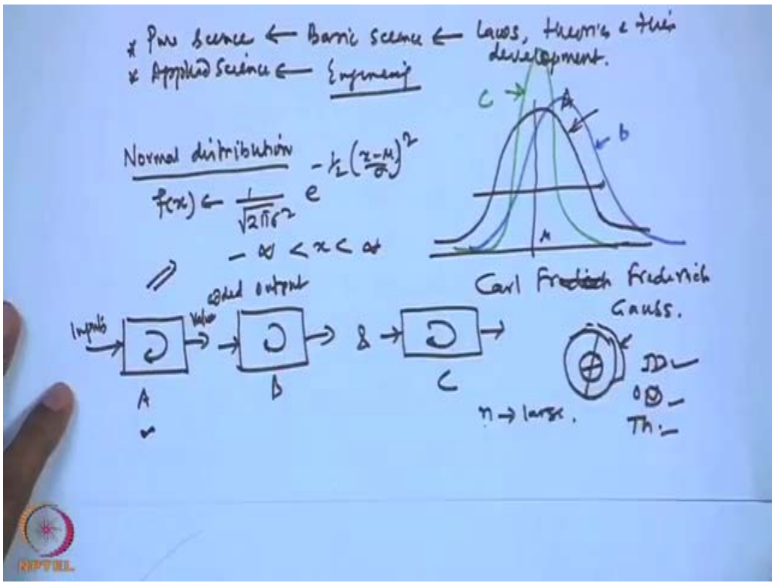
Now, when I talk about applied statistics what do we mean? I am assuming that you have knowledge on preliminary basic statistics for example, normal distribution. If you know normal distribution then also you know the probability density function f x, which is 1 by root over 2 pi sigma square e to the power of minus of x minus mu by sigma square, where x varies from minus infinity to plus infinity. This is the so called this bell shaped curve which is developed by Carl Friedrich Gauss.

So, theoretical development so that mean in development of this type of distributions it is coming under basics. Now, suppose if I want to apply this knowledge to a real life situation, I can find out a situation like this for example, let us there are three processes, process A B and C take certain inputs, convert into value added outputs, value added outputs all cases. Let there are basically three identical machines which is producing steel washers, steel washers will be shape like this where there is inner diameter ID. There is outer diameter OD as usual as there will be certain thickness of this washer so I can say T h.

Now, if you produce a large amount of steel washers that means the number of items produced is large, n is large then the quality characteristic or the characteristics of the steel washer which is important to the people, the customer ID. If you plot you may get this type of distribution, which is normally distributed and where you will be getting mean here. There will be definitely standard deviation for ID.

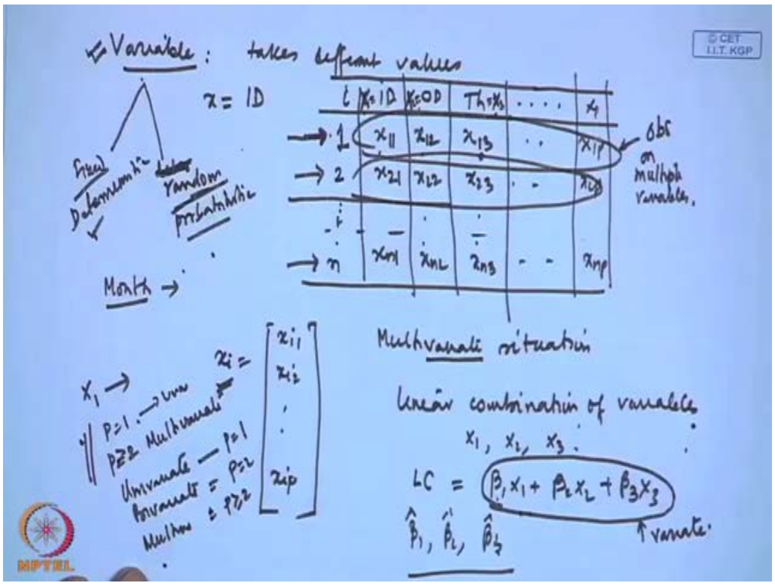
Similarly, for OD similarly, for thickness now then what you are doing by what is applied here? The production process A for example, in this case which is producing steel washers each is converted into a statistical process. In the sense in terms of a distribution like normal distribution, where we are saying that the production process can be interpreted, the behavior of this process can be interpreted like this now in order to get it further clarified.

(Refer Slide Time: 06:45)


* Pure Science ← Basic Science ← Laws, theories & their development.
* Applied Science ← Engineering
Normal distribution
$f(x) \leftarrow \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
$-\infty < x < \infty$
C
A
b
μ
Carl ~~Friedrich~~ Frederich Gauss.
Inputs
Value added output
A
B
&
C
ID
OD
Th
n → large.

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 is this one for production process C, then using these things you will be able to compare. Compare A B and see their performance in terms of mean and standard deviation. There is possibility also to see that whether the mean ID produced by C is equal to that of B or A, this type of comparisons and things possible. So, when we actually when we develop something which will be useful to the society for the mankind, then we say it is applied. Now, come to the second word which is basically multivariate.

(Refer Slide Time: 07:53)


Variable : takes different values
x = ID
Fixed
Deterministic
random
probabilistic
Month →
i
X = ID
X = OD
Th = X3
Xp
1
2
n
x_{11} x_{12} x_{13} x_{1p}
x_{21} x_{22} x_{23} x_{2p}
x_{n1} x_{n2} x_{n3} x_{np}
obs on multiple variables.
Multivariate situation
$x_1 \rightarrow$
p=1 → uni
p≥2 Multivariate
Univariate — p=1
Bivariate = p=2
Multivariate ≥ p≥2
$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{bmatrix}$
Linear combination of variables
x_1, x_2, x_3
$LC = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$
↑ variate.
© CET I.I.T. KGP
NPTEL

Now, in order to understand multivariate we 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 I D, x is I D inner diameter. Then if I produce one item, I stands for the item suppose, first item and the I D value it may take value X 1. When we go for second version it may take X 2. So, if I such way if I go for n washers produced, let X n will come into consideration. So, these are the values so I D takes different values as a result I D is a variable here. Now, in statistics we basically talk about two types of variables, one is fixed variable and the other one is deterministic, sorry 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 probably here but we know all the months. Suppose, what will be the next month is this month is your December next month will be January, it is known with certainty that is a deterministic model, but in this case when you are going to produce a second lot. Suppose, in the second lot even in one lot what is the value of I D for the second item, or second version it is not known with certainty, it is governed through probabilistic distribution.

So, that sense that it is random one, we do not know the value exactly and this value is coming based on certain random experiment. In this case the process which is producing this item so if I go on saying like this then other variable here is O D. Similarly, other one is our thickness, now in order to accumulate more than one variable, we will write this X 1 is I D, X 2 is O D and X 3 is X 1, X 2 and X 3 is thickness the for the first of first item was produced. Then this will be x 11, second one x 2 1 and n one. Similarly, for O D x 1 2, x 2 2 like x n 2, and if I go for the X 3 variable that is observed for first observation, it is x 1 3, second one x 2 3. So, like this x n 3.

So, what we are trying to say here that we are considering three variables X 1, X 2, X 3 which are nothing but the characteristics of the steel washers in this example which has inner diameter, which has outer diameter, which has thickness. Now, if you produce n number of washers then what will happen? Every washers will be having different values for I D, O D and thickness. So, this is my observation, first one is observation number 1, second one observation number 2, like that there is observation number n and you see in observation number 1 if I consider only I D that value is x 1 1, if I consider all three together, observation 1 takes value x 1 1, x 1 2, x 1 3.

So, similarly if you go on increasing the number of variables up to X p then here it will be X 1 p, X 2 p like this X n p. Now, each of these as well as this, these are observations on multiple variables. What do you want to define here? We want to define here multivariate. So, in order to do so we know variable, deterministic variable, probabilistic, that is random variable and this is one example where every observation is measured on several variables. Then when multiple variables come into picture then each observation is a variable vector example, if I take the ith observation here then x i will be x i 1, x i 2 like this x i p.

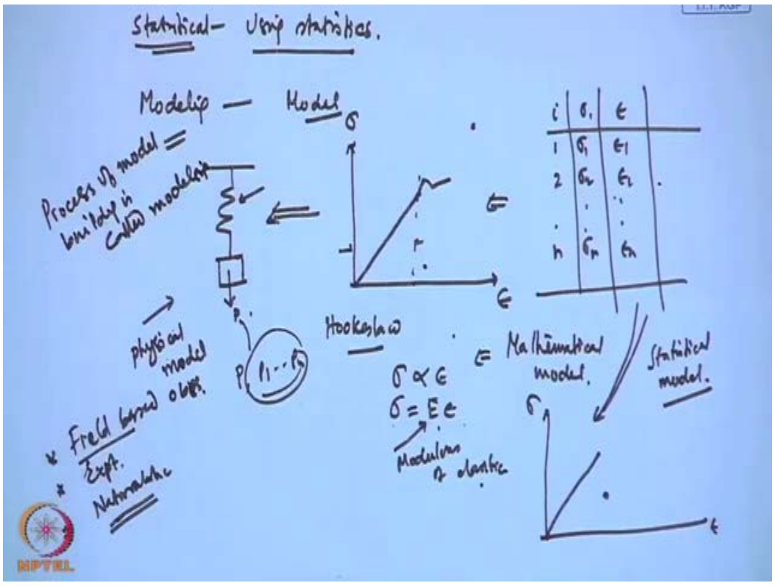
So, it is a variable vector that is ith observation on p variables. So, when we deal with this type of situation where our observations or each of the observations have multiple values in the sense values on multiple number of variables more than 1 then the situation is multivariate situation. Now, we define variable, we define multivariate situation, let us understand what is variate getting me? Instead of saying that x i is like this, if I create something different based on all those observations that there is linear combination of variables.

For example, here in this in our example there are three variable X 1, X 2 and X 3. If I create a combination linear combination L C which is beta 1 X 1 plus beta 2 X 2 plus beta 3 X 3. So, this combiningly will give a quantity or a value or other way we can also say variable which is we are saying linear combination of variable which is variate, this is variate and then what is the definition of variate? Linear combination of variables with empirically written mean weights, that means beta 1, beta 2 and beta 3 will be determined based on observations. There are n observations so we will be able to determine all those variables.

So, linear combination of or weighted linear combination of the variables where the weights are determined empirically that is variate. Now, in this case you can go for one variables, simple one variable that means if I say there are 3 variables, we are going variable p equal to 1 then that will be univariate, when we go for p equal to greater than equal to 2, that is multivariate. That is what multivariate usually in statistics books you will be finding univariate statistics. For example, in terms of normal univariate, normal distribution bivariate, normal distribution multivariate, normal distribution, so all the bivariate is a part of multivariate, we basically talk about when univariate means p equal to 1, bivariate means p equal to 2, multivariate is p greater than equal to 2.

So, this is what is multivariate, by word multivariate we definitely talk something about linear combination of variables where more than one variable is there, and there are multiple observations, not a single observation, n number of observations and weights. We determined empirically based on the X observations n observations that will be collected from the population for which we want to infer something. All those inference we will discuss later. So, third one, the third issue is statistical. Now, what is statistical? By statistical we want to say that it is basically using statistics that is what we want to infer.

(Refer Slide Time: 18:11)



i	σ	ε
1	σ_1	ϵ_1
2	σ_2	ϵ_2
.	.	.
n	σ_n	ϵ_n

So, whatever you are developing something using the statistical tools and taking it, then this development is statistical development. Now, what is statistics? If I say statistics is nothing but collecting, organizing, analyzing, then representing and interpreting. What I mean to say collecting data, organizing data, analyzing data, representing the results and interpreting the results for the population for which the statistical model, or the statistics is used for some purpose, some purposeful work will be served.

So, when we talk about statistical that means we talk about the 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 infer something from of the population and we collect data accordingly we organize the data, we analyze the data, then we find the result and the result we summarize, and based on this summarization these findings we infer about the population so that is what is the word statistical is used.

Now, last two are but very important one is the modeling, if you want to understand then first you understand this model. A model there are many types of model actually very simple one is in our school days. I can remember we talk about the spring balance like this, so what happened this is a spring, a elastic one, a load is attached with this is P and it behave in some way, that behavior if you increase the load, the elongation will be more. If you reduce it will be less.

So, when this is the behavior, this is the spring balance model so to show the behavior of the spring this type so physical model are developed. So, this is one model which is basically a physical model, which is a physical model. Now, same thing when I came to my engineering studies, I found that there is one important concept called or development or theory called Hooke's law, where that sigma he stress developed on the spring.

And the elongation strain developed on it they are modeled in such a manner that there is a relationship like this. This is the range of elasticity, there is another concept called elasticity. So, what I have seen there or we all have seen there that sigma epsilon. So, sigma is E epsilon, where E is young modulus or modulus of elasticity. So, this is what is the theory behind the for elasticity, the area of elastic body when the load is so developed that each will not go to the yield point or beyond yield point, that is elastic zone.

So, for so long the body is stressed within the within the elastic limit, what will happen to the property that if you remove the load then it will recover back to the original position. So, this development is possible because the physics of this particular spring was known and I can say if we, if I known the modulus of elasticity, I will be able to tell the relationship between sigma and epsilon. And that time in engineering mechanics and strength of materials subject we learn on these things, basic mathematical model.

So, in reality you will get different types of mathematical model so that means, what I mean to say here that a physical model, a mathematical model. Now, what you mean by statistical model in this case for example, you take a case I think the inner beginning of this particular study for example, the how did we develop all these things.

So, to experiment I have no idea but suppose you do not know the modulus of elasticity, but you know that say elastic body and you want to find the relationship in that case you can do experiment with P, varying P from P 1 to P n. So, that means you will create n different combinations then you will be getting 0 to n observations and sigma, epsilon values you will be getting sigma 1, sigma 2, sigma n; three epsilon, epsilon 1, epsilon 2 and then epsilon n.

Now, if you plot this what will happen you may get a plot like this, here it is sigma essentially what is the difference between this and this here what I am saying, I am straight way without I when you showed you have shown me this spring balance. Then I immediately say that in elastic body this is the diagram, because the this Hooke’s law is known to me.

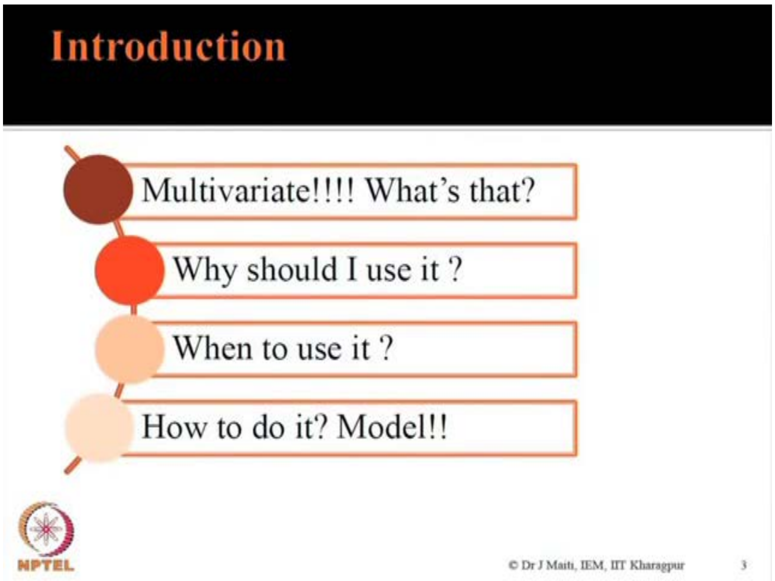
So, mathematics is known to me but in case it is not known I have done several experiments here. And based on this I am trying, I will do plot like this need not the perfect straight line, you will get when you go for the empirical relationship. So, this is what is the empirical 1 model? So, this empirical model when we talk about empirical model like this experiment based or data based models like this, these are basically the statistics, these are all statistical. So, for me this is for all of us, this is our statistical model.

Now, what is modeling? Then modeling is basically you want to get this type of results, it is not that a immediately you will get all this there is a process. The steps I have to understand what is my purpose? I have to understand in one or two full the purpose what are the variables that are affecting there. I have to identify all the important variables, then I have to see that how the data on the variable will be collected.

For example, here I 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, observation someone is interested to see the behavior of a particular animal. So, he cannot do the experiment may be but there are large number of wild animals of that particular species.

So, we can observe that we are just going and observing field based so field based observation our this one is our experiment, sometimes what happened we will go for some naturalistic observations, naturalistic observations which we talk about the wild animal case field based observation. In the production we go suppose, the steel washer case go to the production shop, and see that what is happening there and collect data and accordingly you do some modeling, some naturalistic observations. So, all those type of data collection mechanism comes under empirical modeling and you have to understand all these things. So, this is a process, the process of modeling, the process of model building is called modeling, the process of model building is called of model building is modeling.

(Refer Slide Time: 27:51)



Introduction

Multivariate!!!! What's that?

Why should I use it ?

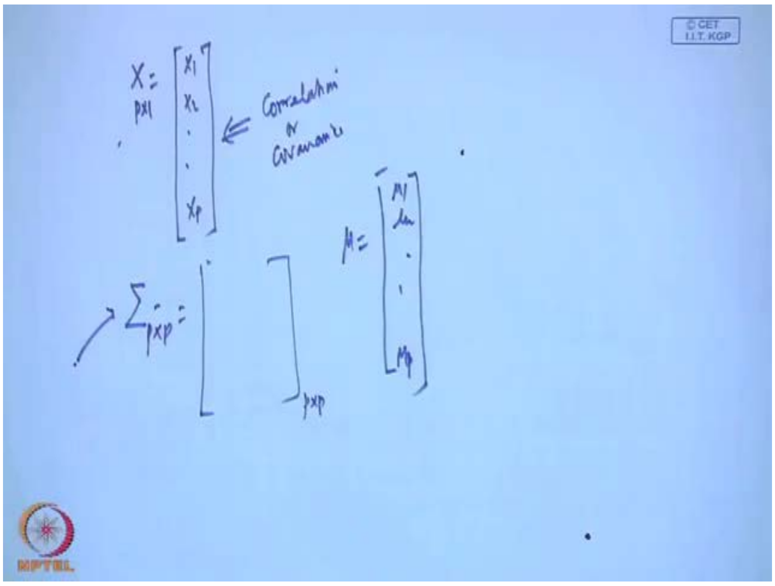
When to use it ?

How to do it? Model!!

© Dr J Maiti, IEM, IIT Kharagpur 3

So, let us see some of the slides now that I told that what is multivariate and then what is discussed, why should I use it and it is a base question and that was should I go for multivariate things? If I can do by some other way, why multivariate? So, they are some key issues which basically will be known to you later on that when we talk about multivariate.

(Refer Slide Time: 28:21)



©CET I.I.T. KGP
$\underset{p\times1}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ X_p \end{bmatrix}$
Correlation or Covariance
$\Sigma_{p\times p} = \begin{bmatrix} & \end{bmatrix}_{p\times p}$
$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_p \end{bmatrix}$

We talk about multiple variables that is p cross 1, if p the number of variables then X 1, X 2 like your X p. Now, there is possibility that these variables are interrelated, there is correlation, one of the easiest way is correlation in between the variables. So, that means you may be get a correlation matrix or other way it is basically the covariance between the variables or covariance. By covariance what I mean to say, if one variable varies there is a possibility that in particular way that some other variable also varies, then there will be covariance and standardized covariance is correlation. This is and in the subsequent lectures so covariance that will be p cross p matrix will come.

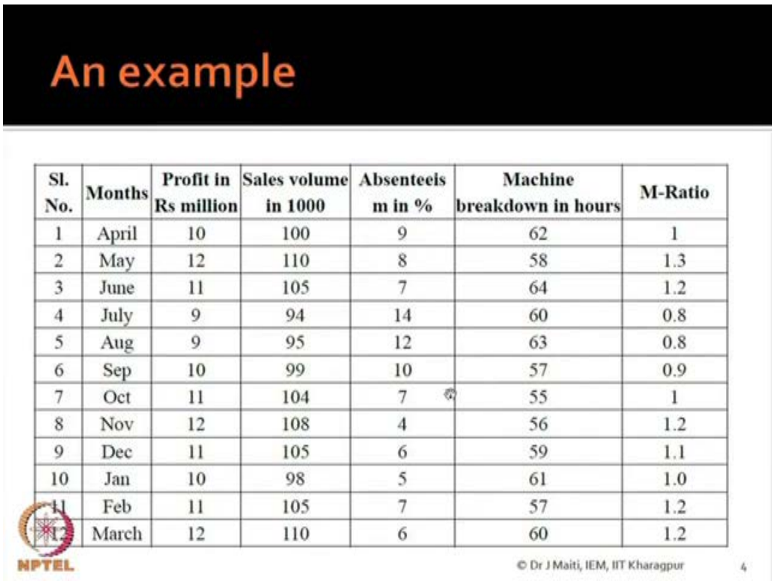
So, all those things so similarly, the mean values for all those variables mu 1, mu 2 like mu p, this things will be there. Now, so my answer to your question is that why should I use it because no physical process or as such any other systems also, which is characterized by multiple variables. They should be analyzed other like their behavior should be analyzed by taking into consideration of all the variables characterizing it.

When these variables consider very, very important for the design development or improvement of the system, for which it is developed. And as none of as it is obvious there will be covariance or correlation between the variables. If I go for univariate analysis we will lose substantially the information about the behavior, because of non-inclusion of the covariance structure.

So, we require to control this covariance structure and in multivariate statistics covariance is a very big issue and which will be found in multivariate distribution. We will be discussing all this covariance things so it is required. For example, for this case like our this one steel washer, this case the steel washer, three variables are visibly controlling its quality, inner diameter, outer diameter and thickness. There is chance that inner and outer diameter will be related, also the thickness in that case the customer will not be able to apply it or fit it to its own situation if there is huge mismatch.

Now, if we control inner diameter or outer diameter or thickness then what will happen? Then correlation structure will not be considered and ultimately we will not be able to satisfy the customer. So, we will be using multivariate statistics or multivariate modeling. When your system is complex in terms of number of variable it may be in conditions like this, the correlation structure is intact in order to extract a those correlation information, you want to extract the pattern from this data that is why you will be using. So, how do I do it? It is through the third models, so these models will be discussed a little later. Now, what is next?

(Refer Slide Time: 32:34)



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

NPTEL

© Dr J Maiti, IEM, IIT Kharagpur

4

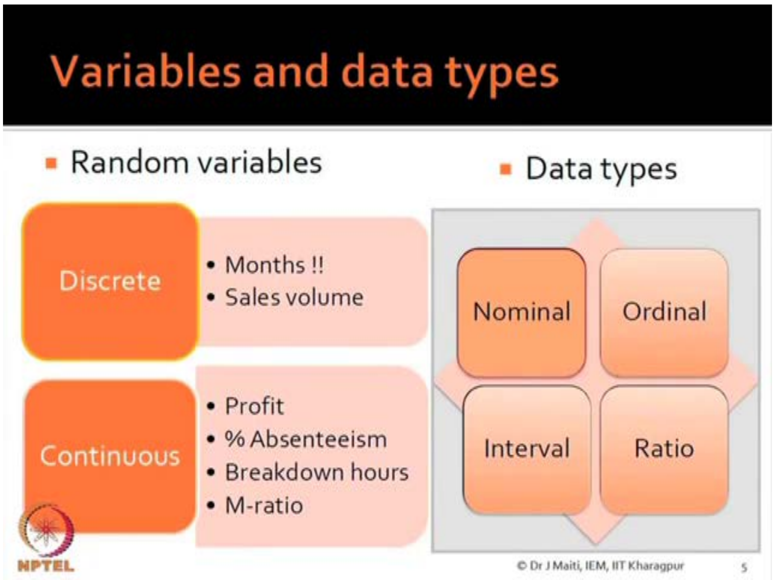
Next one example, here we are saying that a particular company operating may be in a city market and we want to see the organizational health of this company, with respect to profit in rupees million with respect to sales volume in rupees hundred, absenteeism, machine breakdown and M ratio. Actually, this is schemated intentionally first one is profit and sales volume, these are the organizational issue that health if you sell more your profit may be more. And if your profit is more you are healthy in financially, and another issue is absenteeism, if you are paying substantially and if you are taking care the well being of the employee's absenteeism will be less.

If you are maintaining the health of the process here we are saying machine, your machine breakdown will be less. And if you are if you are able to coordinate with customer as well as your supplier and your M ratio, that much M ratio particularly I say marketing ratio will relate 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 we have measured. This is nothing but a case of multivariate situation where each of the row like starting from 1 the first row, these values are talking about multivariate observations for month April.

Similarly, for second these are multivariate observations so there are we have multivariate observations. Now, you may be may be interested to know how profit varies over the months, then it will be univariate one if you want to say that how sales volume vary over the month, it will be also univariate. Now, if you want to know absenteeism varies over the year over month that is also univariate like this.

But if you are interested to see that how the profit and sales volume covary and they are own variation, then you will have to have to consider two variables. And then should be multivariate situation, sometimes you may be interested to know how the sales volume will be dependent on absenteeism and machine breakdown and marketing ratio. Then there must a dependent model and that is the same multivariate issue. So, this is in that shelf what I am talking about multivariate observations.

(Refer Slide Time: 35:29)



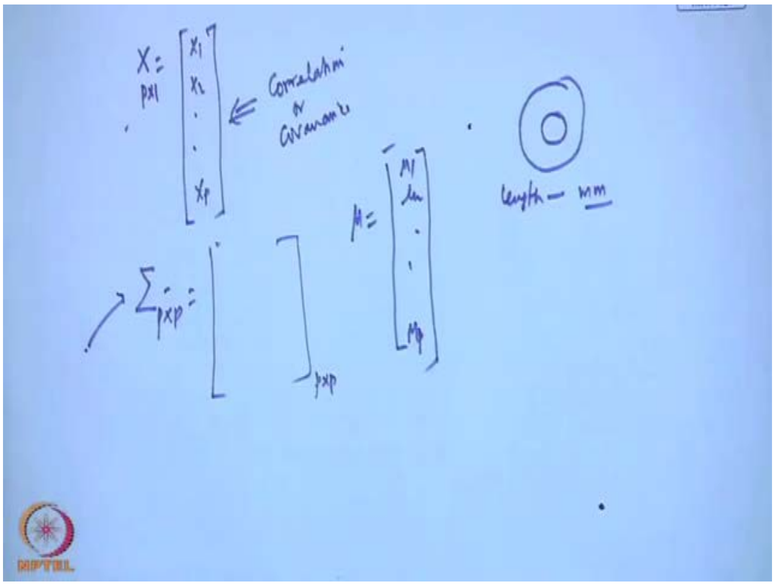
Variables and data types

- Random variables
- Data types

© Dr J Maiti, IEM, IIT Kharagpur 5

So, now we have discussed some of the things, some of the variables and we have seen that we have assigned them some values, but how where from those values are coming?

(Refer Slide Time: 35:46)



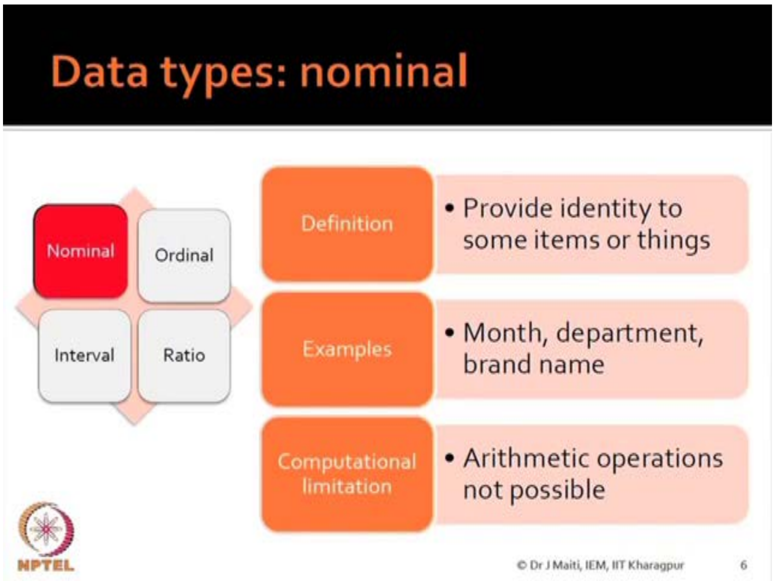
For example, if I say steel washer the thickness that mean be the inner or the outer thickness OS, how it is known? So, you have used some measurement scale to measure this, if I want to say that it may be you have used Vernier caliper to measure the outer diameter, may be used Vernier caliper to measure the inner diameter. So, you have used some instrument and as well as you have there is scale of measurement. In this case the scale is basically length which may be in terms of millimeter.

So, you have to sue some scale of measurement and based on the scale used whatever data you get those data will be of different types. So, you see this line here, the left side we are talking about random variables and right hand side we are talking about data types. I have explained you this random variable earlier, so I will not spend much time here, but you must please understand one thing that in random variable there will be discrete and continuous random variable.

By discrete random variable we mean to say that they will take some counted account values like 0, 1, 2 or something like this or January, February, March something like this, and your continuous case that profit absenteeism breakdown or what is M ratio here? What is that any value is possible? So, please understand one thing here, since volume are coming under your discrete because it is countable one but many countable, such count values can also be considered as continuous in any situations. But any how so there are two types.

Now, your data types I told you that what measurement scale you are using. Based on these data types you will be known, means that data will be having certain properties because data is nothing but information. How much information is available with the data getting me, so did it all depends on what scale you have used to measure this data. So, based on that there are four types of data, one is nominal data, ordinal data, interval data and ratio data.

(Refer Slide Time: 38:17)



Data types: nominal

Nominal
Ordinal
Interval
Ratio

Definition
- Provide identity to some items or things

Examples
- Month, department, brand name

Computational limitation
- Arithmetic operations not possible

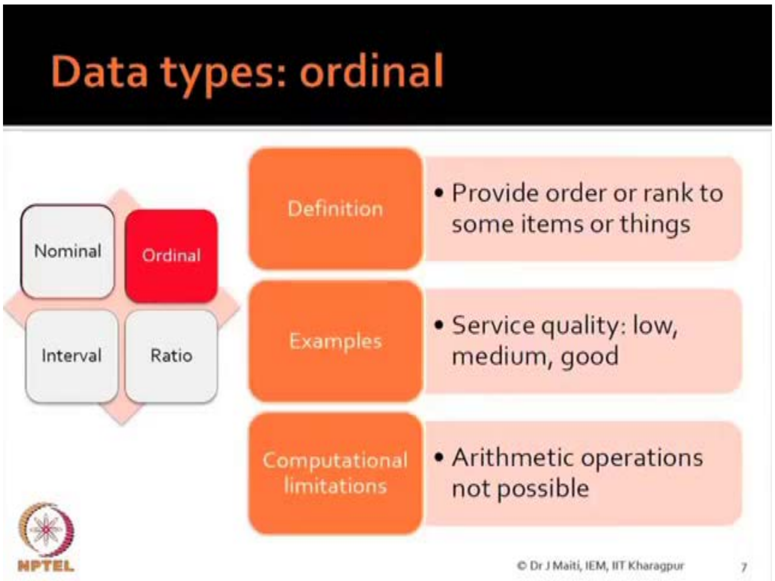
© Dr J Maiti, IEM, IIT Kharagpur 6

Let us discuss something about nominal data. My definition is this provide identity to some items or things is I say the month, the company, small company that is the I should have shown you that they want to do over the different months, what is the status. So, the month is a 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 the particular series.

Suppose, you just think of you are trying to know that some performance or status of the different department of a for example, it so then if I say the department of chemistry, department of physics, department of mathematics, department of computer science, department of industrial engineering and management.

So, all those things and they are basically providing identity but we sometimes require this type of data in our to include in our analysis. So, this is nothing but nominal data. Now, what is the problem with nominal data? Problem with nominal data is that there is huge computational limitations, because you cannot do any arithmetic limitations, you cannot add department of chemistry plus department of physics like this. We cannot say department of chemistry is 1 and department of physics is 2 and accordingly we will add, we cannot subtract, we cannot multiply, we cannot make division also this is the problem.

(Refer Slide Time: 40:05)



Data types: ordinal

Nominal
Ordinal
Interval
Ratio

Definition
- Provide order or rank to some items or things

Examples
- Service quality: low, medium, good

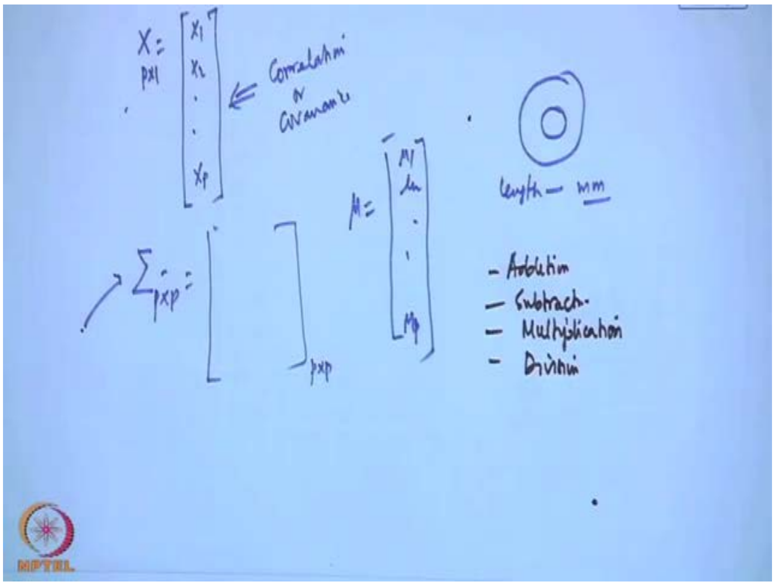
Computational limitations
- Arithmetic operations not possible

© Dr J Maiti, IEM, IIT Kharagpur 7

Next data type is your ordinal data type. What is ordinal data type? Suppose, you just see that you have you have travelled in flight several times may be, or train or some other places or you have gone to the students, and when you have taken food and you might have seen that you are giving a feedback form. They are seeing that they are pleased, they have read the in case of hotel food quality, service quality, room quality all those things in terms of not satisfied.

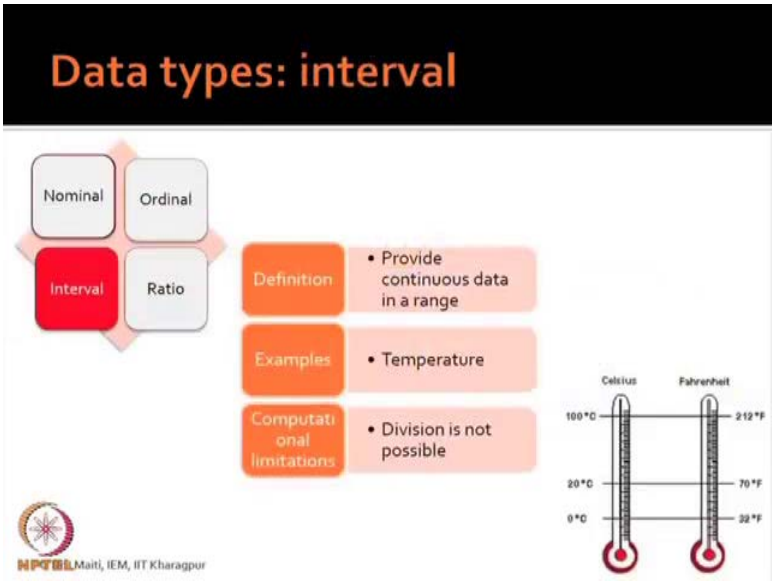
We are totally unsatisfied to extremely satisfied, this type of scale we have used for example, for the food case it is taste wise this very good, good or something like this. So, this type of ordering when order thing is there this is called ordinal data. So, what it does provide some order or rank to some items or things examples, service quality, it is low medium or good and computational limitations.

(Refer Slide Time: 41:17)


$X_{p\times1} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_p \end{bmatrix}$
Correlation or Covariance
$\Sigma_{p\times p} = \begin{bmatrix} & \end{bmatrix}_{p\times p}$
$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_p \end{bmatrix}$
length — mm
- Addition
- Subtrach.
- Multiplication
- Division

Here also we cannot do any arithmetic operations like your addition, subtraction, multiplication and division. You cannot do then what way it is better than our nominal data? It is better than nominal data because here you are getting a order, a rank you are getting. If I say the performance that my student performance is low, average and very good excellent like this, the person who is getting excellent is definitely better than the person or the student who got very good. So, I have a ranking skill here ranking ability with this data. So, ordinal data is rich compared to nominal data.

(Refer Slide Time: 42:18)



Data types: interval

Nominal | Ordinal | Interval | Ratio

Definition
- Provide continuous data in a range

Examples
- Temperature

Computational limitations
- Division is not possible

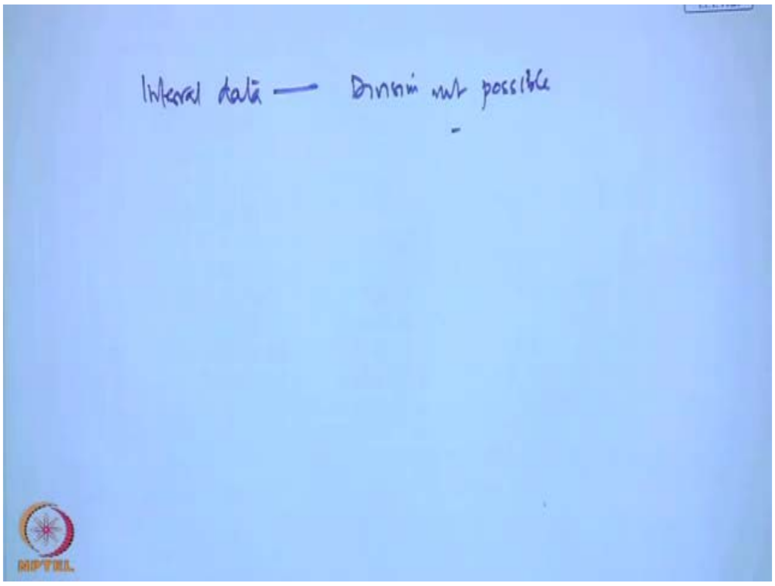
Maiti, IEM, IIT Kharagpur

Next data type I said that interval data, what is interval data? It is basically well understood if we take this example, here temperature we are measuring using two scales, one is celsius, another only Fahrenheit. In developing these two scales, Fahrenheit scale as well as your celsius scale, the reference point is taken at two different points, means locations. It is not the same you getting me so and if you see the horizontal lines here you see that 0 degree centigrade, 20 degree centigrade and 100 degree centigrade. Then the corresponding Fahrenheit will be 32, 70 and 212 Fahrenheit, understanding?

So, there is a range that if I say the difference from 100 to 0 degree you are getting this range, here are also 212 to 32 the corrseponding range is this. So, whether we measure in using celsius scale or Fahrenheit scales we will be getting the equal range. Now, what will happen suppose, I measured temperature today? Today day temperature is 20 degree centigrade to 30 degree and may be day after tomorrow 21 degree, then if I want to do the averaging I can add them and then divided by 3, that 3 days average I will get if I do the same thing in Fahrenheit. Also it is possible I can do that 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 compared to the tomorrows, yesterday’s temperature. Then if I use Celsius scale and if I divide 22 by 20 and then here it will be it may be 70 and other things, then we will find out they are not matching. So, that means interval scale is some scale where you will get a interval data range data and they are all having al, type of continuous properties except and they can do 3 arithmetic operations very easily, addition, subtraction and multiplication. But when you do division, you will find out that when you change, it changes the scale. Ultimately what will happen? You will find that they, so in interval data you cannot go for division.

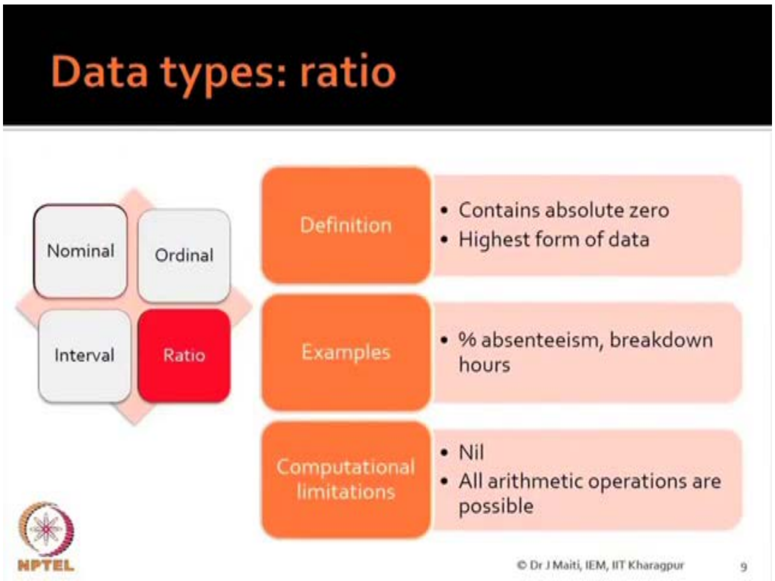
(Refer Slide Time: 45:17)



Interval data — Division not possible

Interval data division is not possible, all other arithmetic operations are possible. Let us go to the next slide.

(Refer Slide Time: 45:34)



Data types: ratio

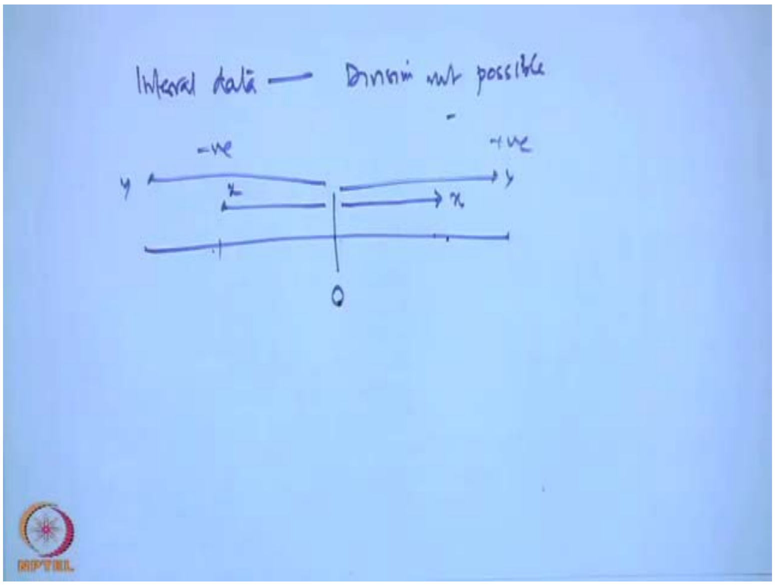
Nominal
Ordinal
Interval
Ratio

Definition	• Contains absolute zero • Highest form of data
Examples	• % absenteeism, breakdown hours
Computational limitations	• Nil • All arithmetic operations are possible

© Dr J Maiti, IEM, IIT Kharagpur 9

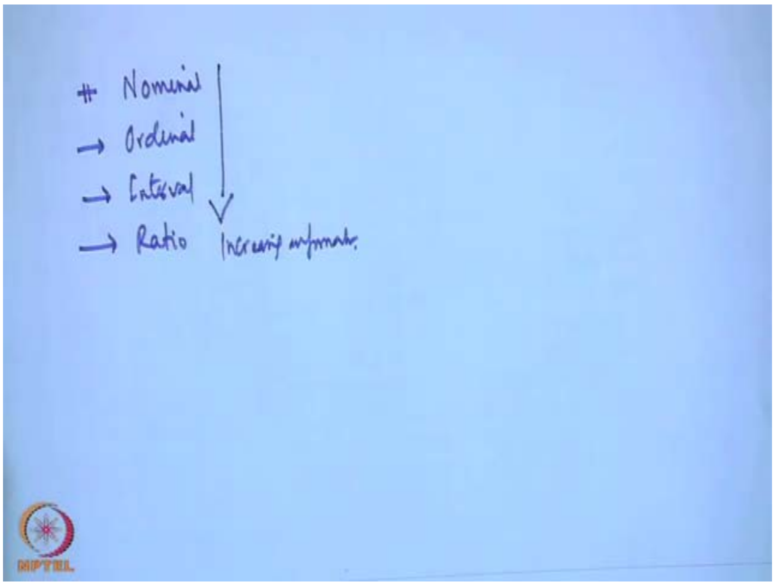
Our slide that is we are talking about ratio, data ratio. Data is something where there is absolute 0 in the scale of measurement.

(Refer Slide Time: 45:46)



This is 0, if I move towards right suppose 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 the same. So, that means if you go in to the left it is that is the same. So, that means if you go in the to the left it is negative, this side it is positive, but there is absolute 0 in between. So, this 0 is the reference point not in terms of the Fahrenheit and centigrade scale that where is the two different definition, it contains absolute 0, highest form of data, sorry. So, ratio data is it contains absolute 0 highest form of data example absenteeism breakdown hours as shown earlier and computational, all arithmetic operations are possible here.

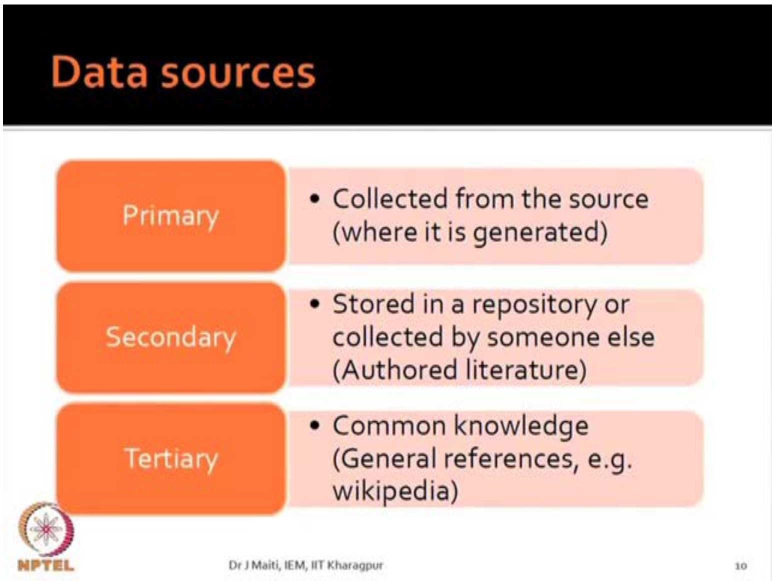
(Refer Slide Time: 47:07)



# Nominal
→ Ordinal
→ Interval
→ Ratio
Increasing information

Now, if I go by the order of information available then definitely your first one is if it is nominal then followed by ordinal, then your interval, then your ratio. Then definitely in order of increasing information this will the, this is the case my best data is this, next best is this, next best is this, next and this is the lowest of information data.

(Refer Slide Time: 47:44)



Data sources

Primary	• Collected from the source (where it is generated)
Secondary	• Stored in a repository or collected by someone else (Authored literature)
Tertiary	• Common knowledge (General references, e.g. wikipedia)

Dr J Maiti, IEM, IIT Kharagpur 10

So, you know that different data types. Now, you know that as you will be applying multivariate statistical modeling, you must require full-fledged data. So, you need to know the data source, primary data collected from the source where it is generated for example, in the case of a steel washer example, if you collect data from the production shop and just going there collecting data or that is what is known as primary data. Suppose, you want to see the behavior of the animals in the jungle go and observe and then accordingly note down and that will be your primary data.

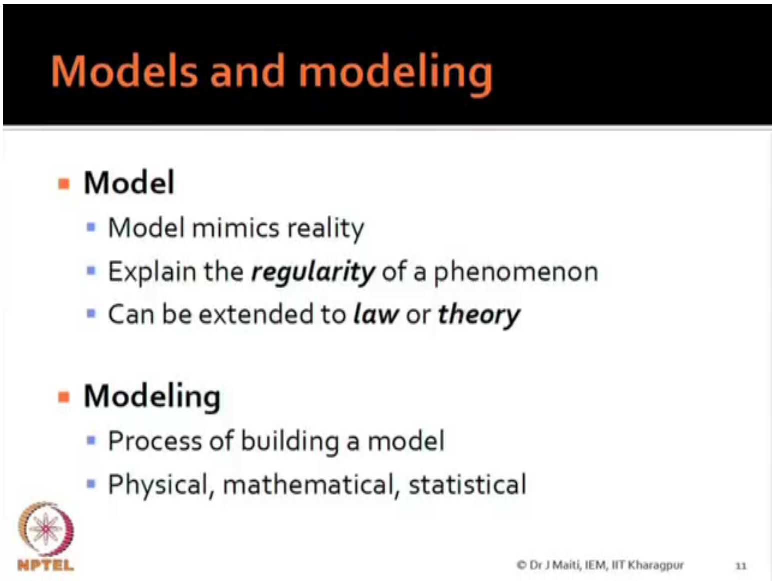
So, for the production that and that example the profit and sales volume case that is also primary data. So, long you are collecting from the source, what is secondary data? Secondary data stored in repository or collected by someone else, you are getting me? So, you are not collecting, it is already there. We have different sources for example, you may get the financial data from some sources.

And suppose company is maintaining records of their production and suppose their maintenance or the health of machines and many things. So, you have not collected so company has stored and you have gone there and collected these things, or it is better that in a literature you studying something in your own area. You found that a paper is there where some data is given.

So, this type of data is secondary but secondary data must have must be authentic, in the sense that reference of the data is available, author references are there, this is that author literature data but this is definitely as it is done by somebody else. It is not primary, there you have to rely on the authenticity of the data collected by somebody else. The tertiary data which is basically a common knowledge type of things.

Suppose, you know you will find many things are there actually when in terms of modeling, modeling when you start with a subject area you start with this that when your knowledge is not very clear, you will start with the tertiary sources. And then slowly you go to the secondary source. Finally, when you do actual work you may go for the primary data sources.

(Refer Slide Time: 50:31)



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

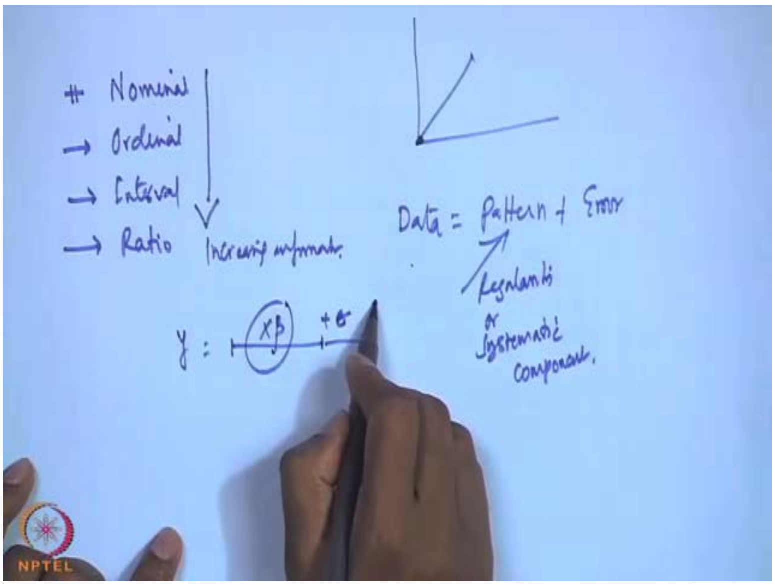
NPTEL

© Dr J Maiti, IEM, IIT Kharagpur 11

I told you earlier for this model, let me repeat this again that model mimics reality that when you develop a model that without considering the reality, the real thing you are not doing the justice. So, the model reality so it should be a, it should have real applications that is what is the meaning. For example, suppose you think of a car which is got with by suppose any they develop, they develop these things. What I mean to say they develop a model simulation model in computer first before going for a developing the car, one after the other manufacturing. The car in the manufacturing shop or there must be some simulation model and means how the car will work.

So, that type of things are known as that means it is a in terms of the reality the car is the real thing. So, your modeling can be so that is simplest example and that the mathematics is related to the elastic behavior of it that is the reality. In statistical sense when we talk about the how sales volume is dependent on other things that is your absenteeism, M ratio and all those things that also is going to talk about the ways, which show actually in statistical sense, a model talks about explain the regularity of a phenomena.

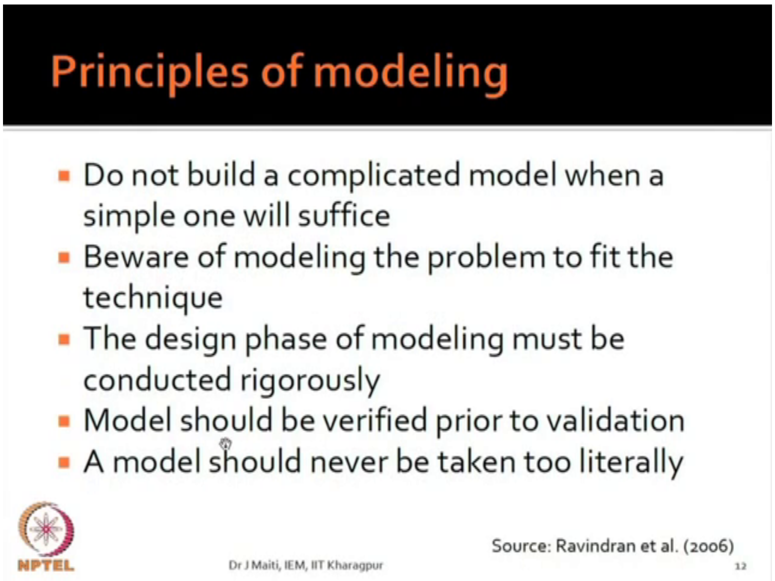
(Refer Slide Time: 52:25)



In Hooke’s law the regularity is so long, it is within the elastic limit. When the load is released, 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 is pattern plus error, this pattern is the regularity pattern or systematic component. So, we must know what is our problem? And accordingly all data if you collect it and you want to extract pattern from this data. In case of prediction 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 variable that is X beta, then plus l will be there. This is my regularity or my data.

So, when you repeat similar that similar development under different situations then what will happen? Then if it performs well under the different situation for which it is developed, the one day we may say it is a law or a theory like Hooke’s law or Hooke’s or this Hooke’s law is this, I left there that elasticity thing. So, we all know that Newton’s laws of motion and we all know that Dalton’s atomic theory and many other things that these are not one day everything is developed and people accept it. It basically developed at test stage verified, validated after several years and then other scientist other that is the researcher, they accepted the fact and then it was applied to different situations and found that it is working. I told you modeling, also process of building a process, physical, mathematical and statistical, this is I have already explained to you.

(Refer Slide Time: 54:52)



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)

Dr J Maiti, IEM, IIT Kharagpur

12

I hope that you got the glimpse of actually the purpose of applied multivariate statistical modeling. Actually, we want to develop empirical model, those empirical models is these are all data based, data based in the sense that they contain you have data. And you are going for building models and you are building models, and to find out the regularity of the data, or the pattern of the data. And show that you will be able to describe the relationships of the population or the behavior of the population or system in consideration. You will be able to establish the strength of the relationship, you will be able to predict something, you may be able to prescribe something also, but when you talk about a statistical modeling.

Usually this is the description and prediction part is description explanation and prediction this three things come into consideration. So, slowly you will be knowing different types of statistical all together and you will be tempted to develop different models, also based on the data whatever available to you but before model, going for modeling or applying any statistical techniques what is happening?

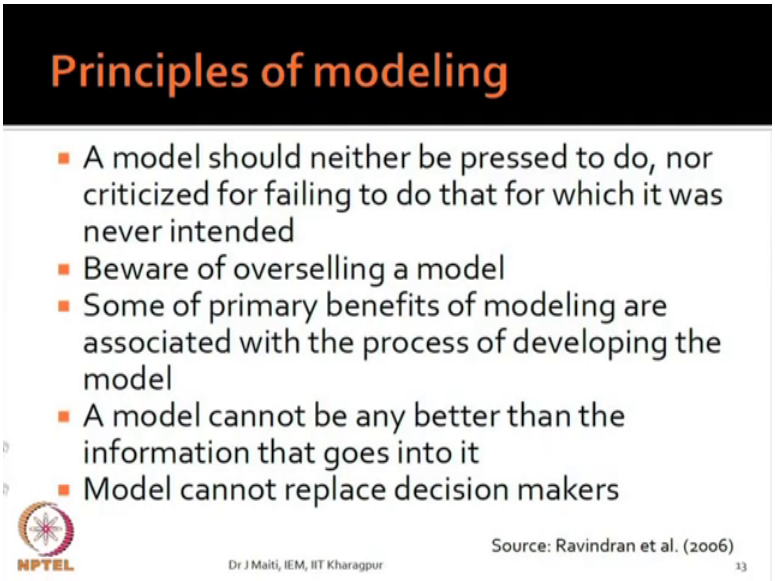
What is we want to say that you have to have some principles in your mind before going for this here. I have just jotted down some of the principles which I have taken from text book by operation research by see what is they said that do not build a complicated model when a simple one will suffice. For example, suppose if I know the mean value of the different lots of steel over mid value of a particular characteristics; for example, the inner diameter of a different your lots produced. And if that suffice my purpose go for mean, or at max you may require the standard deviation of the inner diameter produced by the different processes A B C as I told you.

So, there you may you do not go for may be that covariance structure, many other thing. So, you do not go for if it is needed you go for you are modeling of the problem to fit the technique, many a time I have seen it my case that there is one model which we will be discussing later on known as structural equation modeling. The people are using structural model everywhere where a simple regression model can be. But people are interested to fit the structural equation model.

So, please be little bit of cautious on those things, that model is for problem and model comes from the problem, not to fit a statistical technique. Design phase of modeling must be conducted rigorously and it will discussed later. What do we mean by design phase? Coming under study design model should be verified prior to validation. Verification means suppose you when you collect data you split the data into two halves, one for your training other for test.

What other way I can say? One set for model building, other set 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 take in too literally but many a times what I have found that model there are more many variables, statistics is taken very in very loose end. So, if there are many variables let us find the relationship is there or not this type of or whatever variable is there. Let us find that relationship without considering the purpose.

(Refer Slide Time: 59:19)



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

Source: Ravindran et al. (2006)

Dr J Maiti, IEM, IIT Kharagpur

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 variable of a particular population. Now, later on you want to see that how I want to predict something, see you developed a model to see the pattern strength of relationship not to predict. So, why how can your model will predict which was not intended for, so that is another issue. So, if it fails to do prediction when it was just to understand the covariance structure, then we should not criticize for this nor we should not press the model to do it, beware of overselling a model many a times.

We basically make sure of I can say recommendation based on the model and many of the things basically from common sense, and so that type of selling I prohibited some of primary benefits of modeling are associated with the process of developing the model. So, the see as all we of you are busy in learning multivariate statistics, multivariate modeling. So, do not think that always you will be doing something great with these type of modeling you are learning. So, the learning process when you develop something you know the physics of the problem, may be you know the process through which data is generated, you know how the data to be captured, how the data to be analyzed, what techniques is applicable.

So, this is a entire gamut, so this gamut of process is very, very important. So, very many fits very, very fits you acquire out of it a model cannot be any better than information that goes into it. So, you cannot say that you are using nominal data and you will be basically talking about a model of regression where y variable is nominal. So, you have to have go for some other type of model for that may be your regression. So, the information what you are the quality of information what is fed into that model, that is more important because it if input is not good then output also, you should not expect good.

So, model cannot replace decision maker, getting me? You cannot think that you are your model is superior than you the decision maker, the analyst who has having the system knowledge they will act smart. So, they are more important people, so whatever you develop, whatever you do for, what purpose you are developing, all these things. So, that is in your brain, in it is the root what work there so better than any model. So, in this case what I want to say that you please take all those issues what I have discussed, the principles particularly in this series and accordingly develop the model and today it is up to these. Next class, we will be studying the statistical approach to problem.

Thank you for your patience.