**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

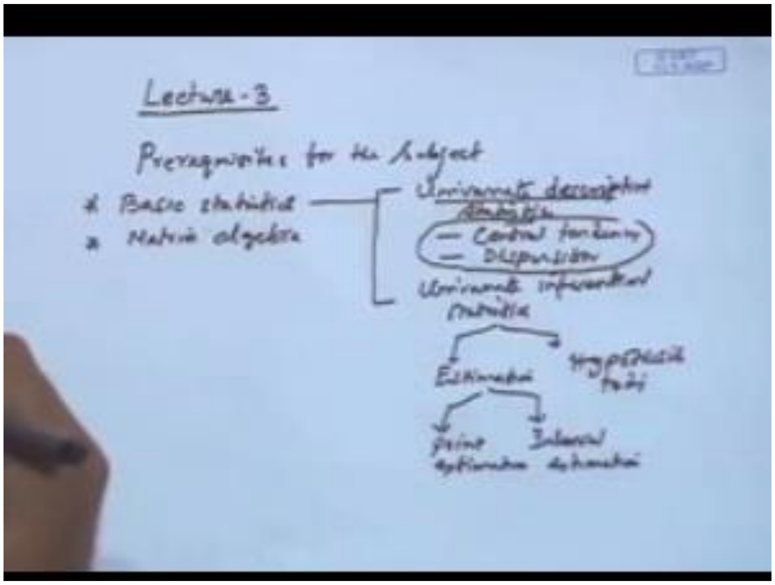
Lecture – 03

Topic

Univariate Descriptive Statistics

Good afternoon last class we have described the multivariate statistical modeling.

(Refer Slide Time: 00:25)



From the purpose as well as different modeling techniques point of view and we ended that lecture with prerequisites for the course. Prerequisites for this courses subject and what we have described here that basics statistics is one the prerequisites and I told you that we have also required to know matrix algebra repeat. Now under basic statistics the univariate descriptive statistics and univariate inferential statistics are important. So again under univariate descriptive statistics usually the central tendency and dispersion measure these two issues are described under descriptive statistics.

Under inferential statistics estimation and your hypothesis testing. Under estimation there will be point estimation and interval estimation okay. Today we will, in this lecture we will describe this one univariate descriptive statistics we will see the content.

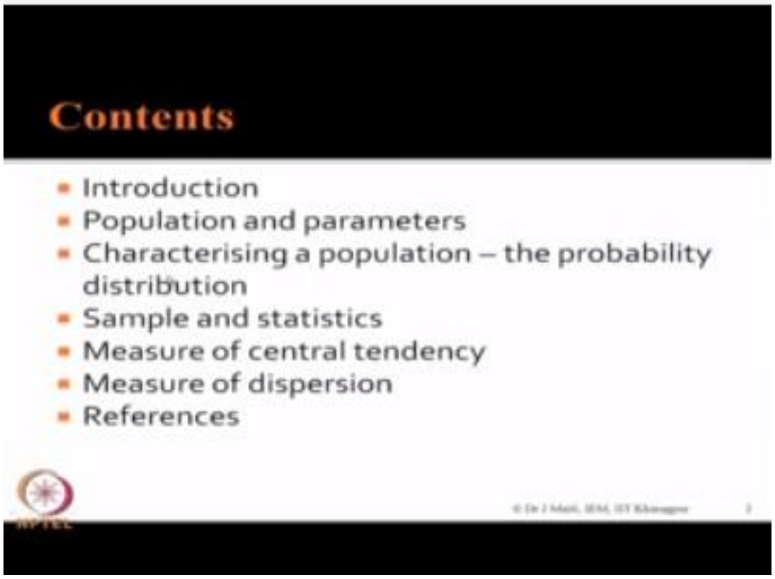
(Refer Slide Time: 03:07)

Univariate Descriptive Statistics

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

Of today’s.

(Refer Slide Time: 03:08)



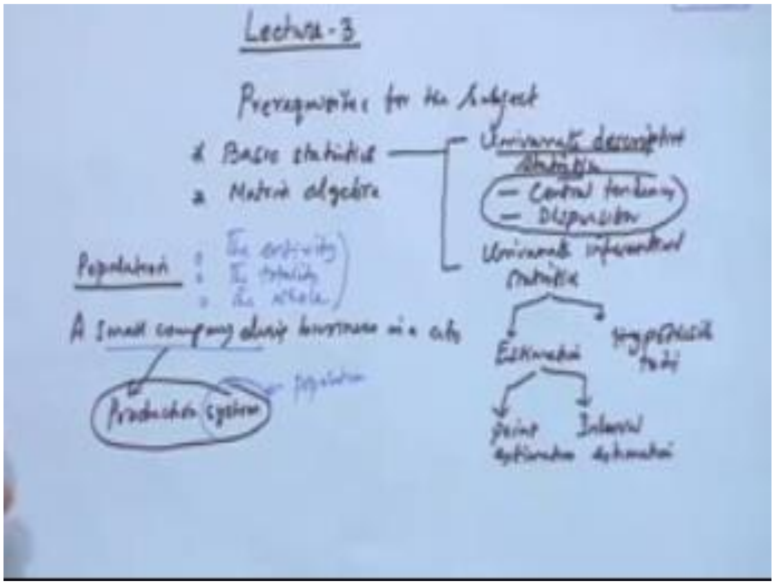
Contents

- Introduction
- Population and parameters
- Characterising a population – the probability distribution
- Sample and statistics
- Measure of central tendency
- Measure of dispersion
- References

© Dr J Maiti, IEM, IIT Kharagpur 3

This lecture we will start with population and parameters, then we will describe probability distribution particularly the normal probability distribution. Then we discuss sample and statistics followed by measure of central tendency, measure of dispersion, and followed by references. Now do you have any idea about population?

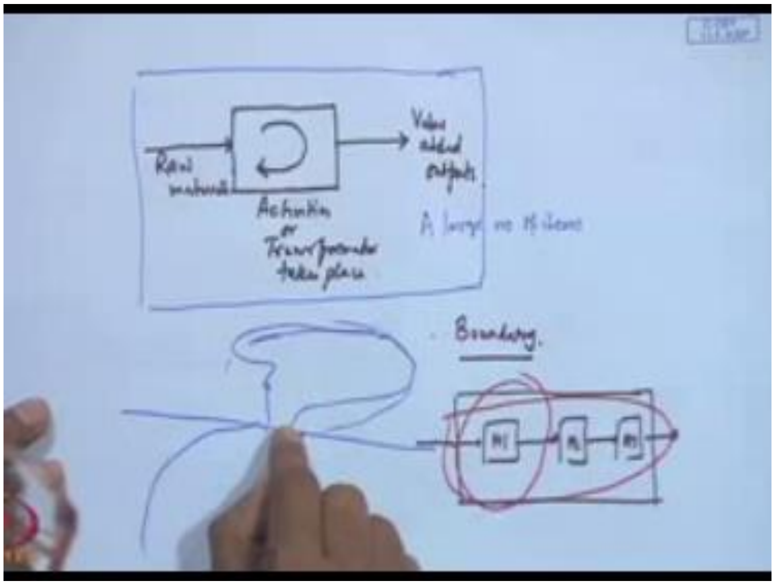
(Refer Slide Time: 03:42)



What do you mean by population? General sense we say that the population of West Bengal, population of India, but in statistics this population has much border sense. For example, in last class we had described one example that is small company doing business in a city. So the company has a production system can it be a population? If you define population from statistics point of view population is the entirety or the totality the whole population the entirety or the totality or within the whole, when you talk about the population of West Bengal that means each and every residence legal residence of West Bengal is considered.

From the production system point of view, this system this work also represent population the way we understand application of statistics. So the system also is unanimous for us it is also population because system can be characterized by different variables. For example, in the, for this company there are profit, sales volume, absenteeism and so many other variables where we have discussed. So these are basically the variables which characterize a population or a system, another word could be for us that is a process.

(Refer Slide Time: 06:19)



A process also we can think in this line also the process can be from our purpose point of view, a process is something where transformation or activities taking place, activities or transformation takes place. For example, you give inputs as a raw materials and the process production process it converts into value added output. Now if we consider the total life cycle of this process, then it will produce a large number of items, so large of items will be produced all items collectively is the entirety the totality or the whole.

So that means with respect to the items produced by these production process we can define population, from, if you go the service sector for example, the health care system or banking system there also you can define population. So essentially if you want to define population and you declare to keep in your mind two things that when I talk about the population of West Bengal suppose this type, this we got let be the portion.

Now the hilly regions population at the hilly region that is different than the population in the west of West Bengal or south of West Bengal. Now for a particular purpose you may be interested to understand what is the educational status of the people of the hilly people of West

Bengal, then what is happening you are making a boundary you are creating a boundary for the system.

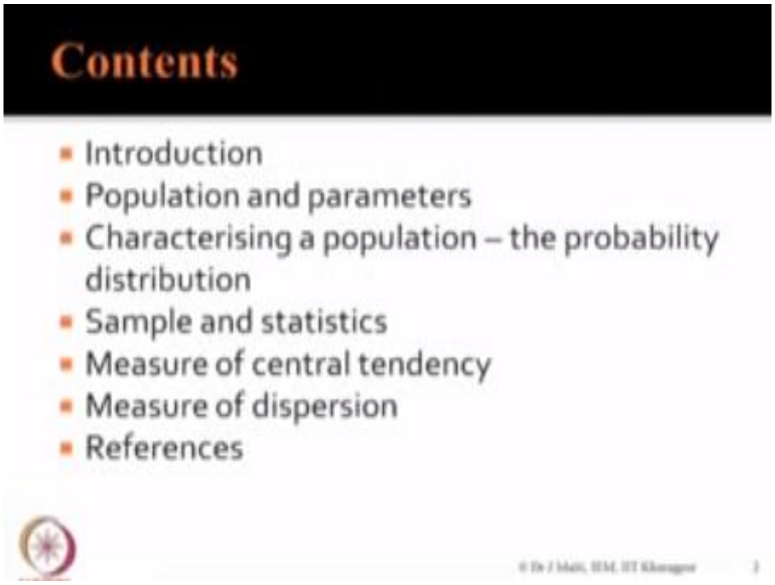
So these boundaries is the hilly region so in that case your population is this hilly region only. Now if you think from the voting point of view suppose the election time so this in all the legal that voters they go for voting, in that case all the voters of the total West Bengal they are the population. So in that sense what is happening that means if you really define, want to define population the boundary is very important getting me.

Boundary in the sense if you go, you come back to again the manufacturing scenario in manufacturing senior you will find out that the total production system may composed of several sub system that for example, this may be machine 1, machine 2 and machine 3 and they are doing different operations, raw material coming here and transform through machine M1 then going to machine M2 some other activities are going on.

Now if you are interested to infer something about machine 1, suppose you want to infer something about machine 1, then your population is this. If you think that there are some common characteristics applicable to all the machines then you may be interested to see the totality including all the machines then your population will be will consider all the machines.

This is very important and unless we understand population there is no usual statistics because statistics is used to infer about the population inference related too many things during in inferential statistics we will be telling you what are the different inferences possible. But for the time being you please understand that when we talk about population we talk about a system or a process, okay. For and why we require to study the process of the population? Because we want to understand the behavior of the process or the system or in terms of that is basically population we want to study the behavior.

(Refer Slide Time: 11:49)

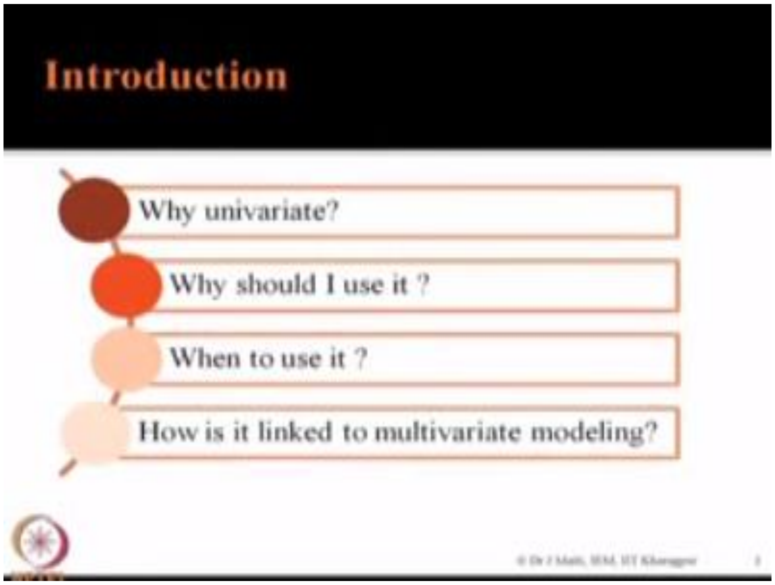


Contents

- Introduction
- Population and parameters
- Characterising a population – the probability distribution
- Sample and statistics
- Measure of central tendency
- Measure of dispersion
- References

© Dr J Maiti, IEM, IIT Kharagpur 2

(Refer Slide Time: 11:49)

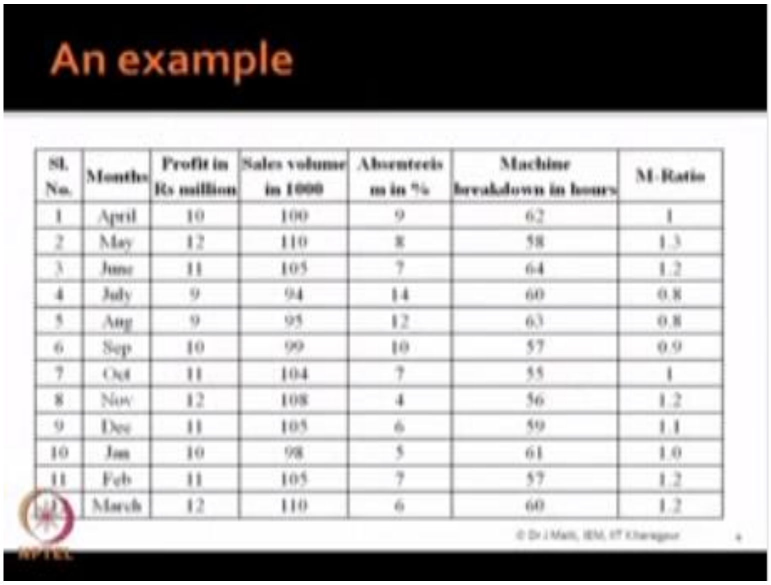


Introduction

- Why univariate?
- Why should I use it ?
- When to use it ?
- How is it linked to multivariate modeling?

© Dr J Maiti, IEM, IIT Kharagpur 3

(Refer Slide Time: 11:50)

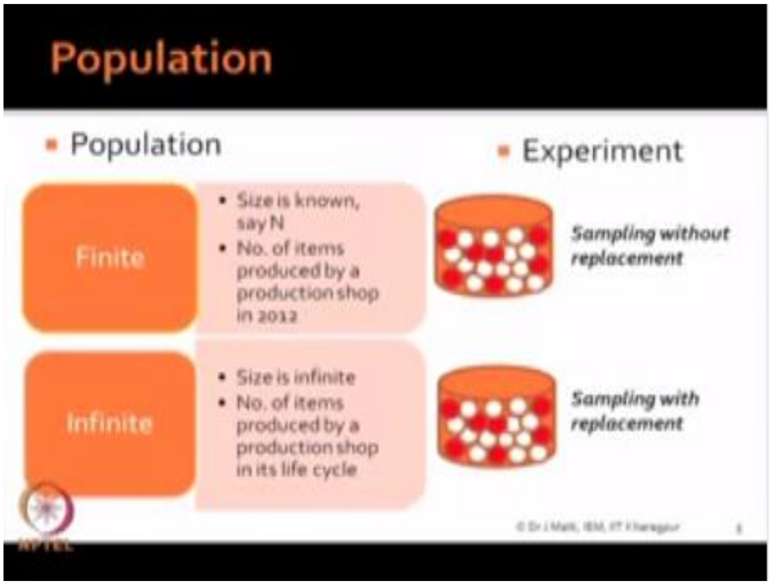


An example

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
[illegible]	March	12	110	6	60	1.2

© Dr J Maiti, IEM, IIT Kharagpur

(Refer Slide Time: 11:52)



Population

- Population
- Experiment

Finite

- Size is known, say N
- No. of items produced by a production shop in 2012

Sampling without replacement

Infinite

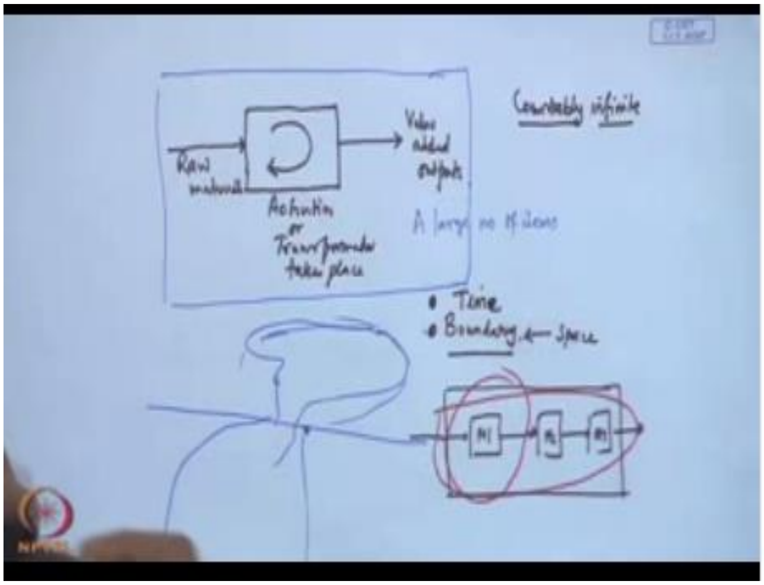
- Size is infinite
- No. of items produced by a production shop in its life cycle

Sampling with replacement

© Dr J Maiti, IEM, IIT Kharagpur

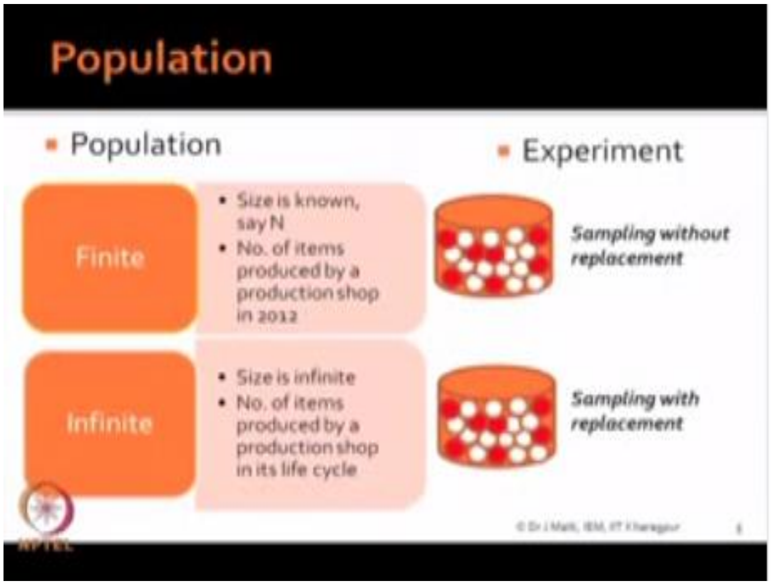
Now if you see the size of population what will happen? Population can be finite, can be infinite, when I am talking about suppose the production of a process for 1 year number of items produced per year and if that is my population, then it is a finite population. So time is another aspect which also define you should define the population.

(Refer Slide Time: 12:30)



So one is the boundary and other one is the time. So in two's this is basically boundary in the sense base boundary and time be the time boundary, if you go for the entire life cycle of a process then what will happen can you count that what are the number of outputs, it is very, very difficult. So if we talk about the entire life cycle total time of the life of the process what will happen the number of items produced will be countably infinite, okay. Whether countable infinite or infinite we will basically define in statistics in two chains.

(Refer Slide Time: 13:20)



Population

- Population

Finite
- Size is known, say N
- No. of items produced by a production shop in 2012

Infinite
- Size is infinite
- No. of items produced by a production shop in its life cycle

- Experiment

Sampling without replacement

Sampling with replacement

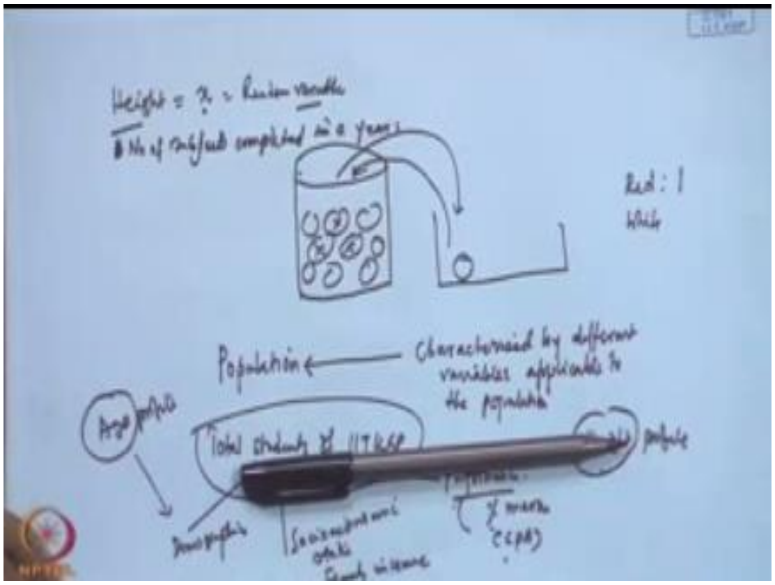
© Dr J Maiti, IEM, IIT Kharagpur

One is that your population will be finite population or infinite population, and finite population mean the size is known which is N for example number of item produced by a production shop in 2012. Infinite population size is infinite, number of items produced that is on the life cycle of thee process that is countably infinite. If you need further explanation as I told you in the last class that experiment random, experiment is the issue.

In statistics deals with random variables and random variables comes from random experiments, basically we generate random variable based on the experiments conducted. So if we do one experiment like this you see this figure inside this if I say this is basically on then, inside this there are red and white balls. Now you pick up one ball next one ball like this one after another without replacement, what will happen?

After some time there will be no ball to pick up experiment will end, this is finite population. Now in other cases see there what we will do in a second experiment you pick up again replace. So what will happen in that case?

(Refer Slide Time: 15:02)


Height = x = Random variable
No of subject completed in a year
Red : 1
White
Population ← Characterized by different variables applicable to the population
Agro profile
Total students of IIT KGP
profile
Demographic
Socio-economic status
Family income
Performance
% marks
(CGPA)

In that case the number of ball will never adjust it, there are so many balls red, white what you are doing you are picking up and finding out whether it is red or white. So either red or white so you are counting that red okay then again you are replacing this in a similar manner you are continuing the experiment you will need a size of the population, what will happen? The number of balls will remain as it is from experimental point of view it will never end, so this is what is infinite population.

In statistics most of the issues what we will be discussed later on we will consider infinite population. So in reality there may not be 100% true that all populations are infinite, but countably infinite populations are meaning and for our practical purposes if I consider this infinite population there is no problem, okay. Population behavior if you measure you required to know that what are the variables that is governing the population.

In this sense characterization population. So population is characterized by different variables, applicable to the population. For example, if we consider the total students of IIT Kharagpur all students of IIT Kharagpur this is my population all students of IIT Kharagpur and the Kharagpur students they come from different their demographics differ they as usual economics panelist socio economics status differ.

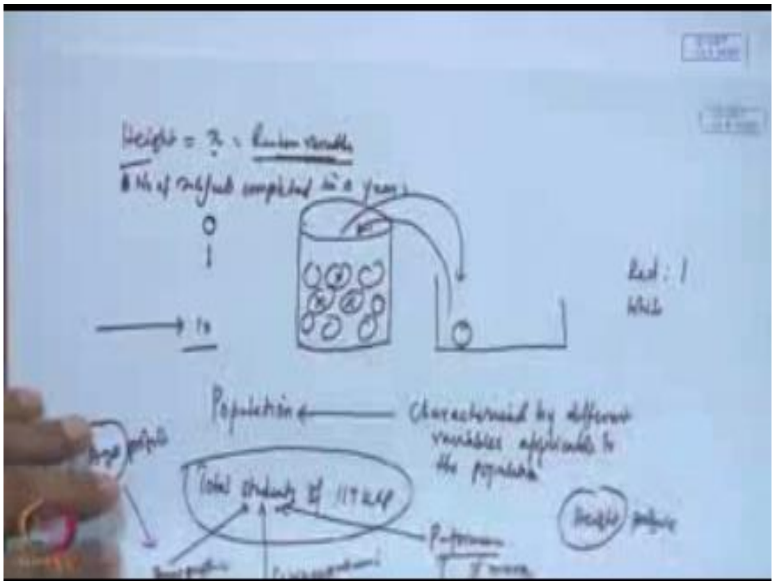
Their performance in the graduation that mean in IIT Kharagpur like jumps that also differ okay. So for performance you may be interested to see that what is the percentage or marks obtained or CGPA cumulative grade point average or somewhere let it do demographics you may be interested to see that what is the age profile, age sometimes we may be interested to know their height profile.

So see age, sorry height, age, percentage or marks, CGPA, under socio economic status family income all are basically coming under these are all variables which characterize the students of IIT Kharagpur. So if we want to understand population not only the space and time boundary, we also required to understand what are the variables that governs the population that is what we see basically.

And if you consider any of the variables for relate height I am writing height is the students and this I am denoting as X which is the random variable let it be, here we are saying random because if we just pick up one student we do not know what is this height but when you measure you will find out some height that’s it.

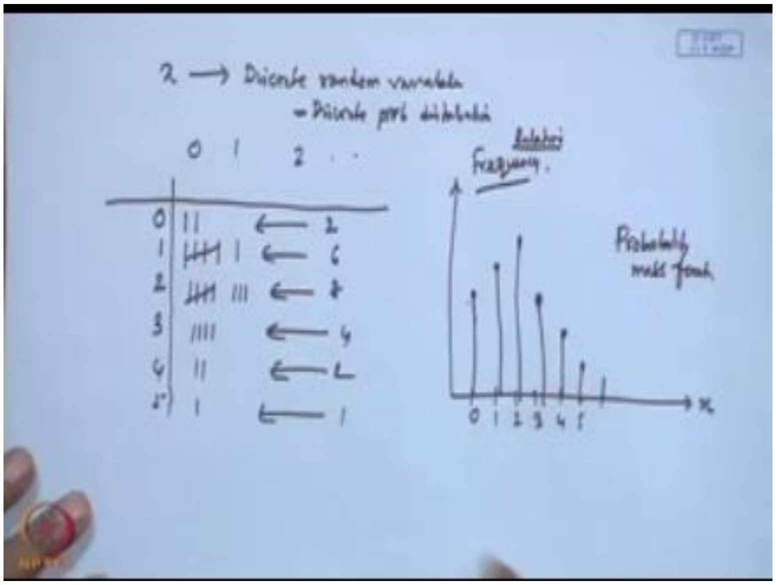
So it’s a this one x each so are you want to characterize IT students in terms of their height, or you may be interested to characterize the students in terms of their number of subjects completed in a year. We will find out that there are the many battle of cases, many students could not complete so in that sense it may show happen that if we count sure that the team subjects to be completed it may show happen.

(Refer Slide Time: 20:23)


Height = ? : Random variable
# No of subjects completed in a years
0
1
10
Red : 1
White
Population
Characterised by different variables applicable to the population
Total students of IIT KGP
Performance
Height

That you will find some student all know come subjects completed some student want or like this maybe up to 10 although it is, it will be every by towards 10 but this is possible okay, should depending on the, I mean what type of variable random variable you have considered and accordingly you require too huge certain probability distribution.

(Refer Slide Time: 20:56)

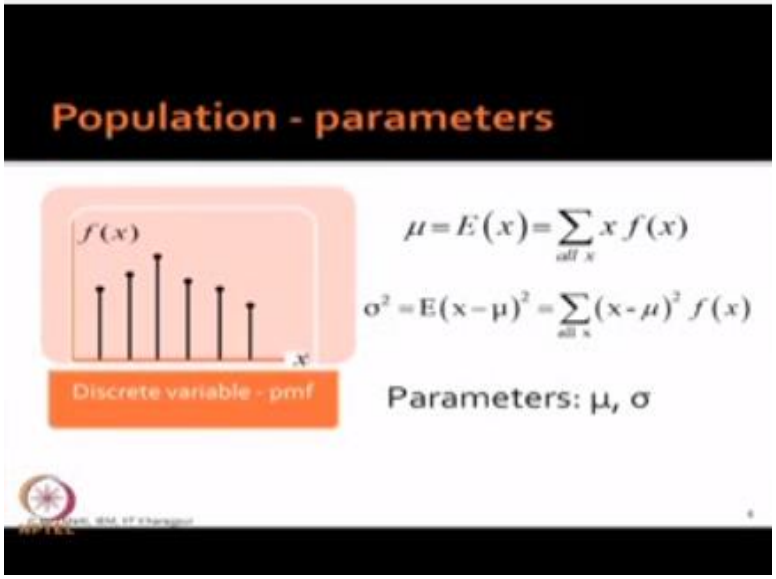


Last class I told you that if the variable is discrete suppose x is a discreet variable discrete random variable then you have to use discrete probability distribution. We discussed last class, but we have not said what are those bowler distribution later on we will see, but what you can see where easily that suppose x is discrete variable it can take value 0 1 2 3 like this then if you make telechart in the sense frequency 0 1 2 3 4 5 like this suppose when you are getting 0 counts your putting 1 like this, then again suppose 0 count and similarly like this what it is this basically we count.

What happen that what is the occurrence of 0 two times this one is 6 times this one 8 times this one 4 times this one 2 times this one, 1 time so if you by characterization what do we mean that I have my discrete random variable which can take different values suppose 0 1 2 3 4 5 and it appears for different times. I think all of you know this is nothing but the frequency diagram, and these frequency diagram if I know that total number and if you divide each of the frequencies by that total then you will be getting relative frequency.

And that relative frequency will give you that empirical or good probability distribution and these distribution is known as probability mass function.

(Refer Slide Time: 23:36)



Population - parameters

Discrete variable - pmf

$$\mu = E(x) = \sum_{all\ x} x f(x)$$

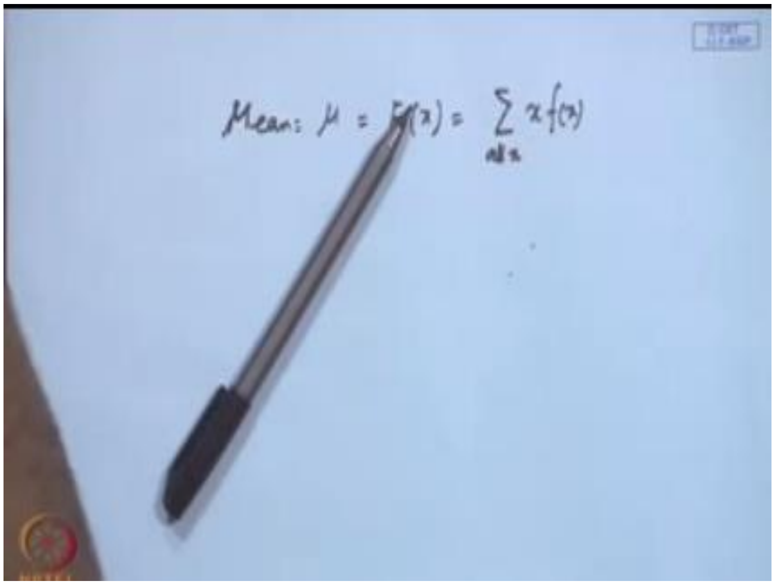
$$\sigma^2 = E(x-\mu)^2 = \sum_{all\ x} (x-\mu)^2 f(x)$$

Parameters: μ, σ

What we have said this pma pr you see that discrete variable when you get this type of plot you basically developing probability mass function. But I told you that we will be considering infinite population. So infinite population mean the totality is not known second thing is that our variable is random what will happen next what value it will assume we do not know. So when you where in the population domain you cannot that value and immediately you will do when we are in the population domain.

Oh! Yes we will get the values when you go for the sampling but at least before sampling you do not have though all those values so what you can do for a particular variable each concerned you can expect something okay what is this expectation suppose we want to know what is the average height of variety students this is nothing but the expected value of x so that expected value of x or the variable we are inter raised decision known as mean at each mean.

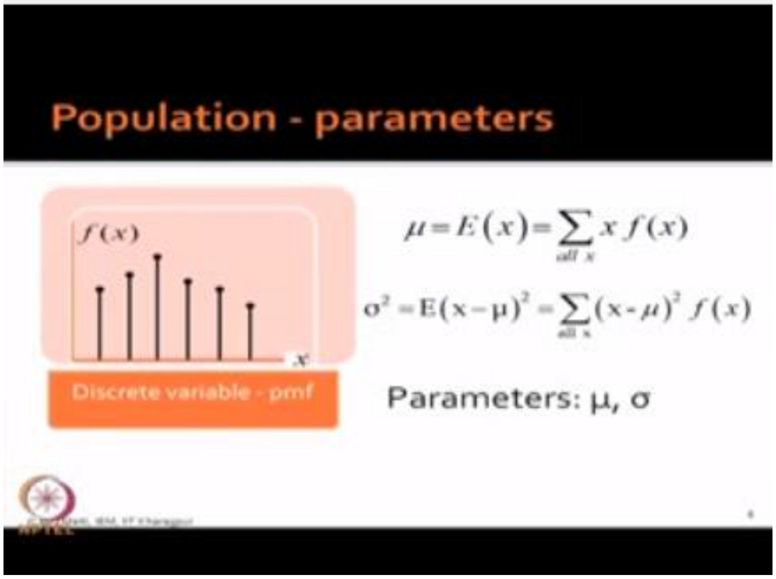
(Refer Slide Time: 25:00)



Mean: $\mu = E(x) = \sum_{\text{all } x} x f(x)$

Mean stands for mean men h expected value of x when your variable x each discrete variable so you will get like this x fx for this for all i all sorry all x what about the maybe the your number for all x if I if you see this example here if you see the example.

(Refer Slide Time: 25:38)



Population - parameters

Discrete variable - pmf

$$\mu = E(x) = \sum_{all\ x} x\, f(x)$$

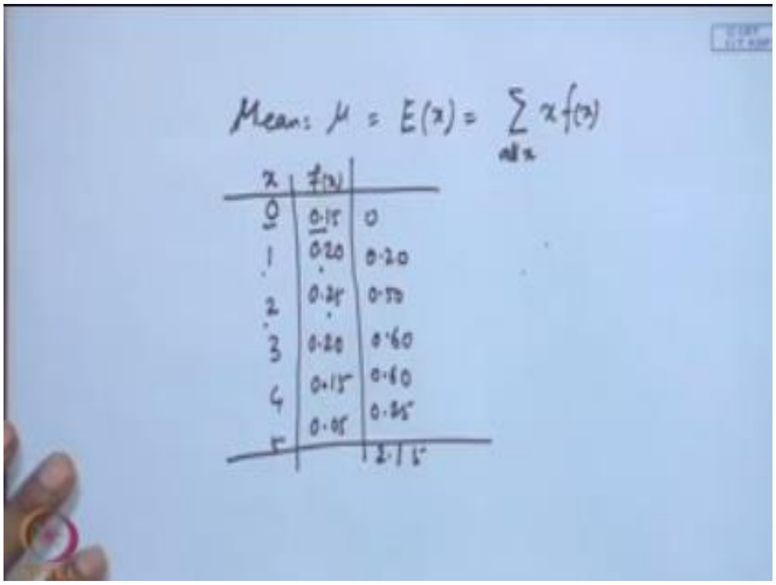
$$\sigma^2 = E(x-\mu)^2 = \sum_{all\ x} (x-\mu)^2 f(x)$$

Parameters: μ, σ

[illegible], IEM, IIT Kharagpur

So we are we are thinking is saying here that x can take this value like this oh there are 1 2 3 4 5 6 values so if I assume that these values are.

(Refer Slide Time: 25:50)



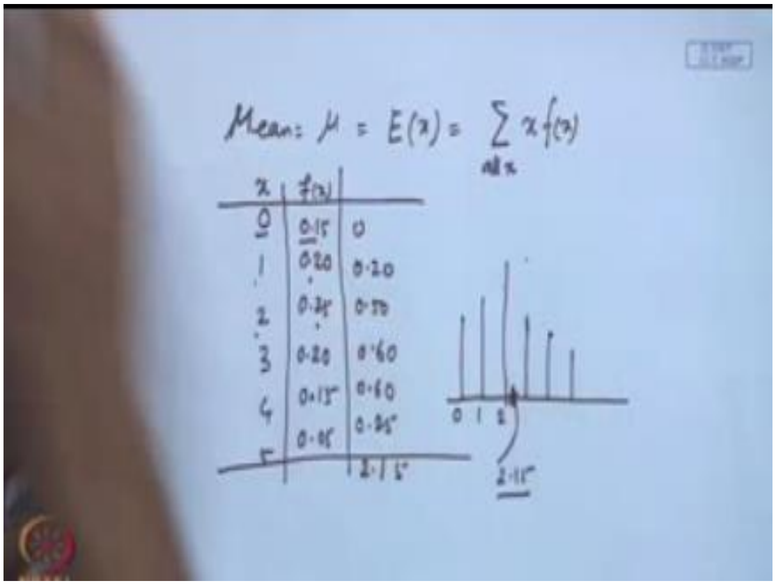
$$\text{Mean: } \mu = E(x) = \sum_{\text{all } x} x f(x)$$

x	f(x)	
0	0.15	0
1	0.20	0.20
2	0.25	0.50
3	0.20	0.60
4	0.15	0.60
5	0.05	0.25
		2.15

Nothing but 0 1 2 3 4 5 and so these are all x values and if I assume that they are density here as discrete they are also the probability values then their probability values is like this suppose the first one is 0.15 second one is 0.20 third one is 0.25 forth one again you can write 0.20 fifth one is suppose 0.15 then what is left the 20, 40, 60 95 so 0.05 then what is your expected value here x into fx we are defined that so 0 x .15 is 0 12.2 each point 2. 2 in 2.25 age. 50 3 in 2.2 is 0.60 so like this gain 0.60 and this should be 0.25.

If you add what you will get you add 5 6 + 2 8 14 19 21 so 2.15 so that means your expected value is if this is 0 this is 1 this is 2 somewhere here so if I draw here I can say that.

(Refer Slide Time: 27:22)

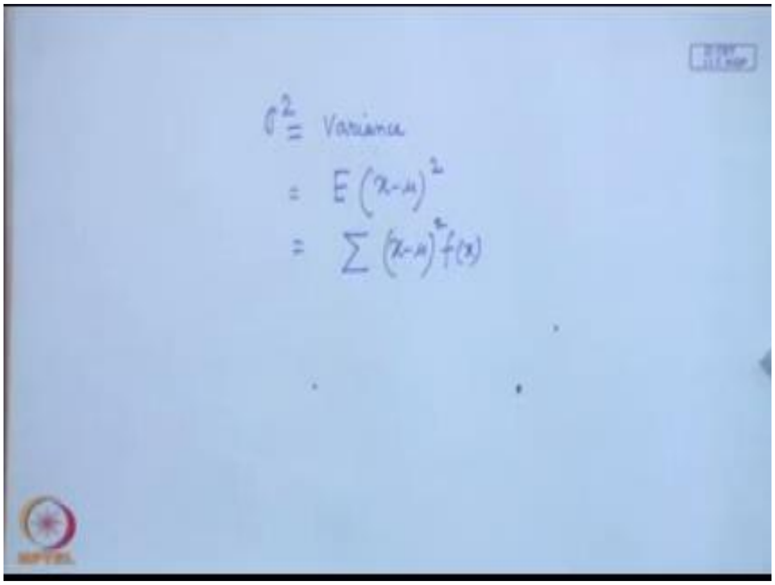


Mean: $\mu = E(x) = \sum_{\text{all } x} x f(x)$

x	$f(x)$	
0	0.15	0
1	0.20	0.20
2	0.25	0.50
3	0.20	0.60
4	0.15	0.60
5	0.05	0.25
		2.15

Suppose this is my 0 value this is 1 values this 1 two values and this one is three values this is your four and then five so this is 0 1 2 some here this is more values 2.15 this is what is expectation but another major here it is here so another one is σ2 so we have we said here your new that is mean we have just said as well as.

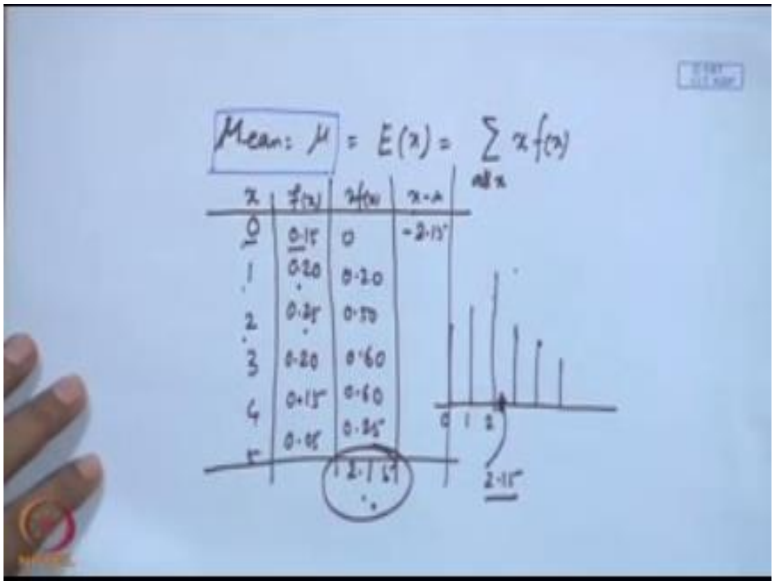
(Refer Slide Time: 28:13)



$$\sigma^2 = \text{Variance}$$
$$= E(x-\mu)^2$$
$$= \sum (x-\mu)^2 f(x)$$

Let there is another major which is σ^2 that is the variants okay what is variance is expected value of x – μ whole square so for this case your discrete case you will write x - μ whole square f$_x$.

(Refer Slide Time: 28:50)



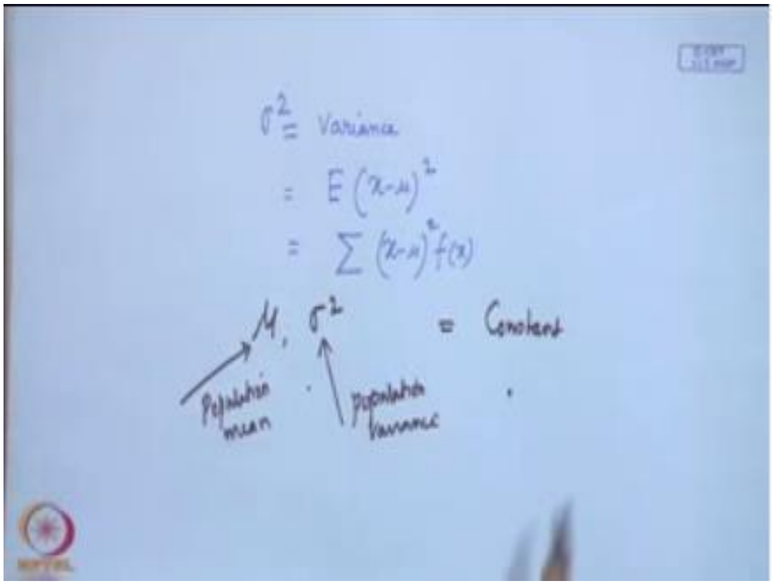
$$\boxed{\text{Mean: } \mu} = E(x) = \sum_{\text{all } x} x f(x)$$

x	$f(x)$	$xf(x)$	$x-\mu$
0	0.15	0	-2.15
1	0.20	0.20	
2	0.25	0.50	
3	0.20	0.60	
4	0.15	0.60	
5	0.05	0.25	
		2.15	

And you have computed here two things what are those things that you computed x then your this one is the x f_x you know the μ value mean value know now you can create x – μ that mean 0 – 2.15 that will be – 2.15 like this we can calculate and again you square it multiply it then add it so you will be getting the σ^2 that is variance square okay then what we have assumed here we assumed that x can take this 5 values only then this is the probability mass function.

So will be the total of this probability values one then what is μ and σ or σ^2 correct long run standard deviation that means you are saying that mean and that mean and standard deviation they will vary for a population probably particular characteristic. No even for small population, it should not vary it is a constant, when you talk about the parameter so actually these are here.

(Refer Slide Time: 30:29)



$$\sigma^2 = \text{Variance}$$

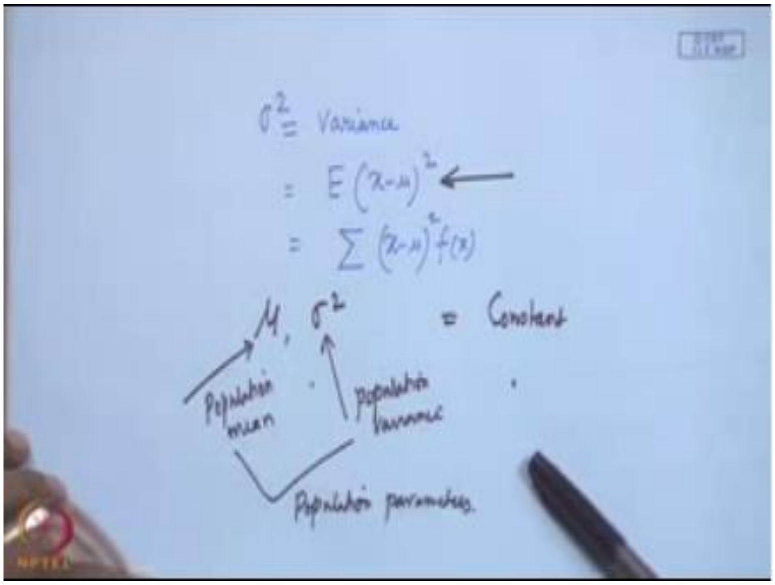
$$= E(x-\mu)^2$$

$$= \sum (x-\mu)^2 f(x)$$

$$\mu, \sigma^2 = \text{Constant}$$

This μ and σ^2 in statically we say population mean and population variance they are constant, the another issue will be they are not known also, here we are assume to a very finite population small size population and we will calculate it something like this, actually a population size in finite you will not get all values of x, you will never get if the suggestion finite, so that mean you cannot measure this.

(Refer Slide Time: 31:30)



σ^2 = Variance

$= E(x-\mu)^2 \leftarrow$

$= \sum (x-\mu)^2 f(x)$

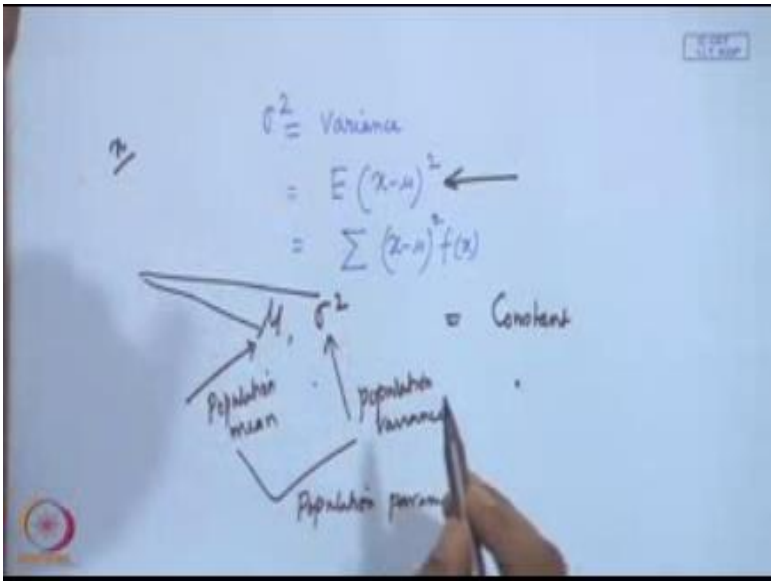
μ, σ^2 = Constant

Population mean
population variance
Population parameters

Or I can just compute this, but only what you can do, you can expect something that is why the expectation term is used here. Now the expectation, okay so if I say population parameter so now you can understand that these two are population parameter it is by saying these two are population parameter please do not consider that there are, there is no other population parameters, these two are sum of the population parameter and many population parameters but okay, these two mean and standard deviation or variance a population parameters. Why we go for population parameter?

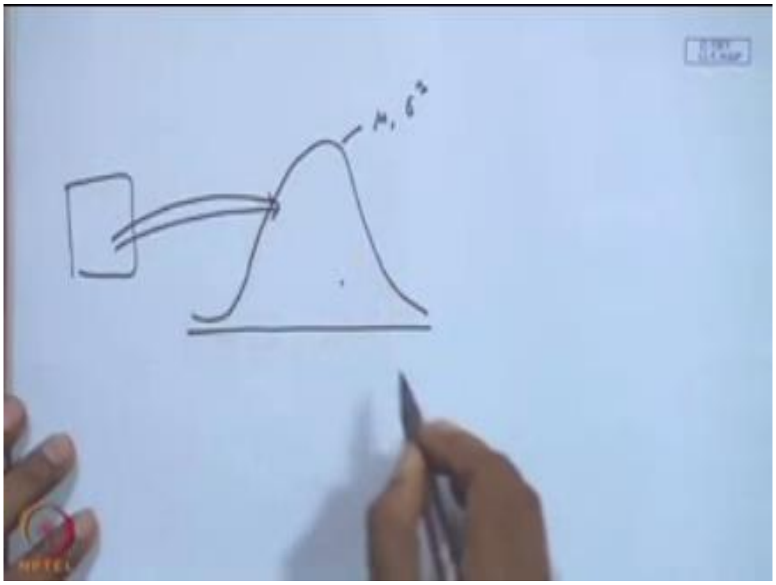
Because see, the lecture is today is this topic is very simple topic, calculation point over that, but understanding point with you we must understand that why we declare population parameter. We equate population parameter because.

(Refer Slide Time: 32:38)



If you know these two parameter and also you know that you are x is random variable and that can follow certain probability distribution and if you know that distribution and with also know the parameters what happen, you do not require to go for that particular system or process for further study, for this particular variable is concerned.

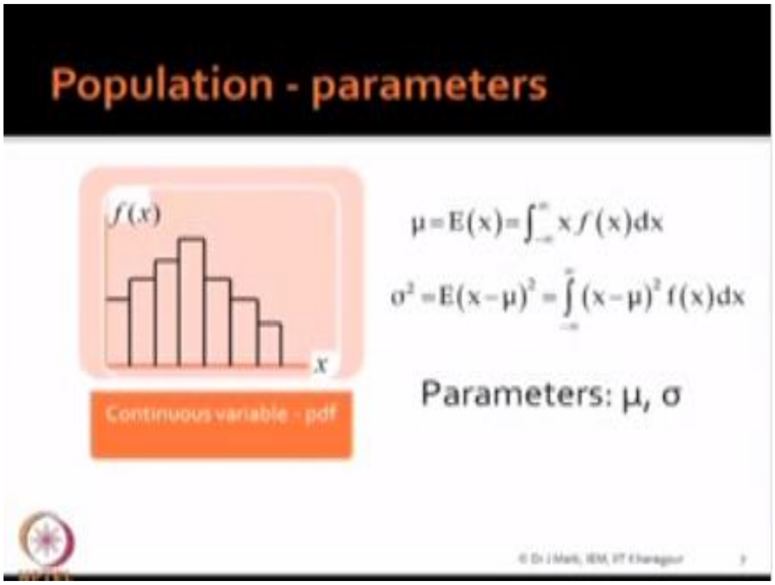
(Refer Slide Time: 33:05)



If I see that the absent is in the top floor for the production sub considered if follows something like normal distribution and we know that it is mean is μ and σ^2 is the variance component that mean I am in a position to this. So if I know the distribution what is actually happening here that real productions of from worker performs point is view here the absent is in is converted to a mathematical equation statically equation, that is the advantage, that means if I know truly I know the what is the probability distribution with respect to a variable and what are the population parameter for that variable I have the distribution with at my hand.

So I do not require to go for there, so long the process we will not change by process will not change what I mean to say that suppose it is a machine, in machine works what time machine condition did to its, so that mean today a new machine its performance now after ten years the machine will not perform at the same level, that mean what happen, the characteristic changes. So long the characteristic is not changing therefore even the distribution itself will enough for you.

(Refer Slide Time: 34:39)



Population - parameters

Continuous variable - pdf

$$\mu = E(x) = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

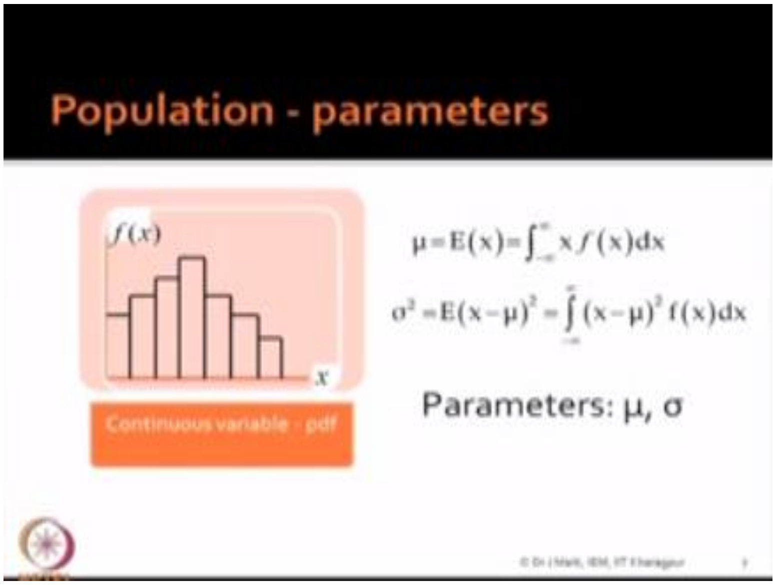
$$\sigma^2 = E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\,dx$$

Parameters: μ, σ

© Dr J Maiti, IEM, IIT Kharagpur 3

Okay, now if your variable discrete, not discrete that is continuous your variable is continuous you see, what is this one.

(Refer Slide Time: 34:49)



Population - parameters

Continuous variable - pdf

$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x) dx$$

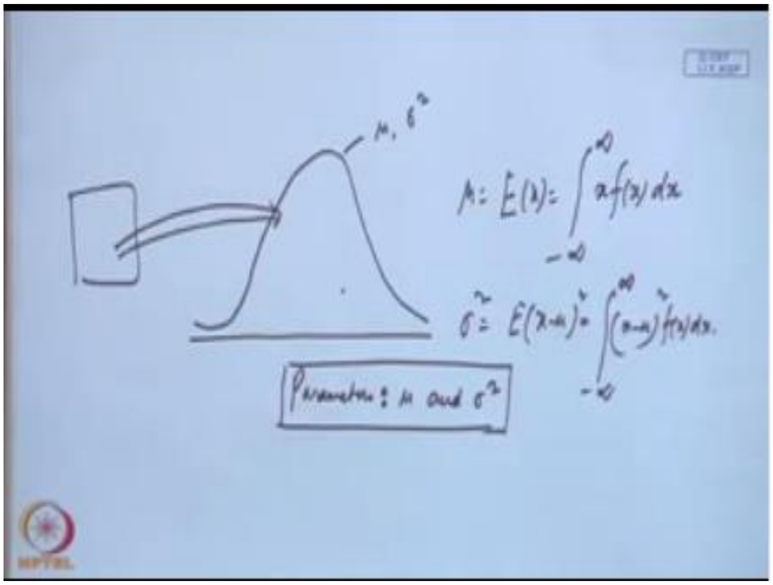
$$\sigma^2 = E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$$

Parameters: μ, σ

© Dr J Maiti, IEM, IIT Kharagpur

Left hand side, this is μ for pdf, this is probability density function. Now why in continuous case we see a probability density function, whereas in the discrete case we see a probability mass function looking this one. So here also, if we know that this particular population it has mean and variance component and where it is in the continuous level you have to use these two equations for expectation. So basically, integration will come into picture.

(Refer Slide Time: 35:37)



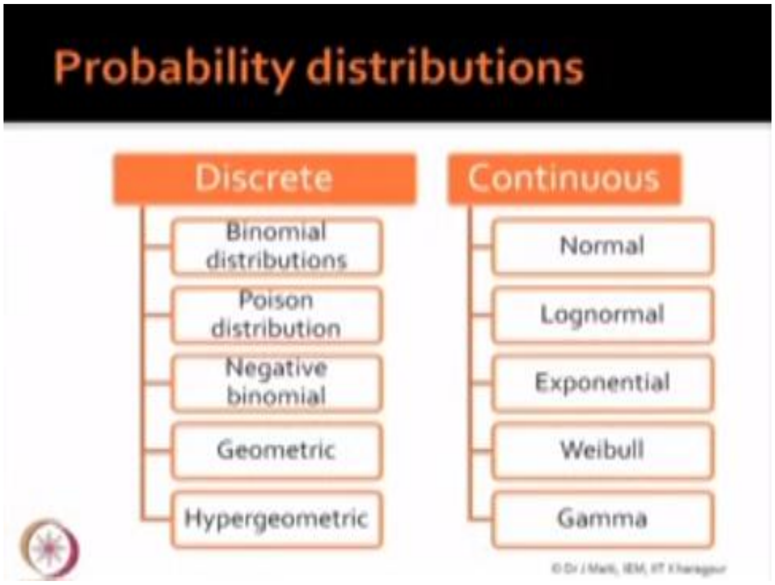
$$\mu = E(x) = \int_{-\infty}^{\infty} x f(x)\,dx$$

$$\sigma^2 = E(x-\mu)^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx.$$

Parameters: μ and σ^2

E(x) this is integration -∞ to +∞ depending on the range for which the variable you define, then f(x)dx and your σ^2 is nothing but again $(x-\mu)^2$ this is infinite to infinite that $(x-\mu)^2f(x)dx$. So I am saying the parameters μ and σ^2 here. I hope that you understand now let what is population and population is characterized by quality distribution, if the random variable has a probability distribution and if you know that for variable the parameters of the distribution you have characterized the population, that is what is known as characterized population in terms of probability distribution. Now there are many probability distributions, you see here.

(Refer Slide Time: 36:56)



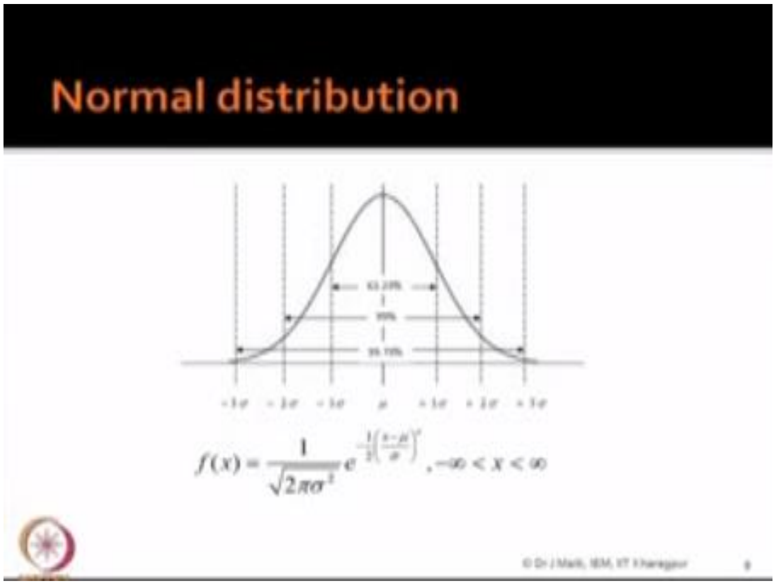
Probability distributions

© Dr J Maiti, IEM, IIT Kharagpur

We have, we can see here that under two heads discrete and continuous under discrete distribution binomial distribution, poison distribution, negative binomial, geometric, hypergeometric many more is series. Similarly under continuous normal, abnormal, exponential, Weilbull, gamma there are so many distributions at probability distributions. We will not discuss all the distributions we will only discuss normal distribution here.

Because in multi variant statistical modeling normality assumptions, this normality assumption is very, very valid to vital one many of the models there is some assume normality of the data definitely at the multi variant level with that is be multi variant normality.

(Refer Slide Time: 37:55)



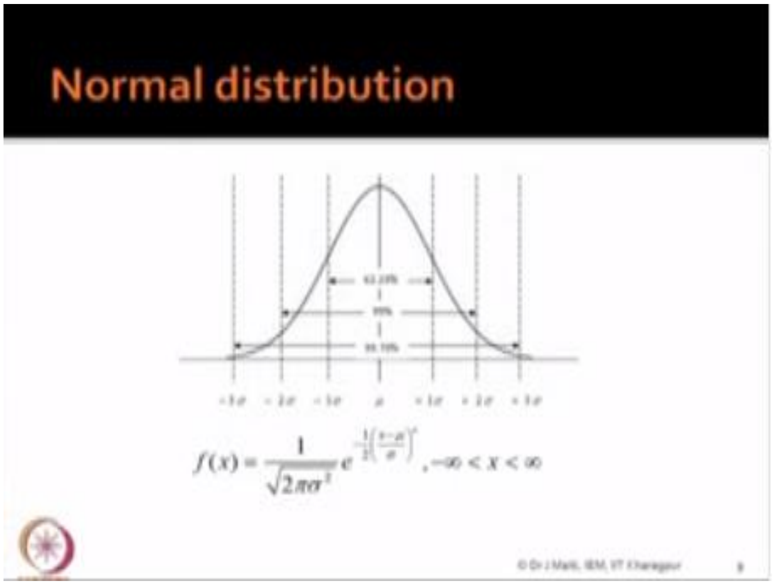
Normal distribution

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

© Dr J Maiti, IEM, IIT Kharagpur 8

So we will discuss only on normal distribution are the distribution you can follow junction book easier you can follow this. and I am sure all of you are familiar with these distribution, this is what is normal distribution, it looks like this, and this μ is the center point here and here that is basically symmetric so maximum number of observations also you will find out along this level and it gradually both side it will gradually reduced and finally after 3σ level it will almost negligible to 0th level like this.

(Refer Slide Time: 38:43)



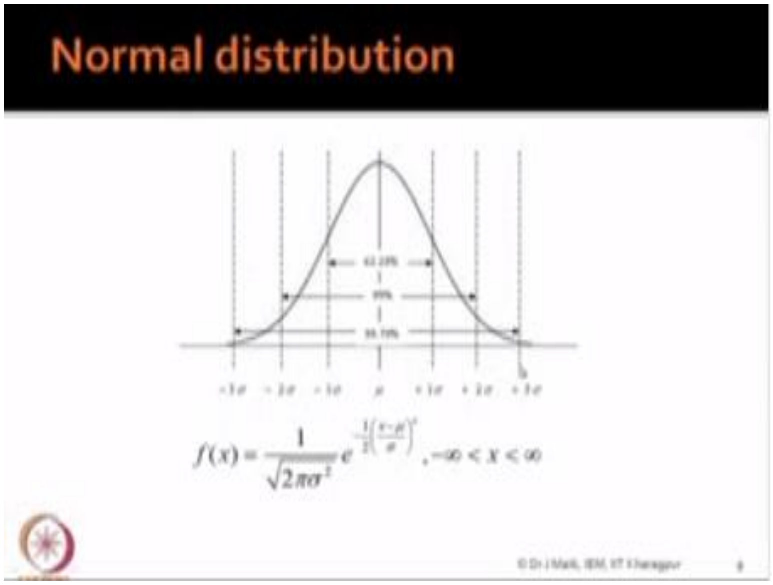
Normal distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < X < \infty$$

© Dr J Maiti, IEM, IIT Kharagpur

Okay, now the, how to read this normal distribution you see that within 1σ ± 1σ, this one is very important -1σ to +1σ 62 to 0.23% observation follows within this, then within ± 2σ level it will be little more than 95%, with more than 95% but not 96, 95 point something and if you consider ±3σ level then your 99.73% observation will fall under this category this zone, okay. Now this is important because suppose you think that you are producing something and you’re the variable of interest follows normal distribution.

(Refer Slide Time: 39:40)



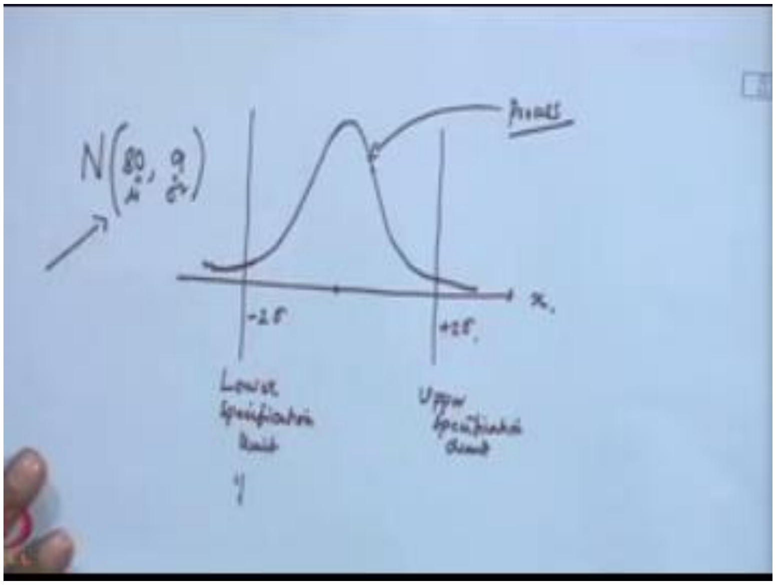
Normal distribution

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

© Dr J Maiti, IEM, IIT Kharagpur 8

Now from datas you are using, you will get like this distribution is like this. What will happen that the speed of this particular variable values within ± the σ level 99.735 of the items produce will fall under within this, now what will happen if the customer will not be interested at the wide range.

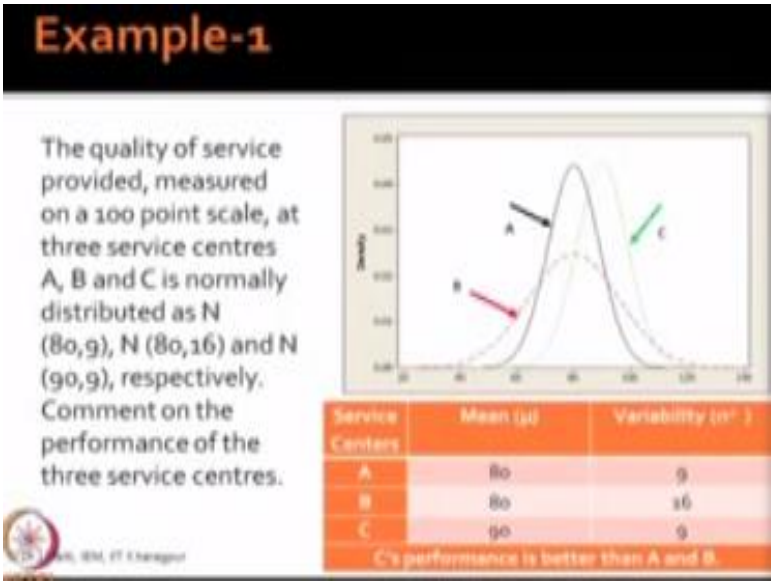
(Refer Slide Time: 40:05)



For example this is my quality characteristics x and this one is the lower specification limit this, I am giving the physical inter division, example this is your upper specification limit then what will happen ultimate label this is the main one but your normal this policy is normal distribution and your distribution maybe like this. So this maybe let it be -2σ, +2σ so that mat what will happen ultimately that 5% almost 5% of your production each reject product.

Because people that customer will not accept it so when I talk about say that characteristics is not the process that mean with respect to this distribution I mean this is your process exactly that this is the soft weapon on in this process you are not going to this up lowed this is the case but you customers is in here and you are producing at this will 5% of your production is not accepted by the customer you want to improve it. Rating me you want to improve it how will do it if just can we explain this figure.

(Refer Slide Time: 41:42)



Example-1

The quality of service provided, measured on a 100 point scale, at three service centres A, B and C is normally distributed as N (80,9), N (80,16) and N (90,9), respectively. Comment on the performance of the three service centres.

Service Centers	Mean (μ)	Variability (σ^2)
A	80	9
B	80	16
C	90	9

C's performance is better than A and B.

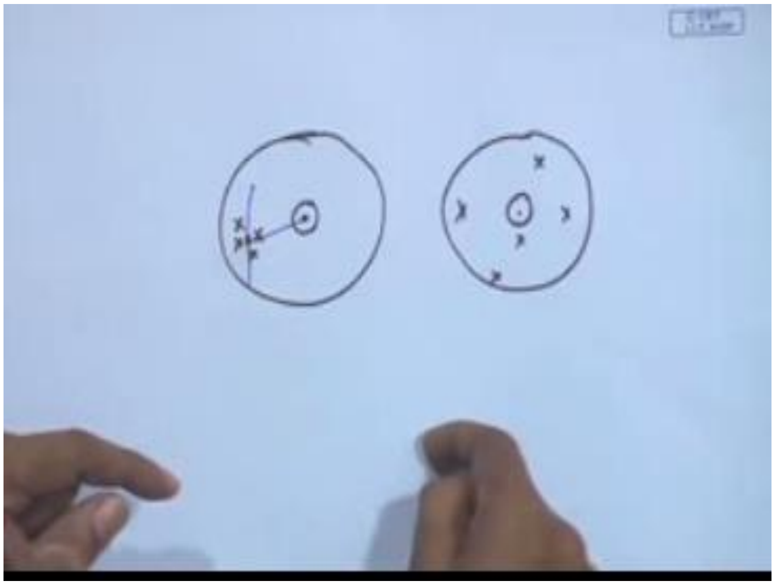
This figure you have see this is again basically as I told you that characteristics process through already distribution this is another example the quality of service provided measured on a 100 point scale at three service center a, b, and c is normally distributed as N 80, 9 80 stands for the main value 9 stand for the variants, because a original notice and what will be following is normally distributed non instant for normal distributed then μ first is μ, that is σ or σ^2 basically so μis this and σ^2 is this.

This is the general notation we will be following all through, then your second process as a b80, 16 third one is 99 and I approved the probability distribution for all the three process a, b, and c and please remember that your variable of inter is it is quality of service provided if I ask you which process is better? You will say c yes or no, yes why mean yes your right so main value is 9D and it is quality of service provided on a 100 point scale you are measuring higher the value the vat of the process.

But Parolee you see the variability is 9 so both from main point of view it is that the higher level and variable D point of view it is at the lowest level compare when you compare the 3 process, now if ask you from compare a and b which one is better a because the variability is low mean at the same level 8 so what is the physical interpretation of this D, physically what happen it is the variability the most difficult parameter.

Very difficult to control variability I am giving you another important good example here suppose use thing you all know that Archie this one.

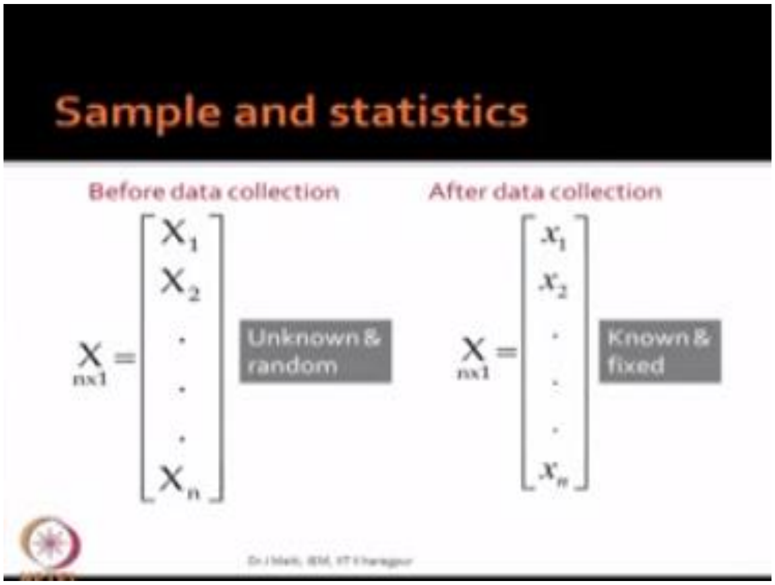
(Refer Slide Time: 44:24)



Suppose this is the bullshey that one our this the gold medal winner what is his name that, I do any how this is the target now someone all shooting are here and someone shooting is like this, this is the bullshey you are the trainer to shooters a and b who by training who will be improve first, first one the position is first one the position level is higher, then the second one what will happen you can shift this main value to this.

But here is variable one when something is variable win for the students point of view some students are very pirating irregular very, very difficult but some students maybe because of some reason very regular were suddenly or they are basically coming late by some few minutes always you can but they are still coming you can motivate them and so this is the physical meaning of the distribution.

(Refer Slide Time: 45:58)



Sample and statistics

Before data collection

$$\underset{n\times 1}{X} = \begin{bmatrix} X_1 \\ X_2 \\ . \\ . \\ . \\ X_n \end{bmatrix}$$

Unknown & random

After data collection

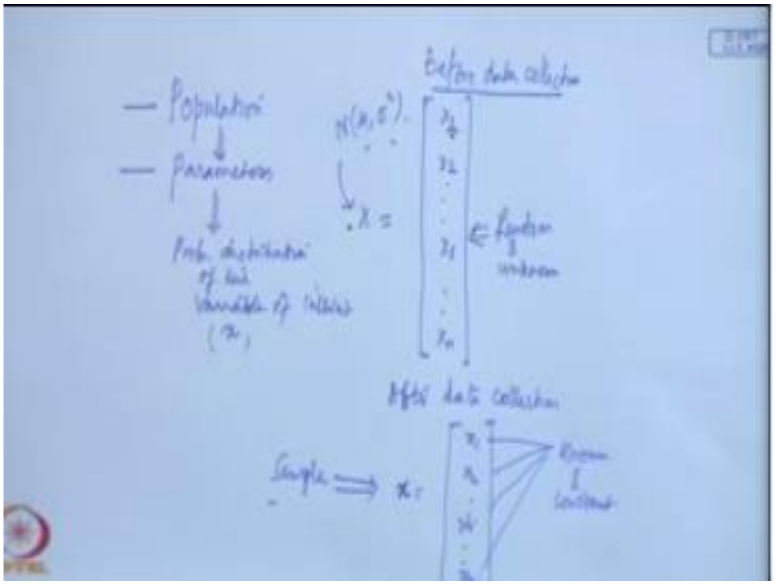
$$\underset{n\times 1}{X} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

Known & fixed

Dr J Maiti, IEM, IIT Kharagpur

Now we will come to the next important concept is called sample and statistic so we have seen so far that population and parameters.

(Refer Slide Time: 46:11)



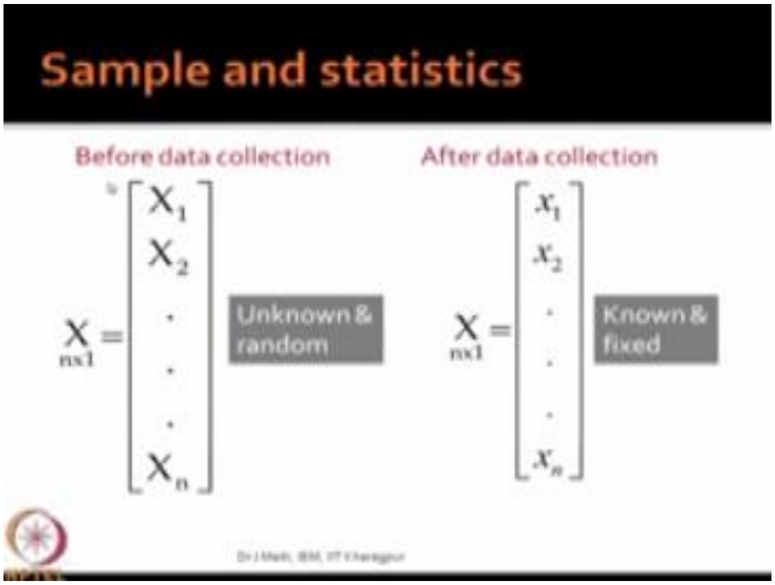
And please remember the random variable is they are abrivaring so population and parameter when you talk about population with deepingly talk about parameters and we talk about in the particularly for this subject the probability distribution also probability distribution of the variable of interest, that is x variable inter is x now see you are planning to collect data what you are thought of that I know my variable that variable is x I want to collect data how many data you want to collect you want to collect n data points.

What you can say about each of the observation what can we expect getting me so if x is normally distributed it is suppose x is normally distributed with μ and σ^2 that means x_1 also normally distributed with mean in σ^2 ’ x_2 is normally distributed with the μbut they are coming from the same population there is no guaranty that when you if you go you will observe some value of x.

Somebody else will go you will observe some different value of x even thought the observation is the first observation fort you it is first for m also it is first. Keeping mind this one this is very important concept you have not collected data this before data collection you are planning that you will be collecting data n data so x_1 to x_n first you can observe x_1 like x_2 like x_n now what is the issue here this all these values are random.

As it they are basically it like random and unknown okay now you third the real collected data after data collection what will happen so you will collect data when I am denoting in terms of small x let at be this small x so x_1 x_2 x_i x_n what are this values known values realized every value is realized known and constant. This is in the population domain this is now sample of sigeon you see this slide here.

(Refer Slide Time: 49:14)



Sample and statistics

Before data collection

$$\underset{n\times 1}{X} = \begin{bmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_n \end{bmatrix}$$

Unknown & random

After data collection

$$\underset{n\times 1}{X} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

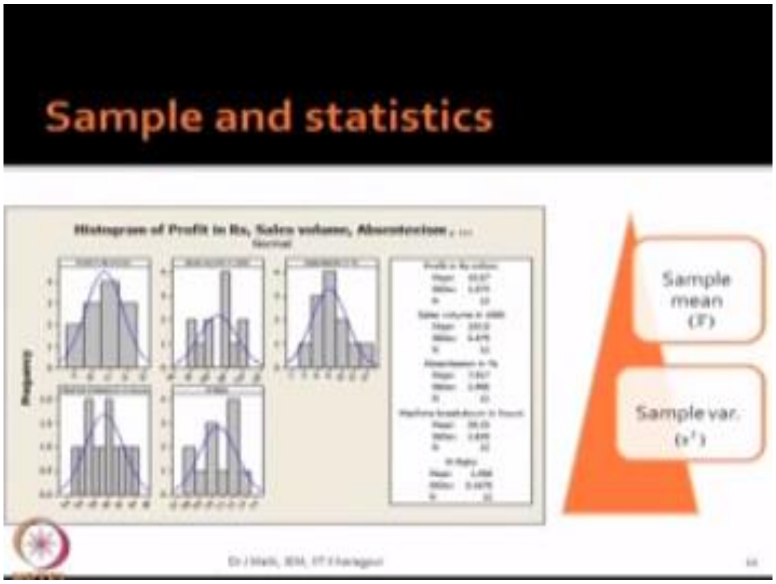
Known & fixed

Dr J Maiti, IEM, IIT Kharagpur

Before data collection you have plan to collect n data point and the unknown and random and when you collect data after collection they already known fixed values. That is the big difference and another important concept you keeping mind that all the observations each of the observation will follow the same probability distribution by same probability distribution you meaning that if it is normal distribution with μ and σ^2 as mean and variants x_1 also a normal distribution with mean μ σ^2 as variants.

Once data you have collected forget all distribution now foxed value there is no random les in the data to you already collected.

(Refer Slide Time: 50:10)



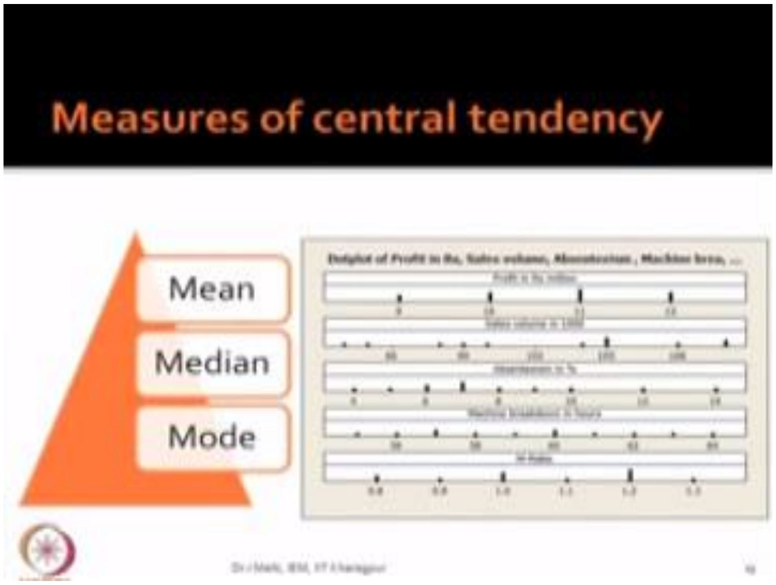
Sample and statistics

Sample mean ($\bar{x}$)

Sample var. (s^2)

The this one can we remember that example last example that small company I can give a name to this I have giving that city camp so late on I will use the what city can company may city can so these are the profit sales volume have cent agents all those things what is basically we have we are correct raging the that city can process in totality in terms of this variables and your sample mean and sample variants these are the statistics with respect to population lean and population variants. Correct

(Refer Slide Time: 51:04)



Measures of central tendency

Mean

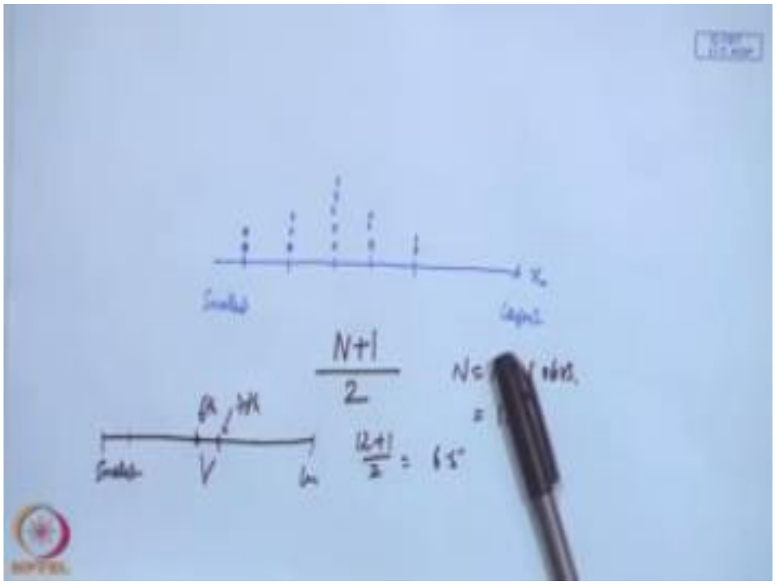
Median

Mode

Dr J Maiti, IEM, IIT Kharagpur

Do you know what is this is dot plot; dot plot is something like this.

(Refer Slide Time: 51:18)

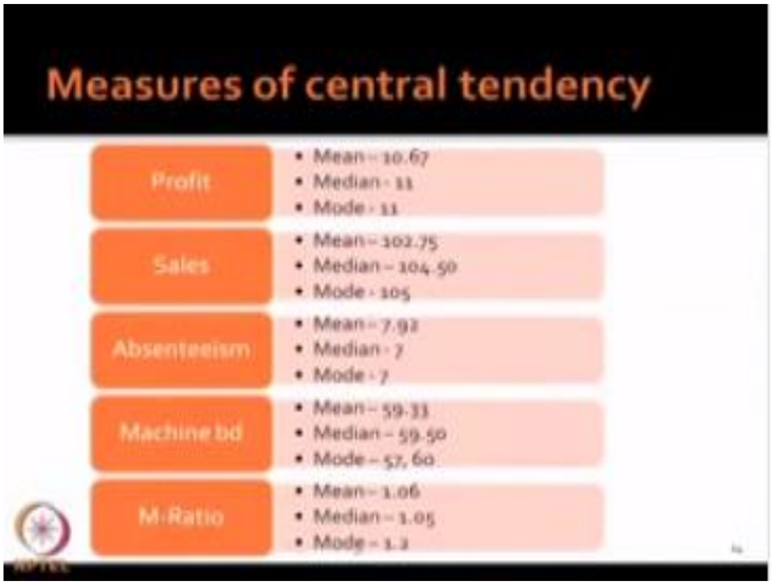


Your variable is x you arrange from the smallest to the largest now this is suppose one value this is second value this is the third value like different values are there suppose this value there is you have observations they are let it be 3 observation here let it be 5 observations this type of plotting you are doing here, so again these suppose this again two values this is let it 2 values like this.

This is known as dot plot. Dot plot again it is similar to gram plot now here you are able to count the number of observation against each of the values of the x so it will help you to find out the mode suppose if I say for this example profit in rupees million you say 9 million Rs cash it is 2 observations for tan it is 3 for 11 it is 4 for 12 it is 3 so that mean the mode of the data points for profit is 11.

Mean you have already seen median is the middle value hoe do calculate compute median may for computation of median and you have to find out the position n +1.2 where n is the number of observations in this particular example there n =12, because 12 months data so 12 + ½ that means 6.5 so 6.5 means when you arrange your data from smallest to the largest you just find out the position then 6 and 7 position suppose this is your 6 position this is you 7 position you take the average of this 2 values 6 position value 7 position value. Okay so and what is the mode? Mode is the value of x for which there are maximum occurrences

(Refer Slide Time: 53:31)



Measures of central tendency

Profit
- Mean – 10.67
- Median - 11
- Mode - 11

Sales
- Mean – 102.75
- Median – 104.50
- Mode - 105

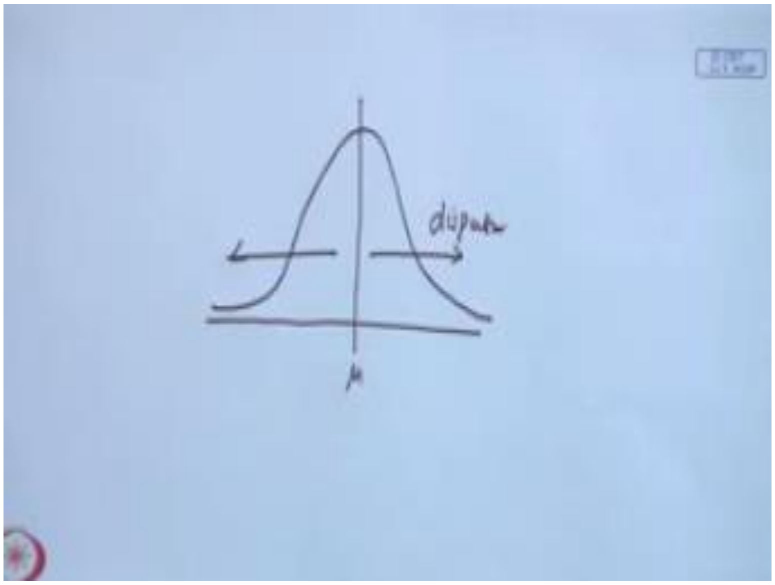
Absenteeism
- Mean – 7.92
- Median - 7
- Mode - 7

Machine bd
- Mean – 59.33
- Median – 59.50
- Mode – 57, 60

M-Ratio
- Mean – 1.06
- Median – 1.05
- Mode – 1.2

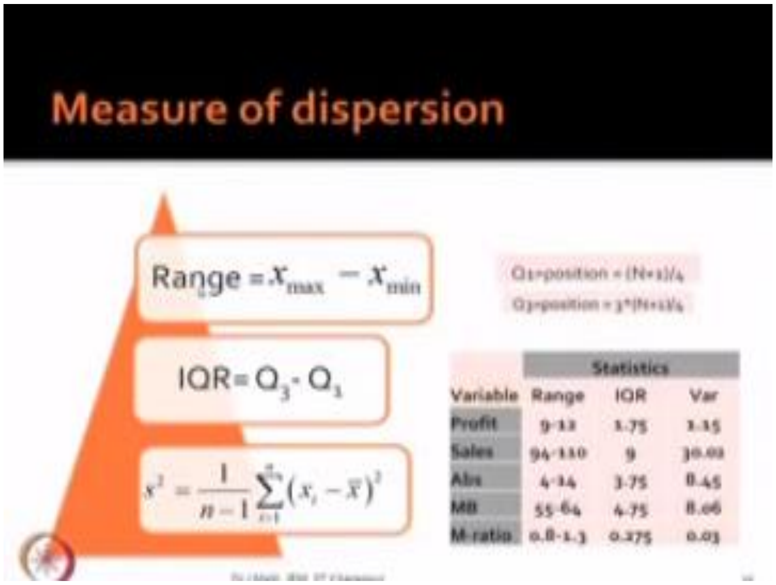
This is the calculation for that data and using excel sheet you can very easily calculated this thing now major of dispersion is this what we have seen that.

(Refer Slide Time: 53:47)



That if the data follows suppose normal distribution then this one is μ and this side how much it is going and this side that is dispersion there are several ways to measure dispersion one is range.

(Refer Slide Time: 54:04)



Measure of dispersion

$\text{Range} = x_{\max} - x_{\min}$

$\text{IQR} = Q_3 - Q_1$

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

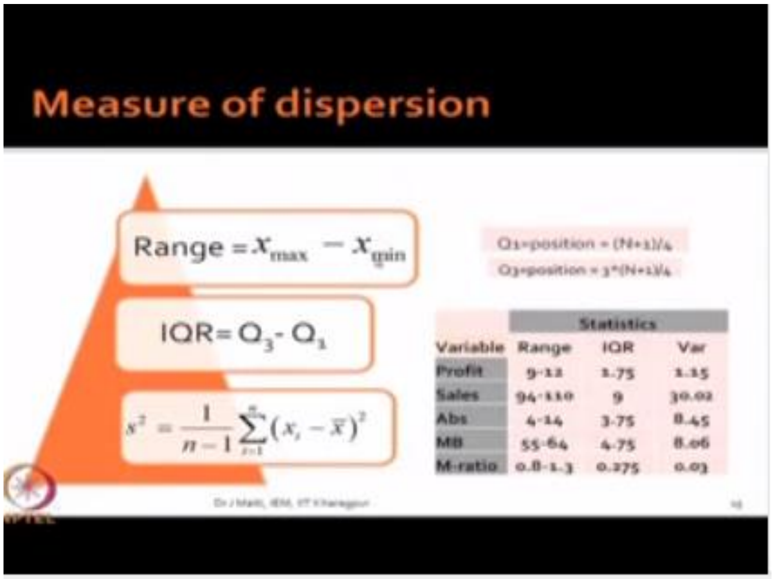
Q1=position = (N+1)/4

Q3=position = 3*(N+1)/4

	Statistics		
Variable	Range	IQR	Var
Profit	9-12	1.75	1.15
Sales	94-110	9	30.02
Abs	4-14	3.75	8.45
MB	55-64	4.75	8.06
M-ratio	0.8-1.3	0.275	0.03

That is minimum maximum - minimum value and other one is inter quartile range which the third quartile – first quartile and this are all position.

(Refer Slide Time: 54:06)



Measure of dispersion

$$\text{Range} = x_{\max} - x_{\min}$$

$$\text{IQR} = Q_3 - Q_1$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Q1=position = (N+1)/4

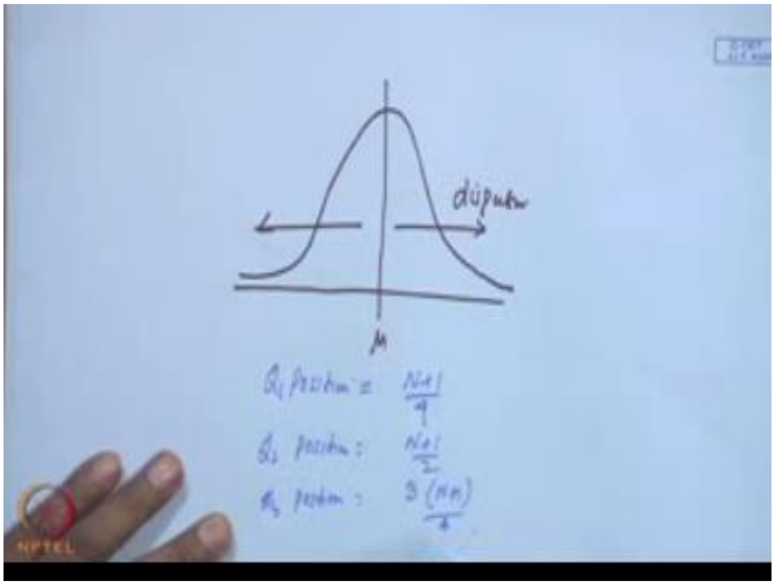
Q3=position = 3*(N+1)/4

	Statistics		
Variable	Range	IQR	Var
Profit	9-12	1.75	1.15
Sales	94-110	9	30.02
Abs	4-14	3.75	8.45
MB	55-64	4.75	8.06
M-ratio	0.8-1.3	0.275	0.03

Dr J Maiti, IEM, IIT Kharagpur

There are several value to major dispersion one is range, that is maximum – minimum value, another one inter quatel range which is the third quatel – first quatel and these are all position. First quatel is basically N+1/4 and your quatel one position Q1 position is N+1/4, then Q2 position is the median which is N+1/2 and Q3 position is, that is third quatel which is 3(N+1)/4. So, all those position values, you have to find out and appropriately have to manipulate the data.

(Refer Slide Time: 54:45)



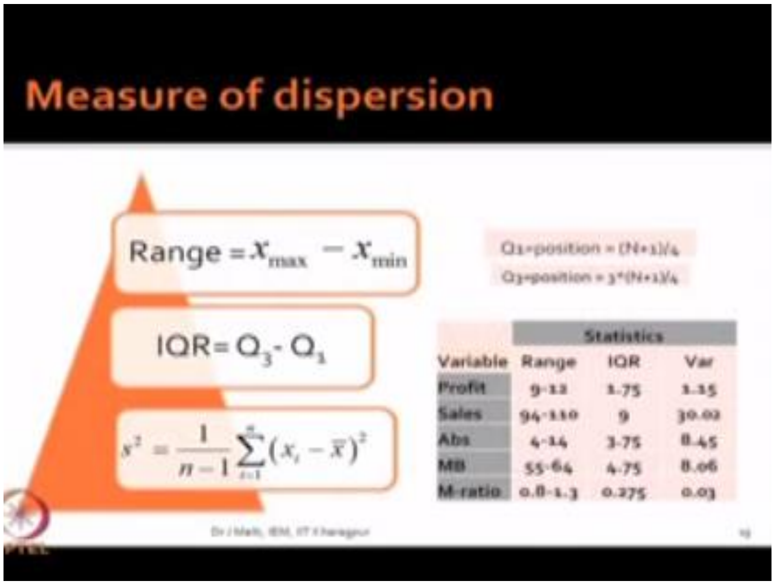
Q_1 position = $\frac{N+1}{4}$

Q_2 position : $\frac{N+1}{2}$

Q_3 position : $\frac{3(N+1)}{4}$

So, if there is two values is coming in the middle then you take the average and if it is not coming middle may be more than the middle, say 7.75 position, suppose 3.75 position. So, accordingly that 7.75 that will be given for the data. Ok? And these are all very simple thing you be able to find out, but these are equally important and you require to know also this.

(Refer Slide Time: 55.25)



Measure of dispersion

$$\text{Range} = x_{\max} - x_{\min}$$

$$\text{IQR} = Q_3 - Q_1$$

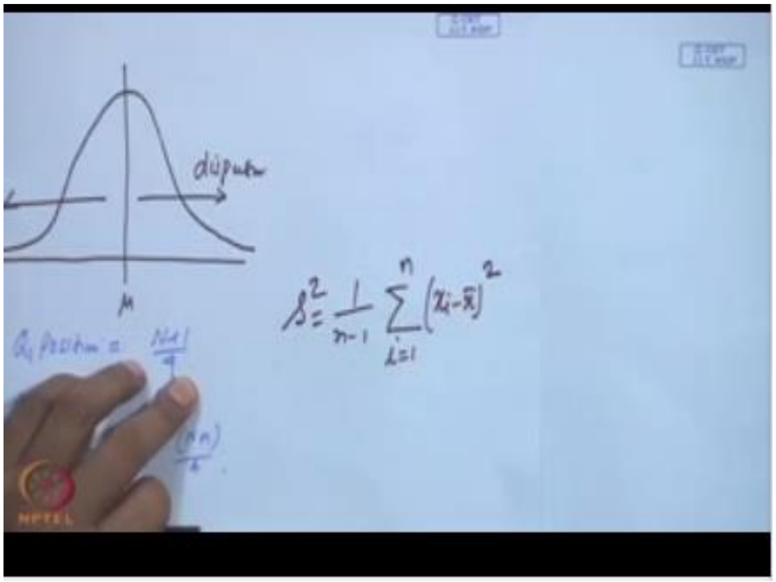
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Q1=position = (N+1)/4

Q3=position = 3*(N+1)/4

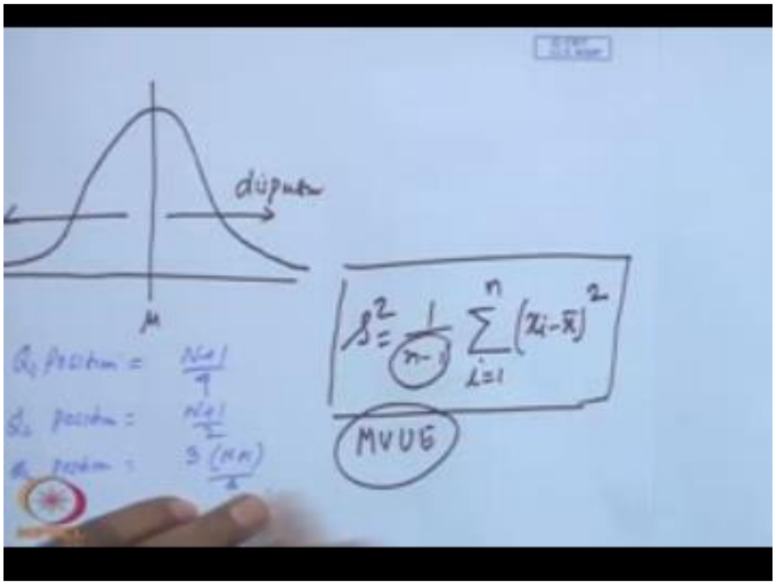
	Statistics		
Variable	Range	IQR	Var
Profit	9-12	1.75	1.15
Sales	94-110	9	30.02
Abs	4-14	3.75	8.45
MB	55-64	4.75	8.06
M-ratio	0.8-1.3	0.275	0.03

(Refer Slide Time: 55.49)


dispersion
μ
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
Q1 Position = $\frac{N+1}{4}$
$\frac{(n)}{4}$
NPTEL

And you the variants also, So yes or no? How to compute variance, statistical sense s^{2}=1/n-1 sum total of I =1 to n then (xi-$\bar{x}$) ^{2,} this is the variability measure. Why this n-1? Minimum variance unbiased estimated.

(Refer Slide Time: 56:20)



Q_1 position = $\frac{N+1}{4}$

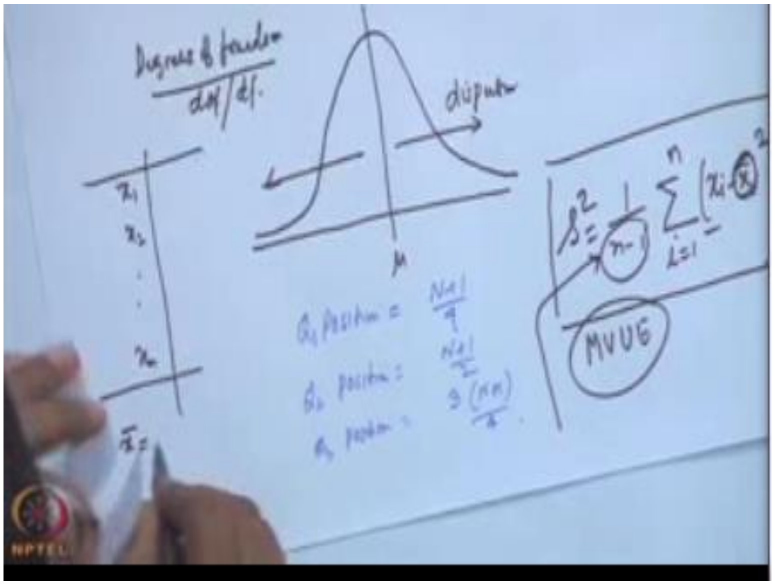
Q_2 position = $\frac{N+1}{2}$

Q_3 position = $\frac{3(N+1)}{4}$

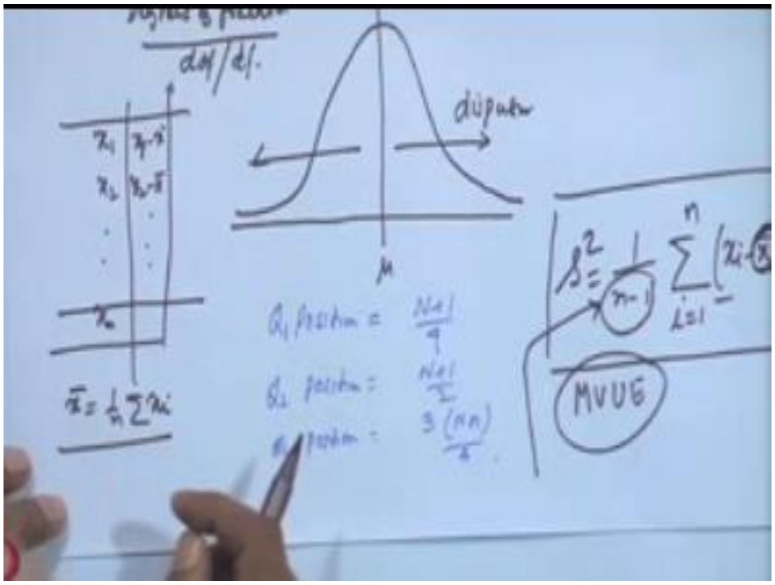
$$\boxed{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Ok any other explanation?

(Refer Slide Time: 57:11)


Degrees of freedom
d.o.f./d.f.
x_1
x_2
x_n
$\bar{x} =$
dispersion
μ
Q_1 position = $\frac{N+1}{4}$
Q_2 position = $\frac{N+1}{2}$
Q_3 position = $\frac{3(N+1)}{4}$
$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
MVUE
NPTEL

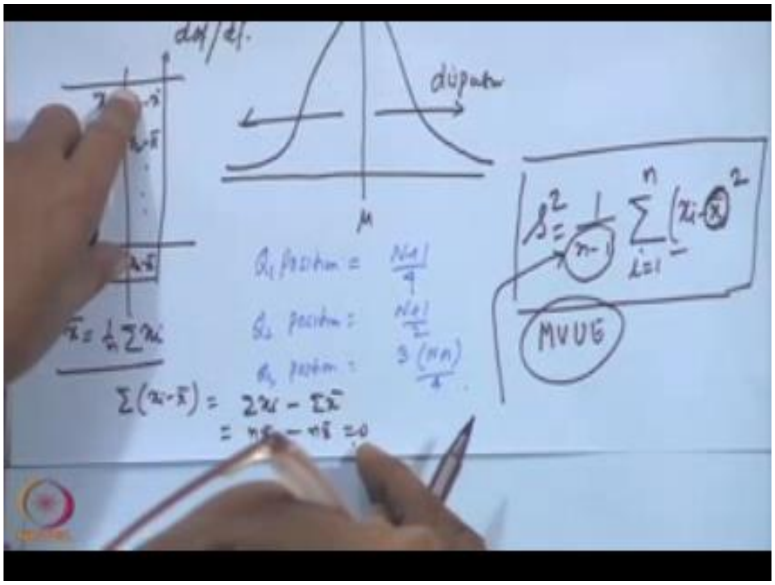
(Refer Slide Time: 57.29)


dof/df.
dispersion
x_1
$x_1-\bar{x}$
x_2
$x_2-\bar{x}$
x_n
$\bar{x}=\frac{1}{n}\sum x_i$
μ
Q1 position = $\frac{N+1}{4}$
Q2 position = $\frac{N+1}{2}$
Q3 position = $\frac{3(n+1)}{4}$
$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}$
MVUE

Now later on we will discussing very much, Very frequently the degrees of freedom, will be using dof or df . Now see in this case this n-1 is coming because of the degrees of the freedom, because you have in data points x1 to xn and when you are computing these variants you require, what you require? You require x̅ to be computed. So, as x̅ is computed with this formulation, So what is happened ultimately here when you find out that xi- x̅ is, that is x1- x̅ and x2- x̅ like this, for the last one you don't require to compute.

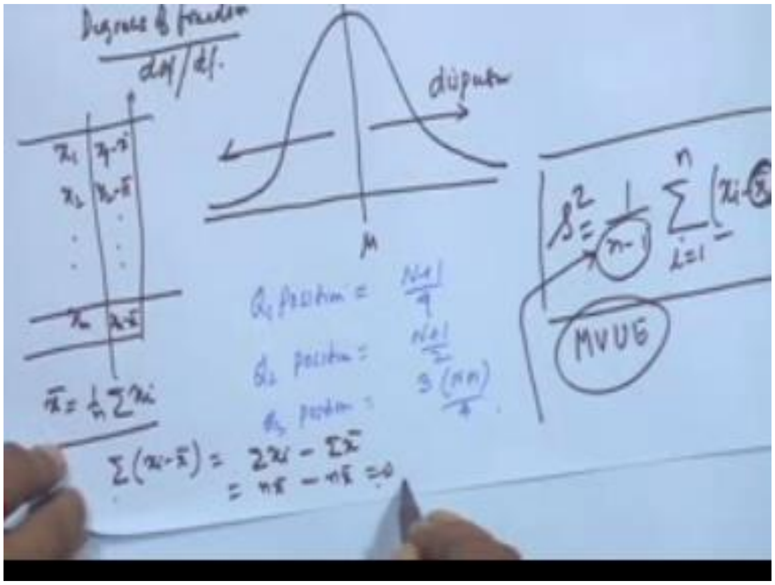
It is automatically computed, So I will write here suppose xn- x̅, what I mean, I mean that suppose if I write like it this sum of xi-x̅, what will be the value? So it will be now, what is mean value, so that mean summation of xi – summation of x̅ this is in nx̅-nx̅ ,so this is zero.

(Refer Slide Time: 58:13)


d.o.f/d.f.
dispersion
μ
Q1 position = (N+1)/4
Q2 position = (N+1)/2
Q3 position = 3(N+1)/4
$\bar{x} = \frac{1}{n}\sum x_i$
$\sum(x_i-\bar{x}) = \sum x_i - \sum\bar{x}$
$= n\bar{x} - n\bar{x} = 0$
$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2$
MVUE

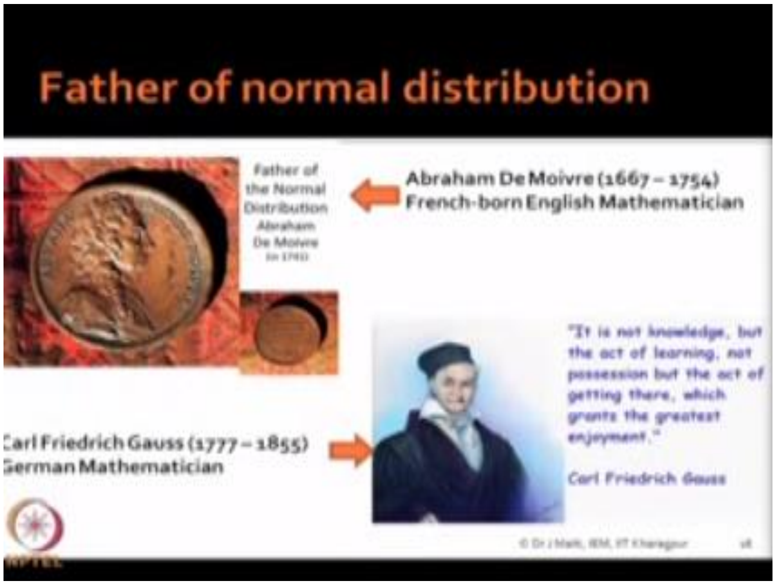
Very simple other way, if you say suppose I have given you one equation, suppose x+y+z=5, what is the degree of freedom here, you see that, if you change the x and y, z cannot be change, it's fixed. Even though the three values are there, but you have two digit of freedom, because I have made,it is made as 5,in this case also the same thing is happening, the sum of all those digits will be zero.

(Refer Slide Time: 58.44)



So you require how many data points you haves s-s^2 in equation n-1,that is the another explanation of yn-1 will be divided while computing this.

(Refer Slide Time: 59.06)



Father of normal distribution

Father of the Normal Distribution Abraham De Moivre (in 1760)

Abraham De Moivre (1667 – 1754)
French-born English Mathematician

Carl Friedrich Gauss (1777 – 1855)
German Mathematician

"It is not knowledge, but the act of learning, not possession but the act of getting there, which grants the greatest enjoyment."

Carl Friedrich Gauss

© Dr J Maiti, IEM, IIT Kharagpur 18

NPTEL

Now I will complete this lecture, See we have told you ,that normal distribution is very very important, and later on for this subject multi varied normal distribution. So we must know that who is the father of normal distribution and you see that Abraham De Moivre ,this French born English mathematician, basically he has given this general form of this normal distribution, in what form you see that 1 by root of power $2\prod\sigma^2$ into the power minus of x minus μ by sigma to the power square.

Okay, Now he is considered the father of normal distribution, but only equation will not become be sufficient, later what happened that Gauss, he is another famous mathematician, statistician also, what he has given the properties, all statistical properties of normal distribution is identified and tested by Carl Fedris Gauss, he is basically that germen mathematician, and if you see this is very interesting, it is not knowledge but the actable learning, not possession but the act of getting their which grants the greatest enjoyment.

So long suppose you do PhD, suppose not getting PhD ,you are thinking that once get PhD, I will be very happy, but it is not true once you get within two three days ,you will find you are the same person, but the run going there the act of going there, that is what is very, very important and this famous people have quoted and we must obey to their all suggestions.

Thank you very much, Next class I will tell you the sampling distribution

NPTEL Video Recording Team

NPTEL Web Editing Team

Technical Superintendents

Computer Technicians

A IIT Kharagpur Production

www.nptel.iitm.ac.in