

Design and Analysis of Experiments
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture – 06
Random Variables and Probability Distribution

Welcome to the sixth lecture of D O E and now within half an hour of time, we will discuss random variables and probability distribution and I hope that some of you know what is random variable and what are the different kinds of probability distribution, but some of you may not be knowing. So, as a result what I think that there must be half an hour discussion on this.

(Refer Slide Time: 00:58)

Contents

- Random variable and its types
- Probability and probability distribution
- PDF, CDF, mean and variance of random variable
- Important distributions

Source: This lecture is prepared primarily based on "Engineering Statistics" by D C Montgomery, G C Runger and N F Hubele, Wiley, 5th Edition, 2013

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us see what is random variable the content of presentation now is random variable and its types probability and probability distribution probability density function cumulative distribution function mean and variance of random variable and important distributions.

(Refer Slide Time: 01:15)

Random variable and its types

- A **random variable** is a numerical variable whose measured value can change from one replicate of the **random experiment** to another.
- A **discrete** random variable is a random variable with a finite (or countably infinite) set of real numbers for its range.
Example: Number of scratches on a surface, proportion of defective parts among 1000 tested, number of transmitted bits received in error
- A **continuous** random variable is a random variable with an interval (either finite or infinite) of real numbers for its range.
Example: Electrical current, length, pressure, temperature, time, voltage, weight

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Let us say theoretically define what is random variable random variable is a numerical variable whose measured value can change from one replicate of a random experiment to another what does it mean; if you recall our example what we are discussing so far that we said that we have done experiment and we got 24 observations and they are all those these 24 observations are randomized.

(Refer Slide Time: 01:45)

L6

Random Variable & Probability Distribution

24 obs

L 1
H 2
H 3

RV

Discrete ← Prob mass fn (pmf)

Continuum ← Prob density fn (pdf)

RV = ~

Physical process

Prob density fn (pdf)

So, the experiment was randomized; that means, that way in the ground clay clutter level abc or 1, 2, 3;

Student: (Refer Time: 02:07).

1, 2, 3. So, ground clutter level 1, 2 3 that is low medium and high and you got 6 data points here, 8 data points here, 8 data points and 8 data points total 24 data points. Those data point; data point the observed Y value, this can based on certain random experiment. So, what we say; that means, in this case the response variable the detection the signal intensity level at the when the operator is able to detect the target this is a random variable.

So, that mean a random experiment the outcome of interest of random experiment is a random variable. Now random variable can be discrete, can be continuous. So, random variable can be discrete 2 types can be continuous suppose the intensity level at target detection it can have any value it is not that 90 or 90 one it can be 90 point something or some like this. So, that is a continuous random variable.

But in some cases you may find you may find that you do not get values in between some values like if I toss several coins, the number of heads is a random variable; random value and it is the variable is also random variable, but if I say what is the how many what is the number of heads; you can say n ; $n + 1$ or 5, tail like this, you cannot say 5.5 or 6.5 or 5; 625. So, these are all counts. So, these are all discrete data discrete random variable.

So, let us see that a theoretically what we were a discrete random variable is a random variable with a finite or countably infinite set of real number for its range for example, number of scratches on a surface we can say that 5 scratches 10 scratches 50 scratches proportion of defective parts among thousand tested. I think it is number of defective parts among thousand tested similarly number of transmitted bits received in error this number some counts are coming; what is continuous random variable continuous random variable is a random variable with an interval either finite or infinite of a real number for its range electrical current you may be you are measuring using the unit ampere.

So, it can be having that ampere 5 to 10 ampere anything possible or 50 ampere anything possible 50.5, 50.05, 50.005; anything possible length overall pressure exerted in a body temperature in the environment time taken to complete the war voltage in a electric circuit weights all those things these are all continuous. So, for discrete I said a number of scratches number of defective parts among these number of transmitted these, but if

we make proportion; what happened? It will; it can be within a certain range, it can have many values. So, you do not write proportion you say number.

(Refer Slide Time: 05:52)

Probability

- **Probability** is used to quantify the likelihood, or chance, that a measurement, described by a random variable, falls within some set of values.
- A probability is usually expressed in terms of a random variable.
- *For example*, if we repeatedly manufacture parts (replicate the random experiment an infinite number of times), and 25% of them will have lengths in the interval (10.8 mm to 11.2 mm), Y denotes the part length and the probability statement can be written in either of the following forms:

$P(Y \in [10.8, 11.2]) = 0.25$ or $P(10.8 \leq Y \leq 11.2) = 0.25$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Then a random variable have a peak; it is interesting characteristics is there; what is this? I say that random variable takes some value, but in advance you cannot tell that what is that value; for example, when you toss a coin you know that 2 outcomes will be there head or tail, but before or while tossing the coin you cannot say whether head or head will come.

If it is a random one and it is a unbiased one, you cannot say it head; will you; will see head. So, that mean outcome is known, but what outcome you will get that is not known outcomes are head or tail in case of toss in case of our this detection level that intensity range maybe a 70 to 120 that will be the intensity level that is also known, but at when an operator detect through the scope whether his intensity level will be 80 or 90; that is not known in advance unless the experiment is conducted.

So, then; what happen because of this nature? So, unless the things are not happened we are not in a position to know, but fortunately what happen these there as there these can be can be estimated or it can be can be known through certain probability distribution because random variable even though you cannot still precisely what is the value, but what happen you may say that what is the probability that the value will be less than 90

or what is the probability that the value will be 90 to 100 or if it is a discrete case. So, what is the; now what is the probability that the number of defective items will be 5?

So, that kind of characterization is possible. So, when you talk about random variable we associate it with a probability distribution now the probability distribution in case of discrete random variable it will be discrete probability distribution in case of continuous random variable it will be continuous probability distribution anyhow let us theoretically define that what is probability? Probability is used to quantify the likelihood or chance that a measurement described by random variable falls within some sets of values a probability is usually expressed in terms of a random variable they are basically Intel interchangeable.

When you say random variable you definitely know that you are basically interested to know the probability distribution of this random variable when you say probability distribution you are basically talking about distribution of random variable. Now let us see one example, if we repeatedly manufacture parts; that means, replicate the random experiment an infinite number of times and 25 percent of them will have lengths in the interval 10.8 millimeter to 11.2 millimeter and if Y denotes the part length and the then the statement the probability statement can be written in the following manner.

That probability Y belongs to 10.8 to 11.2. This interval it will be 0.25 or probability that the Y varies in between 10.8 to 11.2. In terms of $10.8 \leq Y \leq 11.2$ that will be 0.25. This is what is the way; we define probability.

(Refer Slide Time: 09:46)

Probability properties

- $P(Y \in R) = 1$ where R is the set of real numbers
- $0 \leq P(Y \in E) \leq 1$ for any set E
- If E_1, E_2, \dots, E_k are mutually exclusive sets,

$$P(Y \in E_1 \cup E_2 \cup \dots \cup E_k) = P(Y \in E_1) + \dots + P(Y \in E_k)$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, I will quickly give you some of the or a glance through some of the properties for example, if we consider Y is a set of real number, then the Y value; the random variable value, it belongs to that set of real number is equal to 1. Its practical meaning is that the probability value can be maximum of one means the probability value can maximum value; value probability can be one the second probability is 0 less than equal to $P(Y \in E)$ a set of b ; an even E less than equal to one for any set E . So, here what does it mean that probability value can never become negative it will lie in between 0 and 1.

Suppose there are there are k number of mutually exclusive sets representing that may be k number of mutually exclusive events. So, what happened if you take some of them and find out the probability of them, then this is nothing, but finding out the individual probability and summing them. This is what is written here, if E_1, E_2, E_k are mutually exclusive sets probability Y belongs to this equal to probability Y belongs to you want plus probability Y equal to E_2 like this.

So, these are the some; these are some properties which you will be you keep in your mind because you may require sometimes to derive some statistics or some other things as I as I said that random variable and probability distributions are very much close and there are probably the continuous random variable. So, continuous probability distribution and discrete random variable discrete probability distribution; so, this

probability distribution in case of continuous variable is termed as probability density function; probability density function which we say PDF.

In case of discrete we say probability mass function which we say PMF, I expect that you know that probability density and probability mass function if you do not know you go through the engineering statistics book by Montgomery Rancher (Refer Time: 12:15) and you will get a fair treatment on quality distribution for the time being you just understand that this is what is probability pd a probability density function a x axis will be the variable of a random variable of interest here it is Y and Y x E is the density which is function of y.

(Refer Slide Time: 12:35)

Properties for continuous random variables

- ❖ **Probability Density Function**

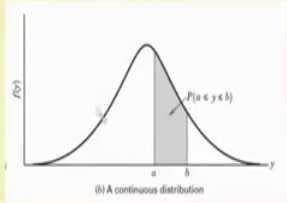
$$P(a < Y < b) = \int_a^b f(y) dy$$
- ❖ **Cumulative Distribution Function**

$$F(y) = P(Y \leq y) = \int_{-\infty}^y f(u) du \quad \text{for } -\infty < y < \infty$$
- ❖ **Mean and Variance**

$$\mu = E(Y) = \int_{-\infty}^{\infty} yf(y) dy$$

$$\sigma^2 = E[(Y - \mu)^2]$$

$$= \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy$$



IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And when I say that what is the probability that that a less than equal to Y and less than equal to b then you are talking about the area under the curve within this range and that is what we have written this. So, this area is computed using this formula a to b f y d y. Now F y; what is the; what is this function F y? That depends on what kind of distribution, it is; if it is normal distribution then with normal parameters we can define f Y those things you will see later another one is the cumulative distribution function. Suppose I consider a value a here and I say that what is the probability that that Y value will be less than equal to a then this is nothing, but this is the area under the curve under this curve less than a and this is written; that means, as we assume that the random variable ranges from minus infinity to plus infinity.

So, that is why we are writing like this the cumulative distribution function. Function is $F(y)$ probability that Y less than equal to a certain value y small y is minus infinity to y $f(y)$ du when this it ranges Y ranges from minus infinity to plus infinity. So, for continuous distribution we use cumulative density function and also cumulative distribution function cumulative distribution function will there will be there for discrete distribution also, but interestingly this is a; this is representing the population the characteristics of the population. In this case, Y the process model the output response variable that is Y .

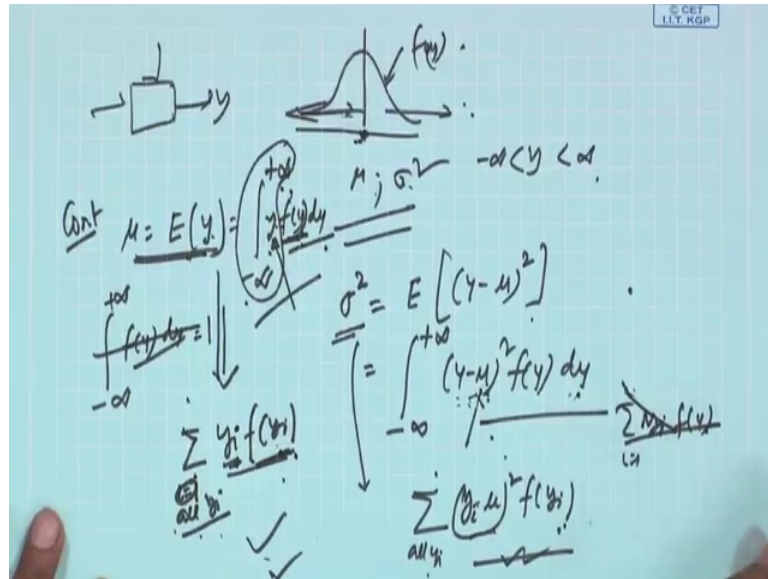
This is nothing, but the it is basically there is a physical process and if I say that Y for Y is like this behaving like this then this is what is happen Y is behaving like this that this is a statistical process you may have converted a physical process to statistical process statistical process in physical process is this one. So, you are giving like this and like this and some x some Y you are getting and statistical process here what happen this Y is this is the $f(y)$ and this side y and this side $f(y)$.

This is one kind of interpretation so; that means, this distribution if you know that mean you know the behavior of Y , when you observe in reality you find out what is the value what earlier values and what will be the next value that also you will be seeing when you get that one, but here what happened using this past data you know that this is the distribution. So, what happened maybe your variable value of value of interest is this and you want that your all the values will be less than this then you wanted to know what is the probability for that the Y value will be less than this.

This may be Y target value, then; this is nothing, but the area under this curve and this we must know because sometimes what happened. Suppose if Y value greater than this, this may be a case that the all those are not will be accepted by the customer then these are all rejects.

So, you want to know what is the probability that things will be rejected this will give you valuable information more interestingly this diagram is giving you the what is the behavior of Y ; whether it is normal abnormal or some exponential or something else that will come in terms of statistical distribution. So, as I told you earlier also that the distribution that or the population is characterized by certain parameters.

(Refer Slide Time: 16:38)



So, if I say my; this population is characterized by this distribution that means this distribution definitely have certain parameters what are those parameters these parameters is mu and the dispersal. Now if I know mu and sigma square then I can I can do my analysis and I can infer about the population. So, so what is in population domain; what is this mu; mu is nothing, but the mean of the population which is expected value of Y this we write in terms of minus infinite to plus infinite. If we consider Y range from minus infinite to plus infinite and then $Y f y dy$ please remember $Y f y d y$ suppose if I ask you what is the value of integration minus infinite to plus infinite $f y dy$. This is the area under the curve this will be one total complete probability total probability is one only it cannot be more, but here one more thing is added here multiplied Y times this.

This is the expected value means I can say that if you do another experiment and you may tell that you may say that the response value will be this. This is the best possible value expecting you are expecting that this is going to happen similarly what will be the expected value of variance that that is sigma square. So, sigma square is expected value of this person with reference to suppose $Y - \mu$ the square we have seen earlier in sample calculus sample statistics calculation $y_i - \mu$ square by $n - 1$ those things we have written this square this is nothing, but again minus infinite to plus infinite $Y - \mu$ square $f y d y$.

So, as it is from minus infinity to plus infinity if and if you know mu and then definitely you will be able to integrate it and get the value suppose if you do not know mu. So, then you cannot calculate a sigma square what is this mean mu is the mu is the mean of the population not mean of the simple mean of the population. So, many a times what happen you call it representative sample compute sample average which will be considered as the estimate of mu in that case you use this and get this one.

But please keep in mind that population parameters are seldom known. Now what will happen?

(Refer Slide Time: 19:47)

Properties for discrete random variables

- ❖ **Probability Mass Function**

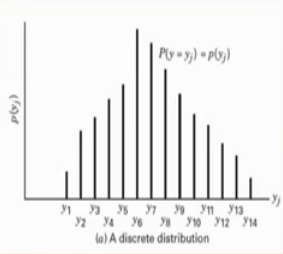
$$f(y_i) = P(Y = y_i)$$
- ❖ **Cumulative Distribution Function**

$$F(y) = P(Y \leq y) = \sum_{y_i \leq y} f(y_i)$$
- ❖ **Mean and Variance**

$$\mu = E(y) = \sum_{i=1}^n y_i f(y_i)$$

$$\sigma^2 = V(y) = E(Y - \mu)^2 = \sum_{i=1}^n (y_i - \mu)^2 f(y_i)$$

$$= \sum_{i=1}^n y_i^2 f(y_i) - \mu^2$$



(a) A discrete distribution

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

If we go for discrete random variable see the discrete random variable case here Y value are all some counts here that mean in the graph. In this graph, the every values this is its location its (Refer Time: 20:03) this location is these in between and y 1 and y 2 nothing is there. So, because if I say y 1 is 50 and y 2 is 52 or 50 sorry; y 2 is 51. So, in between that 50 and 51; in 51.5 is not occurring because of the discrete nature. So, now, in this case what happened what is the probability that y 1 Y value will be y 1, then it this is nothing, but the height of this bar relative height basically of this bar. So, here this side probability of Y j and this side Y.

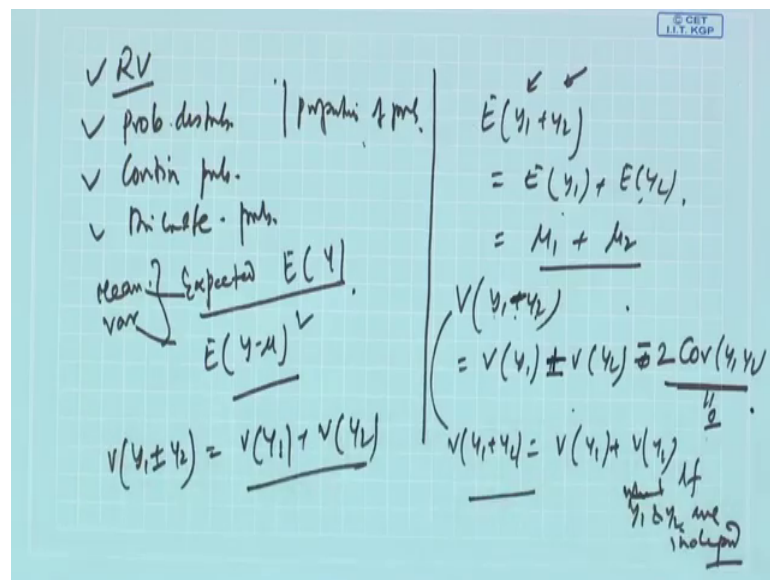
this is giving you the probability mass function not density function I request you to know what is the difference between mass function and density function then a cumulative similarly less than equal to Y here you will be using summation when you are

going for mean and variance you see mean here instead of integration you are using here summation $\sum_{i=1}^n y_i ; f y_i$ here also instead of integration you are using summation. So, analogous to this when I am saying that mean in case of continuous expected value of Y minus infinite to plus infinite $\int Y f y dy$.

So, when it is discrete what you are writing this quantity you are writing like this $y_i f y_i$ and i equal to 1 to n can say all y_i and here you are not writing integration here this integration same part you are writing sum summation here also in case of discrete it will be sum of that all y_i , then y_i minus μ I think this see y_i minus μ square $\sum y_i$ this $f y_i$ this.

So, in place of Y you are using y_i because i is ranging from one to particular value n here also one to particular value n you are writing y_i and function that is PDF of y_i and here also this then this actually if the observations are in identically independent distribution then this will be y_i only a this will be $f y_i$. So, that mean $y_i f y_i$ sum of 1 equal to n if all y_i in identically distributed when all PDF values are same any how, but this is the general formula.

(Refer Slide Time: 23:04)



So, essentially what do you know now you know random variable you know random variable and then there will be probability distribution.

So, you know that continuous probability distribution then you know the discrete probability distribution and you also know that some properties of probability properties of probabilities properties of probability and we have come you have seen that mean and variance these are nothing, but expected values expected value the expected operator you are using depending on if it is mean expectable of Y; if it is very expected value of Y minus mu square this much is known related to random variables.

(Refer Slide Time: 23:47)

Some important concepts of expectation

1. $E(c) = c$
2. $E(y) = \mu$
3. $E(cy) = cE(y) = c\mu$
4. $V(c) = 0$
5. $V(y) = \sigma^2$
6. $V(cy) = c^2V(y) = c^2\sigma^2$
7. $E(y_1 + y_2) = E(y_1) + E(y_2) = \mu_1 + \mu_2$
8. $V(y_1 + y_2) = V(y_1) + V(y_2) + 2 \text{Cov}(y_1, y_2)$

$\text{Cov}(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)]$

9. $V(y_1 - y_2) = V(y_1) + V(y_2) - 2 \text{Cov}(y_1, y_2)$

If y_1 and y_2 are independent, we have

10. $V(y_1 \pm y_2) = V(y_1) + V(y_2) = \sigma_1^2 + \sigma_2^2$

and

11. $E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \cdot \mu_2$

However, note that, in general

12. $E\left(\frac{y_1}{y_2}\right) \neq \frac{E(y_1)}{E(y_2)}$

regardless of whether or not y_1 and y_2 are independent.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, as that expectation is a very very important concept here particularly in u E and later on we will be using different kind of expected expectation. So, I want to give you a elaborate discussion on some of the important concepts where which you will be using in this particular subject suppose c is a constant then what is the expected value of c, it will be c only what is the expected value of Y. if Y is a random variable. it is mu. Now then what will be the expected value of c times Y, it will be a c become constant it will come out it is nothing, but c times expected value of Y which is c mu if c is a constant what is the variability of c, it is 0 what is the variability of Y we have seen that sigma square that is the variance; variance of Y is sigma square variance of c is 0 variance of Y sigma square.

In that case, what is the variance of c; y c times y interestingly, it is not c variance of y, it will be c square variance of y c square sigma square now. Now consider 2 random variable y 1 and y 2 suppose you are summing them up and then you want to know the

expected value of these 2 suppose expected value of y_1 plus y_2 y_1 is one random variable y_2 another random variable, then this will be nothing, but y_1 plus Y_2 . So, if y_1 expected value is μ_1 and y_2 is expected value is μ_2 , then this is the formula. Now if I say; then what is the variance of y_1 plus y_2 . Interestingly it will be variance of y_1 plus variance of y_2 plus 2 times covariance of y_1 y_2 this because we are saying y_1 y_2 two random variable, but we have not said that whether they are dependent or where they are independent or not if they are orthogonal to each other means they are independent means y_1 no way dependent on y_2 or vice versa then the covariance will become 0 for independent variable case this will be variance of y_1 plus y_2 will be the variance of individual random variable.

If y_1 and y_2 are independent. So, other way it will be like this suppose what will happen instead of plus if I write minus here. So, it will not be minus it will always be plus this will be always be plus irrespective of whether it will writing plus minus here, but this will become minus. So, if 2 variable random variables are independent then what will happen variance of y_1 plus minus this is always variance of y_1 plus variance of y_2 irrespective of plus minus sign here variance is additive.

Only if they are dependent to each other there is a covariance; covariance means one value occurrence of y_1 value depends on y_2 value also and vice versa then they say there is a called normally you are going toward if I read more you read more if I read less you read less or if I read more you read less. This kind of relation is there, then you are varying, we too are varying either positively or negatively we are varying that is what is covariance. So, then what will happen this will be σ_1^2 plus σ_2^2 if σ_1^2 is the variability of variance of y_1 and σ_2^2 is a variance of y_2 .

(Refer Slide Time: 28:14)

$E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \mu_2$

$E\left(\frac{y_1}{y_2}\right) \neq \frac{E(y_1)}{E(y_2)}$

Normal distribution
✓ t, z, n, F

$V(y_1 + y_2 + \dots + y_n)$
 $= V(y_1) + V(y_2) + \dots + V(y_n)$
 $= \sigma^2 + \sigma^2 + \dots + \sigma^2$
 $= n\sigma^2$

So, let us see some more things suppose y_1 and y_2 independent then what is the expected value of y_1 times y_2 this will be expected value of y_1 times expected value of y_2 equal to μ_1 times μ_2 , but what will happen if I want to know the ratio division it will never become irrespective of whether y_1 y_2 is independent or not. So, now the last one, you know probability distribution there will be discrete distribution there will be continuous distribution for du we will be interested mostly in normal distribution will be interested to this.

And then there will be certain sampling distribution like t; z distribution t distribution chi square distribution f distribution this understand the sampling distribution will be used, but in general when you talk about probability distribution whether discrete or continuous there are many probability distributions depending on the nature of the variable.

(Refer Slide Time: 29:34)

Important distributions

Continuous probability distributions

- Normal distribution
- Lognormal distribution
- Gamma distribution
- Weibull distribution
- Beta distribution

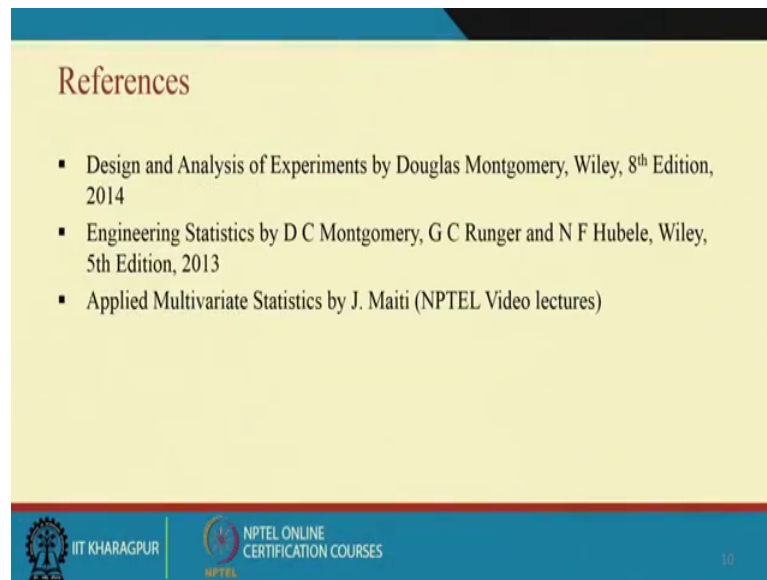
Discrete probability distribution

- Bernoulli distribution
- Binomial distribution
- Poisson distribution
- Exponential distribution

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us see some of the continuous probability distribution normal distribution log normal distribution gamma distribution Weibull distribution beta distribution there will be there will be exponential distribution; Erlang distribution many things many more now under discrete Bernoulli distribution binomial distribution per Poisson distribution exponential distribution will be under continuous distribution. So, I am changing here, fine. So, sometimes these mistakes are good because if you go through this and you find out that you are able to identify what mistake has taken place that gives you more clarity and you will you will be able to keep in mind that exponential distribution will be will be the continuous distribution it is not the discrete distribution.

(Refer Slide Time: 30:38)



So, again I just conclude that that in this lecture it is a very simple one we talk about random variable we talk about warranty distribution we talk about discrete random variables we talked about continuous random variable we talk about expectation from mean point of view variance point of view.

We talk about that different kinds of expected values when the con a constant is multiplied then what will be the expected value in terms of mean and in terms of variance and when there are variables they are related what will happen to their variance of some of them and in that manner.

So, suppose if I say that what is the variance of y_1 plus y_2 dot dot y_n and if I say they are independent then you will definitely say that is variance of y_1 plus variance of y_2 dot dot dot variance y_n . Now in addition, if I say that the variance of all of them are same σ^2 then you say σ^2 plus σ^2 plus σ^2 then this will be n time σ^2 if I assume that all the all the all those y_1 to y_n , all those variable, they have same variance component.

So, in addition you know you seen that there are different kinds of distribution depending on the situation it will be used, but for du in design analysis of experiment we will be mostly relying on normal distribution and other sampling distribution in next class next I think next, but one next class we will discuss in detail the normal distribution very very important and next and the next to next class, we will discuss about the t distribution chi

square distribution f distribution. These are very important and they will be useful in hypothesis testing time and also when in estimation when we develop the confidence interval that time also we get to know what kind of distribution it is and accordingly you will (Refer Time: 32:41) hope you have understood you are free to email me and use the forum effectively.

Thank you very much.