

**Dealing with Material Data: Collection, Analysis and Interpretation**  
**Professor Hina A. Gokhale**  
**Department of Metallurgical Engineering and Material Science**  
**Indian Institute of Technology Bombay**  
**Lecture 32**  
**Special Random Variables II**

Hello, and welcome to the course on dealing with materials data. Presently we are going through sessions on introducing Special Random Variables. What happens is that, when we conduct experiments, the results or the outcome of the experiment sometimes you know by the nature of experiment that, it follows certain underlined distribution and, we are trying to define all those specialized distributions which may occur in many of our experiments and that would make us understand the results of the experiment, the data behavior in a more simplistic manner.

(Refer Slide Time: 01:13)

### Review

- Data from experiments result come from some underlying distribution.
- If data is of countable nature, underlying distributions are known as Discrete Distribution.
  - Discrete Uniform Distribution
  - Bernoulli Trials
  - Binomial Distribution
  - Geometric Distribution
  - Negative Binomial Distribution
  - Poisson Distribution
  - Hypergeometric Distribution

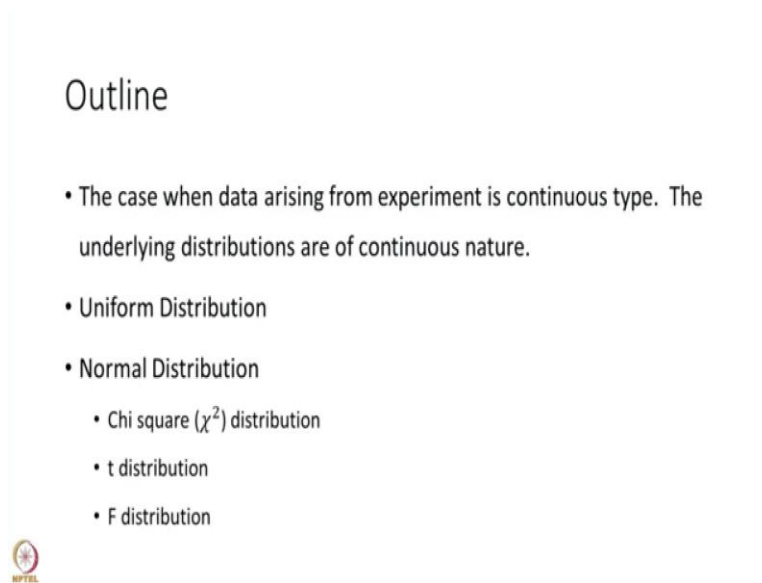


So, what we take in the previous session was that we found that if the data or the result is countable in nature then underlined distribution could be offered discrete nature and they are called discrete distribution. This way we introduced several discrete distributions in the previous session. The discrete uniform distribution, then we introduced Bernoulli trial where there are only two results success or failure. Then we introduced 3 derivatives or the 3 distributions derived from the Bernoulli trails that is, you conduct an independent Bernoulli trails and you count the number of successors that becomes a binomial distribution.

If you conduct, the Bernoulli trials until you counter the first success it is called the Geometric Distribution and, suppose you conduct  $x$  Bernoulli trials until you come across the  $n$ th success in the trials, then it will result into negative binomial distribution. Then we introduced a Poisson distribution, which says that when the number of trials are extremely large it tends to infinity. and, the probability of success or the probability of failure whatever you want to call it tends to 0 in such a way that the multiplication of the 2 that is the product of  $n$  and  $p$  remains constant and we call that constant value  $\lambda$  then it tends to follow Poisson distribution.


Then we also introduce Hypergeometric distribution where you have total  $N$  objects of which  $M$  objects are of a certain kind and you are drawing  $N$  sample of size  $N$  without replacement from it. And, then you are trying to find out there are exactly  $X$  items of the type  $M$  and this is what the distribution it arises is a Hypergeometric distribution. And, at the end we gave an example of 3D atom pro filled ion microscopy in which all this different kind of distributions occur naturally in order to estimate certain probabilities.

(Refer Slide Time: 03:47)



Outline

- The case when data arising from experiment is continuous type. The underlying distributions are of continuous nature.
- Uniform Distribution
- Normal Distribution
  - Chi square ( $\chi^2$ ) distribution
  - t distribution
  - F distribution

 IITM

So, now in this session we would like to have, we would like to have a look at all those distributions, which may arise when the experiment gives out data, which is of, continuous nature. Therefore, the underlying distribution are also of continuous type. We will introduce primarily uniform distribution, normal distribution and in further slides, we will introduce some other distributions such as Chi square t distributions and F distribution, which are derived from the

normal distribution. But, first in this particular one we are going to study normal distribution in details.

(Refer Slide Time: 4:31)

### Uniform Distribution

- A random variable  $X$  is said to follow Uniform distribution,  $X \sim \text{Uni}(a, b)$ , if the pdf of  $X$  is given by

$$f_X(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b \quad : X \sim \text{Uni}(a, b)$$
$$= 0 \text{ otherwise}$$

$$\mu = E(X) = \frac{a+b}{2} \text{ and } \text{Var}(X) = \frac{(b-a)^2}{12} = \sigma^2$$

- Random Number generator for Monte Carlo simulations



So if there is a random variable which takes on value between 2 fixed numbers  $a$  and  $b$ . Then, if the probability that  $x$  takes on value any value between  $a$  and  $b$  is given by

$X \sim \text{Uni}(a, b)$ , if the pdf of  $X$  is given by

$$f_X(x) = \frac{1}{b-a} \text{ for } a \leq x \leq b$$
$$= 0 \text{ otherwise}$$

it is called a uniformly distributed random variable between  $a$  and  $b$ . I have not written the notation, so let us write down the notation in such case it is said that  $x$  is distributed uniformly between interval  $a$  and  $b$ .

$$E(X) = \frac{a+b}{2} \text{ and } \text{Var}(X) = \frac{(b-a)^2}{12}$$

As I have said in the previous sessions these derivations, I will leave it to all of you to try it out yourself. The random number, this particular distributions applications comes in a random number generation, let us see how it happens.

(Refer Slide Time: 06:09)

### Uniform Distribution and Random Number Generation

- $X$  is a random Variable, and let  $F(x)$  denote its CDF  $F(x) = P(X \leq x)$
- Then,  $Y = F(X)$  is also a random variable and is distributed as Uni(0,1)
- Thus  $X = F^{-1}(Y)$
- If a random number  $Y$  is generated from Uni(0,1) distribution then any random variable  $X$  can be calculated once  $F$  and  $F^{-1}$  is known.  $X = F^{-1}(Y)$   
 $\sim F$
- There are pseudo random number generators available using mathematical algorithms.

So if you consider the case that  $x$  is a random variable and I am not saying it is uniformly distributed random variable by the way  $x$  is any random variable and  $F(x)$  denotes its cumulative distribution function. If you want to recall please recall that  $f$  of  $x$  is nothing but probability that random variable  $x$  takes on value less than or equal to small  $x$ . In that case, if we look at this  $f$  of  $x$  the CDF itself has a function of random variable  $x$  then  $Y$  is equal to  $f$  of  $X$ , this is also a random variable.

And, this random variable is distributed as uniform 0,1 this matter can be proved we are not going to cover the proof here, but it can be proved that any CDF of a random variable  $X$  itself can be considered as a random variable. And, in that case it is distributed as a uniformly between 0 and 1 and therefore we can have  $x$  is equal to  $X = F^{-1}(Y)$

Now if a random number generated from 0 and 1, **Uni (0,1)** distribution. Then the random variable  $X$  can be calculated, by finding an  $f$  inverse of the uniformly distributed random variable  $Y$ .

So what it is saying here is let me clarify, it I think I have not clarified it properly. What we are trying to say is that, if a random number  $y$  is generated from the uniform distribution which is here then  $y$  is distributed uniform and, then you take  $X = F^{-1}(Y)$  then you will have, this  $F$  is the CDF of  $x$  The random number, this particular distributions applications comes in a random number generation, let us see how it happens and then you will have a random number with a distribution of CDF  $F$ . And this  $F$  could be any CDF that you are looking for.

How do you generate this Y? With having a uniform random uniform distribution between 0 and 1. When there are number of pseudo random number generators, which generate such values of Y, and therefore, the random number any random number with a distribution function CDF as F can be generated from this. So this is the application of the uniform distribution which is playing a very major role in any kind of a Monte Carlo simulations.

The next distribution we wish to introduce is the most commonly used distribution which is called Gaussian Distribution or normal distribution. This is actually a distribution which was realized first time by Gauss, sometime in 18 centuries if I am not mistaken I might be mistaken on that front. But, Gauss realized that this is interesting function and it call it an error function it arise in this manner.

When you conduct any experiment in exactly identical conditions the results are not always identical there is an error in it, and it found that this error itself is following a certain pattern and that pattern he called as a distribution which came to be known as a Gaussian distribution. It also arises as an error function and we will see the distribution, the relationship of the normal distribution with the error function in the next few slides.


(Refer Slide Time: 10:58)

### Normal Distribution / Gaussian Distribution

- Random variable X is said to follow Normal distribution, if its pdf takes following form
 
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \text{ for } -\infty < x < \infty$$

$$E(X) = \mu \text{ and } \text{Var}(X) = \sigma^2 \quad : X \sim N(\mu, \sigma^2)$$
- Random Variable  $Z = \frac{X-\mu}{\sigma}$  has pdf
 
$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} \quad Z \sim N(0, 1)$$

is called Standard Normal distribution and Z is called standard normal variate with mean 0 and standard deviation 1



So, if a random variable X is following a normal distribution then its probability density function takes on a form of this nature

$$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \text{ for } -\infty < x < \infty$$

The parameters that are used here are mu and sigma and these parameters are the mean values and the variance of this distribution. So, if x follows a normal distribution with mean mu and variance sigma square then it takes on a form of this nature and again I have not written here so let us write down what is notation for that. So in such case you say that x follows a normal distribution with mean mu and variance sigma square.

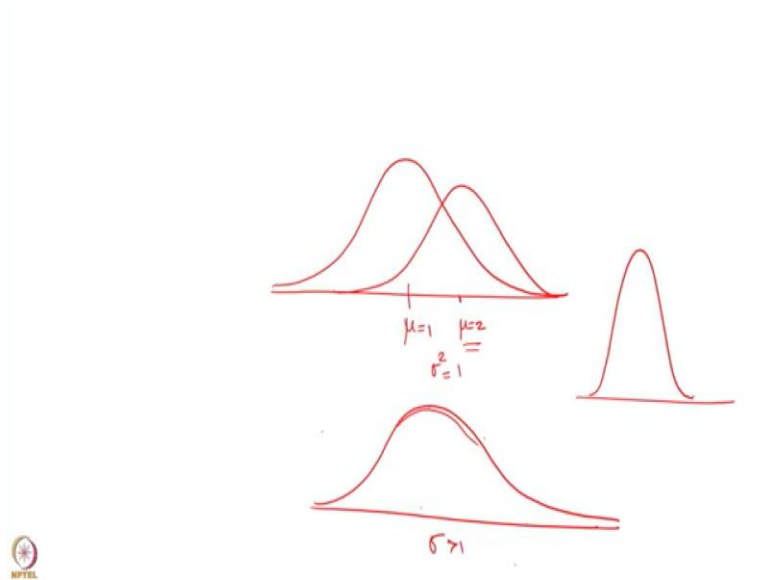
- If you take a random variable z which is  $Z = \frac{x-\mu}{\sigma}$  has pdf

$$f_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

you see that there no parameters and this is also a distributed as a normal distribution with mean. It means that z is distributed as normal with mean 0 and variance 1. This is also called Standard normal variate; it is called as Standard normal variate with mean 0.

So, let us review there is this x we say that follows a normal random distribution, if it PDF takes it form which has a two parameters mu and sigma, mu is it mean value and sigma square is it variance value. If you a take transformed variable z which is x minus mu divided by sigma which we also called Normalization. Then it has a PDF of this nature it is actually a standard normal variate with a mean value 0 and a variance of 1.

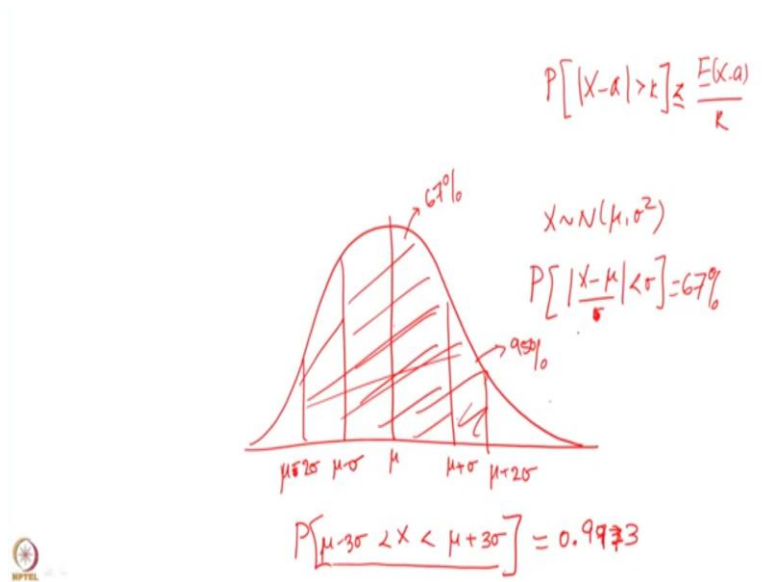
(Refer Slide Time: 13:52)



Now this distribution has a beautiful bell shape, this distribution takes on a very beautiful bell shape means stands in the middle it has a nice bell shape curve. If your mean is 1 it sits here, suppose you have a mean of 2, and the standard deviation is same, say standard deviation of only 1 then it will also look like this.

So, these are the different kind of normal distributions with mean, so with mean the normal distribution moves from right to left it depending on where the mu is. If with respect to sigma you can imagine, if sigma is larger, the distribution spread is larger. This should be sharper this shows that sigma is probably greater than 1 and if the sigma is smaller the distribution becomes even more smaller.

(Refer Slide Time: 15:10)



This distribution also has another beauty in it. If you take this distribution, let us draw it again, let us take a standard normal distribution, or you take any just normal, this is not a very good nicely drawn curve but let it be so, this is mean mu.

If you take  $(\mu + \sigma)$  and  $(\mu - \sigma)$  then this covers about 67 percent of your data, please recall. We have worked with Marko's inequality and then we worked with chebyshev inequality and in both the inequalities our idea was to estimate how much of data lies in a given interval from the mean value of the data.

So here, if it is a normal distribution if you recall that time we mentioned that if you have a normal distribution it gets defined even more clearly. So, Marko's inequality if you recall it says that

$$P[|X-a| > k] < \frac{E(x-a)}{k}$$

So this actually you can have a is equal to 0 also, so this basically says that it gives you an upper bound of number of data points that can lie in this region. Here you make it this says that if it is a probability it is a normally distributed as  $X \sim N(\mu, \sigma^2)$  then it says that

$$P\left[\left|\frac{x - \mu}{\sigma}\right| < 1\right] = 67\%$$

If you take  $(\mu + 2\sigma)$  limit it takes 95 percent of the data and if you take the values of all x lined between  $(\mu + 3\sigma)$  and  $(\mu - 3\sigma)$  then this probability is actually 0.9973. It means that, 0.9973, it means



that more than 99 percent of the data lies between minus 3 and plus 3 limits of the, plus 3 minus 3 of the mean value of standard of a normal random variable with mean  $\mu$  and variance  $\sigma^2$ .

So, we come back to the next issue or, let us clarify what I wanted to say here in one go that, if  $x$  follows a normal distributing with mean  $\mu$  and variance  $\sigma^2$ . Then we have a clear idea as to how many data points lie what percentage of data points lie between  $\mu - \sigma$  and  $\mu + \sigma$  limits which comes to 67 percent of the data would lie if the  $X$  lies between  $X - \sigma$  and  $X + \sigma$ , if it lies between it says  $\mu - \sigma$  and  $\mu + \sigma$ .

If it lies between mean minus 2 standard deviation and mean plus 2 standard deviations it covers 95 percent of your data. And, almost all data 99.73 percent of the data lies between the 3 sigma limits of the mean value  $\mu$ .

(Refer Slide Time: 20:20)


### Error Function and Normal Distribution

- Error Function (Gauss error function) is defined as special function.
- It occurs in partial differential equations describing diffusion, defined as

$$\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt$$

$$= \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

- It can be seen that  $\operatorname{erf}(x)$  describes probability of a normal random variable  $Y$  in the range  $[-x, x]$ , where  $Y \sim N(0, 1/2)$



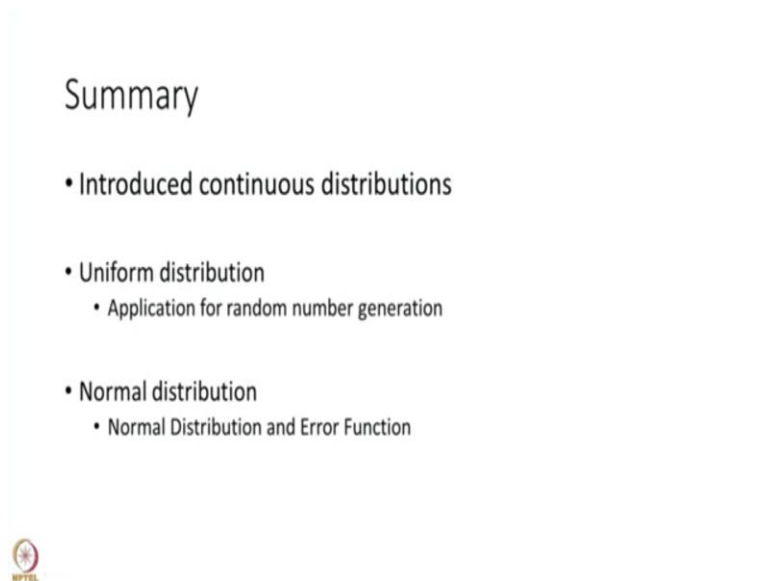
Now we come to the next stage as I said, Gauss error function is generally defined in this format. This occurs in a partial differential equation describing diffusion and as I said it is an error function,

so error function was defined in this manner.  $\operatorname{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-t^2} dt$

$$= \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

And, you can see that this error function has direct relation with a normal distribution, it is if the error function actually describes the probability of normal random variable  $Y$  when it lies between the range of minus  $x$  and  $x$  when  $Y$  is a normal random variable with mean 0 and a variance of  $1/2$ .

(Refer Slide Time: 21:05)



So, now let us quickly summarize, we have introduced first continuous distributions arising from continuous data which comes out of experiments. We introduced two distributions here, one is a uniform distribution with an application for random number generation. And, we introduced a normal distribution and we said that normal distribution has another quality that you can define you can refine the Markov's inequality or chebyshev inequality in a much clear way to say that.

The 1 sigma limit from the  $\mu$  that is  $\mu - 1 \text{ sigma}$  to  $\mu + 1 \text{ sigma}$  contains about 67 percent of the data, plus 2 sigma limit  $\mu - 2 \text{ sigma}$  to  $\mu + 2 \text{ sigma}$  value cover the 95 percent of the data. While 99.73 it means that almost all data is covered with minus 3 and plus 3 limits from the mean value, minus 3 sigma to plus 3 sigma limits of the mean value, please recall. When you talk about the 6 sigma limits you are talking about 0.69999 times data lying within the 6 sigma limits of the data.

If we talk about it in future maybe, you can just understand that is why I am giving you. And, then we established the relationship between the normal distribution and error function. We say it that error function is defined in physics another theory it is defined as if  $Y$  is a normal random variable

with a mean 0 and variance half, then the error function is a probability that the Y lies between the 2 quantities minus x and x.

Thank you.