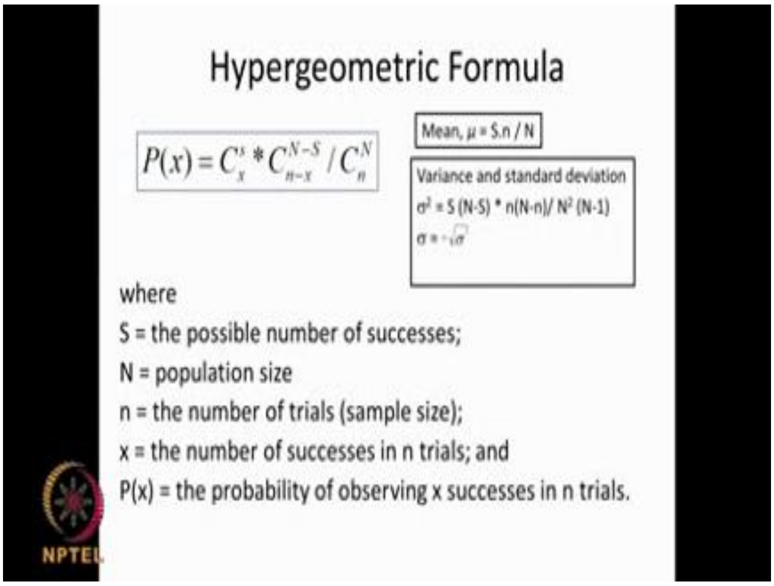
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 31
Hypergeometric/Log normal distributions

Welcome to the course on Biostatistics and Design of Experiments. We will talk about Hypergeometric and log normal distributions. Yesterday I introduced Hypergeometric, let me continue on that. So, Hypergeometric is a distribution, which looks at situations where the population is not very large.

(Refer Slide Time: 00:24)



That means the sample size, are almost in the ball park figure of population, unlike a normal distribution, where the population is very large; almost infinite. So, here sample also is reasonably large comparable to the population. And the probability distribution is given by this

$$P(x) = C_x^s * C_{n-x}^{N-S} / C_n^N$$

$$P(x) = C_x^s * C_{n-x}^{N-S} / C_n^N$$

, where N is a population, n is the sample size, x is the number of successes in the trial, S is the maximum possible number of successes, and this is called the probability of observing x successes. So, the mean is given by

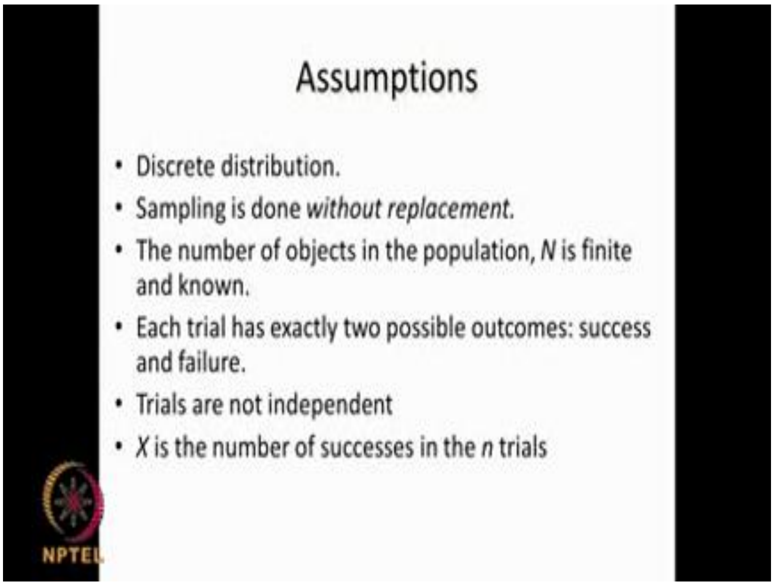
$$\boxed{\text{Mean, } \mu = S.n / N}$$

n is your sample size, N is a population size, S is the possible number of successes, that is maximum that is possible. And the variance is given this formula

$$\sigma^2 = S\,(N-S) * n(N-n)/\, N^2\,(N-1)$$

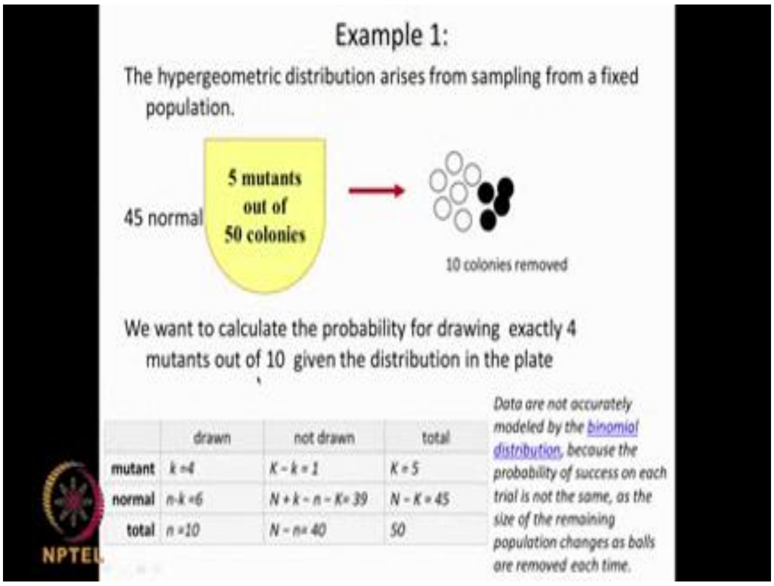
So, this is the variance, and this is the mean of this particular distribution, which is called the hypergeometric distribution. So, as I said here, the sample size is reasonable very close and comparable to the population, unlike the normal.

(Refer Slide Time: 02:00)



So, it is a discrete distribution, because we are taking sample is done without replacement. So, you take it out and there is no replacement. n is a finite and known, each trial has exactly two possible outcomes; success or failure. Suppose I pick up balls from bag, which may have a black or white. So, success-failure, black-white, yes- no; the trials are not independent, X is the number of successes in the n trials.

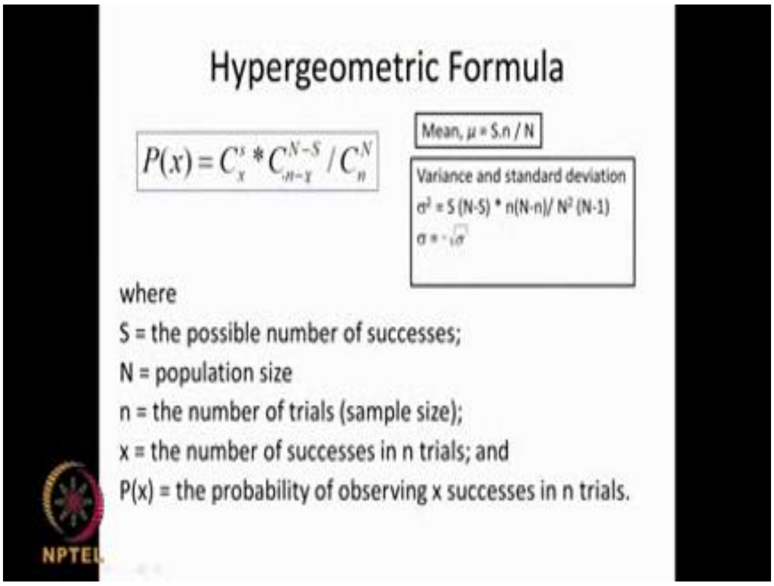
(Refer Slide Time: 02:32)



Let us look at a problem. So, it arises, because of sampling from a fixed population. So, imagine there are 50 colonies, which has 5 mutants, so 45 of them are normal. Ok now I have removed 10 colonies from this lot, and imagine there are 4, exactly 4 mutants out of this 10. So, 6 of them are normal, so we have drawn 10, 4 of them are mutant, normal is 6. Total has 5 mutants and 45 normal, and the total is 50. So, we cannot exactly model by binomial, because the probability of success on each trail is not the same, because we are removing it, and we are not replacing them. Unlike a coin, know we toss heads or tails, then again we toss the coin, so probability is again heads or tails is 0.5 and so on. So, whereas in this case we have removed it, and we are not replacing it. So, the probability keeps changing, so in such situation you think about hypergeometric distribution. Do you understand the difference?

Now, because when you remove for example, these balls or these mutants, what is remaining, changes keep changing. So, the probability of that keeps changing. So, the probability of drawing exactly 4 black balls or 4 mutants in this particular example, can be calculated from that formula. What is this formula?

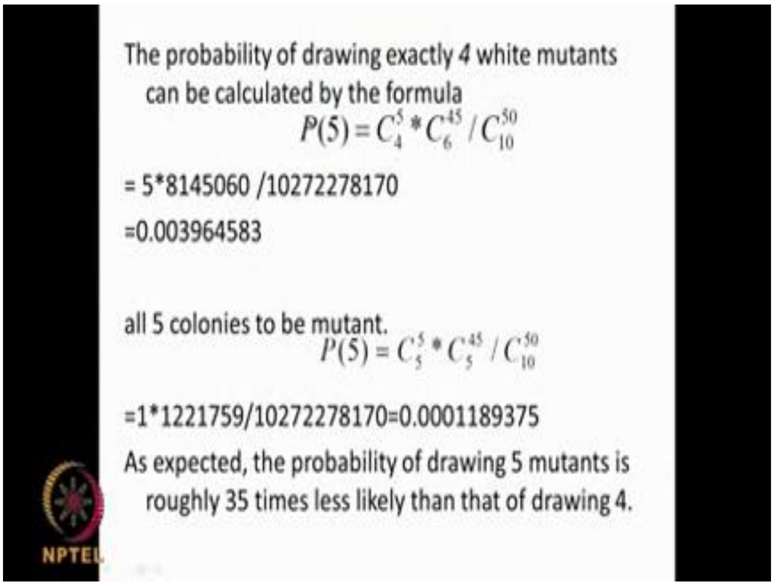
(Refer Slide Time: 04:10)



$P(x) = C_x^s * C_{n-x}^{N-S} / C_n^N$

. So, we just what we have got here in this particular thing is, we have got 4 successes in a sample of 10. The maximum successes possible, is 5 in N of 50. So, N is 50, n is 10, S is 5, x is 4. So, we can substitute in our equation.

(Refer Slide Time: 04:42)

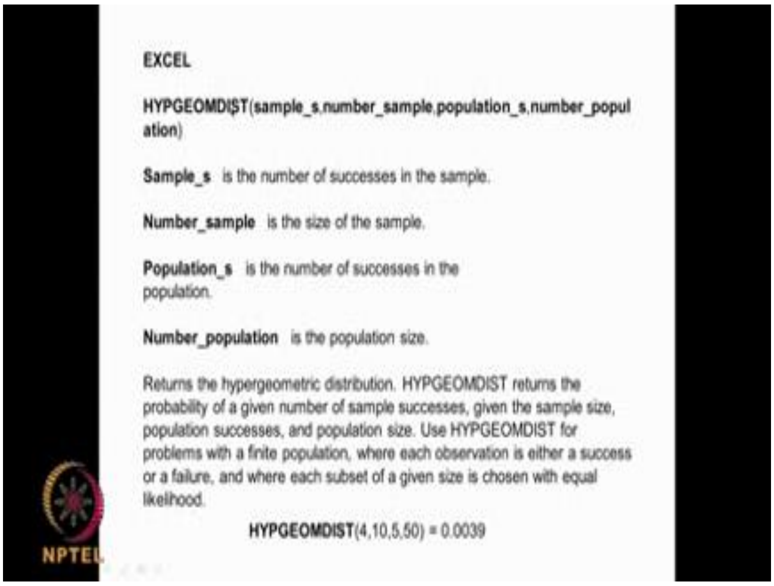


So, we have

$$P(5) = C_4^5 * C_6^{45} / C_{10}^{50}$$

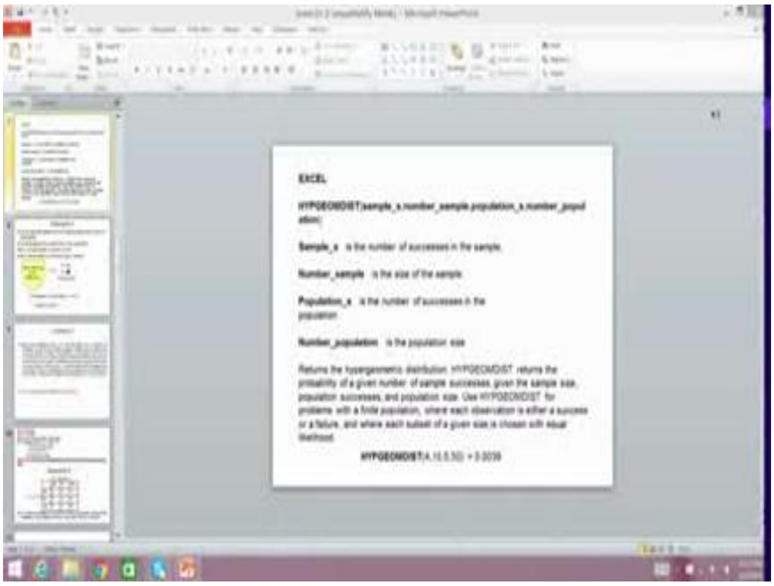
C 5 4, C 45 6, 4 is the mutants you took out from the 10, the remaining is 6 here. So, what is C 5 4? If you remember, the factorial; this is factorial 5 ÷ 4 factorial ÷ 5 - 4 factorial, this one is 45 factorial ÷ 6 factorial ÷39 factorial. This is 50 factorial ÷ 10 factorial ÷ 40 factorial. If you do this C 5 4 will come to be 5, and this particular 45 factorial ÷ 6 factorial ÷39 factorial will come out to be this, and this will come to be this. So, you end up having a probability of 0.0039. What is the probability of getting all 5 as say black or mutant? So, the remaining 5 in that sample of 10 will be normal. So, what we do, we put here 5, this will 6 will become 5, other things will remain same. So, when you do that; of course, C 5 5 is 1, if you remember from your olden days. So, we get out a probability very small 0.00011. If this is as expected, if you want to take out all the 5 mutants, when you take a sample of 10, obviously is much larger than taking out 4 mutants in a sample of 10. Now, we can also do it by excel. There is a function hypergeometric distribution.

(Refer Slide Time: 06:19)



In that we have sample, successes, number of samples. So, this will be like x, this will be like n, this will be like your S, and this will be like your N. Do you understand? This is like your x, this will be like n, this will be like your S, this will be like your N. Let us look at it for this particular problem situation, using this particular excel piece. Let me first get out my excel out of this. So, we do excel.

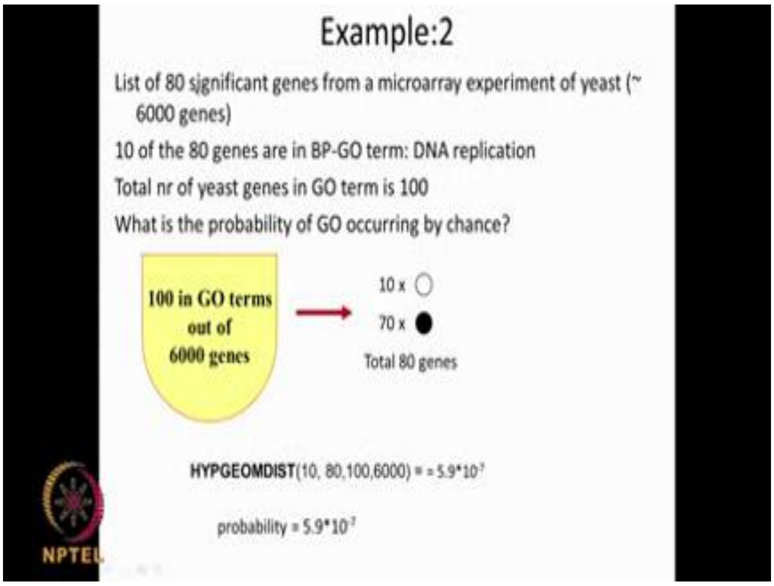
(Refer Slide Time: 07:08)



So, we say hypergeometric distribution I (Refer Time: 07:08)

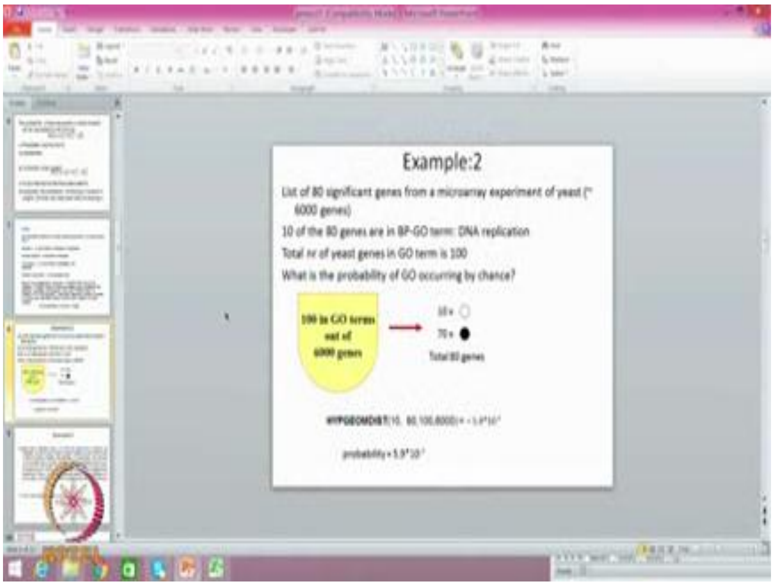
d I s. So, this is like a we are getting 4 mutants or 4 black ball in a sample of 10. The maximum that is possible is 5, the N is 50. So, we substitute and then we should get 0.0039. So, it is got probability of 0.0039. So, if all 5 needs to be black or mutant, same thing, we can do MDIST 5, out of 10, 5 out of 50, 0.30119. So, we can use excel also to do the same thing. It is quite simple with the excel, as you can see here, we give the number of successes. In a sample, there is a maximum number of successes possible and this is the population number. We cannot use binomial as I said we are taking out and we are not replacing; so obviously, the probability may keep changing as you keep drawing mutants. Let us look at another problem.

(Refer Slide Time: 08:44)



So, there are 80 significant genes from a microarray experiment of yeast, and it is got a 6,000 genes. We have taken 80 significant genes. Now 10 of these 80 are in BP-GO term, that is DNA replication, but total number of yeast genes in GO term is 100. So, what is the probability of this GO occurrence by this chance? So, 6000 genes; maximum possible is 100, when we took out 80 in that list, we find 10. So, we will call it 10 successes out of 80, maximum possible is 100. So, N is 6000. So here n is 80, x is 10, n is 80, x is 10, and the maximum that is; S is 100. So, we can get the probability.

(Refer Slide Time: 09:46)



So, hypergeometric distribution quite straight forward, this is HYPGEOMDIST. Then what we got is 10 in a list of, when we took out 80, but maximum possible is 100 out of 6000. So,

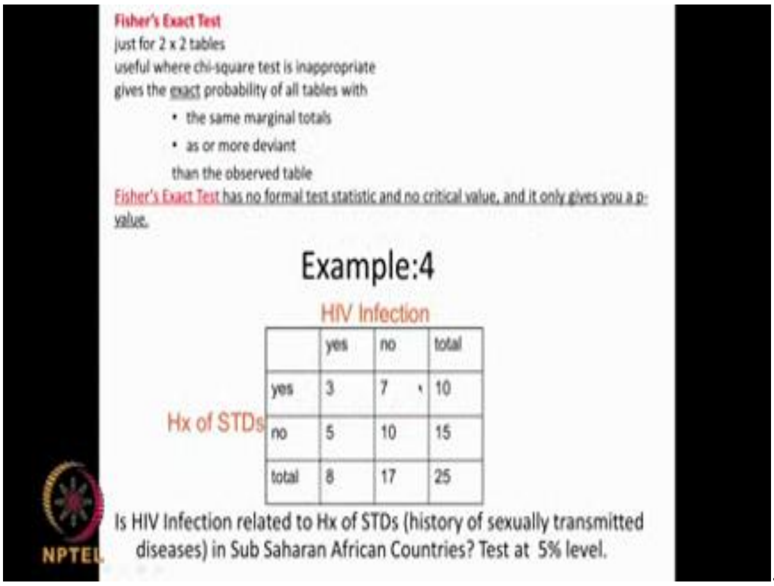
probability = $5.9*10^{-7}$

understand. It is quite straight forward using the excel, but this distribution is useful, where we cannot use binomial, but this also has like yes, no, live, dead and that sort of situation we can use actually.

(Refer Slide Time: 10:29)

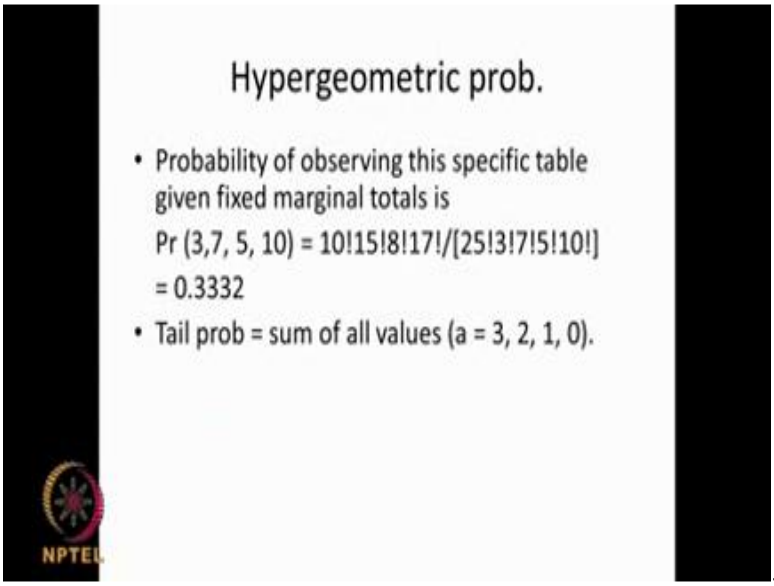
So, we can use in problems, where we actually do not know the situations, unlike we do not. For example, let us take N laboratory mice, N of them, n are males. So, N - n are females, now we do a certain mutation after radiation. So now there are m new mutant mice. You don't know whether in this m, how many males are mutants and how many females are mutant? That is the problem. Then, how do we do this actually? So, there is something called Fisher's exact test, which cannot break it down, but it can give a total probability of this actually.

(Refer Slide Time: 11:20)



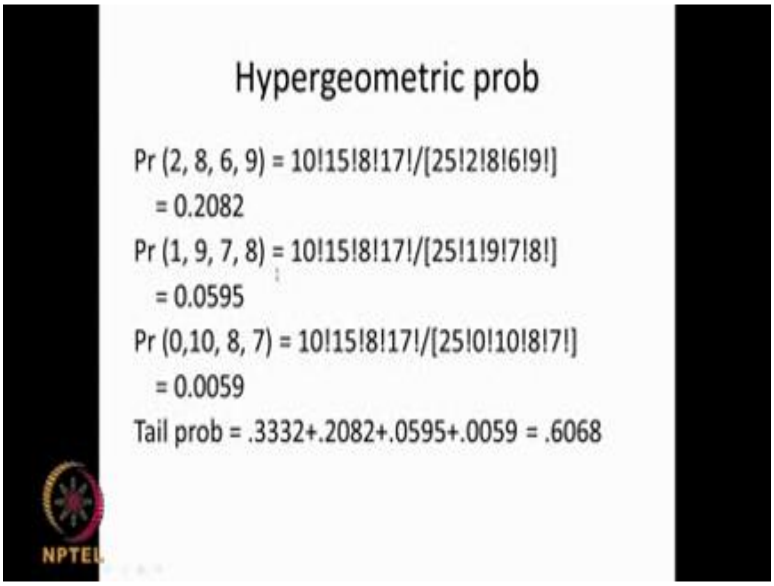
Let us look at one problem. So, there is a 2 by 2 table. Of course, 2 by 2 table contingency we use to use chi square, but here we are going to use Fisher, we will. This Fisher's exact test, it does not do a test statistics and there is no critical value, but it gives you only a p value for observing this type of numbers. So, for different types of tables, it just gives you what is the probability? So, let us look at an example. So, there is a HIV infection, and there is a sexually transmitted in this; that is, the history of sexual transmission in Sub Saharan African Country. So, yes for a history of sexually transmitted disease, no and HIV infection yes and no. So, we find that there are 3 cases, where there is a history of STD and also HIV infection. There is a history of STD but no HIV infection is this much. There is no history of STD, but there is HIV infection. There is no history of STD and there is no HIV infection. So, you have this data. You can find out what is the probability of getting this type of table. So, this Fisher's exact test gives you that sort of numbers. It does not have a test statistics, it does not give you a critical value and comparison and p value and so on actually. So how do we go about doing this? So, we look at it as a hypergeometric distribution. This is a hypergeometric distribution. So, we look at different factorials of creating this type of numbers. So, we can have here 3, if we have 2 then this will become 8, and you can have 1 then this will become 9, or if we have 0 this can become entirely 10 also. So, we will look at each one of them.

(Refer Slide Time: 13:35)



So, probability of 3, 7, 5, 10 is given by 10 ! 15 ! 8 ! 17 !. What is that? 10 ! 15 ! 8 ! 17 ! ÷ 25 !, 25 is the total 3 ! 7 ! 5 ! 10 !, that gives you a probability of 0.3332. So, the probability of seeing this type of contingency table with 3, 7, 5, 10 is this much. Similarly, we can do for, just like 3, we can do for 2. 2 means, as I said if it is 2, this will become, this will become 8 here, if it is 2, this will become 8 and this will become 6, because the total you are maintaining. If this becomes 2 S, S then obviously, this will become 6, and if it is 2 S, S then, obviously, this will become 8. So, again use this same approach.

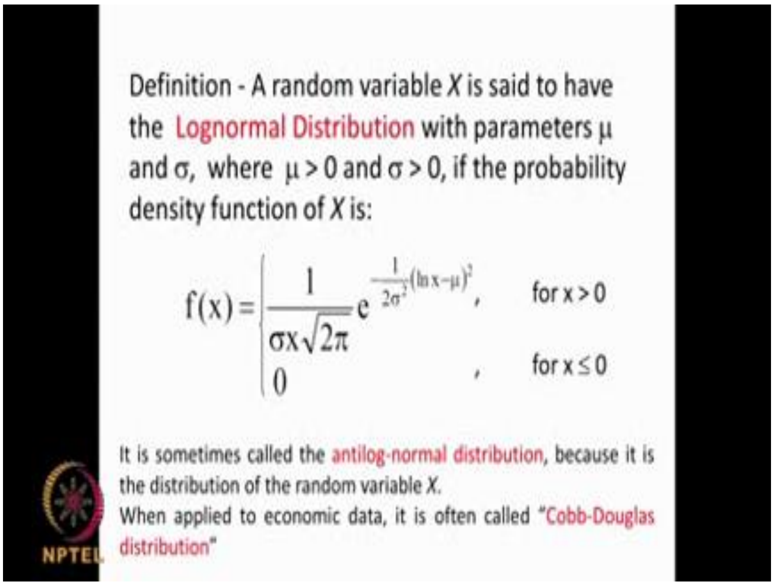
(Refer Slide Time: 14:44)



For a 2, 8, 6, 9 we do 10 ˙! 15 ˙! 8 ˙! 17 ˙!; that is the, these numbers ÷ 25 ˙! 2 ˙! 8 ˙! 6 ˙! 9 ˙!, that is the probability of 0.208. And for a 1, 9, 7, 8; that means, if you put 1, if I put 1 here, this will become 9, if I put 1 here, this will become 7 for a 1, 9, 7, 8. Then the probability is given by 10 ˙! 15, 8, 17. These are the row, sums and column, sums ÷ 25 ˙!, this is a grand total. Then 1, 9, 7, 8 that gives you a probability of 0.0595. So, if you want to see 0, 10, 8, 7; that is 0, 10, 8, 7 means, here 0. So, then this will become 10, this will become 8. So, for a 0, 10, 8, 7 again 10 ˙! 15 8 17 ÷ 25; 0, 10, 8, 7 that gives a probability of 0.0059.

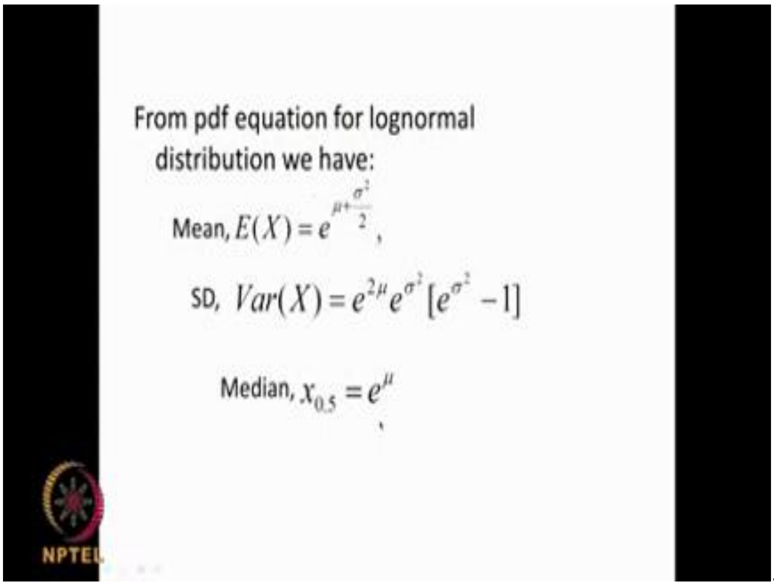
So, a probability of finding this sort of 3, 7, 5, 10 is given like this, probability of finding 2, 8, 6, 9 is given like this, probability of finding 1, 9, 7, 8 is given like this, probability of finding this, is given like this actually, do you understand. So, this Fisher’s exact test, is used in situations where there is no test statistics. We cannot use a chi square, but it gives a probability value here. So, there is no test statistics, there is no critical value, but it tells you what is the probability of observing this type of table. So, you see that the hypergeometric is useful in this type of situations where the n, N is known, unlike the normal distribution, and your n is comparable to N; that n is the sample size. Let us look at another distribution that is called Lognormal distribution. So, it is a logarithmic of normal distribution. So, it is not uniform. it is defined in this form, random variable x is set to have a Lognormal distribution, if it looks like this.

(Refer Slide Time: 16:53)



You can see it almost looks your normal curve, but you have lon x there. For X > zero, if it is zero for X < zero. So sometimes it is called antilog-normal distribution, because the distribution of the random variable X, it is also called "Cobb-Douglas distribution" if you are analyzing economic data.

(Refer Slide Time: 17:26)



The probability function, distribution function for the lognormal. So, the mean is given by

$E(X)=e^{\mu+\frac{\sigma^2}{2}}$,

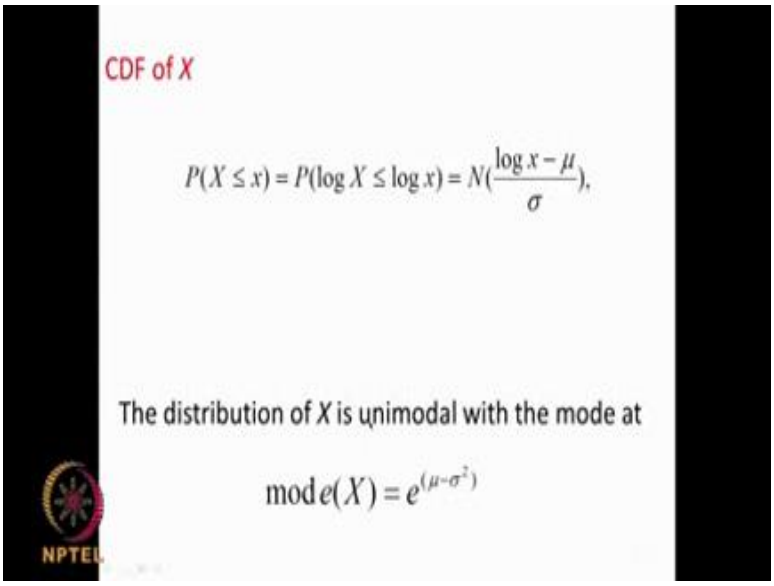
. µ is your mean, and σ is your, and then variance is given by

$Var(X)=e^{2\mu}e^{\sigma^2}[e^{\sigma^2}-1]$

and the median it is at the 50 % point is given by.

$x_{0.5}=e^{\mu}$

(Refer Slide Time: 17:56)



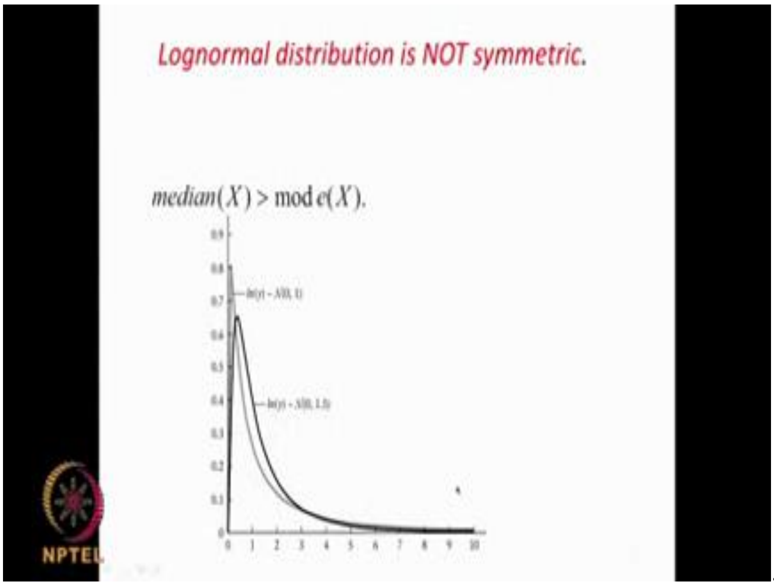
The cumulative distribution function, is given like this;

$N(\frac{\log x-\mu}{\sigma}),$

. Now this distribution is unimodal, and the mode is given by

$\text{mod}e(X)=e^{(\mu-\sigma^2)}$

(Refer Slide Time: 18:12)

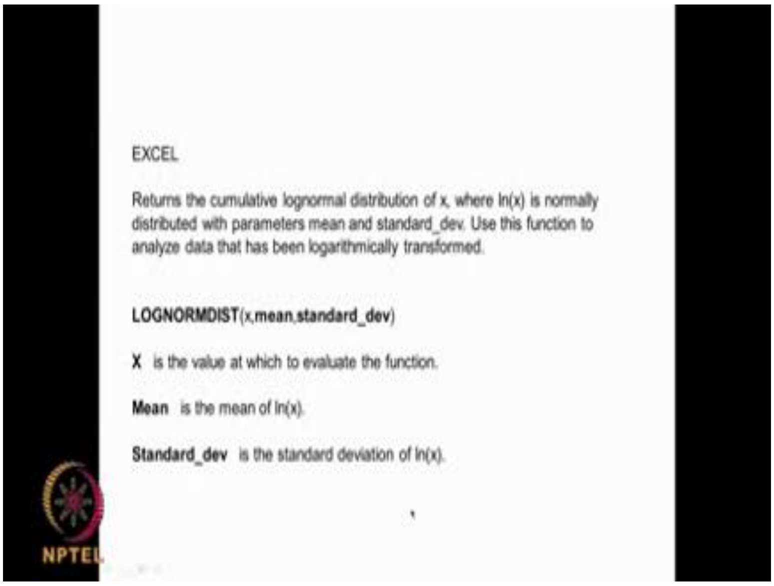


But it is not symmetric, as you can see here it is not symmetric. And this is the mode, mode is

$$\text{mode}(X) = e^{(\mu-\sigma^2)}$$

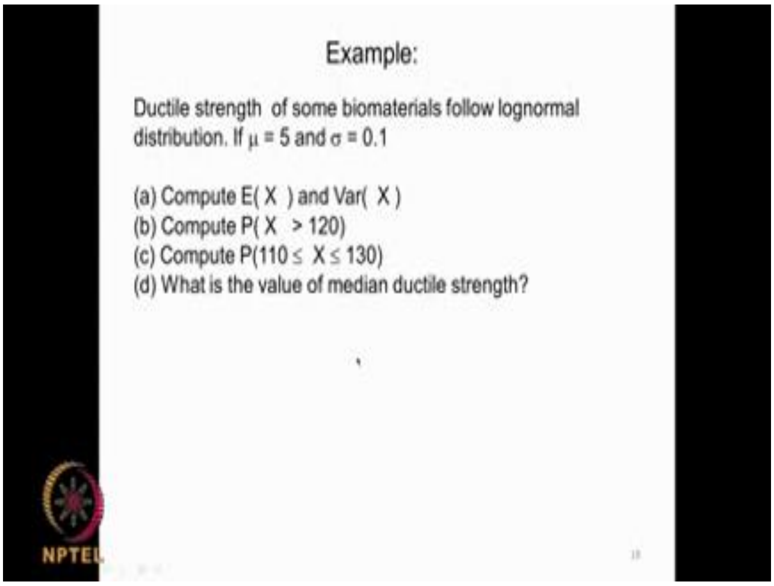
, and the median is e μ. So obviously, median is greater than mode. This is how the graphs will look like a normal distribution.

(Refer Slide Time: 18:34)



Excel also has of command LOGNORMDIST, where you give x is the value at which you want to evaluate the function, mean is the mean of logarithm x. This is the standard deviation of this lognormal distribution. Do you understand? So, we give at which where you want to find, where the mean is given, standard deviation is given.

(Refer Slide Time: 18:55)

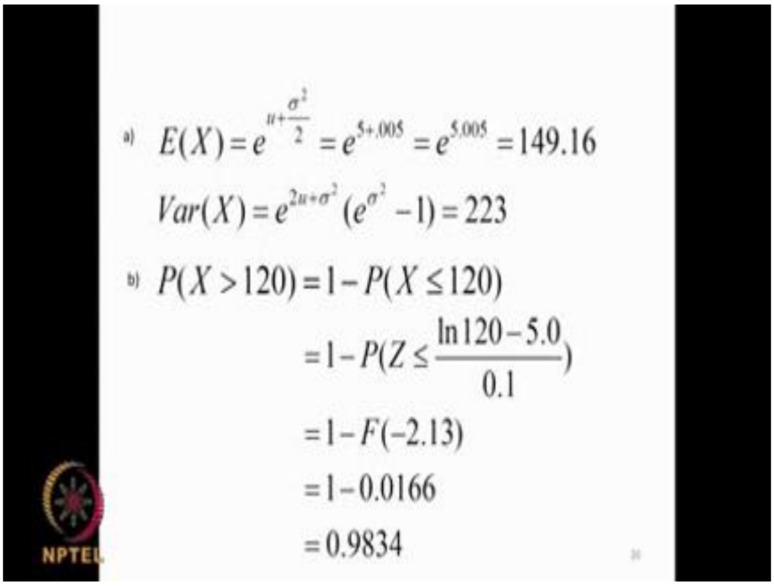


You can look at the problem, which can give you some idea about it. So, Ductile strength of some biomaterials follow lognormal distribution; μ is 5 and σ is 0.5. Compute the e and the variance? What is e? mean,

$$e^{u+\frac{\sigma^2}{2}}$$

, right. So how do calculate?

(Refer Slide Time: 19:19)



Very simple

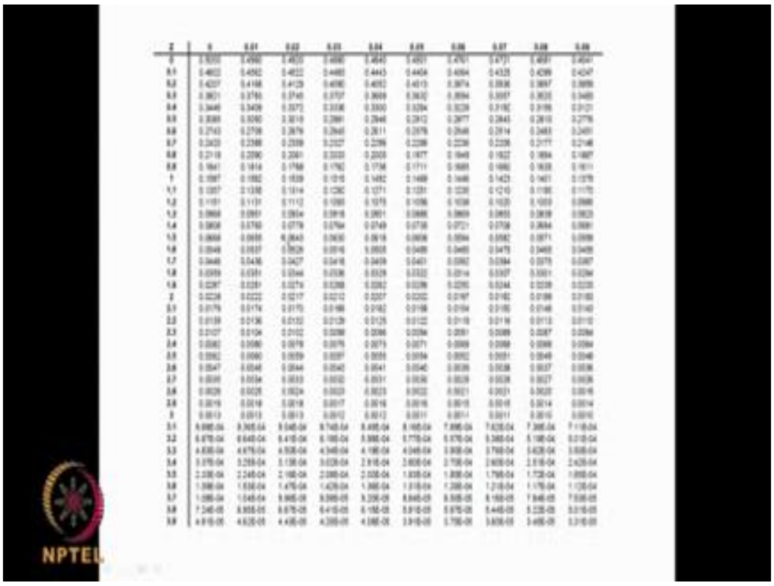
$e^{u+\frac{\sigma^2}{2}}$

. So, μ is given by 5, and σ is 0.1. So, square 8 ÷2. So, the mean is given by this variance, what is the formula for variance? Variance is given like this; right

$e^{2u+\sigma^2}(e^{\sigma^2}-1)$

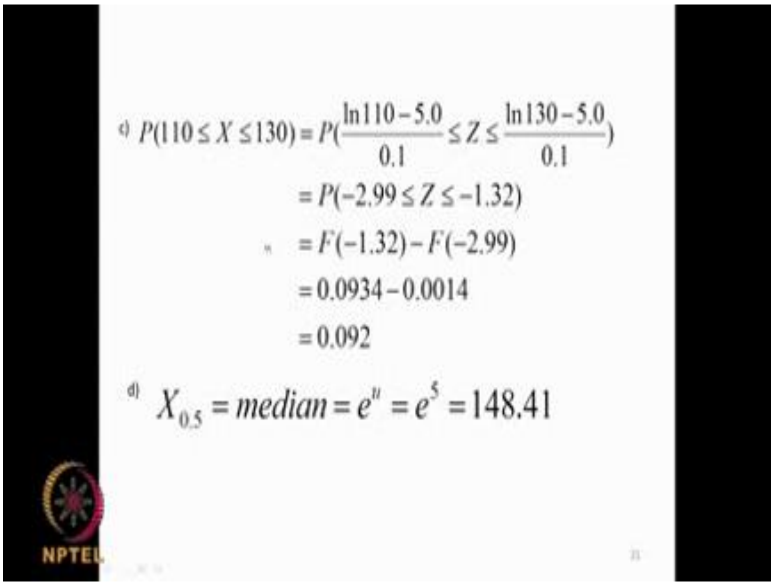
. So, we calculate that and we get a variance of 223. Now, the question is, third, compute the probability, that the data will lie between 110 and 130? So, 110 and 130. Before that sorry, before that this, compute that what is the probability for x greater than 120? So, what do we do. We calculate 1 - x < 120. You all remember that right. And then we calculate the z, if you remember our old z, z how do we calculate? x - μ / σ right. So, x - μ, here we take logarithm x, right 120 for x means logarithm x, μ is given by 5 σ is 0.1. So, we get 2.13.

(Refer Slide Time: 29:40)



So we go to; remember our old z table 2.13, 0.0166, 0.0166. Remember our old z table. So, it is 0.0166, 1 - 0.0166 is 0.983. So, the probability of having x > 120 is 0.9833. So here we have to use the z table. But here remember x - µ / σ is our z, but here we take the logarithm of x; that is the only difference. Now how we do this part, compute for probability for x lying between 110 and 130. This is like just like our z calculation. You remember long time back we did that, only thing is instead of x we use logarithm x.

(Refer Slide Time: 21:32)



So how do you do this, for x between 110 and 130? So, we take logarithm of 110, we take logarithm of 130, μ is 5, σ is 0.1. So, when we calculate this, it comes out to be 2.99, 1.32. So, go to our Z table 2.99, 2.99 is 0.0014 and 1.32 is given by 1.3, remember this, this is our z table 0.0934. This is our Z table if you remember 0.0934. So, we subtract and we end up with 0.092. So, the probability having between 110 and 130 of x, is 0.092. Median, the question is here, what is the value of the median ductile strength? As you know median, we have this equation;

$$median = e^{u}$$

simple. So, what we do $median = e^{u}$, μ is given by 5. So, that gives you 148.41. So, you see all these calculations can be done, instead of x we use lon x. So, we calculate Z, and then we go to our Z table and start getting the probability value. So, it is like that standardized normal distribution.

So, the lognormal distribution is also useful type of function, and you have the excel command also for lognormal distribution, which also plays a very important role as I showed you in example, related to biomaterial. So, we looked at different types of distributions over the past three or four lectures; the Weibull distribution, apart from our, the f distribution, t distribution, normal distribution, binomial. When the Poisson we looked at things like Weibull distribution, like the beta distribution, then lognormal distribution, hypergeometric distribution. So, all these are very useful distributions to have, as you can see many of them are ordinal type data, and in some situations where we cannot use chi square test, then we may have to use these type of (Refer Time: 24:07) distributions and tests for our calculations.

Refer Time: 24:07 inserted slide

a) $$E(X) = e^{u+\frac{\sigma^2}{2}} = e^{5+.005} = e^{5.005} = 149.16$$

$$Var(X) = e^{2u+\sigma^2}(e^{\sigma^2} - 1) = 223$$

b) $$P(X > 120) = 1 - P(X \le 120)$$
$$= 1 - P(Z \le \frac{\ln 120 - 5.0}{0.1})$$
$$= 1 - F(-2.13)$$
$$= 1 - 0.0166$$
$$= 0.9834$$

20

Thank you very much. We will continue further.

Key words - lognormal distribution, variance, sigma square, probabaility, Hypergeometric/Log normal distributions, Weibull distribution