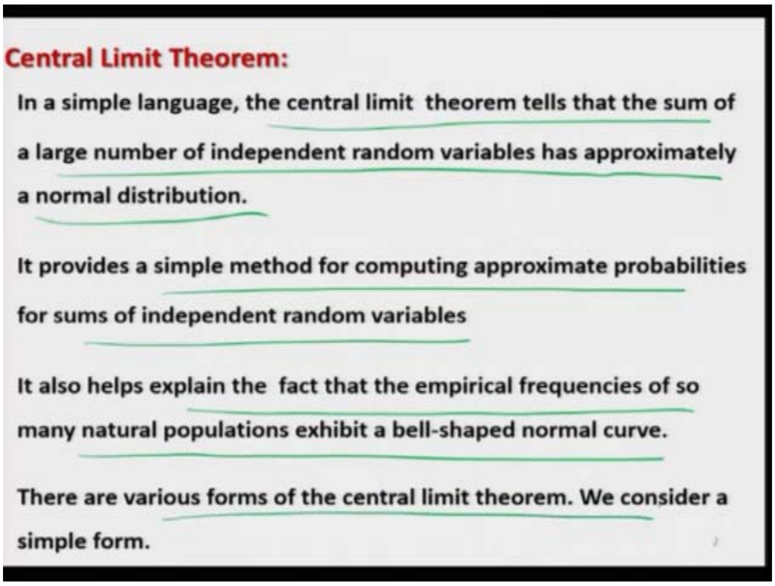
Essentials of Data Science with R Software – 1
Professor. Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur
Lecture No. 54
Central Limit Theorem

Hello friends welcome to the course Essentials of Data Science with R Software - 1 in which we are trying to understand the basic fundamentals of probability theory and statistical inference. So, now, you can see, we have now understood about the role of probability in computing different types of tools for the statistical inference.

So, now, I ask you a very simple question. In case if somehow it is becoming difficult to compute the probability from a given PDF or PMF what do you do? Can you recall that when we had discussed the normal distribution, then we had computed different types of probabilities, how? Just by taking the random variable - its mean divided by standard deviation. And we have seen that, that was a good approach to find out different types of approximations for the probabilities, which are coming from different distribution either discrete or continuous, but they can be very well approximated by that normal distribution.

So, now, in this class today, in the lecture today I will be working on this concept, and we are going to talk about a very important result this is Central Limit Theorem. Well, if you come to the pure statistics, there are several forms of the central limit theorem, which are defined for different types of conditions. Here in this case, I am going to discuss the most simple form of the central limit theorem, and I will try to illustrate that how µch it is useful in real data applications where you are trying to find out the value of approximate probabilities using the normal distribution. So, let us begin our lecture.

(Refer Slide Time: 02:22)



Central Limit Theorem:

In a simple language, the central limit theorem tells that the sum of a large number of independent random variables has approximately a normal distribution.

It provides a simple method for computing approximate probabilities for sums of independent random variables

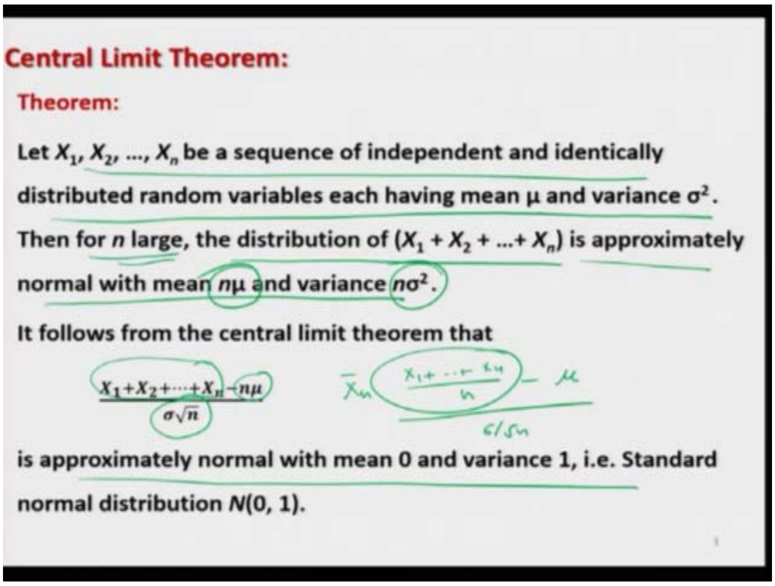
It also helps explain the fact that the empirical frequencies of so many natural populations exhibit a bell-shaped normal curve.

There are various forms of the central limit theorem. We consider a simple form.

So, what is this central limit theorem? In a very simple language, I can say that the central limit theorem tells that the sum of a large number of independent random variables has approximately a normal distribution. And it gives us a very simple method for computing approximate probabilities for the sums of independent random variables. You can recall that at couple of places I had shown you that, if two random variables are independent, then their sum also has got a probability mass function or probability density function with some specific parameters.

And then, in case if you want to find out the probabilities of their sum, that means, you have to first compute their joint probability density function of the random variable $X_1 + X_2$ and then you have to find out the probabilities from there, which is many times difficult. So, the central limit theorem helps us and explain the fact that the empirical frequencies of so many natural populations exhibit a bell shaped normal curve. That is a very useful information for those who are working in real data applications. So, there are various forms of the central limit theorem and we consider here a very simple form.

(Refer Slide Time: 03:37)



Central Limit Theorem:

Theorem:

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2.

Then for *n* large, the distribution of $(X_1 + X_2 + \ldots + X_n)$ is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

It follows from the central limit theorem that

$$\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}} \qquad \bar{X}_n \; \frac{\frac{X_1+\cdots+X_n}{n}-\mu}{\sigma/\sqrt{n}}$$

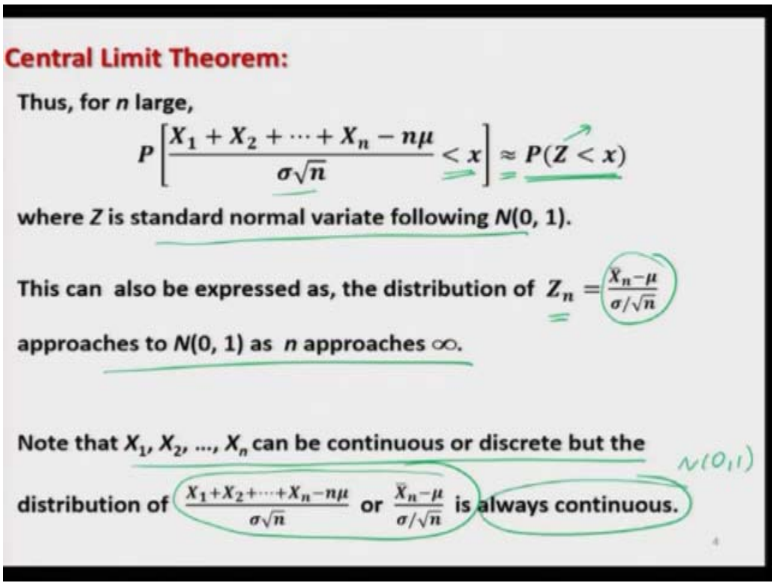
is approximately normal with mean 0 and variance 1, i.e. Standard normal distribution *N*(0, 1).

So, this theorem says that, let X_1, X_2,…, X_n be a sequence of independent and identically distributed random variables. And each of this variable is having a mean μ and variance σ^2, I am not assuming any distribution, or even I am not saying that whether X_1, X_2,…, X_n are the say discrete or say continuous.

Now, for large n, the distribution of $X_1 + X_2 + X_n$ is approximately $N(n\mu, n\sigma^2)$. Then it follows from the central limit theorem that $\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$, this will approximately follow a normal distribution with mean 0 and variance 1 that is your standard normal distribution.

And now, you can see here if you can just write down this thing as $(X_1 + X_2 + \ldots X_n)/n - \mu$, then what will happen, this result will be converted into a sample mean $\bar{X}_n$. So, now, both the options are there, and I will try to explain you.

(Refer Slide Time: 04:55)



Central Limit Theorem:

Thus, for *n* large,

$$P\left[\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}} < x\right] \approx P(Z<x)$$

where *Z* is standard normal variate following *N*(0, 1).

This can also be expressed as, the distribution of $Z_n = \frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$

approaches to *N*(0, 1) as *n* approaches ∞.

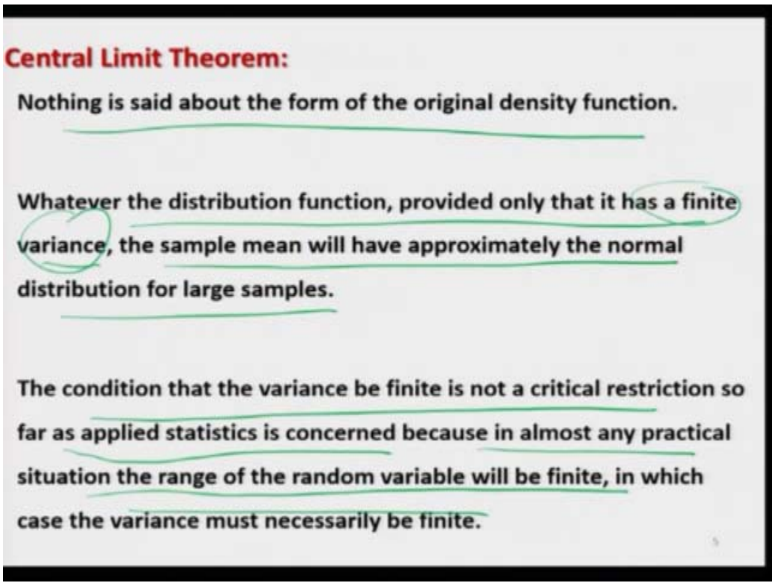
Note that $X_1, X_2, \ldots, X_n$ can be continuous or discrete but the distribution of $\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ or $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$ is always continuous.

N(0,1)

So now, in case if you want to find the probability that $\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ this is less than some quantity say X, then finding out this probability might be difficult, but you can very well approximate it by the N(0, 1) .So, this can be approximated by probability that Z is smaller than X, where your Z is a standard normal variant following N(0, 1).

So, this can also be expressed as the distribution of say here Z_n, which is $\bar{X}$ and - μ divided by σ by root n approaches to N(0, 1) as n approaches infinity. So, now, you can see here in this result, nowhere I have assumed that $X_1, X_2, \ldots, X_n$ are continuous or discrete, but we are saying that this result is valid for actually both. So, this $\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ or say $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$, their distribution is always going to be continuous, which is N(0, 1) and N(0, 1) is a continuous distribution.

(Refer Slide Time: 06:09)



Central Limit Theorem:

Nothing is said about the form of the original density function.

Whatever the distribution function, provided only that it has a finite variance, the sample mean will have approximately the normal distribution for large samples.

The condition that the variance be finite is not a critical restriction so far as applied statistics is concerned because in almost any practical situation the range of the random variable will be finite, in which case the variance must necessarily be finite.

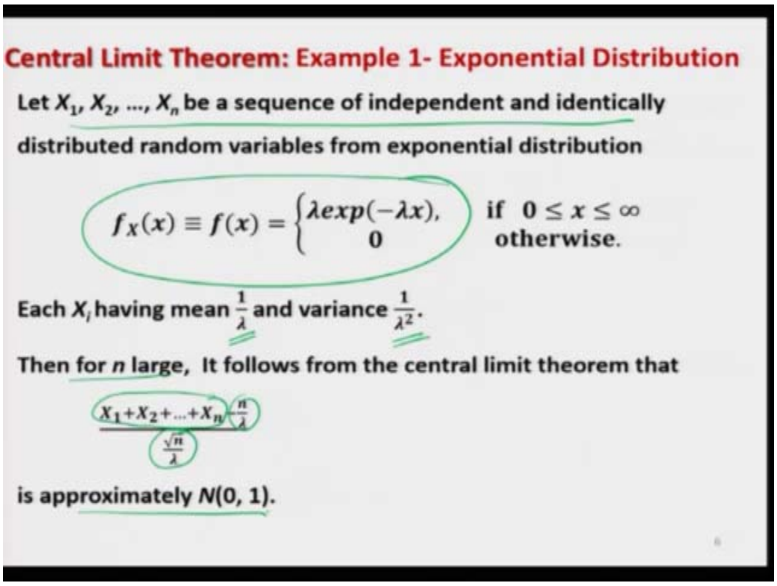
And beside this thing, if you have observed it, nowhere I am saying that X_1, X_2,…, X_n they are coming from binomial or Poisson or exponential or geometric, we have not talked about it. So, nothing is said about the form of the original density function. The only thing what we are trying to assume that we are assuming that the variance is finite.

So, whatever the distribution function be provided that provided only that it has a finite variance. The sample mean will have approximately the normal distribution for large samples. That is the condition that for finite sample, the approximation may not be good, but as you try to increase the value of your sample size or n, this approximation will become better.

Now, in case if you ask me that you are assuming here that the random variable should have finite variant do you think that is it a very difficult assumption? Whatever distributions you have done up to now, means, you have seen that they had finite variance. And similarly, if you try to look for other probability distributions, then most of them have finite variance. There is only one distribution, Cauchy distribution, which has this problem that it does not has the finite variance, but for all other things, there should not be any problem.

Because the condition that the variance be finite is not a very critical restriction so far as we are concerning the Applied Statistics. Because in almost any practical situation, the range of the random variable will always be finite. And in that case, the variance µst necessarily be finite. So, this is not a very difficult condition. So, it is not a very stringent condition that cannot be satisfied in real life. So, we need not to worry that µch and it will give us a good result.

(Refer Slide Time: 08:07)



Central Limit Theorem: Example 1- Exponential Distribution

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables from exponential distribution

$$f_X(x) \equiv f(x) = \begin{cases} \lambda exp(-\lambda x), & \text{if } 0 \le x \le \infty \\ 0 & \text{otherwise.} \end{cases}$$

Each X_i having mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$.

Then for n large, It follows from the central limit theorem that

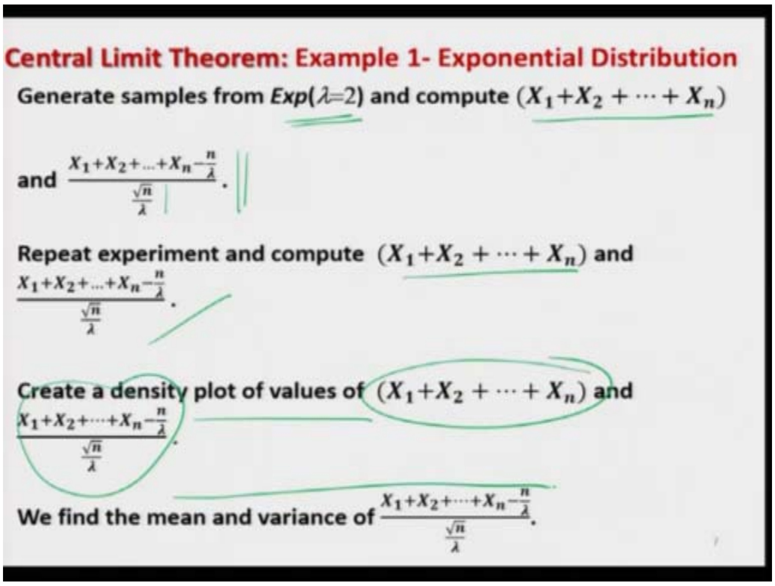
$$\frac{X_1+X_2+\ldots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$$

is approximately $N(0, 1)$.

So now, let me try to demonstrate that how this result will look like by taking an example of exponential distribution. Now, you know what is exponential distribution. So, let X_1, X_2,…, X_n be a sequence of IID random variables from exponential distribution whose PDF is this as n is having a parameter λ. So, each of this X_i has mean $1/\lambda$ and variance $1/\lambda^2$.

So, now, I can say that if I try to consider here this quantity here $\frac{X_1+X_2+\ldots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$, then for large n, this will have approximately a N(0, 1) distribution.

(Refer Slide Time: 08:54)



Central Limit Theorem: Example 1- Exponential Distribution

Generate samples from *Exp*(λ=2) and compute $(X_1+X_2+\cdots+X_n)$

and $\frac{X_1+X_2+\ldots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$.

Repeat experiment and compute $(X_1+X_2+\cdots+X_n)$ and $\frac{X_1+X_2+\ldots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$.

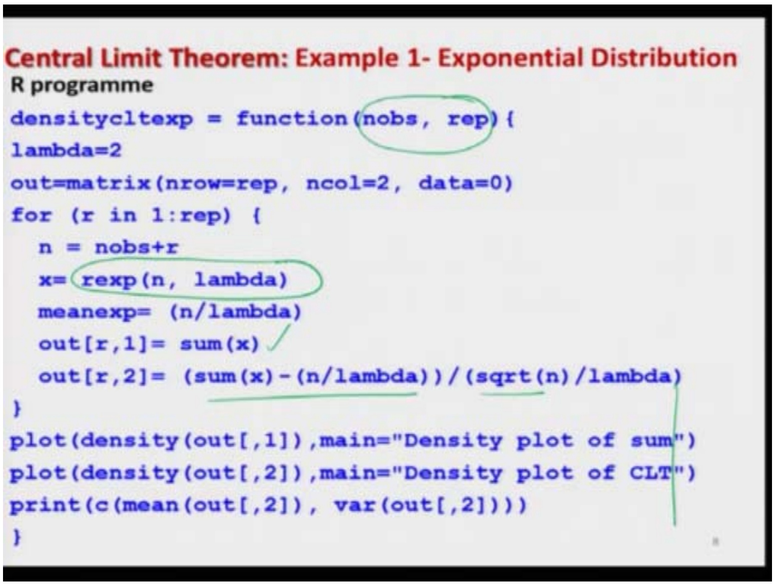
Create a density plot of values of $(X_1+X_2+\cdots+X_n)$ and $\frac{X_1+X_2+\cdots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$.

We find the mean and variance of $\frac{X_1+X_2+\cdots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$.

So, now, I will try to demonstrate this result using the R software, so that I can show you that how it look like. So, what I am going to do here that I will try to generate the sample from this exponential distribution with say λ equal to 2 and then I will try to compute the sum X_1 + X_2 +… X_n and I will try to compute this quantity also that $\frac{X_1+X_2+\ldots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$.

And now, I will try to repeat this experiment and every time I will compute these two quantities for a given sample. And then after that, I will try to create a density plot of both the values which are obtained for this and this and we also try to find out their mean and variance.

(Refer Slide Time: 09:40)



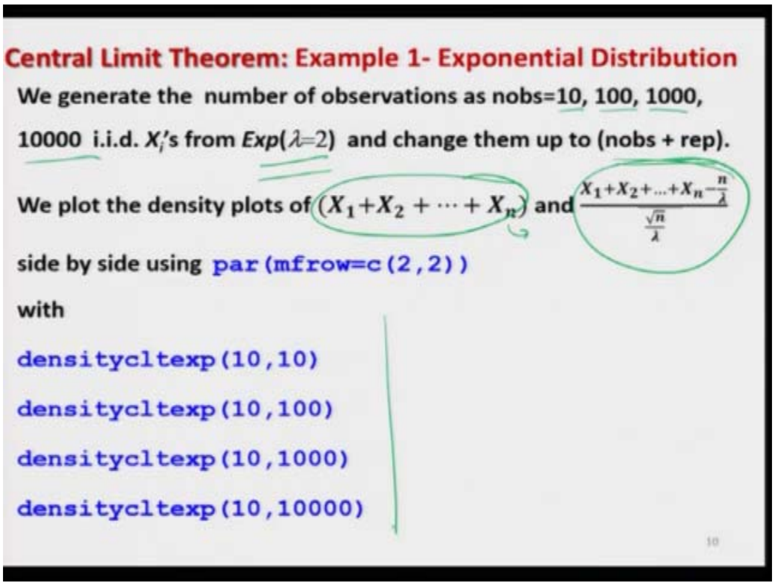
Central Limit Theorem: Example 1- Exponential Distribution

R programme

```
densitycltexp = function(nobs, rep){
lambda=2
out=matrix(nrow=rep, ncol=2, data=0)
for (r in 1:rep) {
  n = nobs+r
  x= rexp(n, lambda)
  meanexp= (n/lambda)
  out[r,1]= sum(x)
  out[r,2]= (sum(x)-(n/lambda))/(sqrt(n)/lambda)
}
plot(density(out[,1]),main="Density plot of sum")
plot(density(out[,2]),main="Density plot of CLT")
print(c(mean(out[,2]), var(out[,2])))
}
```

So, it is not difficult I am now giving you here the entire program, you have the slides also. So, you can simply copy and paste this program and you can just do it. So, you can see here this program is pretty simple. We have only here the number of observations and the number of times you want to repeat it. And then I am trying to generate the observation from this exponential distribution using the command nλ that you know, and then I am trying to compute the sum and then sum - n λ, and then the square root of n divided by λ. And then whatever is the outcome I'm trying to use here in plotting the density curve.

(Refer Slide Time: 10:17)



Central Limit Theorem: Example 1- Exponential Distribution

We generate the number of observations as nobs=10, 100, 1000, 10000 i.i.d. X_i's from $Exp(\lambda=2)$ and change them up to (nobs + rep).

We plot the density plots of $(X_1+X_2+\cdots+X_n)$ and $\frac{X_1+X_2+\ldots+X_n-\frac{n}{\lambda}}{\frac{\sqrt{n}}{\lambda}}$ side by side using `par(mfrow=c(2,2))` with

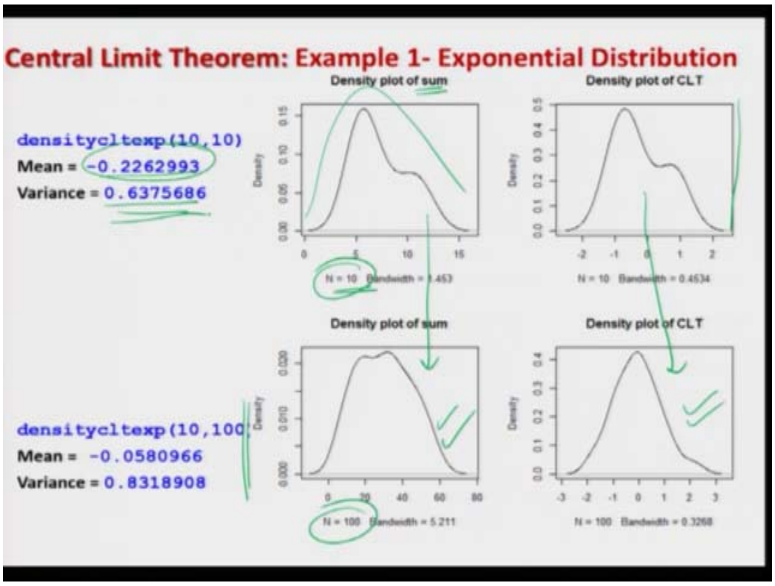
```
densitycltexp(10,10)
densitycltexp(10,100)
densitycltexp(10,1000)
densitycltexp(10,10000)
```

10

And this is the screenshot right? So, well, my basic objective is to make you convinced and to show you that how these things are happening. So, now, we try to generate the number of observation as a 10, 100, 1000, 10000 from exponential with λ equal to 2, and then we change them up to n observation + rep, rep means repetition. So, this n is going to be increasing.

So, we start with n and then continuously the n is going to be increased and every time these two quantities, their sum and then this quantity combined is n λ divided by square root of n λ. This is going to be calculated. And then, means, I have plotted all this curve clearly here.

(Refer Slide Time: 11:07)



Central Limit Theorem: Example 1- Exponential Distribution

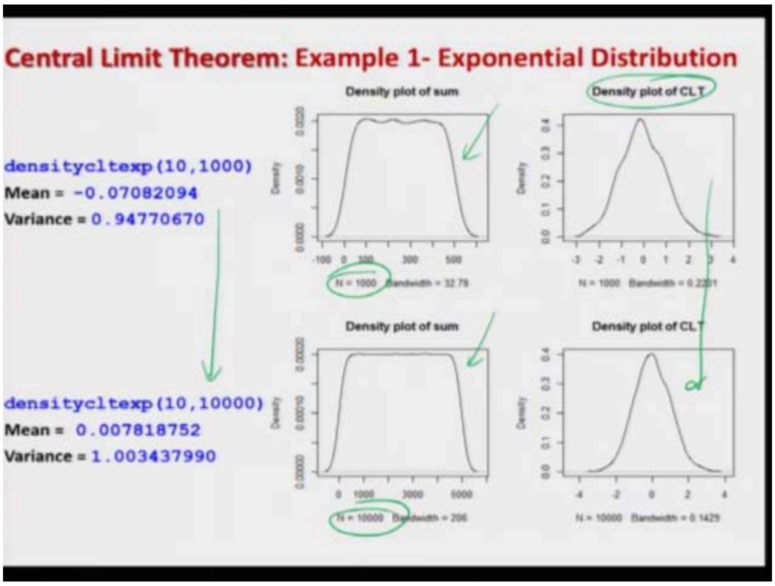
```
densitycltexp(10,10)
Mean = -0.2262993
Variance = 0.6375686

densitycltexp(10,100)
Mean = -0.0580966
Variance = 0.8318908
```

So now, you see that is the outcome. So, when I try to take out here this ken observation, then the plot of the sum that means $X_1 + X_2 +\ldots X_n$ this looks like this one, you can see here the curve. And if I try to plot the central limit theorem, that sum - its mean divided by standard deviation it looks like this. But now, in case if I try to increase the sample size. I tried to increase your samples from 10 to 100. You can see here now try to compare this curve and try to compare discuss what is really happening?

This is becoming a smoother, but this is now becoming more symmetric. And in both the cases if you try to see here that means and variants are obtained here like this and here even you can see that the mean is coming out to be saved - 0.22 which is say means quite close to 0, but still that is not actually 0 and variance is not actually here 1, but if you try to increase the sample size, the mean and variance are going closer to 0 and 1 respectively.

(Refer Slide Time: 12:13)



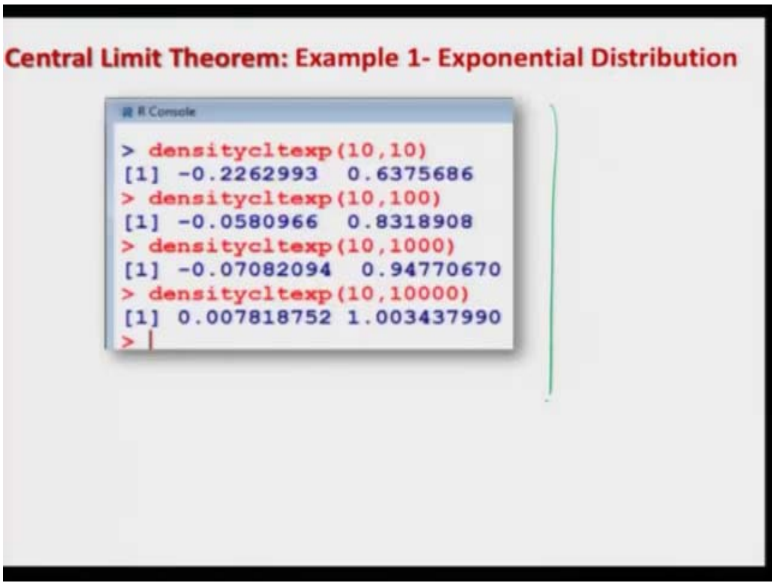
Central Limit Theorem: Example 1- Exponential Distribution

densitycltexp(10,1000)
Mean = -0.07082094
Variance = 0.94770670

densitycltexp(10,10000)
Mean = 0.007818752
Variance = 1.003437990

Now, in case if you try to increase the sample size to 1000 and then to 10,000 then the density plot of the sum will look like this or this, but in case if you try to plot the sum - its mean divided by standard deviation that is CLT it will look like this. And you can see here that as n going to infinity, this curve is becoming more symmetric and more similar to the normal density curve, and even the mean and variance they are approaching towards 0 and 1 respectively. So, this is the basic idea of the central limit theorem.

(Refer Slide Time: 12:49)

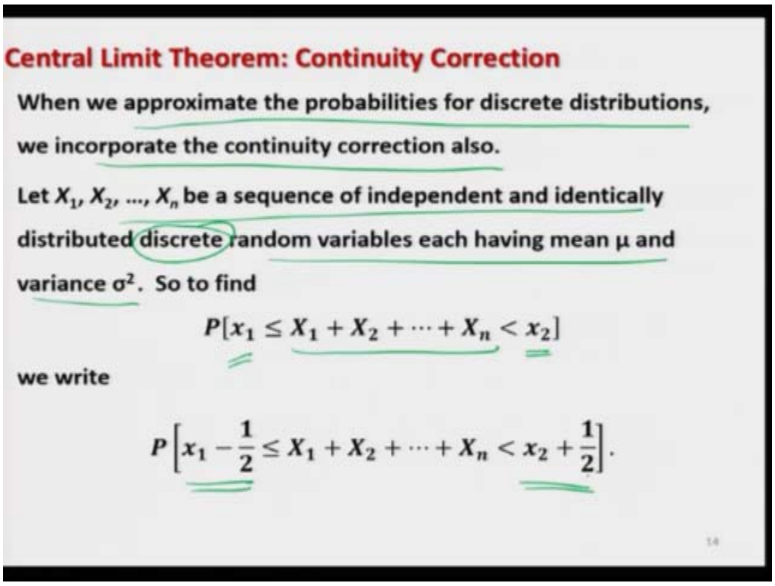


Central Limit Theorem: Example 1- Exponential Distribution

```
> densitycltexp(10,10)
[1] -0.2262993  0.6375686
> densitycltexp(10,100)
[1] -0.0580966  0.8318908
> densitycltexp(10,1000)
[1] -0.07082094  0.94770670
> densitycltexp(10,10000)
[1] 0.007818752 1.003437990
>
```

And you can see here these are the results, which are you have reported you here.

(Refer Slide Time: 12:54)



Central Limit Theorem: Continuity Correction

When we approximate the probabilities for discrete distributions, we incorporate the continuity correction also.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed discrete random variables each having mean μ and variance σ^2. So to find

$$P[x_1 \le X_1 + X_2 + \cdots + X_n < x_2]$$

we write

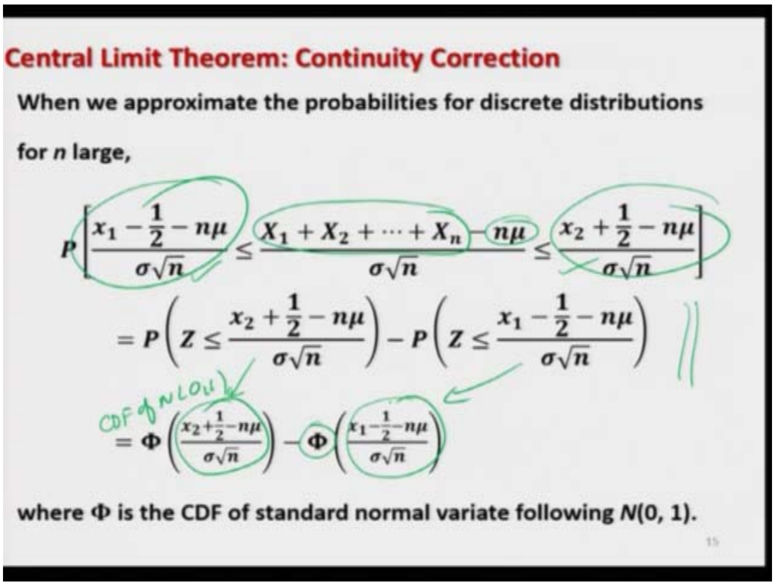
$$P\left[x_1 - \frac{1}{2} \le X_1 + X_2 + \cdots + X_n < x_2 + \frac{1}{2}\right].$$

14

Now, do you remember that when we did the normal distribution, then we also had discussed this aspect continuity correction? And there I had explained you that when we are trying to approximate a discrete distribution by a continuous distribution, then there is a need of continuity correction. And I had explained you what is this how it is obtained. So, now, I will not explain it again, but I will try to use it here, because in the central limit theorem also when you are trying to approximate the probabilities for a discrete random variable, then you need to apply the continuity correction also.

So, and we approximate the probabilities for discrete distribution, we incorporate the continuity correction also. So, let X_1, X_2,…, X_n be a sequence of IID, but discrete random variables and each having a mean μ and some variance σ^2. Suppose, we want to find out this probability that $X_1 + X_2 + X_n$ they are lying between two numbers X_1 and X_2. So, now, I have to do the same thing that we have discussed earlier that I have to subtract 1/2 in X_1 and I have to add 1/2 in X_2 and then I will try to compute the probability after standardization.

(Refer Slide Time: 14:10)



Central Limit Theorem: Continuity Correction

When we approximate the probabilities for discrete distributions for n large,

$$P\left[\frac{x_1-\frac{1}{2}-n\mu}{\sigma\sqrt{n}}\le\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}\le\frac{x_2+\frac{1}{2}-n\mu}{\sigma\sqrt{n}}\right]$$

$$=P\left(Z\le\frac{x_2+\frac{1}{2}-n\mu}{\sigma\sqrt{n}}\right)-P\left(Z\le\frac{x_1-\frac{1}{2}-n\mu}{\sigma\sqrt{n}}\right)$$

CDF of N(0,1)

$$=\Phi\left(\frac{x_2+\frac{1}{2}-n\mu}{\sigma\sqrt{n}}\right)-\Phi\left(\frac{x_1-\frac{1}{2}-n\mu}{\sigma\sqrt{n}}\right)$$

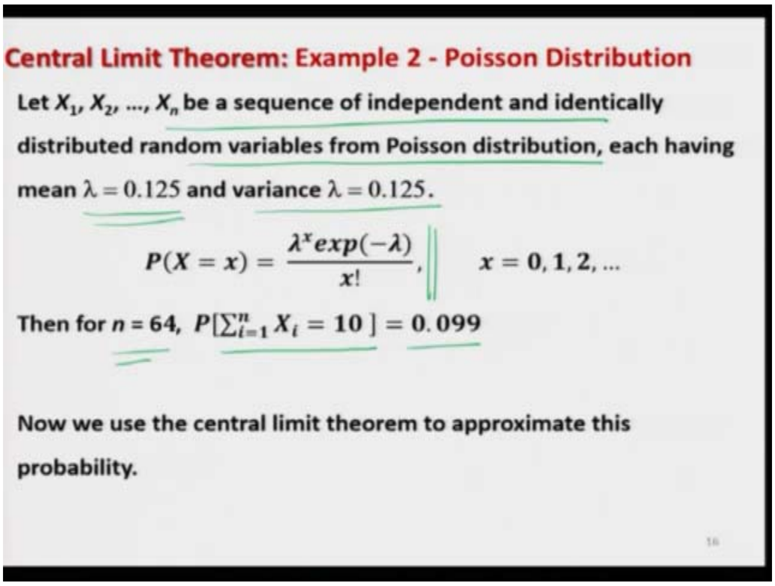
where Φ is the CDF of standard normal variate following $N(0, 1)$.

15

What we try to do we try to simply write down here $\frac{X_1+X_2+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ and the same thing I try to do on the both sides and this probability can we obtain the probability of here this quantity - probability of this quantity very straightforward algebra that now, you know and this is nothing but this can be obtained by the CDF of normal 0, 1, and this is also the CDF of N(0, 1) .

So, both these probabilities can be obtained at these two point $\frac{x_2+\frac{1}{2}-n\mu}{\sigma\sqrt{n}}$ and $\frac{x_1-\frac{1}{2}-n\mu}{\sigma\sqrt{n}}$.

(Refer Slide Time: 14:52)



Central Limit Theorem: Example 2 - Poisson Distribution

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables from Poisson distribution, each having mean $\lambda = 0.125$ and variance $\lambda = 0.125$.

$$P(X=x)=\frac{\lambda^x exp(-\lambda)}{x!}, \qquad x=0,1,2,\ldots$$

Then for $n = 64$, $P[\sum_{i=1}^{n} X_i = 10\,] = 0.099$

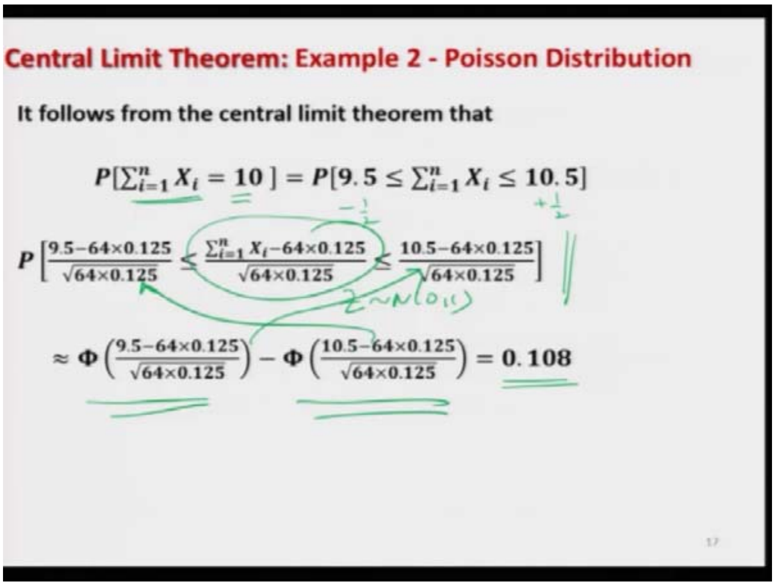
Now we use the central limit theorem to approximate this probability.

16

So, now, let me try to give you here one more example from the Poisson distribution where we try to apply this continuity correction. So, let $X_1, X_2, \ldots, X_n$ be a sequence of IID random variables from Poisson distribution each having a mean say λ is equal to 0.125 and variance is equal to 0.125. Because in the case of Poisson distribution, the mean and variance both are the same, and its probability mass function is given by this quantity.

So, for example, if I try to see here, therefore, n equal to 65. Suppose, we want to know the probability that some of these X_i is equal to 10 is equal to 0.099 that we can obtain the R. Now we try to approximate the same probabilities using the central limit theorem.

(Refer Slide Time: 15:40)



Central Limit Theorem: Example 2 - Poisson Distribution

It follows from the central limit theorem that

$$P[\textstyle\sum_{i=1}^{n} X_i = 10] = P[9.5 \le \textstyle\sum_{i=1}^{n} X_i \le 10.5]$$

$$-\tfrac{1}{2} \qquad +\tfrac{1}{2}$$

$$P\left[\frac{9.5-64\times0.125}{\sqrt{64\times0.125}} \le \frac{\sum_{i=1}^{n} X_i-64\times0.125}{\sqrt{64\times0.125}} \le \frac{10.5-64\times0.125}{\sqrt{64\times0.125}}\right]$$

$$Z \sim N(0,1)$$

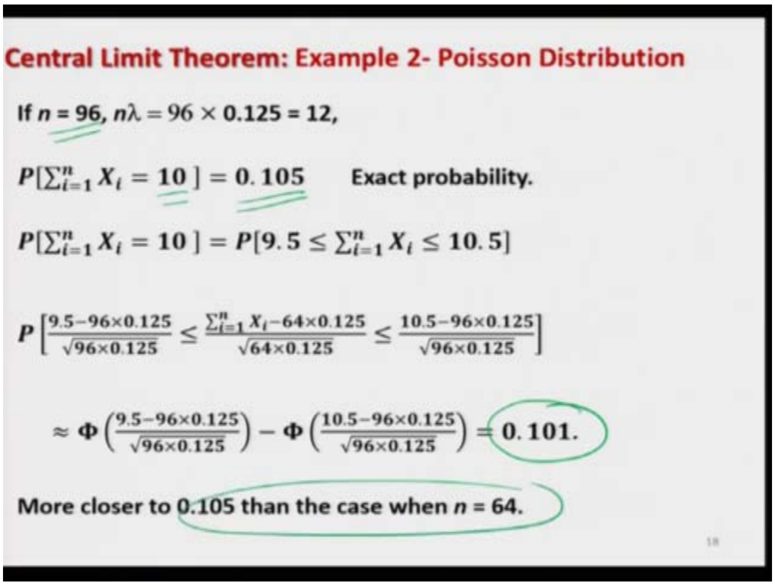
$$\approx \Phi\left(\frac{9.5-64\times0.125}{\sqrt{64\times0.125}}\right) - \Phi\left(\frac{10.5-64\times0.125}{\sqrt{64\times0.125}}\right) = 0.108$$

17

So, what I try to do here that, I try to write down here $\sum_{i=1}^{n} X_i = 10$ and this can be obtained by applying the continuity condition that will try to say subtract 1/2 and add 1/2 on the left and writer limits and then you try to standardize it that just mean summation X_i - its mean divided by standard deviation and you get here this expression and this probability can be can be written.

Because now, this is here you are here is Z that is the standard normal variant following a N(0, 1) distribution, and this probability can be written in the form of CDF as CDF at this point and the CDF at this point. And if you try to compute this probability from the R software, this will come out to 0.108 this is not difficult at all.

(Refer Slide Time: 16:32)



Central Limit Theorem: Example 2- Poisson Distribution

If $n = 96$, $n\lambda = 96 \times 0.125 = 12$,

$P[\sum_{i=1}^{n} X_i = 10] = 0.105$ Exact probability.

$P[\sum_{i=1}^{n} X_i = 10] = P[9.5 \le \sum_{i=1}^{n} X_i \le 10.5]$

$$P\left[\frac{9.5-96\times0.125}{\sqrt{96\times0.125}} \le \frac{\sum_{i=1}^{n} X_i - 64\times0.125}{\sqrt{64\times0.125}} \le \frac{10.5-96\times0.125}{\sqrt{96\times0.125}}\right]$$

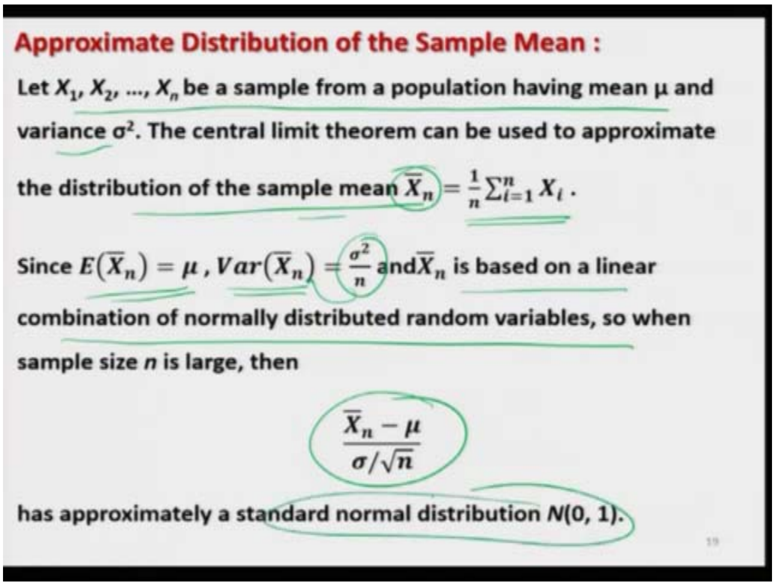
$$\approx \Phi\left(\frac{9.5-96\times0.125}{\sqrt{96\times0.125}}\right) - \Phi\left(\frac{10.5-96\times0.125}{\sqrt{96\times0.125}}\right) = 0.101.$$

More closer to 0.105 than the case when $n = 64$.

18

So, now, in case if you try to take here, suppose here n is equal to 96 and you try to compute the probability that $\sum_{i=1}^{n} X_i = 10$ this will come out to be 0.105. But, if you try to use it on the basis of the central limit theorem, this will come out to be here 0.101. So, you can see here, this is more closer to 0.105 than the case when we had n is equal to only 64.

(Refer Slide Time: 17:04)



Approximate Distribution of the Sample Mean :

Let $X_1, X_2, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Since $E(\overline{X}_n) = \mu$, $Var(\overline{X}_n) = \frac{\sigma^2}{n}$ and$\overline{X}_n$ is based on a linear combination of normally distributed random variables, so when sample size n is large, then

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$$

has approximately a standard normal distribution $N(0, 1)$.

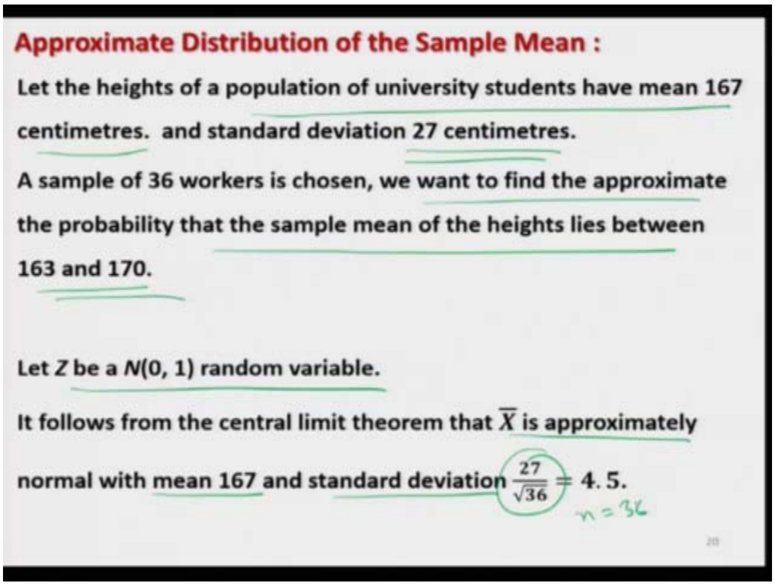
19

So, you can see that as you are trying to increase the sample size, the probabilities are becoming more clear. So, now, we try to come to one more result, where I am trying to give you the approximate distribution of the sample mean, because, now, in the last lecture, we have seen that sample mean is playing a very important role, when we are trying to draw different types of statistical inferences.

So, like this $X_1, X_2, \ldots, X_n$ be a sample from a population having a mean μ and variance σ^2. Now, the central limit theorem can be used to approximate the probability distribution of the sample mean say$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Now, we know we already have actually proved that expected value of $\overline{X}_n$ is equal to μ and variance of $\overline{X}_n$ is σ^2/n which depends on this here n.

And we also know that this $\overline{X}_n$ this is based on a linear combination of the normally distributed random variable. So, when sample size is larger than$\frac{\overline{X}_n-\mu}{\sigma/\sqrt{n}}$, this will has this will also have an approximately normal distribution with mean 0 and variance 1 that is the standard normal distribution N(0, 1) .

(Refer Slide Time: 18:19)



Approximate Distribution of the Sample Mean :

Let the heights of a population of university students have mean 167 centimetres. and standard deviation 27 centimetres.

A sample of 36 workers is chosen, we want to find the approximate the probability that the sample mean of the heights lies between 163 and 170.

Let Z be a $N(0, 1)$ random variable.

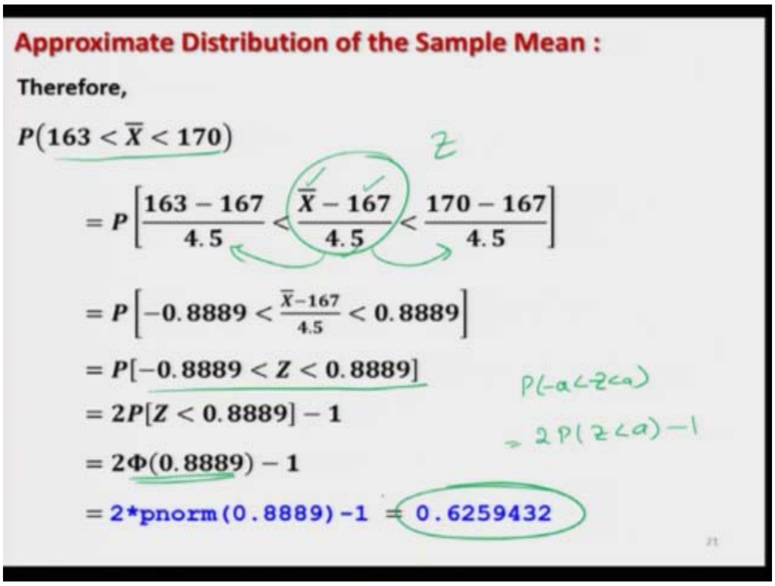
It follows from the central limit theorem that $\overline{X}$ is approximately normal with mean 167 and standard deviation $\frac{27}{\sqrt{36}} = 4.5$.

$n = 36$

So, now, let me try to give you an example of this result, suppose the height of university student was measured and it was found that it has got a mean of 167 centimeter, and a standard deviation of 27 centimeter. Now, a sample of 36 workers is chosen from that population. And we want to find the approximate probability that the sample mean of the height lies between 163 and 170. That is what we want to find out.

So, now, using the standard notations let Z be a N(0, 1) random variable and it follows using the central limit theorem, that $\overline{X}$ is approximately normal with mean 167 and a standard deviation 27 divided by the square root of 36, which is the value of here n equal to 36, and this will come out to be here 4.5.

(Refer Slide Time: 19:08)



Approximate Distribution of the Sample Mean :

Therefore,

$P(163 < \bar{X} < 170)$

Z

$$= P\left[\frac{163-167}{4.5} < \frac{\bar{X}-167}{4.5} < \frac{170-167}{4.5}\right]$$

$$= P\left[-0.8889 < \frac{\bar{X}-167}{4.5} < 0.8889\right]$$

$$= P[-0.8889 < Z < 0.8889]$$

$P(-a<Z<a)$

$$= 2P[Z < 0.8889] - 1$$

$= 2P(Z<a)-1$

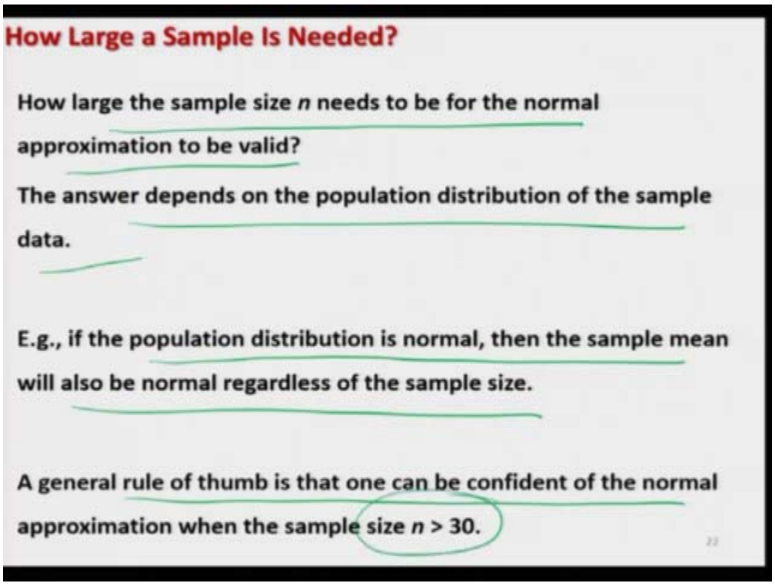
$$= 2\Phi(0.8889) - 1$$

= 2*pnorm(0.8889)-1 = 0.6259432

Now, in case if you want to compute this probability that $\bar{X}$ is lying between 163 and 170, you can just standardize the sample mean by writing a sample mean by - its mean divided by standard deviation and you try to do the same operation on both the sides. And you can simply solve it after this this will become a standard normal variant. And this probability will be equal into finding out the probability between - 0.8889 and + 0.8889.

So, you have done a result that probability that is lying between - a and + a. then this can be written an as a twice for probability that less a - 1. So, I can use that result to directly here and can find out the value of the CDF at this point from the R software, so that you know how to find out the value of the CDF in the standard normal distribution. So, this value comes out to be here 0.625 and so on.

(Refer Slide Time: 20:12)



How Large a Sample Is Needed?

How large the sample size *n* needs to be for the normal approximation to be valid?

The answer depends on the population distribution of the sample data.

E.g., if the population distribution is normal, then the sample mean will also be normal regardless of the sample size.

A general rule of thumb is that one can be confident of the normal approximation when the sample size $n > 30$.

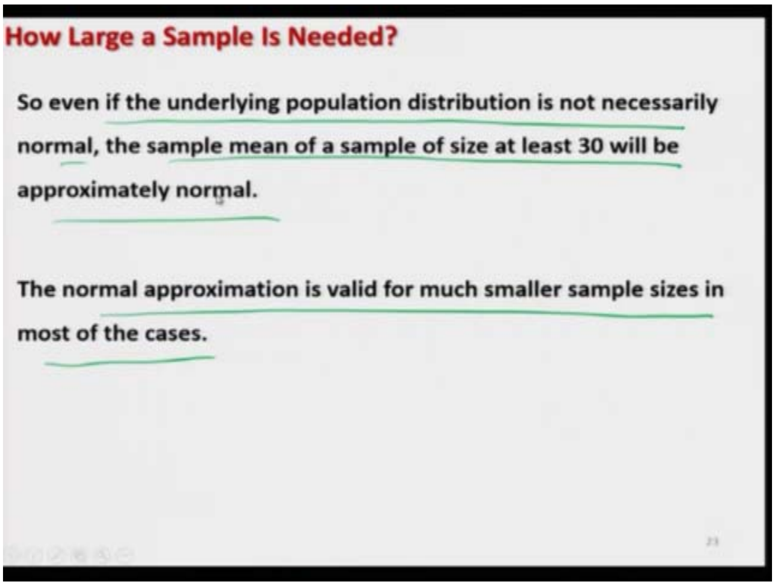
So, you can see here it is not difficult to obtain such complicated probabilities without any problem. Now, let me try to address the last topic of this lecture. At many times in real life, we always have a question that, what should we the sample size? Means, how many observations should we draw, so that we can take a reasonable conclusion?

So, now, the question I am trying to phrase here is that how large the sample sizes needs to be for the normal approximation be valid for example. In all these cases, we are trying to approximate the probabilities using the normal distribution and we are saying at least theoretically that as n goes to infinity this will work, but, in practice, somebody will always like to know, what is this n for which it is really going to work?

So, this answer depends on the population distribution of the sample data that is there that you have seen means, I was always trying to do an experiment and I was trying to show you the results after every probability density function and probability mass function that I will try to generate the observation from that distribution and then try to compute its mean and variance. And then, I was trying to show you that as n is increasing the estimated value of mean and variants are converging towards the theoretical mean and variance.

So, now, for example, if I say the population distribution is supposed normal, then the sample mean will also be normal regardless of that sample size then there is no issue. A general rule of thumb is that one can be confident of the normal approximation when the sample size is n is greater than 30. Now, you know what is the reason.

(Refer Slide Time: 22:00)



How Large a Sample Is Needed?

So even if the underlying population distribution is not necessarily normal, the sample mean of a sample of size at least 30 will be approximately normal.

The normal approximation is valid for much smaller sample sizes in most of the cases.

Means, this reason will become more clear as we move forward, but do you remember the discussion in the t distribution lecture try to think about it. So, in such a case, even if the underlying population distribution is not necessarily normal, the sample mean of a sample of size at least 30 will be an approximately normal. And the normal approximation is valid for µch smaller sample size in most of the cases, but it depends on say different types of parameters, which you are going to involve or what is the probability density function probability mass function and so on.

So, now, let me come to an end to this lecture that was a very short and simple result. But what I am trying to address here that is very important, because in data science, you are dealing with complicated distributions. And sometimes there can be a combination of discrete as well as continuous random variables also and you are always interested in computing different types of probabilities.

Well, without probabilities, you cannot do the statistical modeling that we have understood. So, now, whenever you are trapped, whenever you want to find out the value of the probability. Yes, you cannot find out the exact probability, but, if you try to use this central limit theorem possible, you will get a very good approximation, and this is going to work very well in practice.

Now, regarding the sample size issue, my experience says that the sample size required depends on many things. For example, even if you are trying to take the N(μ, σ^2), in case if the σ^2 value is very, very low possibly, you will need a smaller sample size. But in case if your σ^2 is very high, then you will need a µch larger size.

So, the same thing can happen in other type of probability density functions also. Also in real life, whenever you are trying to deal with the real data, it will be difficult many times to know exactly what is the probability density function of probability mass function so you are simply trying to approximate or assume that this is the distribution, which is going to represent that true data up to a great extent. So, that is a sort of approximation.

So, under those type of situation this CLT is going to help you a lot. So, now, I would stop here and I will request you that you please try to think, try to look into books and you will find different versions of the central limit theorem under different types of condition. For example, here we have assumed that the random variables are identically and independently distributed. Supposed they are not independent then these things are not going to be valid.

So, for that you have to look into the books and find out the correct version. But now I am sure that after so µch of training, it should not be a very difficult task for you to learn and understand those things. So, you try to revise, have a look and I will see you in the next lecture with more topics. Till then, goodbye.