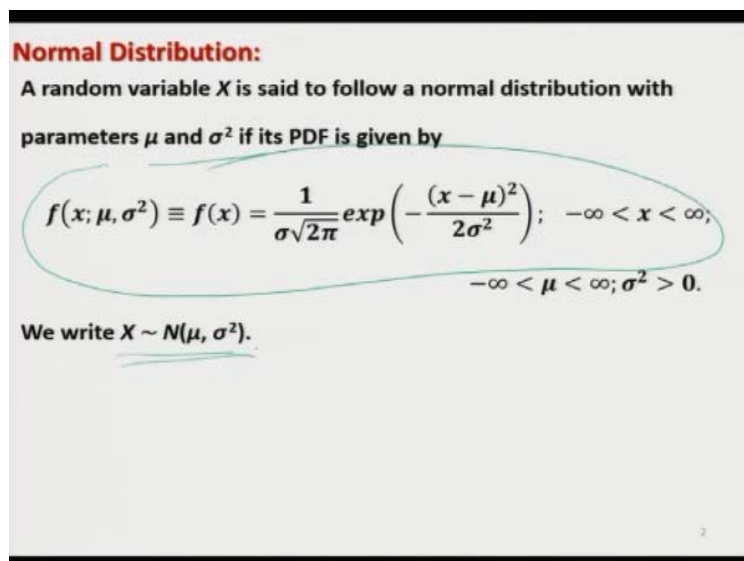


Essentials of Data Science with R Software- 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur
Lecture 41
Normal Distribution – More Results

Hello friends, welcome to the course Essentials of Data Science with R Software and now you can see that in the last two lectures we had a discussion on normal distribution and now, I am sure that you must have understood the importance of this probability distribution and why it has become so popular.

So, in this lecture we will continue on the same lines and I will try to show you that why this normal distribution became more popular and I will try to illustrate here some very important results which are very useful when you are trying to use statistics or probability or statistical inference in bigger data sites, particularly your data science. So, let us try to understand these results one by one.

(Refer Slide Time: 1:12)



Normal Distribution:
A random variable X is said to follow a normal distribution with parameters μ and σ^2 if its PDF is given by

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

$-\infty < \mu < \infty; \sigma^2 > 0.$

We write $X \sim N(\mu, \sigma^2)$.

So, now just for your quick review a random variable x is said to follow a normal distribution with the mean μ and σ^2 if its PDF is given by this function and we write it at x follows a normal $\mu \sigma^2$ like this.

(Refer Slide Time: 1:27)

Normal Distribution: Distribution of the Arithmetic Mean

Assume that $X \sim N(\mu, \sigma^2)$.

Consider a random sample $X = (X_1, X_2, \dots, X_n)$ of independent and identically distributed random variables X_i with $X_i \sim N(\mu, \sigma^2)$.

Then, the arithmetic mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) + \text{Covariance} = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

where $Cov(X_i, X_j) = 0$ for $i \neq j$. $Cov(X_i, X_j) = E[(X_i - \bar{X})(X_j - \bar{X})]$

Now, I come to our main job. Now, here we are trying to find out the distribution of the arithmetic mean, it is very important, because you will see that one of the basic objectives in statistics and data science is to compute the or to find out the central tendency of the data and for that we have to compute value statistics like arithmetic mean, median, more geometric mean, harmonic mean or anything else and all these things they are going to be based on the X_1, X_2, \dots, X_n . So, that is why we need this result over here.

So, we assume here that the random variable x is following a $N(\mu, \sigma^2)$ distribution and now we consider here a random sample. Now, this random sample is here X_1, X_2, \dots, X_n , so you have to be just be careful here that I am using here this symbol here X . So, once I say this is a random sample that means all the observations are independent and identically distributed and every observation is coming from a population $N(\mu, \sigma^2)$. Remember every observation is coming from the population which have got the same mean and same variances.

Now, in case if you try to compute here the value of arithmetic mean of this observation, so that is going to be simply $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, and now the result which I am trying to explain you here is that this sample mean will follow a normal distribution with mean

μ and σ^2 by n . So, you can see here the arithmetic mean \bar{X} will also have the same distribution as of the original observation.

The mean of sample mean and the observations, that is the same as μ , you can see here μ and here μ but the variance is changing, for the observations the variance was σ^2 and for the sample mean the variance here is \bar{X} . So, variance is σ^2/n . So, now you can see here that the variance is σ^2/n , that means this is smaller than the variance of X , that is a very important property, but now first we try to understand how are we getting this result.

So, suppose if you want to find out expected value of \bar{X} that will become here $E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$. and if you want to find out here the variance of \bar{X} that will become here the $Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i)$ plus there will be some covariance terms also but covariance will become 0 because these observations are independent, what is called this quantity C o v this is actually here the covariance.

Up to now we have not covered it but very soon we are going to cover it to understand it but at this moment you can assume that this is some statistical measure, and that is a very simple thing if you want to define here the covariance between here X_i and X_j this is simply here expected value of $(X_i - \bar{X}) X (X_j - \bar{X})$. So, you can see that this is a quantity like variance, as the variance is trying to indicate the variability of the observations. So, covariance is going to indicate the covariability of the observation, but do not worry we will consider it in more detail.

So, once these observations are independent then this covariance term becomes 0 and that is what I have used here. so, now you can see here this quantity becomes $\frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$.

(Refer Slide Time: 5:54)

Normal Distribution: Sum of Normally Distributed Random Variables

Assume that $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2, \dots, n$,

X_1, X_2, \dots, X_n are independently distributed random variables, (not necessarily identically distributed) and

a_1, a_2, \dots, a_n are real numbers, then

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

Normal Distribution: Distribution of the Arithmetic Mean

Assume that $X \sim N(\mu, \sigma^2)$.

Consider a random sample $X = (X_1, X_2, \dots, X_n)$ of independent and identically distributed random variables X_i with $X_i \sim N(\mu, \sigma^2)$.

Then, the arithmetic mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$$E(\bar{X}) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n Var(X_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}$$

where $Cov(X_i, X_j) = 0$ for $i \neq j$. $Cov(X_i, X_j) = E[(X_i - \mu)(X_j - \mu)]$

And you can also see here that because this \bar{X} this is a linear function of the normally distributed random variable. So, this is also going to follow the normal distribution. So, I can see here the distribution of \bar{X} is normal with mean here μ and σ^2/n . And in case if I try to extend this result that instead of considering here only $\sum_{i=1}^n X_i$, in case if I try to consider here $\sum_{i=1}^n a_i X_i$ where a_1, a_2, \dots, a_n are some real numbers, then this result can be extended to that this linear function of normally distributed random variable.


Summation i goes from 1 to n $a_i X_i$ will follow a normal distribution with the mean which is $\sum_{i=1}^n a_i \mu_i$ and variance $\sum_{i=1}^n a_i^2 \sigma_i^2$. So, that is the result which possibly you will be using many times.

(Refer Slide Time: 7:06)

Normal Distribution: Normal Approximation to the Binomial Distribution

Normal distribution can be to approximate binomial probabilities for cases in which n is large.

An illustration is provided where the area of each bar equals the binomial probability of x .



Notice that the area of bars can be approximated by areas under the normal density function.

Because a continuous normal distribution is used to approximate a discrete binomial distribution, a modification is needed which is referred to as a continuity correction.

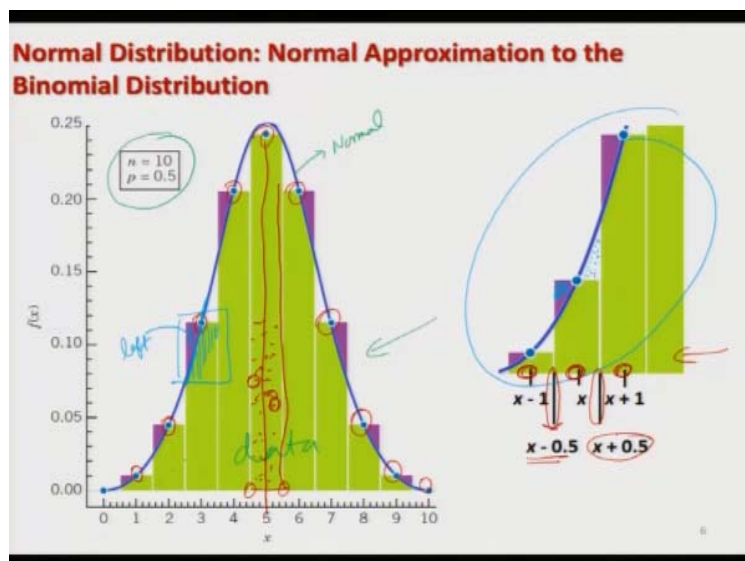
Now, I come to one more aspect, you know we have done the binomial distribution whose probability mass function is $\binom{n}{x} p^x (1-p)^{n-x}$. Now, you can imagine that computation of n choose x will become difficult if the value of n or say x they become very high and, in those cases, you can also believe or you must have experience that computing the probability like $\binom{n}{x} p^x (1-p)^{n-x}$ etc., they are computationally expensive and they are not so state forward thing, these are not so straightforward concepts that can be applied directly.

So, the question comes that when you are trying to deal in a situation where the binomial probabilities are difficult to compute, then what should you do? In such situation the normal distribution comes to our help and normal distribution helps in approximating the binomial probabilities. So, that is what I am going to now discuss.

So, normal distribution can be used to approximate the binomial probabilities for cases in which the n is large and just for the sake of example, we can consider the histogram in which each of the bar is going to indicate the binomial probabilities. Now, suppose if that the area under each of the bar that can be approximated by a normal density function which is a smooth curve actually like this one, then how to get it done? One thing what you have to keep in mind that normal is a continuous distribution and whereas binomial is a discrete distribution.

So, obviously a basic issue comes to our mind that how a continuous random variables distribution can approximate the probabilities of a discrete random variable probability mass function. So, well we can do it but for that we need to first understand a basic fundamental concept that is needed to approximate the binomial probabilities by the normal distribution and for that we try to make some sort of a small adjustment in the computed values which is called as here continuity correction.

(Refer Slide Time: 10:13)



So, what is this thing first we try to understand. Well suppose, this green color is the data that you are going to obtain and now you will get here a histogram which is here. Now, you are trying to approximate this histogram by a normal curve, so this normal curve here is given by this blue color. So, and these values, the values on which the histogram has been created they have been generated from a normal distribution with n equal to 10 and

p is equal to 0.5, and now if you try to create this normal curve or in the first step we try to create here a smooth curve, for that what we try to do? We try to first mark the midpoints of the bars of this histogram like this one, and then we try to join them by a smooth curve, what is here indicated by the blue color curve.

Now, but when you are trying to do it you can see here that the area under the curve will be only that area which is under the blue color curve. Now, what is really happening is the following, if you try to look at suppose the quickest part, the curve is passing through here like this one, so area under the curve will be only this one, but this is the part which is actually left and if you try to magnify it, it will actually here look like this one.

So, what is really happening, that you can see here that this part is ignored but on the other hand, there is here some part which is unnecessarily included because that was not covered by the histogram and in histogram actually you assume that all the observations which are lying inside the bar they are assumed to be concentrated at the midpoint, for example, if you try to look here, in this particular bar, what is happening? This is something like 4.5 and 5.5, so all those values which are lying between 4.5 and 5.5 they will be scattered here like this, but what are you assuming? That all these values they are assumed to be concentrated at the midpoint 5.

So, ideally this value should have which is here, now looks to be on this line and this value on the left-hand side of the line it also comes to the center of the value and even those points which are closer to the line they are also on the line and those points which are away from the line they are also on the point. So, now what to do? So, what we try to do here, that we try to consider the two points on the left-hand side and right-hand side of the value x , for example, if you try to look into this picture here, this is here x , this is here $x - 1$ and this is here $x + 1$. Now, in between there is a point here like $x - 0.5$ and just after this there is another point here $x + 0.5$. So, we try to consider such values and we try to make here a sort of continuity correction.

(Refer Slide Time: 14:24)

Normal Distribution: Normal Approximation to the Binomial Distribution

If X is a binomial random variable with parameters n and p , then

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable.

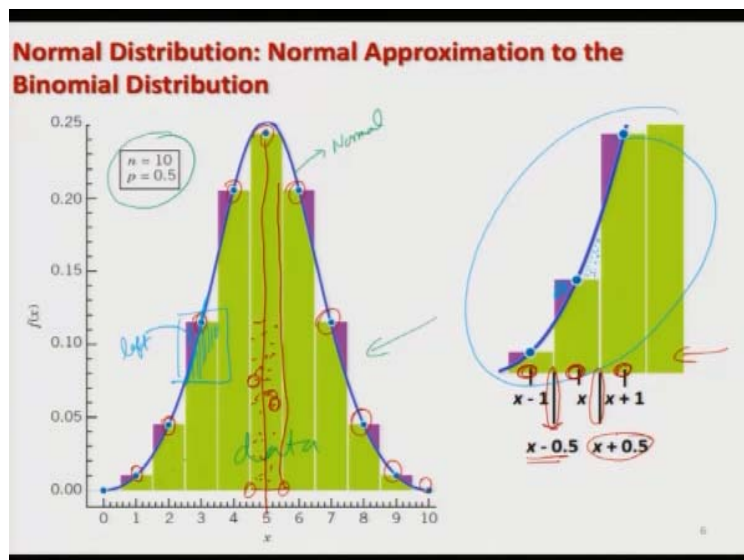
To approximate a binomial probability with a normal distribution, a continuity correction is applied as follows:

$P(X \leq x) = P(X \leq x + 0.5) = P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$

and $P(X \geq x) = P(X \geq x - 0.5) = P\left(Z \geq \frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$

The approximation is good for $np > 5$ and $np(1-p) > 5$.

Handwritten notes:
 $X \sim B(n, p)$
 $\text{mean} = np$
 $\text{var} = npq$
 $q = 1-p$
 $\sigma^2 = npq$
 $\sigma = \sqrt{npq}$
 Normal
 Bin
 Bin



How we try to do it? I will just try to now illustrate, while I try to show you that how the binomial probabilities can be approximated by the normal probabilities. The concept is very simple, you simply have to standardize the variable actually you already have computed such probabilities but now because you are trying to approximate the probabilities of a discrete distribution by a continuous distribution, so this continuity correction concept is required.

So, now let us try to first understand what we are trying to do here, let X be a binomially distributed random variable with parameter n and p , so in this case you know that if X is binomial np , then its mean is given by np and variance is given by $np(1 - p)$ or you had written here as a npq . So, now I try to standardize it X minus its mean and p divided by σ which is σ is something like here, $\sigma^2 = npq$, so this becomes here square root of $np(1 - p)$.

So, now the statement is that this quantity Z , this is approximately a standard normal random variable that we can do without any problem, we have learnt this technique that you try to subtract the observation by mean and divide by the standard deviation. So, to approximate this binomial probability with the normal distribution, now we try to make a continuity correction before we try to use it.

So, you can write down here you want to find out probability that X is less than equal to x , but now we try to find out here probability that X is less than or equal to x plus half and this half is coming from here, that all values are here but you are trying to include this part also.

So, and then you simply try to make it here $\frac{X - np}{\sqrt{np(1-p)}}$ and this you try to do on the both hand sides. So, this will transform the first variable as Z and the second variable will become here $\frac{x + 0.5 - np}{\sqrt{np(1-p)}}$ and similarly, if you want to compute the probability like X greater than or equal to x , then in this case what you have to do?

You have to make a continuity correction like this, that you are going to compute probability X greater than or equal to $x - \text{half}$, so and then doing the same transformation on the both the side that $(x - \mu)/\sigma$ on both the sides we get here probability that Z is greater than $\frac{x + 0.5 - np}{\sqrt{np(1-p)}}$.

And this is how you can compute the probability of your Z which is going to approximate the probability of a binomial distribution, because you can see here this is here binomial and this is here normal and this is here binomial and this is here normal. And this approximation is actually good for those situations where you can see that the value of np here is greater than 5 and np into $1 - p$ is also greater than 5.

(Refer Slide Time: 18:01)

Normal Distribution: Normal Approximation to the Binomial Distribution

For a binomial variable X , $E(X) = np$ and $Var(X) = np(1 - p)$.

Consequently, the expression is the formula for standardizing the random variable X .

Probabilities involving X can be approximated by using a $N(0,1)$.

The approximation is good when n is large relative to p .

Write the probability and then adjust by the 0.5 correction factor.

disc X : original s.v.
 Z : function of X

Normal Distribution: Normal Approximation to the Binomial Distribution

If X is a binomial random variable with parameters n and p , then

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable.

To approximate a binomial probability with a normal distribution, a continuity correction is applied as follows:

$P(X \leq x) = P(X \leq x + 0.5) = P\left(Z \leq \frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$

and $P(X \geq x) = P(X \geq x - 0.5) = P\left(Z \geq \frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$

The approximation is good for $np > 5$ and $np(1 - p) > 5$.

*$X \sim \text{Bin}(n, p)$
mean = np
var = npq*

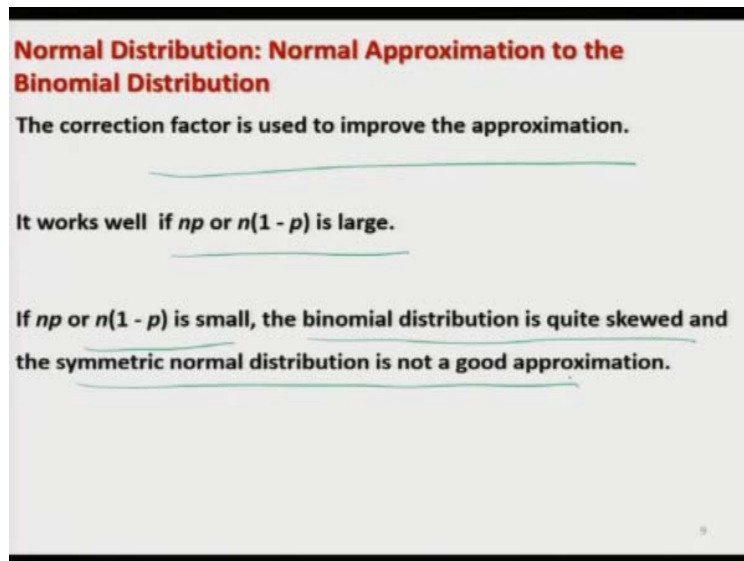
Bin *Normal*

And so, what are we trying to do here? That we have a binomial variable with mean np and variance $np(1 - p)$. so, we have standardized the variable here X by its mean and standard deviation and now the probabilities involving X can be approximated by normal $0, 1$. And once again I will say that this is X here is discrete and the other quantity on the right hand side this is here, Z here is normal but you have to remember one thing you can see here there are two quantities here X and here Z , so one thing you have to see here X is the original random variable and Z here is the function of X .

So, definitely sometime people get confused how one can approximate the probabilities of a discrete random variable by a continuous random variable, it is not like that, you are trying to use different types of probability density function, the X is going to be discrete no issue, but the distribution of its function, that will be continuous, there is no conceptual issue outcome such conceptual problem in this thing.

So, many times in the examination people ask this question that how can you approximate the probabilities of a discrete distribution by a continuous one. So, you have to simply say these are two different distribution as soon as you try to make $(X - \mu)/\sigma$, this is a new random variable and it's a distribution does not remain as binomial but it is changed to normal, that is why you are doing it. So, now we have understood that this approximation is good when n is large related to p and what are we going to do, simply write the probability and then adjust by the factor 1 by 2 correction factor.

(Refer Slide Time: 20:03)



And then this correction factor is used to improve the approximation and this works if np or $n(1 - p)$ is large. If there is small actually then the binomial distribution is quite skewed and the symmetric normal distribution is not skewed because this is symmetric, so the approximation will not be good.

(Refer Slide Time: 20:28)

Normal Distribution: Normal Approximation to the Binomial Distribution

Example: The number of bits are received in a digital communication channel. Assume that the number of bits received in error follows a binomial random variable, and assume that the probability that a bit is received in error is 1×10^{-5} . If 16×10^6 (16 million) bits are transmitted, then the probability that 150 or fewer errors occur is obtained as follows:

So, now let me try to take one example to show you the difference in the probability that you are going to see when you try to use it. Suppose, the number of bits received in a digital communication channel and are assumed that there will be some error in the receiving of this signal and that error is assumed to follow a binomial random variable and assume that the probability that a bit is received an error is 1 into 10^{-5} , and if suppose 16 million bits are transmitted, then the probability that 150 or less or fewer errors occur is to be obtained.

(Refer Slide Time: 21:11)

Normal Distribution: Normal Approximation to the Binomial Distribution

Example:

Let the random variable X denote the number of errors is a binomial random variable.

$$P(X \leq 150) = \sum_{x=0}^{150} \binom{16 \times 10^6}{x} (10^{-5})^x (1 - 10^{-5})^{16 \times 10^6 - x}$$

This probability is difficult to compute. Use the normal distribution to provide an approximation for this probability.

11

So, now how to do it? So first we try to use here the binomial distribution and then we try to show you the normal approximation. So, in this case you want to compute that X is going to follow a binomial distribution and you want to compute the probability that X is less than equal to 150.

So, that we know, this can be commuted as $\sum_{x=0}^{150} \dots$ and this is your here 16 million choose X into the probability $p(1 - p)$ like this one. You can see that here it is not an easy job to compute such a function, so what we try to do? We try to use the normal distribution to approximate this probability.

(Refer Slide Time: 21:52)

Normal Distribution: Normal Approximation to the Binomial Distribution

Example:

$$P(X \leq 150) = P(X \leq 150.5) = P\left(\frac{X-160}{\sqrt{160(1-10^{-5})}} \leq \frac{150.5-160}{\sqrt{160(1-10^{-5})}}\right)$$

Handwritten notes: Continuity correction, mean, sd, N(0,1), Z

$$= P(Z \leq -0.75) = \text{pnorm}(-0.75) = 0.2266274$$

Handwritten note: CDF(-0.75)

Those binomial probabilities that are difficult to compute exactly can be approximated with easy to compute probabilities based on the normal distribution.

Normal Distribution: Normal Approximation to the Binomial Distribution

Example:

Let the random variable X denote the number of errors is a binomial random variable.

$$P(X \leq 150) = \sum_{x=0}^{150} \binom{16 \times 10^6}{x} (10^{-5})^x (1 - 10^{-5})^{16 \times 10^6 - x}$$

Handwritten note: Bin

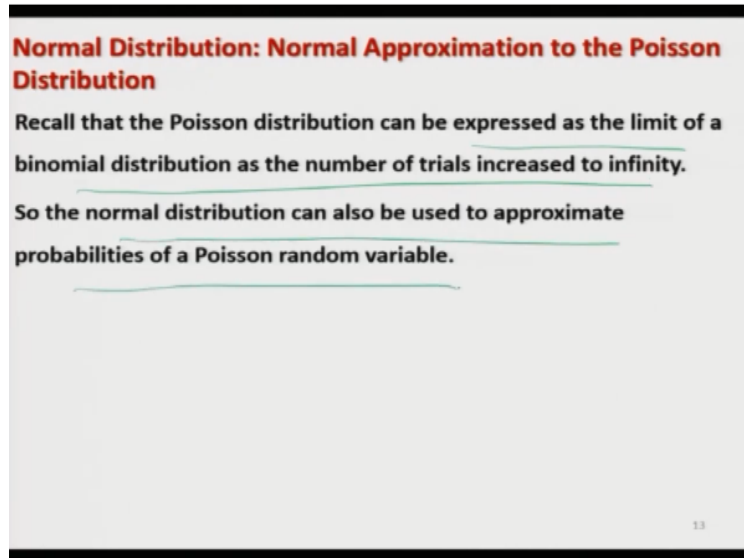
This probability is difficult to compute. Use the normal distribution to provide an approximation for this probability.

That is now very simple, you want to compute probability X less than equal to 150, so that will be now probability X less than or equal to 150.5 because of continuity correction. And then this will be X minus its mean divided by standard deviation on both the sides and now this quantity becomes here Z which is following a normal 0, 1.

So, now this becomes here if you try to simplify this quantity, this comes out to be - 0.75, so you simply have to find out the value of CDF at - 0.75 and this can be obtained by the command `pnorm` in the R software and if you try to find it out this will come out to be

0.2266274. So, you can see here this probability was very difficult to compute but now using this normal approximation you can at least approximate do it well that is the advantage.

(Refer Slide Time: 22:58)



Now, I come to one more topic and I try to do the similar exercise with the Poisson distribution. The Poisson distribution can be expressed as the limit of a binomial distribution as the number of trials are increased to infinity. So, this normal distribution can also be used to approximate the probabilities of a Poisson random variable.

(Refer Slide Time: 23:24)

Normal Distribution: Normal Approximation to the Poisson Distribution

If X is a Poisson random variable with $E(X) = \lambda$ and $Var(X) = \lambda$, then

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \sim N(0,1)$$

is approximately a standard normal random variable.

The same continuity correction used for the binomial distribution can also be applied.

The approximation is good for $\lambda > 5$.

How? We are just going to do the same thing what I have just applied in the case of binomial distribution. So, if I say that X is a Poisson random variable with parameter λ , that means mean is λ and variance is also λ , then $\frac{X - \lambda}{\sqrt{\lambda}}$ that is standard deviation, we will have an approximate standard normal distribution and that is going to be a standard normal random variable. So, this will have an approximate normal 0, 1 distribution. The same concept of continuity correction that we use in the case of binomial distribution that can also be used in this case without any problem and this approximation is good for λ greater than Φ

(Refer Slide Time: 24:15)

Normal Distribution: Normal Approximation to the Poisson Distribution

Example: Assume that the number of radioactive particles passing through a hole follows a Poisson distribution with a mean of 1000.

The probability that 950 or fewer particles enter in the hole is

$$P(X \leq 950) = \sum_{x=0}^{950} \frac{\exp(-1000)1000^x}{x!}$$

This is difficult to compute. So we use normal approximation.

$$P(X \leq 950) = P(X \leq 950.5) \approx P\left(Z \leq \frac{950.5 - 1000}{\sqrt{1000}}\right)$$
$$= P(Z \leq -1.57) = \text{pnorm}(-1.57) = 0.05820756$$

Well, these things have been observed by comparing the two probabilities that under what type of condition either binomial or this Poisson probabilities will give us a good approximation. So, now let us try to take some example and try to do the same thing, you do not have to do anything else, you have to do the same type of algebra what you have learned.

So, assume that the number of radioactive particle passing through a hole follows a Poisson distribution with mean 1000. So, λ is 1000 and now we want to compute the probability that 950 or fever particles enters into the hole, so this probability can be computed by probability X less than equal to 950 which is x equal to 0 to 950 and this is the PMF of Poisson distribution, $\frac{\exp(-\lambda)\lambda^x}{x!}$.

Now, this is difficult to compute, so we try to use here the normal distribution, so we try to write down here probability that X less than 950 and then we try to apply here the continuity correction, because we are trying to approximate the probabilities of a discrete distribution by a continuous distribution.

So, this will become here X less than or equal to 950 plus half which is 950.5, and then you simply try to write down the value of $\frac{x-\lambda}{\sqrt{\lambda}}$ and this value will come out to be

probability Z is less than equal to -1.57 and this can be computed in R without any problem by the command `pnorm(-1.57)` and this value can be obtained as 0.05820756 . You can see here that now you can compute such complicated probabilities very easily using the normal distribution.

So, now we come to an end to this lecture and with this I will also stop with the normal distribution, well there are many more results about the normal distribution but definitely taking all results in this course is difficult because then we will lose some more topic towards the end.

So, now in this chapter on probability density function I will consider only one more probability density function. Well, there is a long list of probability density function, in fact, we have books, one book is only on the probability mass function, one book is only on the probability density function and so on.

So, I will try to restrict somewhere and I believe that once you have understood so many probability mass functions and probability density function if you have to understand any other probability function that is not difficult for you and even implementation in the R software is not difficult for you that much now I am confident. The only thing is this you have to look into the books, that is the job which I cannot do for you but without books what you will do in your life, that is my simple thing.

So, why do not you try to look into books, try to see what are some other different types of probability mass functions, probability density function and try to see under what type of conditions, under what type of experiments they can be used. That is the most important part for you to learn and in data science this is very important because you are trying to deal with very huge data sets, the data is just coming, coming and coming you have no idea.

So, these types of approximation this type of probability density function etc., they are the only ways which can rescue you, they can help you, they can give you different types of guidelines, solution, numbers, interpretation, results etc.. So, you try to look into the books, try to practice and I will see you in the next lecture till then good bye.