**Multivariate Procedures with R**

**Prof. Shalabh**

**Department of Mathematics and Statistics**

**IIT Kanpur**

**Week – 05**

**Lecture – 23**

## Univariate Normal Distribution: Theoretical Properties

Hello friends, welcome to the course Multivariate Procedures with R. You can recall that in the last couple of lectures we had talked about random variables and associated probability, probability density function, probability mass functions in univariate case, bivariate case, multivariate case etcetera. Now, we are going to consider a popular probability density function which is normal distribution. But I think I must clarify one thing that there are many many probability density function and probability mass functions binomial distribution, Poisson distribution, geometric distribution, beta distribution, gamma distribution etcetera. But we are going to consider here only the normal distribution. And the reason for this is that we are going to use the normal distribution and multivariate normal distribution in the forthcoming lectures when we will be doing the multivariate procedures like classification analysis, discriminant analysis etcetera.

And since the objective of this course is not to teach you the probability and probability distribution, so I am not doing it, but they are described in another course which is on the essentials of data science one. So, if you want to have more details on different types of probability distribution, probability mass function, probability theory and other aspects you can look into this in detail courses or you can consult any other book. So, now the question comes here what are we going to do? So, as I discussed in the last lecture about PDF that the probability density functions, they are going to describe the distribution of the probabilities under different type of situation. And this situation depends that how the process is happening.

We have no control over the process. The process is happening in the natural way. We have to simply observe the phenomena and based on that we have to take some observation and based on that we have to see the probability. And different type of secretuations require different types of probability function, probability density function

or probability mass functions. So, among those the normal distribution is one distribution which is very popular.

It is popular because of its nice properties, nice statistical properties and most of our statistical inference procedures they are based on the basic assumption of normality or the normal distribution that our data is drawn from a normal distribution. So, that is why it is important for us to understand about the normal distribution, its important properties and how are we going to implement it in the R software. Well, I am not saying at all that other probability density functions and probability mass function are not important or they are not useful. They are equally useful, but we are going to use the normal distribution. So, that is why I want to prepare this background.

In case if I need any other distribution or if I feel the need of any other probability density function or probability mass function at any point of time I will explain you. So, with this objective let us begin our lecture. So, we are going to consider here the univariate normal distribution and we are essentially going to consider its theoretical properties basically right. I will try to give you example also, but mainly we are concentrating on the theoretical property and yeah, I will not be giving you many proofs, but I will simply be giving you the details of the final result. So, the normal distribution is one of the most important distributions used in statistics and this distribution is also called as Gaussian distribution right.

And the most widely used model for the distribution of a random variable is a normal distribution. You will see that very often in most of the books they always write let X1, X2, … Xn be a random sample from normal distribution, let the random errors follow normal distribution etcetera. So, now the question is what is this normal distribution and how the probability density function of normal distribution looks like right. So, a random variable X is said to follow normal distribution with parameters μ and σ if its probability density function is given by like this.

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}; -\infty < x < \infty.$$

And the range of X here is between - infinity and + infinity, the range of $\mu$ is between - infinity and + infinity and the range of $\sigma^2$ is between 0 and infinity. So, here you can see here this μ and σ they are the parameter right. And these parameters try to explain all the properties of the probability distribution and other properties of the random variable. And our objective in real application is to know the values of the parameters of probability density functions based on the random sample or the data set what we observed during the experimentation. So, we express this normal distribution as that $X \sim N(\mu, \sigma^2)$.

This is a standard notation right. And you know I am not going to give you here the mathematical derivatives or mathematical derivation of different properties of normal distribution, but I will tell you the important properties of the random variable X. So, the mean of X that follows the normal distribution, is $\mu$. So, this is called as expected value of X is $\mu$. And the variance of X is $\sigma^2$.

Well, you can do some mathematical calculation and you can obtain it very easily and if not you can look into any standard book on statistics and you will find the all the details to that how to find out these things. Now, the next question comes what this $\mu$ and $\sigma^2$ are trying to indicate. So, this value this $\mu$ that is expected value of X equal to $\mu$ expected value of X means this is actually the mean. The mean of X is $\mu$ actually this determines the centre of the probability density function. And this quantity $\sigma^2$ which is the variance of random variable X this determines the width of the probability density function right.

Now, there is another variant of normal distribution which is very useful and popular it is standard normal distribution. In case if I try to choose $\mu$ to be 0 and $\sigma^2$ to be 1 then X is said to follow a standard normal distribution. So, if you try to put here this $\mu$ here to be 0 and the $\sigma^2$ here to be 1 then you will get the probability density function of standard normal distribution. So, now you can see here that now both the parameters are known and so the variable random variable lies between - infinity and + infinity. And we write it as like this X follows normal (0, 1) right.

So, now we try to look at the structure of this probability density function that how it looks like. And it is important to understand the structure of this PDF and it will help us in computing different types of quantities. And it will make us understand that whenever we want to do something how we have to adjust different types of issues. For example, if I want to compute a particular type of probability by understanding this curve we can compute them easily. So, that kind of we show you in the next slides right.

So, if you try to see here this line where I am moving my pen this is the density of normal distribution right. So, you can see here that this density curve this density curve is symmetric and bell shaped. What do you mean by bell shaped? You know the bells are like this and yeah it is here like this. So, this structure is similar to the normal curve right.

And if you try to look at this middle value where I am trying to move my pen in red color if you try to look here.

This is the point here at which this curve attains the maximum value right. So, this point is here actually the mean $\mu$. And you can see here that this if you try to divide this curve into these two parts from in on this side and on this side. I will try to remove these lines so that I can make it better and neat presentation right. Then you can see here that this density is symmetric around $\mu$.

Whatsoever be the distribution in this area this is the same in this area right. So, this is what I mean to say. And now if you try to understand here if you look at this point here $\mu$ and then if you try to see this spread that how the values are spread. Then this spread is trying to give you the value of variability that is $\sigma^2$. And so this point here is $\mu$ - sigma and this point here is $\mu$ + sigma which are called here as say inflection points right.

And in case if you try to divide the entire density into these two equal parts then this part is 50 percent and this part on the right-hand side it is also 50 percent of the total area right. So, now let me try to give you here an idea that by changing the value of $\sigma^2$ how the curve will look like. So, as I said that $\sigma^2$ is the variability. So, if I try to make here a line and suppose the mean is here at 0 right. So, then this curve belongs to a value of $\sigma^2$ is equal to 0.5 and what can we write here curve number 1, 2 and here 3. Curve number 2 here is normal 0 1 where the value of $\sigma^2$ is 1 and in the curve number 3 the value of $\sigma^2$ here is 2. So, what you can see here this spread from this central value this spread of curve number 1, this spread of curve number 2, this is 1 1, 2 2 and the spread of here curve number 3 right. You can see here that as the value of sigma square is increasing the spread of the values around the mean that is becoming more and the shape of the curves changes like this that become more flatter right. So, a higher value of σ indicates a flatter density and a lower value of $\sigma^2$ indicates a higher concentration around the mean V is $\mu$ or say 0 for example, I have taken here, but the mean is $\mu$ right ok.

Now, I try to give you here one more situation which we will encounter when we will try to develop some multivariate procedures in the classification problem. You can see here I am drawing here the curves, please try to look at the number which I am trying to give

here. Suppose this is your here curve number 1 and this is same as this curve. So, this is your here normal density curve where sigma square is equal to 1 right. So, both are similar, but the issue is that their means are different.

The mean for the first curve is at 5 and the mean of the second curve is at 15. And even if you try to see here in the first curve the mean remains here at 5, but the variance is changing in the first curve it is $\sigma^2$is equal to 1 and in this curve which I am trying to highlight this $\sigma^2$ is here 4 right. And in case if you try to compare here the curve number 1 and this here curve number 2 which I am trying to write down here. If you try to compare this curve and this curve you can see that they have got the same mean, but different variances. And in case if you try to compare curve number 1 and curve number 3 which is here like this then you can see here, they have got the same variance, but different means.

So, sometime many students ask what is the relationship between mean variance if mean happens to be this what will happen to the variance etcetera etcetera. So, this curve is trying to give different types of answer to such questions ok. Now, another important aspect the cumulative distribution function of X following the $N(\mu, \sigma^2)$. So, by definition you know that the CDF is given by this expression $F(x) = \int_{-\infty}^{x} \phi(t)dt$. Well, I can inform you here that a $\phi$ is usually try to indicate the normal PDF.

In many books you will find this symbol. So, I have just used it here like this. And the CDF of a normal distribution is indicated by $\Phi(x)$ like this. So, that is a standard symbols and notation to denote the CDF of normal distribution. Now, in case if you try to write down the expression for this here $\phi(t)$ then you will see that there is no explicit formula to solve the integral, but it has to be solved numerically using some computational method.

That is why the values of this quantity for different value of X have been tabulated. And in many books the tables for this CDFs are presented right. And it is a very important concept because when we will try to deal with test of hypothesis or many other things you will see that this CDF is going to play an important role anyway. Earlier people used

to do it from the table, but now we are going to compute all the things using the R software. Now, an important result that in case if X is following the $N(\mu, \sigma^2)$.

And if you try to take any two constant a and b which are not equal to 0 and if you try to define a linear transformation like as Y equal to a + b X then Y will also follow a normal distribution with mean here a + b $\mu$ and variance b square $\sigma^2$ right. So, expected value of Y is equal to expected value of a + b X a is a constant. So, expected value of a will remain as a and this will become here b into E(X) which is equal to $\mu$. And the variance will become here like this $b^2$ variance of X which is $b^2\sigma^2$. So, many times you will see that we would like to make a linear transformation from a given variable and in those case cases and situation this result will help us.

Now, another very important result which has several applications. If X is following a normal distribution like $N(\mu, \sigma^2)$, then in case if you make a transformation like $(X - \mu)/\sigma$ that means X - mean divided by standard deviation. Then the distribution of Z will be standard normal that is $N(0,1)$ and this is a very popular transformation which is called as Z transformation. And the advantage here is that the random variable X is dependent on the choice of parameters μ and $\sigma^2$, where a Z is independent of the choice of the parameters and its mean is known to be 0 and its variance is known to be 1. So, this result helps us in finding different probability statement about X in terms of probabilities of Z that we will try to see with the in many application that it is going to be useful.

And many times this concept is also used in normalizing the data. You can recall that when we were considering the turn off phases in graphics, then there was a constraint that the data has to be normalized. So, this is one way by which we can make the normalization. Normalization has an advantage that the new variable becomes independent of the unit. For example, you can see here X may have a unit, but when you are trying to subtract it by μ it will have a unit, but when you are trying to divide it by sigma, then this becomes unit free.

There are several advantages, but anyway this is one among them. Now I would like to show you that by looking at the graphic, how can you understand different results about the computation of the CDF of $N(\mu, \sigma^2)$. So, you can see here if you try to look at this

graphic, this is here 0. So, on the left hand side it will take negative values and on the right hand side it will try to take positive values. So, now it will be our convention that with X I will be including the $N(\mu, \sigma^2)$, the random variable associated with $N(\mu, \sigma^2)$ and with Z this will be the normal distribution with 0 and radius 1.

So, if you try to see the CDF of Z is given by here this definition, but with this you cannot solve. So, we try to compute it by some numerical methods. Now the first result $\Phi(-z) = 1 - \Phi(z)$. If you try to see here what is here phi, if I try to create here this probability density function and if I see here this here is the point here - Z and then this area, this is actually here the phi Z. And now X follows $N(\mu, \sigma^2)$ and Z follows $N(0,1)$ both are normal.

So, both are symmetric also. Symmetric means if I try to take this suppose distance and at the same distance if I try to make here a line on the right hand side then this point will become here + of Z and then this shaded area will also be equal to $\Phi(z)$. So, that is what this result is trying to tell you that in case if you want to find out this $\Phi(-z)$ this will be the area over here and the total area of the curve this is actually here 1. So, this is actually here 1 - this curve and this area here like this. So, that is a very simple rule of integral, but it is very helpful when we try to compute different values of in the R software or from the normal tables. So, $\Phi(-z) + \Phi(z) = 1$ for all z and $\Phi(0) = 1/2$ that is obvious $\Phi(0)$ is the value if I try to highlight here you see here this area.

So, the total area under the curve is 1. So, half of the area will be 1 by 2 the straightforward result. So, now in order to compute this type of probabilities the tables are available for phi z and you have to look at the value of z and the corresponding value of $\Phi(z)$ will be available. For example, if you try to see this I am just giving you here the one sample such computation. So, here are the values of here z and based on different values of here, the values are given here like this. But you can see here if the value of z here is 0 and the probability here is 0.5.

So, that means the values have been computed from - infinity. And yeah one thing I forgot to tell you on this side this will be here - infinity and this side it will be over here + infinity. So, these are the table which are available in most of the book, but anyway we

are not going to use here the table, but we will try to see later on that how we can compute them in the R software. But in order to understand that computation we need to understand this rule.

So, if you try to see here this is here the point 1.5 and if I try to say here $\int_{-\infty}^{1.5} \phi(z)dz$. So, this will be over here $\Phi(1.5)$ and that you can obtain from the table without any problem, right. And similarly as you have seen if you try to take here the center part here as a μ then if you try to go either in the left side or in the right side of the μ this is $\sigma$ will be increasing and so if you try to see here this is the area which is here $\mu - \sigma$ and this is the area on the μ + sigma. Right. So, these limits $\mu - \sigma, \mu + \sigma$ are there and similarly you have here $\mu - 2\sigma, \mu + 2\sigma$ in which this area is covered. And similarly this is here area between $\mu - 3\sigma, \mu + 3\sigma$. So, actually the area covered between $\mu \pm \sigma$ is 68 percent of the total area.

The area covered between $\mu \pm 2\sigma$ it is 95 percent of the area and the area covered by $\mu \pm 3\sigma$ is 99.7 percent of the total area. And if you try to see the way I am telling $\mu \pm \sigma$, $\mu \pm 2\sigma$, $\mu \pm 3\sigma$ actually these limits are used in different places. For example, in the quality control we always talk of 1 sigma limit, 2 sigma limit, 3 sigma limit. So, these limits are actually the 1 sigma, 2 sigma and 3 sigma limits.

For example, this $\mu \pm \sigma$ is the 1 sigma limits, $\mu \pm 2\sigma$ is the 2 sigma limits and similarly $\mu \pm 3\sigma$ is the 3-sigma limit. And they are very useful you will see it later on. And now I am just trying to show it here graphically that ok if you try to see here this is my here mean 0, then this area is here - infinity. So, - infinity to here - A this green area is given by here alpha for example, right.

So, this area here is alpha. And similarly, if you try to see here P(Z >a). So, Z is greater than a. So, a is here like this. So, this area will also be alpha because probability that Z greater than - alpha and probability that Z and this alpha they are going to be the same. And similarly, if you want to find out the area between -a and a.

So, this is your here - A, this is your here + A and this green area is going to show this area. So, if you try to see that if you try to say that the whole area is 1 and suppose this

shaded area which I am trying to shade here on the right-hand side and left-hand side. Suppose this is alpha by 2 alpha by 2, then the area in the middle part will be 1 - alpha, right. So, similarly if you try to see into this term number 1. So, if this area is here alpha, then this area will be 1 - alpha.

And similarly in the term number 2 here, this area is going to be 1 - alpha, right. So, this is for your understanding. And as I shown you earlier you can see here probability that Z greater than equal to A which is here this probability is equal to 1 - 1 means the whole area - this area. So, this is here 1 - probability of Z less than equal to A, right. Similarly if you try to see what is this area probability Z greater than - A.

So, this is here your -a. So, now this is here Z. So, this is here green area is your probability Z > -a. And if you try to see here if this is here a, then the probability that Z < a is here this green area, right. So, you can see that both have the same value, right. Now in many situations what will happen that sometime you are given probability distribution $N(\mu, \sigma^2)$ and you want to compute different values. In order to know those values it is important that you should know the value of μ and $\sigma^2$, but that is pretty difficult.

So, in order to solve this problem one option is that we try to standardize them or we try to normalize them by $(X - \mu)/\sigma$ . Now this becomes here $N(0,1)$ and somehow if I can convert X into $(X - \mu)/\sigma$ then we can solve the problem. For example if I want to compute here P(X <=b), right. So, you can see here I just simply try to write here like this X <=b. I try to subtract μ on both the side and I try to divide on both the side by sigma.

So, this is what I have done here. Now this $(X - \mu)/\sigma$ this is now here Z and this $(b - \mu)/\sigma$ that will be some known value. So, now you can just look at this value from the table of CDF. So, this is how you can compute the probability without knowing the relationship. You simply have to just know the value of μ and sigma and then correspondingly you can just this will become a constant and you can find out its value from the table.

And similarly if I say here probability X greater than equal to A. So, this I can express a 1 - probability of X less than A. Now X less than equal to A it is like I try to subtract on both the side by μ and divide it by here $\sigma$. So, now this will become here say 1 - phi of A

- μ upon sigma. So, these different results will help you in various types of computation and calculation. Similarly if you want to know here the value of probability of X <= b.

The same thing I can do here I can adjust the terms with respect to b and a and b terms with respect to μ and sigma. So, now the central part just becomes here Z and then you can find it by here $\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)$ . And similarly if you want to find out this value here probability that X lying between -a and a. So, this will be simply your here $\Phi(a) - \Phi(-a)$ and $\Phi(-a)$is your here $1 - \Phi(a)$. So, this will come out to be here $2(\Phi(a) - 1)$.

Now we come to an end to this lecture and you can see that we have considered here the univariate normal distribution and we have considered its theoretical properties. These simple different rules will help you later on in the computation of different types of probabilities. So, now it is your turn trying to just do this exercise yourself and try to understand these properties. The next question for us is that how to compute these quantities in the R software that we are going to handle in the next lecture. So, you try to practice it and I will see you in the next lecture till then goodbye.