Essentials of Data Science with R Software - 1
Professor Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur
Lecture No. 26
Moments and Variance

Hello friends welcome to the course Essentials of Data Science with R software 1, in which we are trying to understand the basic concepts of probability theory and statistical inference. So, you can recall that in the last lecture, we had introduced a concept of expectation and we had seen that in case if you try to take a particular value of the function of random variable as a x, then expectation of x is going to indicate the central tendency of the probability function of a random variable x.

So, now I would try to continue on the same lines and we will try to consider different types of functions of random variable which have got a particular form and those forms are going to give us some statistics, some tools which are very important for the data analysis. For example, whenever our data comes data has some hidden information. Now, the problem is this how to express that information, which is understandable to us, which is usable for us, which is useful for us.

So, now in case if you try to see whenever the data comes, data is coming from some process and we assume that that process can be modelled through some probability density function or probability mass function depending on whether the random variable is continuous or discrete. Now, if you can imagine if you can guess that there will be a some curve or some graph of that data that is going to represent the probability function.

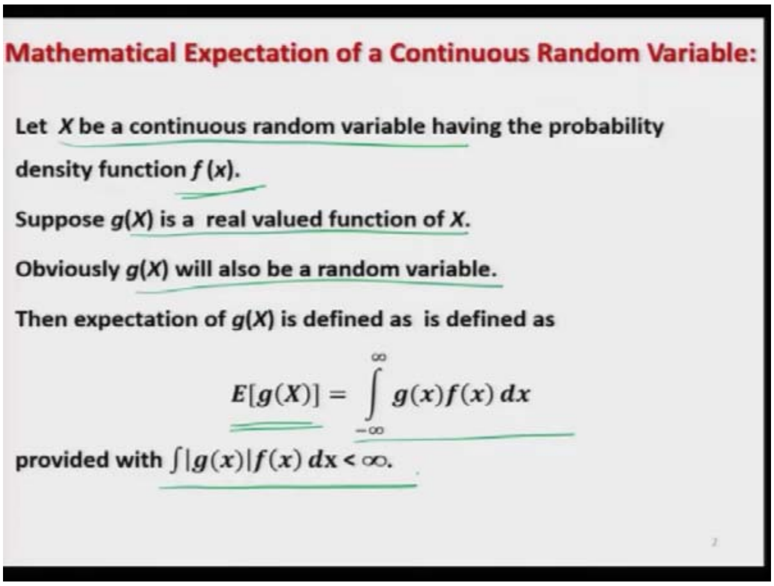
Now, how that probability function looks, whether the curve is symmetric, what is the hump of the curve, what the spread of the curve etc., etc., different types of information are there and in turn, these qualities are going to help us a lot in the data science that whenever we are trying to get the real data without looking into the data, in case if we try to compute these quantities, they are going to reveal the information which is contained inside the data.

We do not need to ask the data, but we have to simply use that tool on that data, that tool will give us some value and we have to understand what is the interpretation and what is the meaning of that value. So, we intend to develop some statistical tools which can provide us the hidden information contained inside that data and for this we are going to introduce here a concept of moments.

These moments are also a special type of expectations of some certain types of function of random variable, but they will give us different type of information to us and in turn when they are trying to be implemented on the real data and when they are computed, they will give us anonymous information which is possibly, not possible for us to observe it by looking at, because they are going to give us the information in a quantitative way.

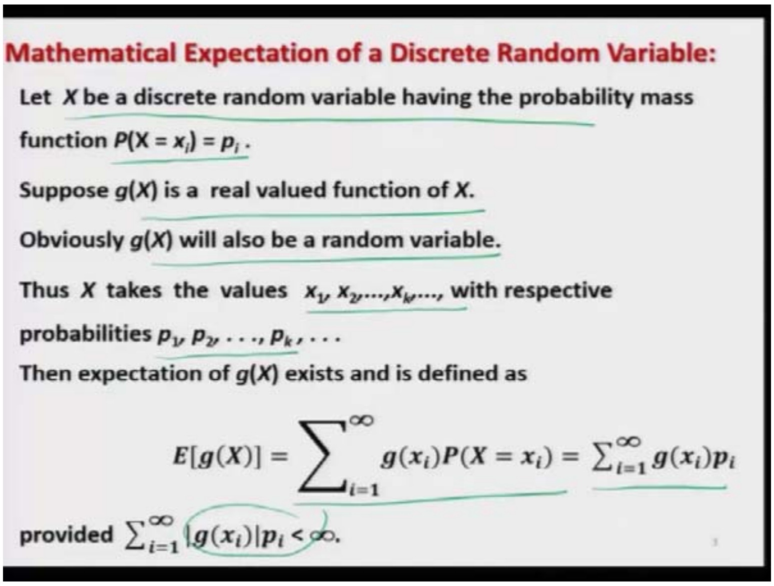
So, now let us begin this lecture and our aim here is in this lecture is to introduce the concept of moments and from there once again I will try to take some special type of functions and we will try to develop these tools, in this lecture and in the forthcoming lectures and all these tools which I am going to develop here, I will try to develop it from the theory point of view and then also I will try to show you that how they can be implemented on the real set of data and how are they going to be interpreted. So, let us just begin our lecture.

(Refer Slide Time: 04:22)



So now, just for a quick review, that how we had to have defined the expectation of a function of a random variable I have these two slides. So, first we try to consider the random variable to be continuous, random variable which has got a probability density function, say f(x) and suppose g(X) is a real valued function of the random variable. So g(X) will also be a random variable and the $E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$ and obviously provided this integral exists for which we have a conditioned that $\int |g(x)| f(x)\,dx < \infty$.

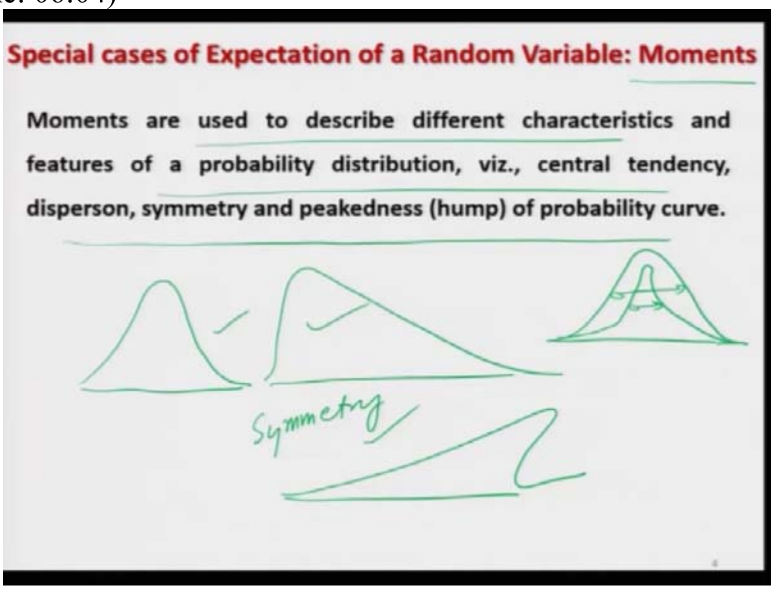
(Refer Slide Time: 05:05)



So, similar definition also holds for the discrete random variable. So, let a capital X be a discrete random variable with the probability mass function, say probability of probability of X equal to x_i is equal to p_i and suppose g(X) is a real valued function of the random variable X. So, this is also a random variable, which and in this case, the random variable X takes the value x_1, x_2, …, x_k and the corresponding probabilities or even p_1, p_2, p_k respectively.

Then, in this case the expected value offer g(X) exists and it is defined as expected value or g(X) equal to $E[g(X)] = \sum_{i=1}^{\infty} g(x_i)P(X = x_i) = \sum_{i=1}^{\infty} g(x_i)p_i$, for which we have a condition that the provided $\sum_{i=1}^{\infty} |g(x_i)|p_i < \infty$.

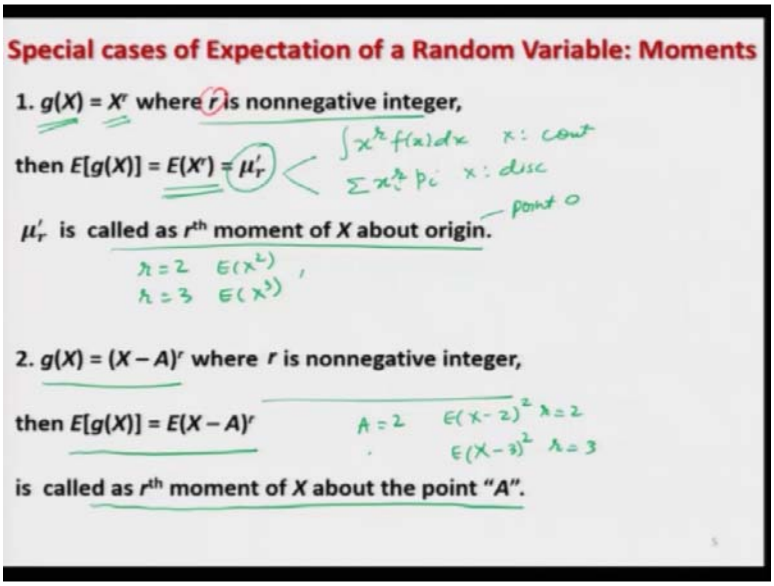
(Refer Slide Time: 06:04)



So, now we have got these definitions of expectation of a random variable now, we try to come under the special cases where we are going to choose a special forms for this g(X) and we try to define the concept of moments. So, these moments are used to describe different characteristics and features of a probability distribution like a central tendency, dispersion, symmetry, peakedness etc.

For example, if you try to say the data is like this one or the data is like this type of curve or data is having this type of curve something like this. So, this types of different characteristics or even if you try to take asymmetric curve over here, whether this is also symmetric and this is also symmetric. So, now this part is going to indicate the variation that I am going to introduce and this, this and this they are trying to indicate the symmetry of the curve. Whether the curve is symmetric or not. So, these types of characteristic of the probability function probability distribution they are going to be revealed through the use of moments.

(Refer Slide Time: 07:20)



So, let us try to define an IP inform you that we have different types of moment or like as a raw moment, central moment, factorial moment, absolute moment etc. But we are going to understand here, only some of them which are more useful for us from the data science point of view. So, now we will get me take the first special case of the function of random variable g(X).

So, let $g(X) = X^r$, where r is some non-negative integer. Now, in this case, the expected value of g(X) will become here expected value of X^r and this is indicated by μ'_r that is a standard symbol in statistics and this μ'_r is called as the rth moment of X about origin, origin means the point 0 that we know. So, this is a particular quantity.

Now, if you try to see, if I asked you how are you going to compute it, that is simply going to be $E(X^r)$. So, that is going to be integral X to the power of r, X dx in case if X is continuous and it is going to be replaced by summation X to the power here r, say here i pi and in case if X is discrete.

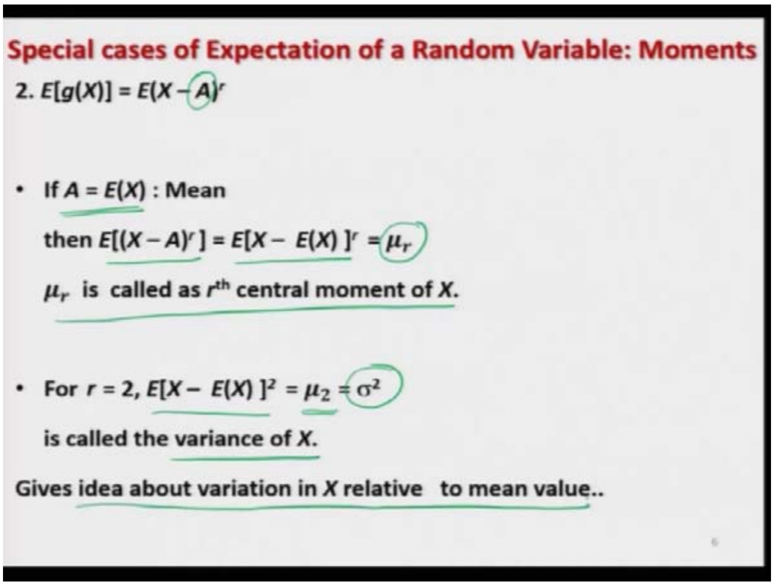
So, you can see here computing moment is very simple, the only thing is you have to understand the meaning of this notation that expected value of X^r and then you have to simply see whether it is continuous or discrete and then if you are using a continuous random variable, then you have to use the integral sign and in case if you are trying to use the discrete random variable, you have to use the summation.

And based on that you can define the quantity and this is called as a rth moment of the random variable X about origin. So, if you try to take here suppose r equal to 2 you will get your $E(X^2)$ if you try to take your r equal to 3, you will get your expected value of X cube and so on and if you try to see these quantities are not difficult, either that is discrete or continuous random variable, you can easily compute them, that will require some mathematics, some algebra, but I am sure that you can solve it.

And if not, then you can use some numerical techniques if you are dealing with a complicated functions and you can approximate it very well. Now, the second special function of a random variable I am considering here is $g(X) = [X - A]^r$, where r is a non-negative integer and A some point, some value some constant. So, now in this case the expected value of g(X) is going to be now, $E[X - A]^r$ and this is called as rth moment of X about the point A.

Suppose if I take A equal here 2, then $E[X - 2]^r$, now in case if you try to take your r is equal to here 2, then it becomes your $E[X - A]^2$ and similarly if you try to take here r equal to 3, then this quantity become $E[X - A]^3$ and now you can compute it without any problem. So, this quantity is called as rth moment of X about the point A.

(Refer Slide Time: 11:03)

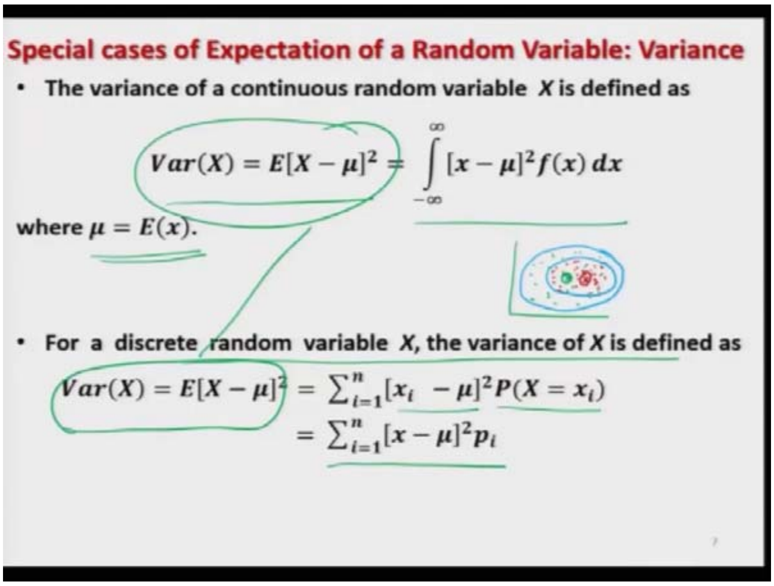


And now, in case if you try to choose here especial values for this A, then we get here different types of quantities, which are very useful in data sciences in and helps us a lot in describing the behaviour of the data. For example, if I try to take here A = E(X), that is the mean of X, then $E[X - A]^r$, this becomes the $E[X - A]^r$ and this is indicated by μ_r and this μ rth central moment of X.

Central moment means you are trying to measure the moments around the mean that is all. Means earlier, you have defined it for the about any point A, now you are trying to define data about the mean of the distribution and this has got some nice function which are very useful in finding out the information hidden inside the data.

For example, if you try to take here say r is equal to 2. Now, in case if I try to choose here r equal to 2, then I get here $E[X - A]^2$, which is equal to here, the second central moment of X and that is measured around the mean expected value of X and this is indicated by sigma square, that is a standard symbol in statistics and this particular quantity is called as the variance of X. This various gives us an idea about the variation in the X relative to the mean value. So, that is a very important quantity that I will try to show you when we try to work on the real data.

(Refer Slide Time: 13:02)



And now, in case if you try to see let us try to first concentrate on this quantity variance and let us try to understand how it goings to be computed for a continuous and discrete random variable. So, in case if X is a continuous random variable, then variance of X is defined at $E[X - \mu]^r$, where mu is the expected value of X that is the mean and so that is going to be simply $\int_{-\infty}^{\infty} [x-\mu]^2 f(x)\,dx$.

And if X is a discrete random variable, no issue that is the basic definition remains the same you can see here this thing and here this thing, they are the same, but the only thing is this now, it has to be computed using the summation, so that will become here

$\sum_{i=1}^{n}[x_i \; - \mu]^2 P(X = x_i)$, which is the probability mass function and this will become a simply $\sum_{i=1}^{n}[x - \mu]^2 p_i$. So, this is how you can compute the variance of X and if you try to see here, what is the meaning of this thing I can just show you here by giving you here a brief idea, for example if you have two data sets, one is being plotted like this one. So, you can see here this is concentrated somewhere here.

Now, there comes another data set and suppose this is like this. So, this value is concentrated somewhere around here, but you can see this circle, the circle in which all these values are contained, for the red the circle is like this and for the green the circle is like this. So, you can see here, that the values which are indicated by the green colour or which are plotted by the green dots, they are more scattered.

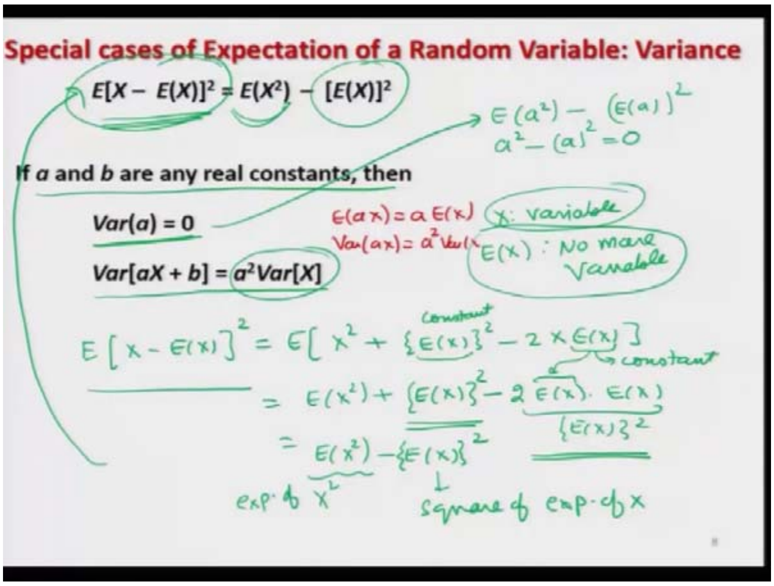
And the values which are plotted by red coloured dots, their scatteredeness is just smaller than the green colour. So, this scatteredness property is measured by this quantity variance and that is a very important quantity in statistics, that we always try to find out that decision rule in which you can have the minimum variability and if you try to understand the basic concept of the variance, I can take a two very simple sentences and they will give you an idea of what is actually the property of variance.

Suppose you have two friends and whenever you call them, the friend number 1, that friend comes just 5 minutes earlier or 5 minutes later. So, some time that can be 3 minutes, sometime can be 2 minutes, sometime can be 1 minute, sometime can be 5 minutes. So, earlier or later both and there is another friend who comes sometimes half an hour earlier or say half an hour late.

So, that friend can come say 29 minutes earlier or 29 minutes later or 5 minutes earlier or 5 minutes later and so on. So, if you try to see the nature of the data of the time of arrival for both these friends, what do you see, the friend number 1, that fellow has a variation only of 5 minutes either earlier or later, but for the friend number 2, the variation goes from 30 minutes from say earlier or later? So, do not you think that you will see simply that the variation in the time of different number 2 is much bigger than the friend number 1? So, this is actually the idea of the variation.

So, sometimes you always use a statement like suppose in the morning it is extremely cold, then in the afternoon, it becomes hot and suddenly in the late afternoon it becomes windy and then in the night it starts raining. So, you say the weather is varying a lot. So, try to see you know about this concept of variation you have used this term variance or say variation and you are using it in day to day life. The only thing is this possible, you never thought about that how to measure this quantity. So, now we are trying to discuss those things. So, let us come back to our slides.

(Refer Slide Time: 17:51)



So, now some properties of this variance because we are going to use it. So, one thing I can show you, I am just trying to take this a small algebra, so that you can be confident that whatever I am writing here, they are based on some stronger mathematical foundation they are, but they are very simple, they are not difficult.

So, if you try to see if I try to solve this quantity $E[X - E(X)]^2$, so using very simple algebra, I can write down here as say, $E(X^2) - [E(X)]^2$. So, you can now you have learnt the rules of expectation that this will become here $E(X^2)$.

And then as soon as you try to write down expected value of X, this does not remain as a variable. Well, X is a variable, no issues. But if you try to take any moment of the random variable that can be expected value of X or X square or any function, then this is no more variable. You have to remember one thing because you can see here expected value of X is the mean of the entire population of the probability function of the random variable X.

So, the population mean is always going to be constant. Well, that is a different thing that we do not know the value of the population mean. But it does not mean that it is a random variable that is known to us and then means how to handle this phenomena that when you are trying to take different types of sample which are going to give you the different values of the mean, how to interpret it, how to handle it that we are going to discuss in the entire course.

But this is the main concept where many times the student get confused that why the expected value of X is going to be a fixed value and whereas capital X is a random variable. So, now you can see here that this is here a constant value. So, now you know expected value of a constant is a constant. So, this will become here expected value of X whole square and minus, now you know that this is expected value of X. So, once again this is constant.

So, this I can write down here to expected value of X into expected value of X, this constant part is coming here and then you are trying to operate the expectation operator on the random variable X. So, now you can see here this becomes here expected value of X whole square and now this is your expected value of X square minus twice of expected value of X square and this will give you here expected value of X square minus expected value of X whole square.

So, what you have to be very careful that here the first term is the expectation of X square and because it is here the square of expectation of X, sometimes the student get confused with the two symbols and notation but you can see here that these are two different quantities and this is precisely what I have written here.

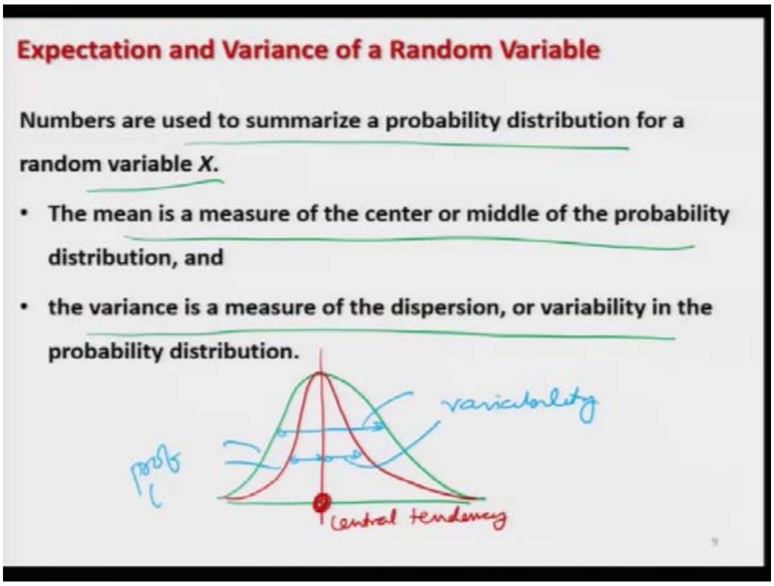
So, you can see here that many times when we want to find out the variance, then instead of finding out expected value of X minus expected value of X whole square this expression, we try to compute this expression $E(X^2)$ and then we try to compute the expected value of X and then we try to find out the value of the variance.

And in case if you assume that small a and small b are any real constants, then variance of a constant that is 0, you can see here that in case if you friend is coming always in time no variation mean exactly at 5pm for example 5 dot 00. So, now, there is no variation all the values are constant. So, variance of a constant is 0 and if you try to verify this quantity, you can simply use this expression over here then you can write down here this is going to be expected value of a squared minus expected value of a whole square.

So, expected value of a square is a square and expected value of a is going to be a and this is whole square which is going to be here 0. So now, I am just trying to convince you that whatever expression I am writing here, they are not difficult and you can find them very easily and if you are facing any problem or if you want to have the proper derivation possibly I will suggest you that you can have a look into any of the book.

And similarly, under property of variances that $Var[aX + b]$ that is the linear function then this is going to be $a^2 Var[X]$. So, you can see here when you are trying to find the expectation of say here aX, then this is the expected value of X is independent of a, because a is a constant. So, this I can write down here a into expected value of X, but in case if you try to find out the variance of aX and so this is going to be a square time variance of X, this is the simple rule what you have to understand and you will see that we will be using this type of manipulation very oftenly.

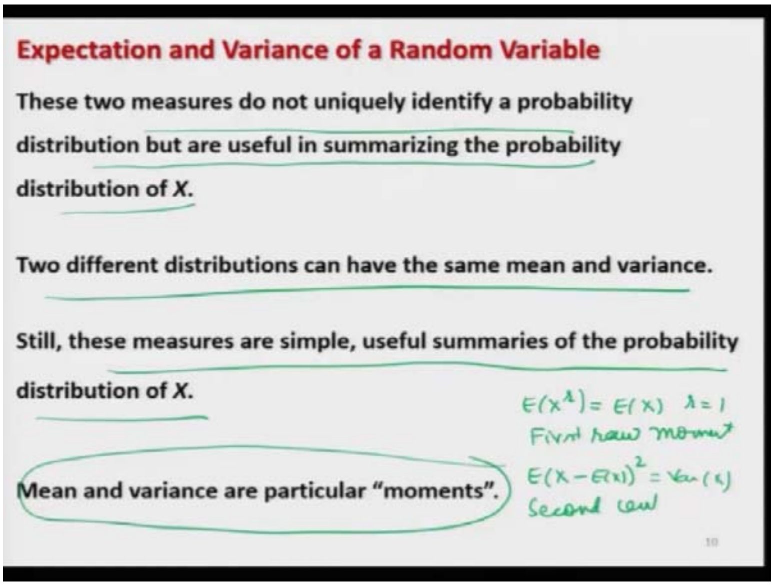
(Refer Slide Time: 23:34)



So, now you can see here that we have got some numbers as data and they are trying to summarize the probability distribution for a random variable X. For that in case if you try to find out the mean, then the mean is a measure of the centre or middle part of the probability distribution and variance is a measure of the dispersion or variability in the probability distribution. For example, if you try to make here two curves like this one and say here, one more curve here like this one, then possibly you can see here that the, this is here the central tendency that most of the data is it is centred around this point.

But if you try to see here, this part in the green and red curves, they are different. So, this part is going to indicate the variability. So, this is how and yeah I mean. So, obviously these two are some probability curves. So, this is how we try to interpret the statistical properties from this mean and variance.

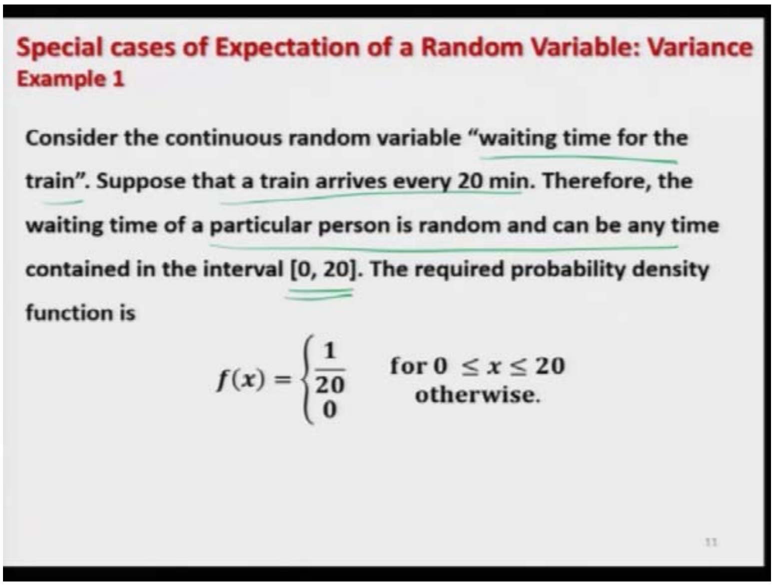
(Refer Slide Time: 24:56)



But these two measures, do not to uniquely identify a probability distribution, but they are useful in summarizing the probability distribution of X. So, what I am trying to say here that if you have got a two sets of values of mean and variance, then by looking at the values, you cannot determine that what type of probability functions has given those value or you are cannot identify that the corresponding random variables are following which of the probability function.

Two different distributions can have the same mean and a variance also and still these measures are very simple useful and summaries of the probability distribution of X are provided by these two function. So, and you can see here that mean and variance both they are particular type of moments. Because if you try to take here expected value of X to the power of r and if you try to take a expected value of X, where when r is equal to 1, this is a first raw moment and in case if you try to find out the first central moment that is expected value X minus expected value of X whole square this is your here variance of X. So, this is your here second central moment.

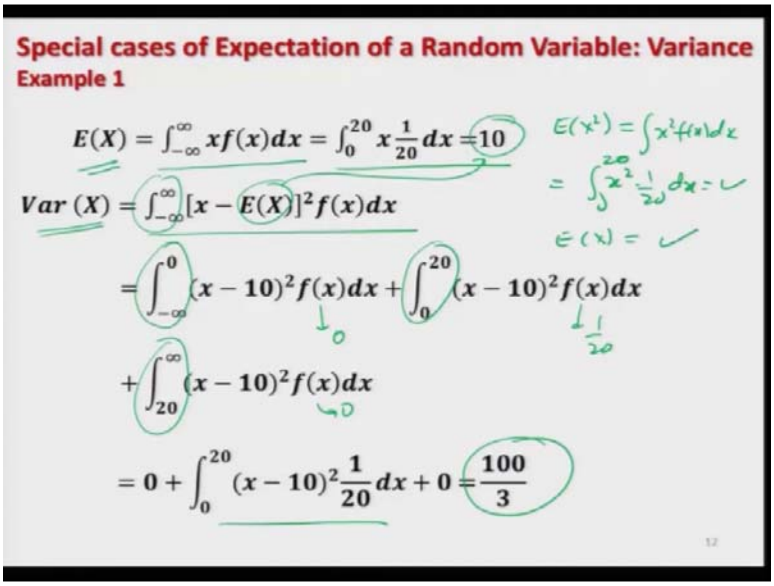
(Refer Slide Time: 26:31)



Now, look at me try to first take some example here. So, that I can show you that how are you going to compute these quantities with the probability functions and then in the next lecture also I will try to show you that how you can extend these quantities over a real dataset also.

So, let me try to take here the first example, this is the same example that we had considered earlier that we have a continuous random variable which is a describing the waiting time for the train and suppose the train arrives after every 20 minutes, then the waiting time of a particular person will also be random and that is going to be any value between 0 and 20 in the closed interval 0 to 20. So, in this case, we had found earlier the probability density function as f(x) equal to 1 upon 20 if x the line between 0 and 20 and 0 otherwise.

(Refer Slide Time: 27:33)

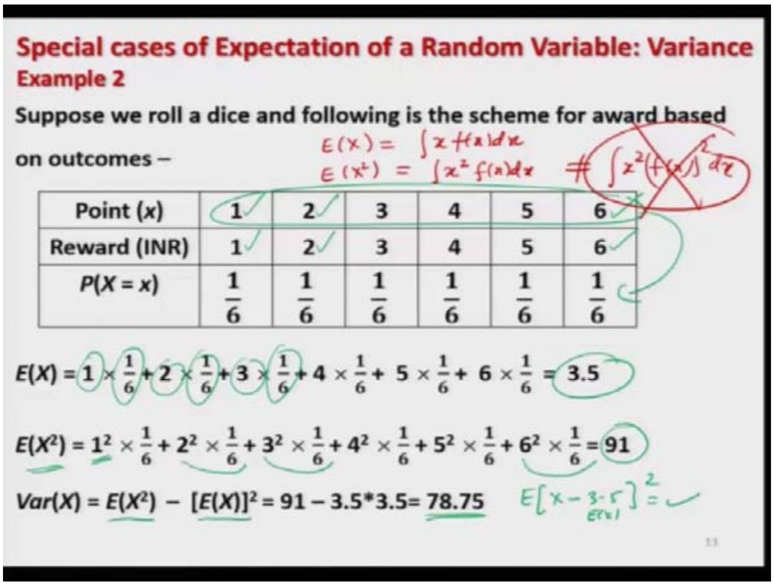


Now, we try to find out its mean and variance. So, you see, it is very simple that $\int_{-\infty}^{\infty} xf(x)dx$ and which will become here $\int_0^{20} x\frac{1}{20}dx$ and in case if you try to solve it, this will come out to be here 10. Now, if you want to find out the variance of X, then what you can do you simply use the basic fundamental definition that this is going to be $\int_{-\infty}^{\infty} [x - E(X)]^2 f(x)dx$.

Now, expected value of X you already have found to be 10. So, I can replace here by 10 and then this limit minus infinity to plus infinity can be divided in 3 parts depending on the range of the probability density function. So, probability density function is defined in the range of 0 to 20. So, we try to divide it into 3 parts minus infinity to 0, 0 to 20 and 20 to infinity and we try to compute the same function. So, obviously, in the first and third case f(x) is going to take the value 0. So, there is no doubt and in the second case f(x) is going to take the value 1 upon 20.

So, if you simply try to substitute and solve it, you get here this integral and this value comes out to be 100 upon 3, the second option is that you can compute here expected value of x square that is $\int_{-\infty}^{\infty} x^2 f(x)dx$ that will be here basically here $\int_0^{20} x^2 \frac{1}{20}dx$, you can find out the value of this and you already have found the value of expected value of X from there you can compute the variance as $E(X^2) - [E(X)]^2$. That will also give you the same answer there should not be any problem.

(Refer Slide Time: 29:23)



Now, I try to take means another case related to the discrete random variable and try to show you that how these things can be can be obtained. So, once again, I will try to take the same example that we had considered earlier. Well, I am trying to do continue the same example so that you can actually understand that if you got such a data, how you are trying to retrieve different types of information.

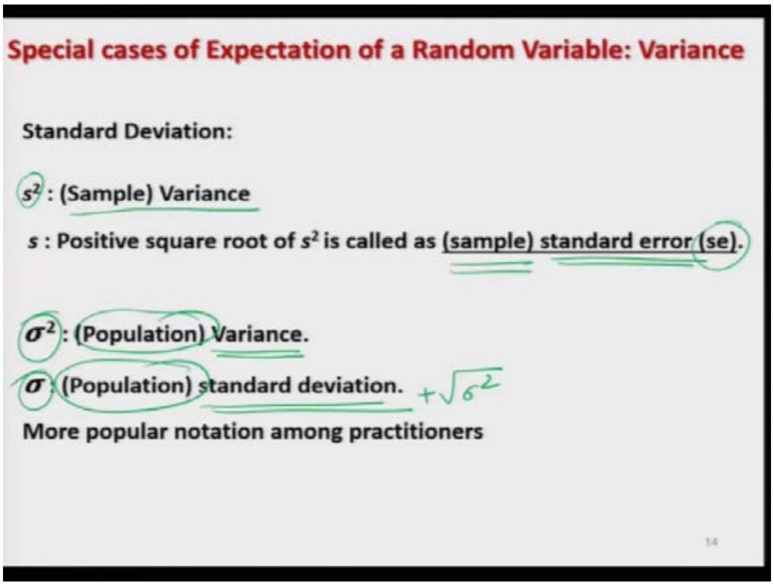
So, in this example, we roll a die and we get the points here 1, 2, 3, 4, 5, 6 and the probability of observing any the point is 1 upon 6 and the reward associated with these numbers is that in case if somebody gets a number 1, the person gets rupees 1, if somebody gets said number 2 the person gets rupees 2 and similarly, if somebody gets a number 6, again the person wins rupees 6.

So, now, in this case, we already had found the expected value of X that means, the means the reward and the probability. So, the reward here is 1 with probability 1/6, reward is 2 with probability 1/6 and similarly, reward is 3 with probability 1/6 and so, on, if you try to find out the expected value of X comes out to be 3.5 and similarly, if you want to find out expected value of X square, so that will simply become here 1 squared that is the value of the reward which is happening with the probability 1/6 plus 2 square into probability for 1/6 plus 3 square into probability 1/6 and this will come to continue up to 6 square into probability 1/6.

And in case if you try to simplify here, this will come out to be here 91 and if you try to find out the variance of X that will become here simply expected value of X square minus expected value of X whole square. So, you have obtained these two value and this value will come out to be 78.75. Well, in case you also try to obtain it directly from expected value of X minus 3.5 whole square which is the value of your expected value of X and if you try to solve it, you will get the same value, no doubt about it.

One thing which I would like to warn means, all the students that when we are trying to find out here expected value of X, then this becomes your x fx dx and when we are trying to find out here expected value of X, this becomes x square into fx dx and which is not equal to something like x square f of x whole square dx remember one thing do not make this type of mistake. Sometimes people try to square the probability function also, no you simply have to square only the value of the random variable. So, I hope you will not make any such mistakes.

(Refer Slide Time: 32:21)



Now, I tried to come on the implementation part of this variance. So, there are two quantities we have what we try to use in data science or in real world application that now we have understood what is the variance that we are trying to indicate by the quantity sigma square, now in case if you try to take the positive square root of sigma square then this is called as standard deviation.

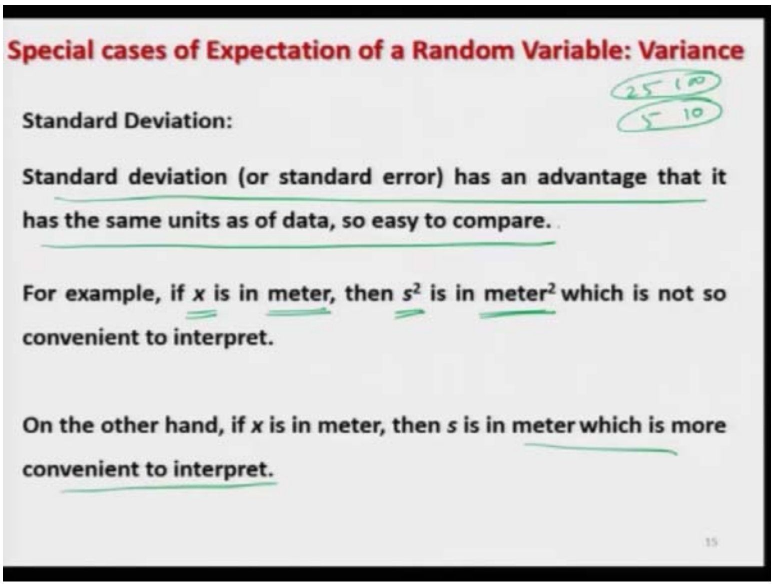
So, this is the positive square root of sigma square and if you try to see this sigma and sigma square they are related to the population, but now you are trying to get here are simple true. Now, sample will give you different values how I can convince you suppose in your class there are 50 students and you want to find out the average height of the students. Suppose, you try to take a sample of size 5 that means, you can choose any 5 students out of the 50 students.

Now, if you try to choose any particular sample and try to find out the arithmetic mean of the heights, then do you think that if you try to take a different type samples of the student or different choices of 5 students out of 50 students, are you going to get exactly the same value? No, there will be some variation.

So, when we are trying to compute this variance on the basis of the given set of data, then this is called as sample variance and we try to indicate by here something like a s square and when we try to take the positive a square root of a squared then this is called as standard error, which is actually based on that sample and the short form of standard error is se, so in practice, you will see that people always try to specify the mean and standard error for both.

So, one thing what you have to keep in mind that a sigma square is the quantity or standard deviation is the quantity which are based on the population value, whereas a sample variance or the standard error these are these are the similar quantities, but they are based on the observation, so you can always estimate them on the basis of the given set of data.

(Refer Slide Time: 34:45)

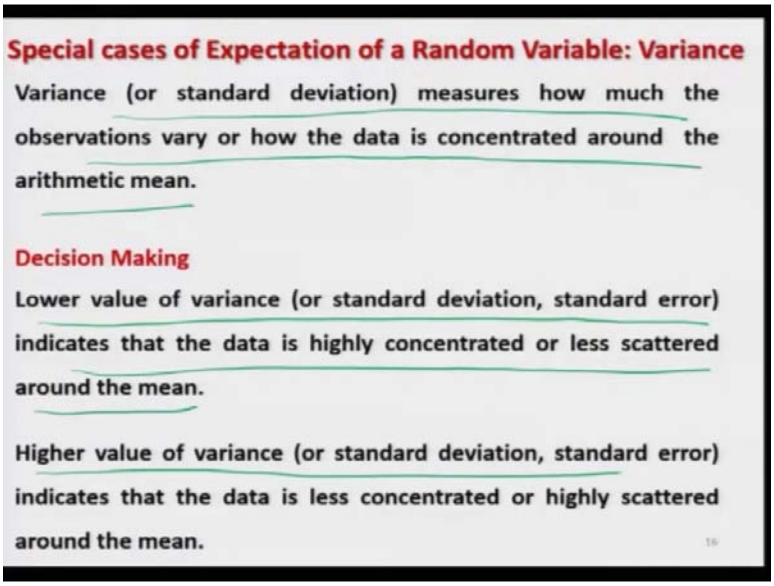


Now, one question comes here that why do you need to say this standard error and sample variance. So, you know the reason is very simple from the, on the application point of view when you are trying to measure some random variable or you are trying to get the data on some random variable that will always have some unit, for example in case if your random variable x here is height, then the values are obtained as a small x and they are going to be for example, in meter.

So, now, they say s square will be meter square. Now, in case if you have to compare two values say here say 25 and 100 so, what do you think what is a better quantity or which is more convenient to understand whether 25 and 100 or the square root 5 and 10 if you try to compare 25 by 100, it looks like the difference between the two values is very high 75 units, but if you try to compare 5 and 10 it appears that the difference here is only 5 units.

So, when you are trying to measure the random variable in certain units, so many applied workers, they always preferred to measure the variance also in the same units. So, the standard deviation or standard error has an advantage that it has the same unit as of the data. So, it is easy to compare and it is definitely easy to interpret compared to the square of the values. So, that is why people try to use this standard error more often in practice than the sample variance. So, this should not be any problem.

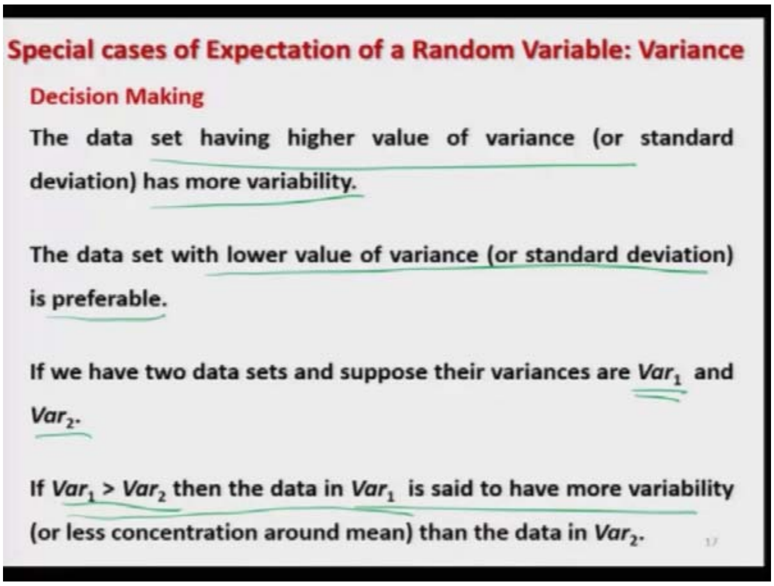
(Refer Slide Time: 36:35)



Now, the thing is to think is this, how are you going to make a decision on the basis of given set of data, this variance or standard deviation or sample variance or standard error, they are going to measure how much the observation vary or how the data is concentrated around the arithmetic mean.

So, in this case, if you want to compare two data sets, you simply try to compare their variances you try to compute their variances and try to see the whose value is lower. So, lower value of variance or standard deviation or standard error, indicates that the data is highly concentrated or less scattered around the arithmetic mean or that mean value and higher value of the variance or standard deviation or standard error is indicate that the data is less concentrated or highly scattered around the mean value.

(Refer Slide Time: 37:36)



So, what will happen to that means any data who has got a higher value of variance has more variability and the data with the lower value of the areas or standard deviation has lower variability. So, any value with the lower variability is always preferable. So, in case if you have two data sets and suppose their variances are found as variance 1 and variance 2. So, in case if variance 1 is greater than various 2, then the data in variance 1 is said to have more variability or less concentration around mean than the data in variance 2 and we will prefer the data set which has got the smaller variability.

How because, if you try to see, you will always select to depend on a friend who is more dependable. How do you define the dependability, in which the variation is less? Suppose you do not go to the class and they and there are two friends, one friend will always share the class notes with you and that friend will always be willing to help you to compensate the teaching in the miss class and there is another friend who will sometimes help otherwise, if he or she will say no and so on.

So, on whom you can depend more in case if you are not going to the class on any occasion due to any reason, definitely you are going to depend upon more on that friend which gives you the nose or which explains you what happened in the class today all the time, that is more dependable because it has less variance.

So, whenever we are dealing with the statistics, we always try to look for the variability and we always try to prefer a rule or a data set which has got the smaller variance and variability is the backbone of all statistical inferences. Whenever we want to make a decision, we always have to interpret in some real data language and there it is very important to specify that what is the variability associated with my decision that is based on certain statistical tool.

So, this variability is one of the basic fundamental thing in the data science, whenever you are trying to work with a huge data set, whenever you are trying to apply some statistical tool, finally you will be interested that whatever rule I am going to get whatever outcome I am going to get, whether this has less variability or more variability, if the variability is more you will try to change your statistical tool also and finally you would like to have a decision which has got less variance.

So, this was a lecture based on that theoretical concept and in the next lecture, I will try to show you something more and I will try to continue with the same topic that is a special cases of expectation. So, you try to practice it, try to have a look in the books and try to read it and I will see you in the next lecture till then, goodbye.