Essentials of Data Science with R Software - 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture No. 25
Expectation of Variables

Hello, friends. Welcome to the course Essentials of Data Science with R Software 1, in which we are trying to understand the basic concepts related to the probability theory and statistical inference. Now, I asked you one simple question, what do you expect from this course? Many expectations. Let me try to make this question very simple. Suppose, we play a game. And I toss a coin and if it comes head, then you will give me 100 rupees; and in case if there is a tail, then I am going to give you 100 rupees. Would you like to play? Hopefully you will say, yes.

Now, I ask you, that suppose in case it comes head, then I will give you only 5 rupees and in case if it come tails, then you are going to give me 100 rupees. Will you play? Are you going to play with me? Yes, is in the first case your answer will be yes; and the second case your unsaid will be, no. My question is how do you concluded that what is yes and what is no and why is yes and why is no in these two cases?

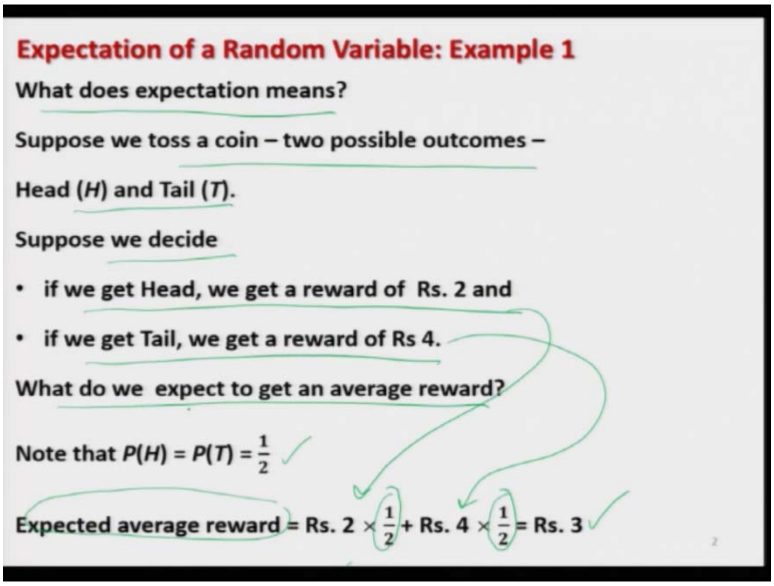
Take a pause and then think about it, but I will continue with my lecture. What you have done that you just calculated what type of profit or loss to do you expect by playing the game. And in the first case, you said well, the profit or loss they have the equal impact, but in the second case, the impacts are different. And possibly you computed that in the first case possibly if you try to play this game for a couple of times, possibly you may win and if not win you may not lose.

But in the second case you just computed within your mind very fast that if you try to play with this game, then you expect a loss. And on the other hand, if I say that suppose if there is a head, then I will give you 1000 rupees; and if there is a tail you will give me only 2 rupees. Then would you like to play? You will certainly say, yes; because you expect that you are surely going to win or the chances that you win are very, very high.

But believe me, I am not going to play with you this type of game, but I want to ask you how you computed and how you came to this decision, that what do you expect? I am sure that you cannot think, what mathematical calculations you have done inside your mind which have given you this expected value. And if you try to see the probability of X getting a head or a tail, that is the same 0.5. The only, the value of the reward is changing.

So, do not you think that you are trying to play with the probability and the value of the reward? And you are trying to do some mathematical manipulations inside your mind because you are very good in computing the profit and loss. This is the topic which I am going to take up today. I will try to show you, what is called expectation of a random variable. And that is possibly going to answer not only this question, but many other questions you will see. So, let us begin our lecture and try to see.

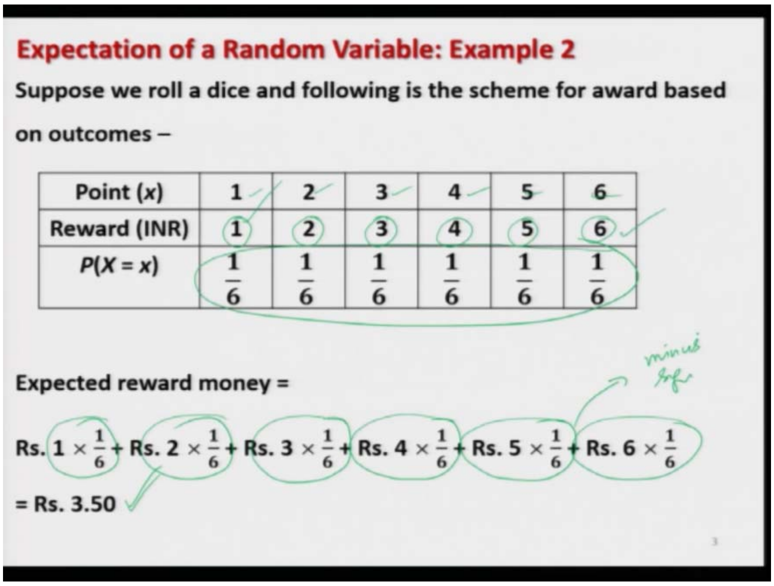
(Refer Slide Time: 04:20)



So now, the first question comes, what does expectation means? So, let me try to take a very simple example as I did, suppose we toss a coin there are two possibilities getting a head getting a tail. And suppose we take a decision that if we get a head, if we get a reward of rupees 2, and if we get a tail, we get a reward of rupees 4. What do we expect to get on an average reward? Possibly, you know the answer but now my question is, how are you getting that answer?

So, we know that the probability of head and tail they are 1/2 each. What do you have done? Now I can show you. You said in case if you are getting a head whose probability is 1/2, you are getting a reward of rupees 2. So, you multiply 2 into 1/2. And in case if you are getting a tail whose probability is 1/2, then you are getting a reward of 4. In both the cases, there are actually rewards. So, there is no loss, there is only a profit. So, but still, you will try to see that on an average, how much profit are you going to get. And if you try to solve this, this will come out to be rupees 3. So, this is the expected average reward.

(Refer Slide Time: 05:40)



And if you try to see many times you go to in some fares, where you have to play a game and you always try to make such calculation, but how are you going to do that you did not knew. So, now, to convince you, let me try to take here one more simple example. Suppose, we roll a dice and following points are observed 1, 2, 3, 4, 5 and 6 and the probability of observing these points is for 1/6 each.

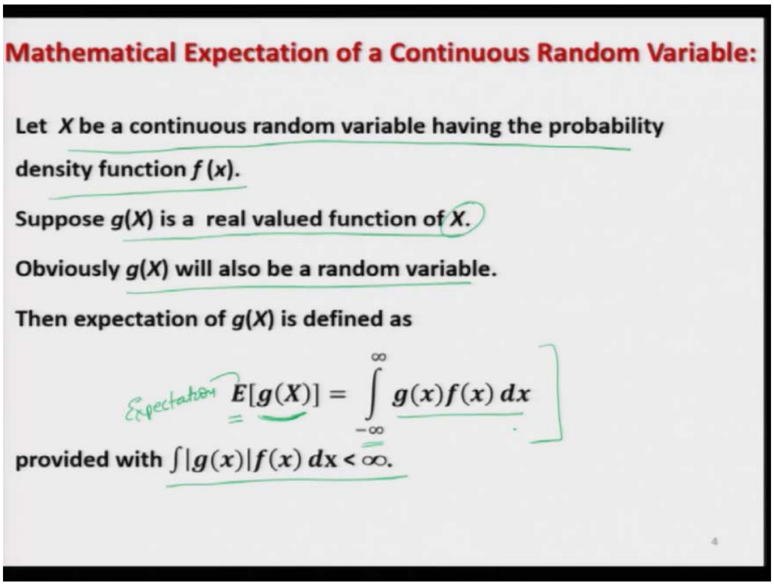
And now, just for the sake of understanding, we try to assume that if we get to the point number 1, we get 1 rupee as a reward. And in case if you get 2 points, then the reward is 2 rupees. 3 points, 3 rupees; 4 points, 4 rupees; 5 points, 5 rupees and 6 points, 6 rupees. These are the rewards associated with the total number of points which are going to be observed on the upper face of the dice.

So, now, what do you expect? Means, if you try to play the game for some time, for a couple of throw of dice, then what do we expect at the end what are you going to get? So, the expected reward money can be found as like this, rupee 1 into probability 1/6 plus rupee 2 into probability 1/6 rupees 3 into probability 1/6 rupees 4 into probability 1/6 plus rupees 5 into probability 1/6 plus rupees 6 into probability 1/6. And if you try to solve it, this will come out to be rupees 3.50.

So, this means that in case if you try to play this game for some number of throws, then on an average we expect that you will get 3.5 rupees as a reward. But you will always see that the limits of rewards are between 1 rupees and 6 rupees. And if there is a loss possibly you can say that it means that this plus sign are going to be changed by minus sign, means if I say that if you get 1, 2 and 3 then you get a profit of 2 rupees, and if you get a say this 4, 5, 6 possibly you get a loss of 3 rupees.

So, anyway this is how you can means compute such expected reward. Well, I am not asking you to play this these types of game related to the money, but these examples I have taken because as soon as I associate the money or profit or loss you understand very quickly.

(Refer Slide Time: 08:05)

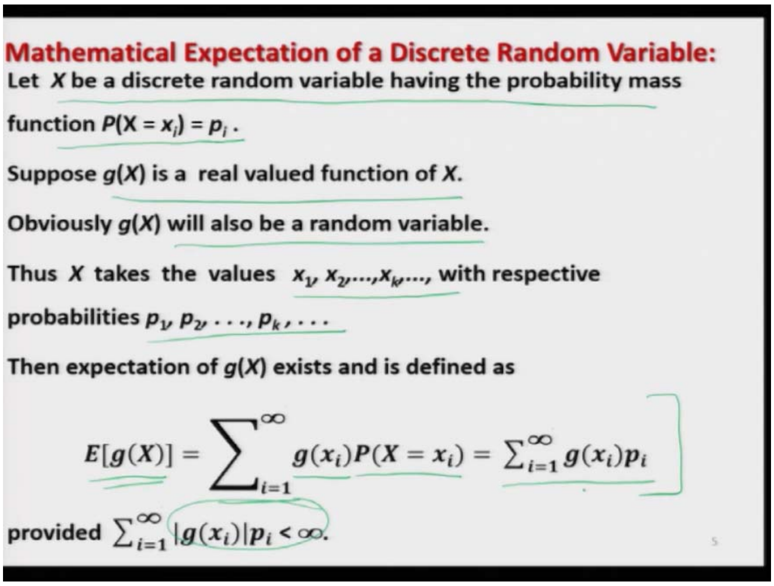


So, now, we try to define this concept of expectation from a statistical point of view using our mathematics. So, now, you know that there are two types of random variables that we are considering, one is continuous type and another is discrete type. So, we will try to define the expectation of a random variable for both the cases. Well, there can also be a combination of the discrete and continuous random variable, but we are not discussing that part.

So, let capital X be a continuous random variable having the probability density function as f(x). And suppose g(X) is a real valued function of X. So, obviously, g(X) is going to be a function of a random variable X. So, g(X) will also be a random variable. Now, the expectation of this g(X) is defined as and it is actually indicated and written as E, this means expectation of g(X), whatever is the variable or variable function that is written inside the bracket or parentheses, whatever you want.

So, this is the range of Ω that is $\int_{-\infty}^{\infty} g(x)f(x)\,dx$. And means provided with this integral exists. So, we have this condition that $\int |g(x)| f(x)\,dx < \infty$. So, this is how we define the expected value of our function in case of a continuous random variable.

(Refer Slide Time: 09:33)

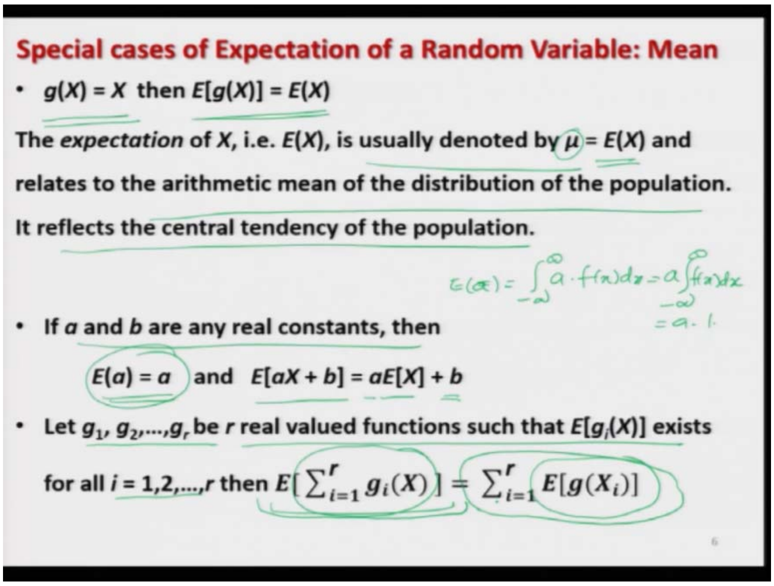


And similarly, we can also define the expectation of g(X) in case of discrete random variable. How? Let us see. Let capital X be a discrete random variable having the probability mass function probability of X equal to small xi is equal to pi. And now, suppose that a g(X) is a real valued function of X. So, this is also going to be a random variable. And now suppose this random variable X takes the values X1, X2, Xk, … with respective probabilities that p1, p2, pk, …..

Now, the expectation of g(X) exists and it is defined as expected value of g(X) is equal to $\sum_{i=1}^{\infty} g(x_i)P(X = x_i) = \sum_{i=1}^{\infty} g(x_i)p_i$. And obviously, if this summation has to exist for which we have a condition that that $\sum_{i=1}^{\infty} g(x_i)P(X = x_i) = \sum_{i=1}^{\infty} g(x_i)p_i$, that is less than infinity. So, this is how we try to define the expectation in case of a discrete random variable.

And believe me, these are very simple things. You may not really think that how simple are they going to be and how useful are they going to be, when you are trying to deal with the data science. That is my promise to you. This mathematical definition may not look so attractive, but I will make sure that this will become the most beautiful attraction of this data science.

(Refer Slide Time: 11:01)



Now, let us see how? Suppose, I doubt try to take different types of forms of the g(X). And let us try to see how they help us. So, the first case I take here in this lecture only, and then I will take some more forms of g(X) in the forthcoming lectures. So, be prepared with the definition of this expected value of g(X) in both discrete and continuous random variable cases. So, in case if I simply say take g(X) equal to X then expected value of g(X) will become expected value of X. What is this expectation of X? Expectation of X is usually denoted by value here μ, the Greek letter. And why this is mu? This will be clear to you when we try to come to some more lectures.

So, μ is equal to E(X) that is the standard notation that has become very popular among the user. And this is related to the arithmetic mean of the distribution of the population. And it reflects the central tendency of the population, you know the mean value, central value. So, this expectation of X is nothing but your mean value. And in case if your a and b are some real constant, then we know that expected value of a is going to be a, because if you want to find it out, you can simply take it here respective value of a is equal to some minus infinity to infinity a times fx dx.

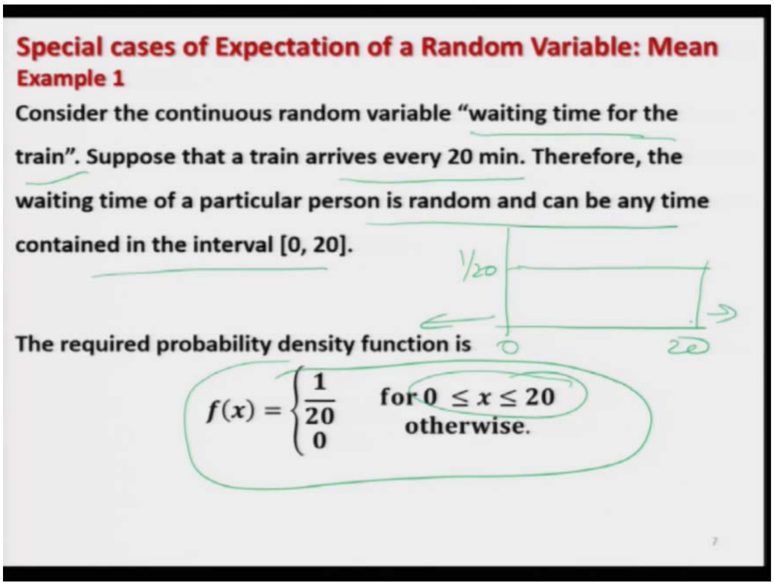
And this is going to be a times $\int_{-\infty}^{\infty} xf(x)dx$. And this is a PDF. So, this is going to be here a

into 1. This integral is going to be 1. So, this will become here a. And similarly, you can prove for the discrete case also. And similarly, if you try to write down here $E[aX + b] = aE[X] + b$. You have to understand these rules very clearly, they are going to help you a lot when you are trying to understand the basic rules of statistics. And this expectation is going to create the soul of the statistics.

So, similarly, there is one more result. Well, I will skip the proof here, but the proofs are extremely simple and if you wish you can find in any statistics book. So, $g_1, g_2, \ldots, g_r$ be r real valued functions such that $E[g_i(X)]$ exists for all i that 1, 2 or, that is the expectation value of all the functions, they do exist. Then $E[\sum_{i=1}^{r} g_i(X)] = \sum_{i=1}^{r} E[g(X_i)]$.

This means the expectation of our sum of the function of a random variable is equal to the sum of the expectation of the function of the random variable. This is what I am trying to say. Just be careful, try to listen to my voice again and again. I will repeat it once again, the expected value of the sum of function of random variable is equal to the sum of the expectation of the function of the random variable. So, these things are going to be very useful.

(Refer Slide Time: 14:12)

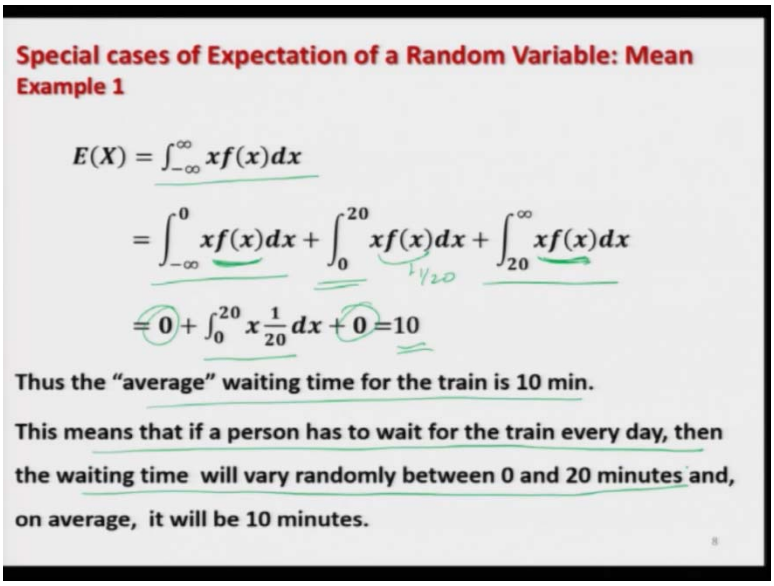


Now, let us try to consider the same example that we considered earlier about the waiting time of the train. And now let us try to find out its mean value. So, we had considered earlier an example in which the random variable is defined as the waiting time for the train. And the train arrives at a railway station after every 20 minutes. So, the waiting time of a particular person is going to be random, and can be any value between 0 and 20. So, we had obtained earlier the probability density function of this x which was obtained here to say

$$f(x) = \begin{cases} \frac{1}{20} & \text{for } 0 \le x \le 20 \\ 0 & \text{otherwise.} \end{cases}$$

.

(Refer Slide Time: 14:54)



So now in case if you try to find out the expected value of here x, the mean value, that means this will come out to be by definition $\int_{-\infty}^{\infty} xf(x)dx$. Now, this PDF is going to take the value

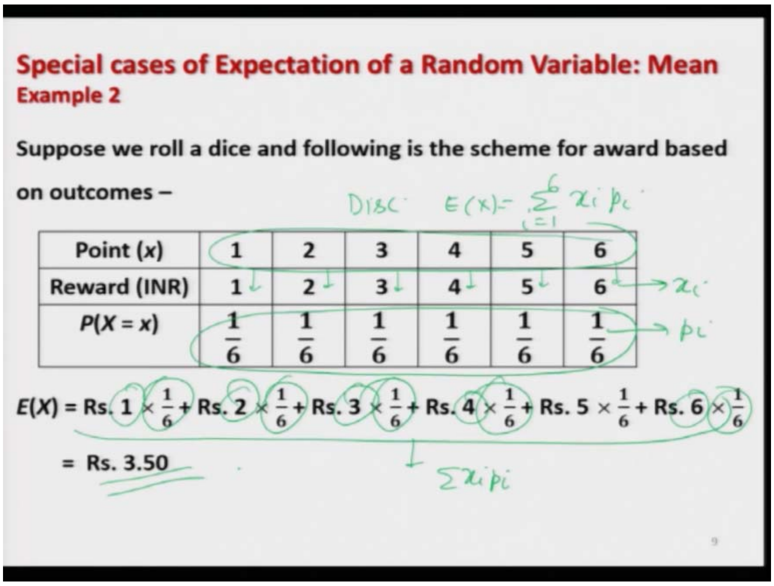
only between 0 to 20. So, if you try to plot it here, this will be somewhere here somewhere here 1/20. But what will happen? This value is going to be between 1 and 20 only between 0 and 20, before that and after that, this is going to be simply 0.

So, I can divide this integral minus infinity to infinity into three parts, minus infinity to 0, 0 to 20 and 20 to infinity. So, you can see here that the value of the this f(x) and this f(x) is going to be 0. So, we get here 0 and 0 like this. And but value of this f(x) is going to be 1/20. So, we get here $\int_{-\infty}^{\infty} xf(x)dx$, and if you try to solve, this will come out to be here 10.

So, what is the meaning of this? The average waiting time for the train is 10 minutes. So, if somebody comes at 5 minutes before the train or 15 minutes before the train or or if it tries to come to the railway station continuously for say large number of time, then on an average the person will have to wait for 10 minutes.

Or if somebody asked you a very simple question that I am going to catch that train, how much time do you expect I will have to wait at the station? Without even knowing the exact time of the train, you can tell very easily that on an average you expect that you may have to wait around 20 minutes. That is the meaning of this expectation value of X. So, you can see here that how useful it is. So, this means that if a person has to wait for the train every day, then the waiting time will vary randomly between 0 and 20 minutes and on an average, it will be 10 minutes.

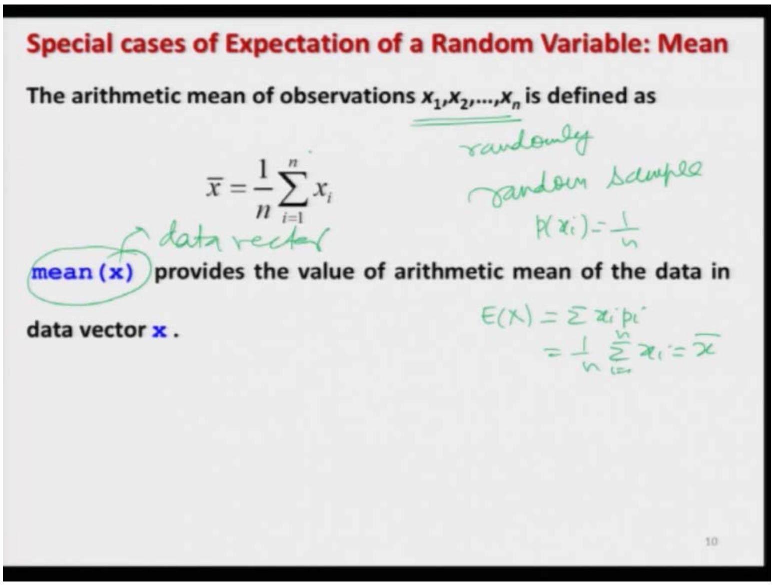
(Refer Slide Time: 16:49)



Now, let us try to take the same example that we did earlier to motivate you. So, now you can understand it here very quickly that we have here a dice on which we can observe the point 1, 2, 3, 4, 5, 6 and probability of each of the point is 1/6. And we have decided to give a reward based on the number of points observed which is same as the number of points observed that is, if you get to 1, you will get rupees 1; if you get two, you get 2 rupees; if you get 3 numbers, then you get 3 rupees; if you get 4 numbers, you get rupees 4; if you get 5 numbers, you get 5 rupees; and if you get 6 number, then you get 6 rupees.

So, now you can see here, this is a discrete random variable. So, this is so, so expected value of X is going to be simply here summation i goes from 1 to 6 xi into pi. So, now these are the values of here xi and these are the values of here pi. And this is the same expression that you obtain earlier, without the knowledge of how to find out the expectation. And this is nothing but your summation xi pi. x_1 into p_1 plus x_2 into p_2 plus x_3 into p_3 plus x_4 into p_4 up to x_6 into p_6. And this value comes out to be here 3.50, that means if you try to play the game repeatedly, then on an average we expect a profit of rupees 3.50. So, this is going to give you the mean value.

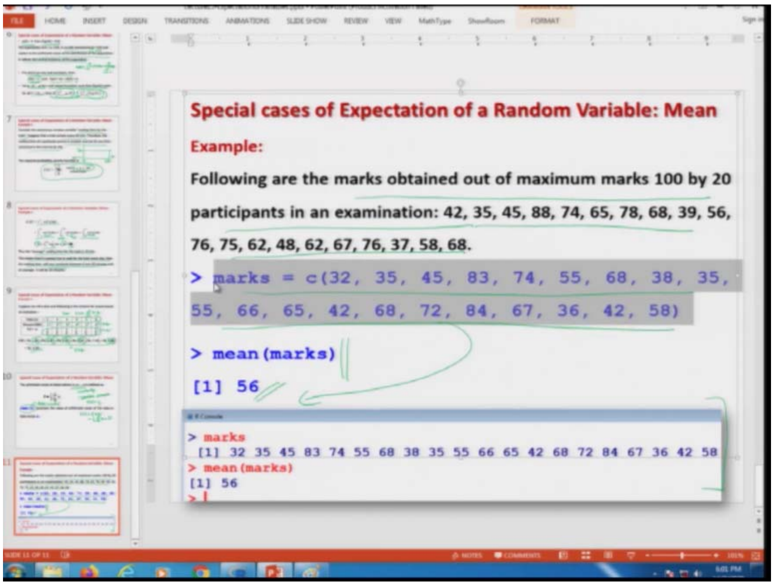
(Refer Slide Time: 18:18)



And if you try to assume that the probabilities of each and every event is the same. Suppose if you try to draw some observations from some population, which are indicated here as x_1, x_2, x_n suppose, they are drawn randomly. This is a random sample. So, in a random sample, the probability of every event is the same. So, probability of observing x_i will simply be here 1 upon the total number of observation 1/n. And if you want to find out expected value of X in this case, what will happen here? $\sum_{i=1}^{\infty} x_i p_i$ and p_i here is 1/n. So, this becomes here, i goes from here $\frac{1}{n}\sum_{i=1}^{n} x_i$ and this is nothing but you know, this is your arithmetic mean $\bar{x}$.

And this plays a very important role and if you want to compute the arithmetic mean in the R software, we have a command here mean of x, x is going to be the data vector. As simple as that. So, you see, that will give you an idea about the average value, what is the expected value of X?

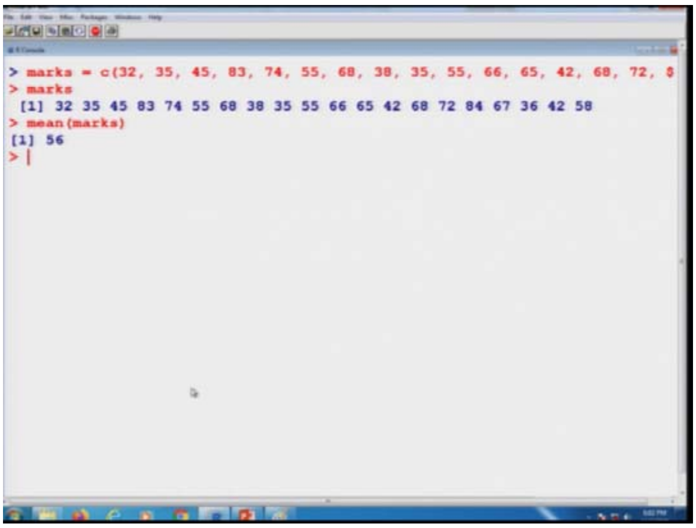
(Refer Slide Time: 19:29)



Well, in case if you have a proper PDF or a proper probability mass function, you can compute it very easily on the R software, using numerical integration techniques also if you want to solve an integral. So, now let me try to take a very simple example and try to show you how this mean function works in the R software.

Suppose, I take a very simple example that that the following are the marks obtained out of 100 by 20 participant at any exam, like this one, and which I have stored in a data vector whose name I have given as marks, and I tried it inside the R console. I get here a value of here 56 and this is here the screenshot of the same thing. So, just by computing this thing you can see here that you can obtain the average value of this X.

(Refer Slide Time: 20:22)



So, now, let me try to show you this thing on the R console also. So, now, let me try to first copy this data into the R console, you can see here it is here like this. So, you can see here these are the marks. And if you want to find out here the mean, mean will give you here the mean of this marks, which is here 56.

So, now, we come to an end to this lecture. So, now, if you try to see what I have done. First, I have given you the concept of expectation, and then I have taken a very special function or the form of the g(X). And then I have shown you that expected value of X is going to give you the idea about the central tendency of the variable. What does this mean? X has got a probability distribution function f(x) or probability mass function, say this p_i. This pi's may not always be the same. They may be different.

But whatever is the central tendency of the data, if the data is originating from that probability function, the expected value of X is going to give us the idea about that value. And if you try to see whenever the data comes, you are always interested in the average value, you are not interested in the actual values. You have lots of data, you are expected to inform your company that what will be the average sale in the next month. You are not interested in what is happening on a particular day.

In case if you are trying to handle customers on a shopping website, you want to know their average behaviour, means average spending, average number of times what they visit the website. Means you have seen that if you are not visiting a website for a long time, you get a sort of message or email that, I have not seen you for a long time, you come here we are giving you a coupon of this money and etcetera, like those. Do you think that is there anybody who is sitting behind the computer to send you that offer? Those things are happening automatically.

So, in short, they can compute that what was the average time what a customer has spent in the last month on this website, but this time, this time has gone down drastically. So why not to attract the customer. For that, what you have to do? Simply find the expected value of X. And similarly, there are many more characteristics which can be obtained on the basis of given set of data, just by changing the form of g(X).

In this lecture, I have talked only about one tendency of the data, central value. And when I started this lecture, I had not given you this idea and possibly you also did not expect that I am going to conclude it in such an application way. So, try to think about it, try to take a data, try to compute the different types of mean central tendencies etc. And I will see you in the next lecture with some more choices of g(X). So, you try to practice it, try to have a look and I will see you in the next lecture. Till then, goodbye.