

Essentials of Data Science with R Software - 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur
Lecture No. 22
Random Variables - Discrete and Continuous

Hello, friends. Welcome to the course Essentials of Data Science with R Software 1, in which we are trying to understand the topics of probability theory and statistical inference. So, up to now, you see we have discussed various definition, concepts and topics related to the probability theory. And we were always interested in counting the events and then trying to find out the probability. But, now, you have to think in a very different way.

Suppose, we are trying to conduct an experiment, and we want to compute the probability of certain event. So, what is your first job? Your first job is to collect the observations on what is your objective you want to find. Suppose, if I say I want to know the probability of head and tail. So, you will toss the coin and you will record head, tail, head, head, tail, tail and so on.

And now, we have simply said, after this you can find out the probability. But my question is how? Can you really add head and tail? Or if I say, if you roll a dice, we will have numbers 1, 2, 3, 4, 5, 6. They are the indicator variable. They are taking values in six different categories, which are obtained by counting the number of points on the upper side of the dice.

So, now, the question is this, how would you like to integrate all this process in a simpler way so, that anybody can understand it very easily. So, whatever is the process and whatever is our objective, based on that, we can define some objective which can convert this qualitative values into a quantitative value.

For example, I can say suppose you toss the coin say in a sequence of 5 times, 5 times, 5 times, 5 times, 5 times and so on. And then you try to count the number of heads in every trial. So, now, in this case, every trial is consisting of 5 continuous flips of the coin. And then we are trying to count the number of heads. So, now, I can say, let there be a random variable, which is indicating the number of heads in the continuous 5 flips of the coin, that is all.

So, now, in this case, what can happen? That I can define a random variable. And that random variable can be defined like as number of heads in the 5 continuous trials of the coin.

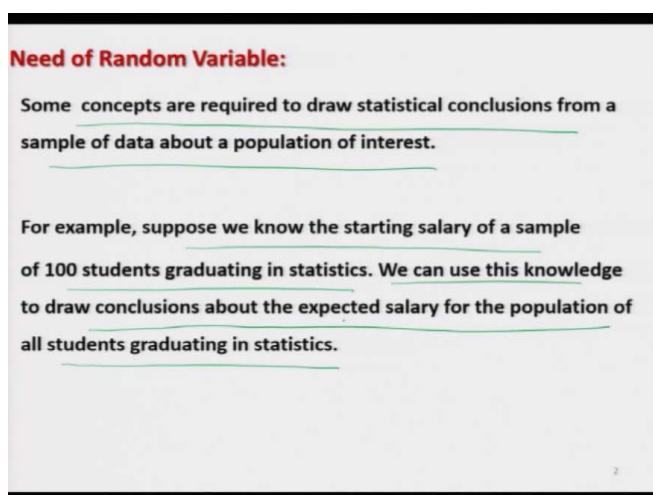
So, now, what will happen? Earlier you had something like head, head, tail, tail, tail or head, tail, head, tail, head, but now, you can define it in a quantitative way. Your X is going to take value 1, 2, 3, 4 or 5. Can you take a value 6? No, because you are flipping the coin only for 5 times.

So, now, you can see, I can convert all my qualitative information into a quantitative way. Now, what is my advantage? Once in statistics you get the information in some quantitative terms, you can apply any of the statistical tool. So, that is what now, we are going to start. We will try to formerly understand different types of concepts, definition which will formerly helps you in formulating our real problem into a statistical rule as a statistical language with which you can apply the statistical rule.

And what is your objective? Means, ultimately you want to draw some conclusion that, what is really happening inside the data, what type of information that data is containing. So, that can be obtained from such a setup. So, that is what we are going to start today. We are going to start a topic on univariate random variables.

Now, as soon as I say univariate, that means there can be bivariate, trivariate or in general multivariate. But to begin with, to understand the concept, it is very important for us to understand the concepts in a univariate way, what is really happening, what is the interpretation, what is the algebra and then we try to convert it into a bivariate, trivariate or multivariate way. So, we are going to talk about how are we going to define the random variable and what are the associated properties, definition, concept risk, etc. in this lecture and in the some forthcoming lecture also. So, let us begin our lecture.

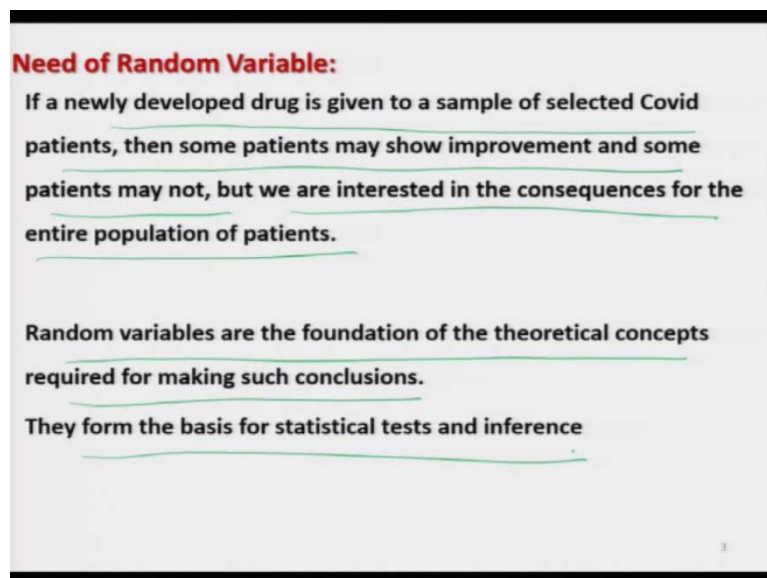
(Refer Slide Time: 05:33)



So, now, we are going to talk in this chapter about random variables and we are going to talk about two different particular type of random variables which are called as continuous and discrete random variables. So, now, the first question comes, why do we need a random variable? What is the need of random variable? So, we know that whenever we are trying to work in the data sciences or decision sciences, some concepts are required to drop statistical conclusion from a sample of data from a population of interest. Because, you see in statistics we are always working only on a sample, which is drawn from a population. And as we have discussed, we assumed that this sample is a representative sample.

Like for example, suppose we know the starting salary of say 100 graduating students from a college in the subject is statistics. Now, we can use this knowledge to draw different types of conclusion about the expected salary for the population of the all students graduating in statistics. So, what is really happening? We have a sample of only 100 students, out of say 5000 students from the college. But now, in case if you ask that, what will be the expected salary if a student graduate from this particular college? Possibly just using these 100 observations, we can get an answer.

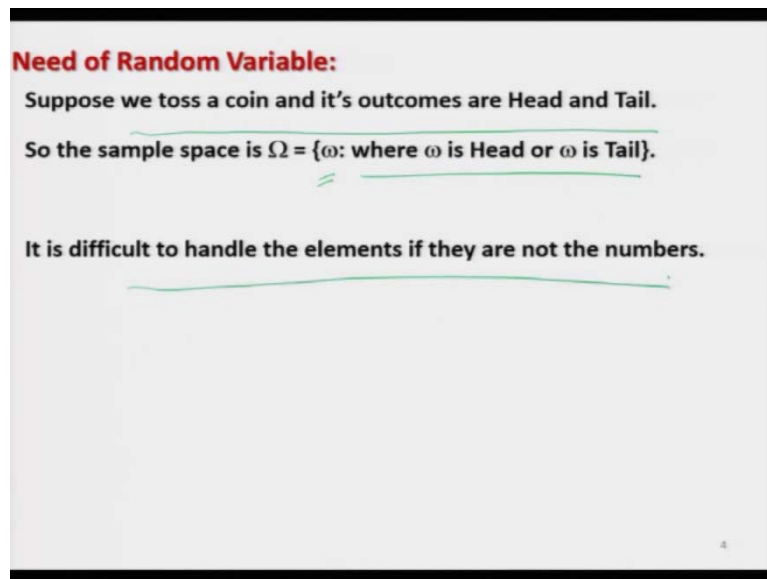
(Refer Slide Time: 06:50)



Similarly, if a drug is being developed for say COVID patients, then what do we know? We simply try to give the medicine to some patients and then we try to see whether the medicine is working or not. But what is your conclusion? What do you really want to know out of that? Think about it. So, in case if a newly developed drug is given to a sample of selected COVID patient, then some patients may show improvement and some patient may not.

But are we interested only these values or are we interested only in this small group of patients? Or are we interested in the consequences for the entire population of the patients all over the world? So, now, in order to draw such conclusion, we need to have some statistical concept and that are going to lay the foundation of the statistical inference, which is our ultimate objective. So, random variables are the foundation of theoretical concept required for making such conclusion and they form the basis for the statistical test and statistical inference.

(Refer Slide Time: 08:01)



So, now, I already had given you this example, but now I can explain you formally. Suppose we toss a coin and its outcomes are recorded as head and tail. So, now, your sample space of this event is going to be some Ω , a ω , where ω is the head or ω is the tail. But now, if you try to say, if I want to know what is the average number of heads coming in the trial of 20 flips.

So, now, you will have 20 values of heads and tails, and how will you know that what is the average value or the average number of heads which are coming in the 20 flips? You cannot, and it is difficult to handle the outcome, because these elements they are not the numbers, they are some head, tail, head, tail and so on.

(Refer Slide Time: 08:51)

Need of Random Variable:

We can solve this issue as follows:

Let X be a function such that

$$X(\omega) = \begin{cases} 1 \rightarrow \text{if } \omega \text{ is Head} \\ 0 \rightarrow \text{if } \omega \text{ is Tail} \end{cases} \quad \hookrightarrow \text{HT} \rightarrow 1, 0$$

Thus X is a real valued function defined on Ω which takes us from Ω to a space of real numbers $\{0, 1\}$.

X is called a random variable.

So the head is denoted by 0 and tail is denoted by 1.

Sample space $(\Omega) = \{0, 1\}$

So, now, in order to solve this issue, one simple option is that, we can define a function a mathematical function, say capital X. And this function is defined, which is based on the outcomes of the experiment, which are indicated by a ω . And this function takes the value 1 in case if ω is head and takes value 0 if ω is tail. Well, if you want you can interchange it also, means you can say that it takes value 1, if ω is tail; or it takes value 0, if ω is head, that will not make any difference.

But that depends on you or the say the choice of the experimenter that how the function is defined. So, now you see what is really happening now? You can see that here, this function X that is now playing a very important role and it is really helping you. This is a real valued function which is defined on Ω , the sample space. But it is taking us from Ω to a space of real numbers 0 or 1. Because Ω consists of say head and tail and now, I am saying that this head and tail they are going to be represented by 1 and 0 or 0 and 1 depending on your choice. This x is actually called a random variable, in a very simple language.

So, in this case head is going to be denoted by 0 and tail by 1. And your sample space now Ω will be converted from head and tail to 0 and 1. So, this Ω will be consisting of two values 0 and 1 which are going to indicate the presence of head or tail in the outcome of the experiment.

(Refer Slide Time: 10:28)

Need of Random Variable:

This can be simulated in R by the sample command by drawing one observation between 0 and 1 by simple random sampling with replacement.

```
sample(c(0,1), size=1, replace = T)
```

Popn (circled around c(0,1)) *SRSWR* (circled around replace = T)

6

Now, in case if you really want to understand that, why do I call it as a random variable? Why not a variable? So, you know that when you are trying to toss a coin, you do not know the outcome unless and until the experiment is completed. You just know either the outcome is going to be head or the tail. But you do not know that after you toss the coin with 100 percent surety what is going to come. That we already have done and if you remember, we had used R software to sample from 0, 1 or say 1, 2, 3, 4, 5, 6 in the case of a dice to understand that the samples that you draw every time they are changing.

Just for your recollection, means I can show you here that we had used the sample command for drawing of head or a tail. And every time I draw it, it will give us a new value. For example, if I try to draw one value out of 0 and 1, so this is going to indicate my here population. And now I am saying a size is equal to 1, that means I need only one observation out of 0 or 1. And then replace equal to True is trying to say that it is drawn by simple random sampling with replacement.

(Refer Slide Time: 11:48)

Need of Random Variable:

The outcome of this experiment is observed as follows:

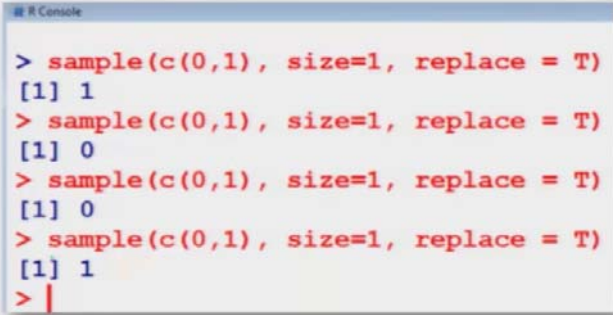
```
> sample(c(0,1), size=1, replace = T)
[1] 1 ✓

> sample(c(0,1), size=1, replace = T)
[1] 0 ✓

> sample(c(0,1), size=1, replace = T)
[1] 0 ✓

> sample(c(0,1), size=1, replace = T)
[1] 1 ✓
```

Need of Random Variable:

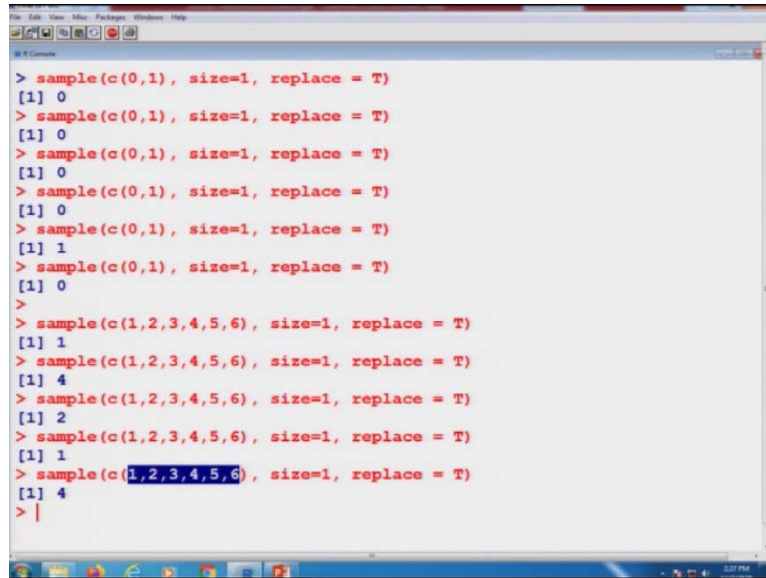


The screenshot shows an R console window with the following text:

```
# R Console
> sample(c(0,1), size=1, replace = T)
[1] 1
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 1
> |
```

So, now in case if you try to execute it on the R console. For example, I am giving you here some outcomes you can see here that when I try to conduct it, means every time I am getting a different outcome. For example, if I repeat it once, I get here 1, then I get 0, then I get 0, then I get here 1. And you can see here, this is the screenshot. So, you can believe on me that whatever outcome I have shown you here they are really going to work.

(Refer Slide Time: 12:23)

A screenshot of an R console window. The window title is "#Console". The console shows a series of commands and their outputs. The first set of commands is `> sample(c(0,1), size=1, replace = T)`, which is repeated six times, resulting in outputs of 0, 0, 0, 0, 1, and 0. The second set of commands is `> sample(c(1,2,3,4,5,6), size=1, replace = T)`, which is repeated six times, resulting in outputs of 1, 4, 2, 1, 1, and 4. The cursor is at the end of the last line.

```
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 0
> sample(c(0,1), size=1, replace = T)
[1] 1
> sample(c(0,1), size=1, replace = T)
[1] 0
>
> sample(c(1,2,3,4,5,6), size=1, replace = T)
[1] 1
> sample(c(1,2,3,4,5,6), size=1, replace = T)
[1] 4
> sample(c(1,2,3,4,5,6), size=1, replace = T)
[1] 2
> sample(c(1,2,3,4,5,6), size=1, replace = T)
[1] 1
> sample(c(1,2,3,4,5,6), size=1, replace = T)
[1] 4
> |
```

But instead of believing on me, let me try to show you this thing on the R console also, so that you can see that how this randomness is coming into picture, which is very important for you to understand. So, you can see here, when I try to execute this command, this time it is giving me 0, now I try to repeat it, is again giving me 0, it is again giving me 0, it is again giving me 0, now it is giving me 1. Now you tell me what it is going to give you? 0 or 1? Any answer? No, you cannot tell me unless and until I press here enter, you can see now I am getting here 0. So, this is what I mean that the values are random.

Now, similarly, if I try to conduct the experiment with a roll of a dice, and I try to write down the number as a 1, 2, 3, 4, 5, 6. And suppose we are trying to roll a dice and then I am trying to see what is coming. So now I can simulate it with the same command by changing the population from 0 and 1 to 1, 2, 3, 4, 5, 6. And now you tell me what will be the answer? It is coming out to be 1.

Now I try to repeat it, I get now here 4. I again repeat it, I get here 2. Now, you tell me, I am throwing my dice once again and now what do we expect? We do not know unless and until I press here enter, that is equivalent to throwing the dice. So, I am getting here 1 and now once again you know what are you going to get? No. So, this time you are getting here a 4. So, you can see here every time you are getting 1, 4, 2, 1, 4. But all of these numbers they are coming from says this set 1, 2, 3, 4, 5, 6. And this is what we mean when we try to say that this variable is a random variable that means X is my number which is coming on the upper face of the dice.

(Refer Slide Time: 14:02)

Need of Random Variable:

In any random experiment, we are interested in the value of some numerical quantity determined by the result.

We are not interested in all the details of the experiments.

These quantities of interest that are determined by the result of the experiment are known as *random variables*.

So, now, you can understand that whenever we are trying to conduct a random experiment, we are interested in the outcome of the experiment. But we want to convert it in some value. We are more interested in the value of some numerical quantity determined by the result. We are not interested in all the details of the experiment that, at what time you did, or whether it came head or tail, we want every information in terms of numerical values. So, these quantities of interest that are determined by the results of the experiment are called as random variables.

(Refer Slide Time: 14:46)

Need of Random Variable:

Since the value of a random variable is determined by the outcome of the experiment, we may assign probabilities of its possible values.

For example

$$P(X = 0) = P(\text{Head}) = 1/2$$

$$P(X = 1) = P(\text{Tail}) = 1/2$$

$P(X=0) = 1/3$
 $P(X=1) = 2/3$
 Tail

We may therefore view X as a random variable which collects the possible outcomes of a random experiment and captures the uncertainty associated with them.

Now, if you try to see, whenever we are trying to conduct an experiment, or a random experiment, we do not know what is going to be the outcome. So, now how to take care of

this phenomena? Because you see, whatever values you are going to get, they are going to indicate the nature of the phenomena, that how the process is working, how the things are happening. So, what we can do, that when we are trying to toss a coin, the outcomes are head and tail that we already have converted by indicating by 0 and 1. But now, in the toss of our trial, whether you are going to get head or tail, how to formulate this process? Because we are not 100 percent confident.

So, one option is this with every outcome, we can associate a probability. And then whatever are the possible values of the variable that it can take, for each value of the variable, we can associate one probability. And that will indicate the process in a much better way than just by saying head or tail or 0 and 1. So, what we can do, that we can assign probabilities to the possible values of the experiment.

For example, if I say probability equal to X equal to 0 is indicating the probability of a head and I can say that this is equal to $1/2$. Probability X equal to 1 is indicating the probability of a tail in the toss of a coin. So, this is equal to $1/2$. So, now looking at this expression, can you conclude that we have got an experiment where we have two outcomes, head or tail and they have got the equal probability of occurrence?

On the other hand, if I suppose, right probability that X equal to 0, this is equal to suppose $1/3$ and probability of X equal to 1, if suppose here $2/3$. So, what do we infer from here? This is indicating that the event X equal to 1 has higher probability of occurrence. So, there are more chances that if you try to toss a coin, then tail will come. So, this is how you can change your statement, and you conveyed what is really happening in the truth. So, now, we may therefore, view as a random variable which collects the possible outcomes of the random experiment and captures the uncertainty associated with them.


So, now, you can see that the concept of this random variable is very helpful for us to understand the statistical phenomena, and it is associating a numerical value with the outcome of a random variable as the letters, as well as it is trying to associate a probability with that value; so, that you can have a complete picture of the phenomena that what are the possible outcomes and how they are going to work.

(Refer Slide Time: 17:56)

Need of Random Variable:

For example, suppose we toss two dice and suppose we want to study about the sum of the points on the upper face to be 7.

Our interest will not be in the individual data points (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) or (6, 1).



These quantities of interest that are determined by the result of the experiment are known as *random variables*.

So, now, suppose it was two dice and suppose we want to study about the some of the points on the upper face that and we want to know what is the the process in which the sum of the two numbers on the upper faces of the two dice, they are coming out to be 7. So, in this case, our interests will not be in the individual data point.

For example, out of those 36 outcomes, there are six values 1, 6; 2, 5; 3, 4; 4, 3; 5, 2 and 6, 1 which you try to sum, they will give you the value 7, say 6 plus 1 is 7, 5 plus 2 is 7, 4 plus 3 is 7 and so on. And we are trying to count these numbers, the numbers which are on the upper surface of the dice. So, these quantities of interest that are determined by the results of the experiment are known as random variables and their values.

(Refer Slide Time: 18:52)

Need of Random Variable:

$P\{X = 2\} = P\{(1, 1)\} = 1/36$

$P\{X = 3\} = P\{(1, 2), (2, 1)\} = 2/36$

$P\{X = 4\} = P\{(1, 3), (2, 2), (3, 1)\} = 3/36$

$P\{X = 5\} = P\{(1, 4), (2, 3), (3, 2), (4, 1)\} = 4/36$

$P\{X = 6\} = P\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} = 5/36$

$P\{X = 7\} = P\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} = 6/36$

$P\{X = 8\} = P\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} = 5/36$

$P\{X = 9\} = P\{(3, 6), (4, 5), (5, 4), (6, 3)\} = 4/36$

$P\{X = 10\} = P\{(4, 6), (5, 5), (6, 4)\} = 3/36$

$P\{X = 11\} = P\{(5, 6), (6, 5)\} = 2/36$

$P\{X = 12\} = P\{(6, 6)\} = 1/36$

Handwritten notes: $\text{Min} = 2$, $\text{Max} = 12 (6,6)$

For example, in this case, you can see here, if you try to see the probability, that is you get the value 2 as a sum of the number of points on the upper surface of the dice, this is there is only one event, 1 and 1 that can occur on each of the upper face of the dice. So, it will give you 1 plus 1 equal to 2, so, there is only one point. So, this is going to be $1/36$. And similarly, if you try to go for the outcome of 3, then 1, 2 and 2, 1 they are the two possible outcomes. So, the probability is going to be $2/36$. And similarly, if you try to go for 4, there are 3 possible outcomes, some really for 5, there are 4 possible outcomes.

And similarly, you can see here for each of the number, because the minimum number is 2, and the maximum number will be 12. That means 6 and 6 occurs. So, now you can see the probability or the distribution of the probabilities of the values of random variable in this random experiment. So, you can see here that the maximum probability is here $6/36$, which means the maximum chances are that the sum will come out to be 7. Then the next probability is $5/36$ that is observing the value 6 or 8. And similarly, you can conclude for all other values.

So, you can see here this simple experiment has been associated with two numerical values on the outcome, and then you have defined the random variable which was the objective of your study and based on that you have associated the corresponding probabilities. And now, looking at the values of the probabilities, one can very easily decide that whether the some event is going to occur or not or what is the distribution of the probabilities among the different values which have random variable can take.

(Refer Slide Time: 20:46)

Need of Random Variable:

In other words, the random variable X can take on any integral value between 2 and 12 and the probability that it takes on each value is obtained.

Let Ω represent the sample space of a random experiment, and let R be the set of real numbers.

Then the random variable X is assigning one and only one number to each element $\omega \in \Omega$, $X(\omega) = x, x \in R$, i.e. $X: \Omega \rightarrow R$.

For example, in $P(X = 2) = P\{(1, 1)\} = 1/36$, the random variable X assigns $X = 2$ to element $\omega = (1, 1)$.

Handwritten notes on the slide include: $\omega \in \Omega \rightarrow X = 2$, $(1, 1) \rightarrow (2, 3, \dots)$, and $(1, 2) \rightarrow \dots$

Now, in this case, as I said, the random variable is going to take the values whose minimum value is 2 and the maximum value is 12 and its probability are given. So, now, let us try to indicate this phenomena in a statistical way through the help of definition of random variables and sample space and corresponding probability. So, let Ω represent the sample space of the random experiment. And let R be the set of real numbers. Now, you can see this random variable is going to take here the values like 2, 3, 4, 5. So, do not you think that I can associate here a set of real numbers, the numbers are going to be between 2, 3, 4, 5 up to 12.

So, let R be the set of real numbers, then the random variable X is assigning one and only one number to each element of this Ω which are indicated by ω . And as you had done earlier that X of head or X of tail, you can also define here $X\omega$ is equal to a small x and this x is small x that means, the x in the lowercase alphabet. What is the meaning of lowercase alphabet and uppercase alphabet? This I will try to discuss soon. And this small x that is belonging to the set of real numbers and now, x is a mapping from Ω to R , that whatever are the values here they are obtained as 1, 1; 1, 2 etc., etc., etc., they are now translated to a set of real numbers like as 2 and 3 and so, on.

For example, if you are trying to say that X is going to take the value 2 that means, this is the probability of occurrence of 1 and 1 and whose probability is $1/36$. So, now, if you try to see this 1, 1 this is your ω which is belonging to Ω . And now, this is being transported to X equal to 2, which is now a real number. So, you can see here through the help of random variable you can transport the sample space or the values of the sample space to some real number.

(Refer Slide Time: 23:10)

Need of Random Variable:

$P(X=2) = P\{(1, 1)\} = 1/36$
 $P(X=3) = P\{(1, 2), (2, 1)\} = 2/36$
 $P(X=4) = P\{(1, 3), (2, 2), (3, 1)\} = 3/36$
 $P(X=5) = P\{(1, 4), (2, 3), (3, 2), (4, 1)\} = 4/36$
 $P(X=6) = P\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} = 5/36$
 $P(X=7) = P\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} = 6/36$
 $P(X=8) = P\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} = 5/36$
 $P(X=9) = P\{(3, 6), (4, 5), (5, 4), (6, 3)\} = 4/36$
 $P(X=10) = P\{(4, 6), (5, 5), (6, 4)\} = 3/36$
 $P(X=11) = P\{(5, 6), (6, 5)\} = 2/36$
 $P(X=12) = P\{(6, 6)\} = 1/36$

Handwritten notes:
 Min = 2
 Max = 12 (6,6)
 + = 1

So, now, you have seen here that in this example, you have obtained the probability of observing 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 and if you try to sum all these probabilities, 1/36 this this this this this this this this this and this all the probabilities, their sum will come out to be 1.

(Refer Slide Time: 23:30)

Need of Random Variable:
 In other words, the random variable X can take on any integral value between 2 and 12 and the probability that it takes on each value is obtained.
 Since X must take on some value, we must have

$$1 = P(\Omega) = P\left(\bigcup_{i=2}^{12} \{X=i\}\right) = \sum_{i=2}^{12} P\{X=i\}$$

$P\{X=2 \cup X=3 \cup \dots \cup X=12\}$

14

Because this is the probability of Ω which was always equal to 1, that we have discussed in the earlier lectures. So, now, you can see here that the earlier knowledge which you obtained that is now going to help you in further developments. And you want to find out here the probability of say happening of 2, 3, 4 up to 12.

So, this I can express by using the notation probability of union of X equal to small i where i is going from 2 to 12, like this, that that is equal to X equal to 2 union X equal to 3 union up to here X equal to 12. And then this is the probability of this thing. So, this is going to be because they are going to be disjoint events, two events cannot happen together. So, I can write down this probability as the sum of individual probabilities. And that is always going to be 1 that you can verify from this example.

(Refer Slide Time: 24:25)

Random Variable:

Let Ω represent the sample space of a random experiment, and let R be the set of real numbers.

A random variable is a function X which assigns to each element $\omega \in \Omega$ one and only one number

$X(\omega) = x, x \in R, \text{ i.e. } X : \Omega \rightarrow R.$

15

So, now, I can express all these things in a one-line definition, very simple way. And for that, I need here two things, one is Ω and there is capital R. So, let Ω represent the sample space of the random experiment and let capital R be the set of real number, then our random variable is a function X which assigns to each element ω belonging to Ω one and only one number, that is $X(\omega)$ is equal to a small x.

So, small x is going to be a number which belongs to the set of real numbers. So, X is a function from Ω to R. So, now you can see here this is pretty simple if you want to understand it. Now, I can give you this information, you can use here that here I am trying to use two symbols capital X and a small x. What is the meaning of this thing?

(Refer Slide Time: 25:22)

Random Variable:

It is a convention to denote random variables by capital letters (e.g. X) and their values by small letters (e.g. x).

If X is height of students, then $x = 168$ Centimetre is the value of X.

Similarly, $x_1 = 170$ centimetre indicates the first value of X and $x_2 = 180$ centimetre indicates the second value of X.

16

So, now, let me try to take one very simple example and then I can explain you very easily. Suppose you want to measure the heights of 5 students. So, let random variable X is representing the heights of the students. Now, you call the first student and ask the height. Suppose this is 168 centimetres. Now you call, so, this is the value of here value of your height, height of first student. And then you try to call the second student and you try to measure the height of second student. So, suppose this comes out to be here, suppose 170 centimetres.

And similarly, you will try to call all the 5 student and finally, you will have all the values of heights of the 5 students. So, now, how to indicate this numerical values? And how to express this statement? These are pretty long statement. It is not possible to write them again and again, height of first student, height of second student and and you suppose there are 1000 students, you have to write height of the 999th student and height of one thousandth students.

So, one simple option is that I am indicating here with this by capital X the height, so, I can indicate the value of the height by small x . So, now, this 168 centimetre which I am taking here, this is actually small x . This 170 centimetre I am taking here this is also a small x . Now, both of them are a small x , how to determine what is my first observation, what is my second observation? So, I can do here whatever is my first observation, I can write down here x_1 and whatever is my second observation, I can write down here x_2 .

So, now, you can see here that x_1 and x_2 they are indicating the value of the variable capital X for the first observation and second observation. And similarly, means, I can have here 5 observations like x_1, x_2, x_3, x_4, x_5 which are going to be some numerical values of X . So, in statistics, we have a convention that by the capital letters we try to denote the random variables and their values are indicated by a small alphabet, a small letter, lowercase alphabets.

So, as I give you this example, here that if X is the height of the student, then small x equal to 168 is the centimetre is the value of the X . And similarly, say x_1 equal to 170 centimetre, this indicates the first value of X and x_2 equal to 100 centimetre indicates the second value of X . So, whenever I try to write down here let x_1, x_2, x_n be a random sample, you can understand that these are n values which are the values of the random variable X , that is our understanding, as simple as that.

(Refer Slide Time: 28:33)

Discrete Random Variable:
 Random variables whose set of possible values can be written either as a finite sequence x_1, x_2, \dots, x_n or as an infinite sequence x_1, x_2, \dots are said to be *discrete*.

A sample space is discrete if it consists of a finite or countable infinite set of outcomes.

For instance, a random variable whose set of possible values is the set of nonnegative integers is a discrete random variable.

Handwritten examples:
 $1, 2, 3, 4, 5, 6$
 $2, 3, \dots, 12$

So, now, the way we are going to represent these values is as follows, random variable whose set of possible values can be written either as a finite sequence x_1, x_2, x_n means, you know that only for example, in the case of say here coin, you have only two options. So, there will be only means, suppose you toss two times, so, there will be only 2 outcomes. And if you try to toss three times, there are going to be 3 outcomes. So, we can write down here x_1, x_2, x_3 .

But similarly, this sequence can be an infinite sequence also and, in that case, we can write down here x_1, x_2, \dots . So, those random variables whose set of possible values can be written either as a finite sequence x_1, x_2, \dots, x_n or as an infinite sequence x_1, x_2, \dots are said to be discrete. And the corresponding random variable is called as discrete random variable.

For example, if I say the toss of a coin or say, let me take better example, say that the number of points on the upper side of the dice when it is rolled, it will take value here. So, you have 1, 2, 3, 4, 5 and 6, but if you try to toss here two dice and, and if you try to find out, find out the sum of the numbers on the upper faces, the possible values will be 2, 3, 4 up to here 12.

So, you can see here that these values are finite or countable infinite. You cannot say that there are 2.3 number of heads or you cannot say that the value of the outcome on the upper face of the dice is 3.4. So, a sample space is discrete, if it consists of finite or countable infinite set of values. For example, a random variable whose set of values possible values is the set of non-negative integer is a discrete random variable.

(Refer Slide Time: 30:38)

Discrete Random Variables:

Example: A customer care phone contains 30 external lines.

At a particular time, the system is observed, and some of the lines are being used.

Let the random variable X denote the number of lines in use.

Then, X can assume any of the integer values 0 through 30.

When the system is observed, if 5 lines are in use, $x = 5$.

Handwritten notes on the slide: "0, 2, ... 30" above the second paragraph; "16.7" and "20.5" circled and crossed out above the third paragraph; "x = 5" circled above the fourth paragraph.

For example, let me try to take an example to explain you this concept in more details. Suppose, there is a customer care office and that office has 30 external lines for that phone and at a particular time, the system is observed and some of the lines are being used, that how many people are calling and how many lines are actually being used to answer those calls.

So, let the random variable capital X indicate the number of lines in use. Then now, if you can see here there are 30 lines. So, the first possibility is that none of the line is going to be used, nobody is calling. Or one possibility will be that only one customer is calling, so, only one line is used. Or second possibility is means two customers are calling, two lines are used. And up to now, this number can go from here 0 to 1, 2 up to 30 only, because there are only 30 lines. So, if there is a thirty first customer the customer will be waiting and then as soon as the telephone line gets free, the customer is going to be entertained.

So, in case if we try to observe this value or the values of capital X , which can take, then X can assume any of the integer value between 0 and 30, it cannot be 20.5. No. Or it cannot be 16.7. No. And for example, if I say that at a given time, there are five lines, which are being used, this information can be indicated by writing a small x equal to 5. So, you can see here this is our symbolic representation.

(Refer Slide Time: 32:15)

Discrete Random Variables:
Example:
A batch of 1000 machined parts contains 20 that do not conform to customer requirements.
The random variable is the number of parts in a sample of five parts that do not conform to customer requirements.

19

Similarly, another example a batch of 1000 machine part contain 20 that do not conform to the customer requirements. Now, in this case, the random variable is the number of parts in a sample of 5 parts that do not conform to customer requirements. So, somebody is going to take a sample of 5 such parts and the number of defective parts will be counted, defective in the sense, that they are not conforming to the requirement of the customer.

(Refer Slide Time: 32:46)

Continuous Random Variable:
Suppose a dimensional length is measured such as vibrations, temperature fluctuations, calibrations, cutting tool wear, bearing wear, and raw material changes.
In practice, there can be small variations in the measurements due to many causes.
In an experiment like this, the measurement is represented as a random variable X and it is reasonable to model the range of possible values of X with an interval of real numbers.
The model provides for any precision in length measurements.

20

So, now, I try to take one more aspect. So, up to now, we have considered the random variable which is going to take the values like which are say 1, 2, 3, 4, 5, 6 etc. and they cannot take a value like 1.2 or 2.3 like this. So, now, I try to take the other part, where the

random variable can take the value like 2.3 and 3.5 also. So, we try to consider the concept of continuous random variables.

Suppose our dimensional length is measured such as a vibrations, temperature fluctuations, calibrations, cutting tool wear, bearing wear, raw material changes, etc., there can be many many such example. And now, suppose we are trying to see the variation. So, in practice what will happen? There can be small variation in the measurement due to many reasons, many causes. So, in such cases, in such experiments, in case the measurements are represented by a random variable X , capital X , then it is reasonable to model the range of possible values of capital X with an interval of real numbers.

For example, the variation in the say temperature fluctuation can be between say, 1 degree and say 5 degrees and this any value between 1 and 5 degrees. For example, 1.1, 1.11, 1.112, 1.113 or 3.1, 3.2, 3.3 any value can be taken by a random variable. So, this type of model provides an information about the precision in length measurement.

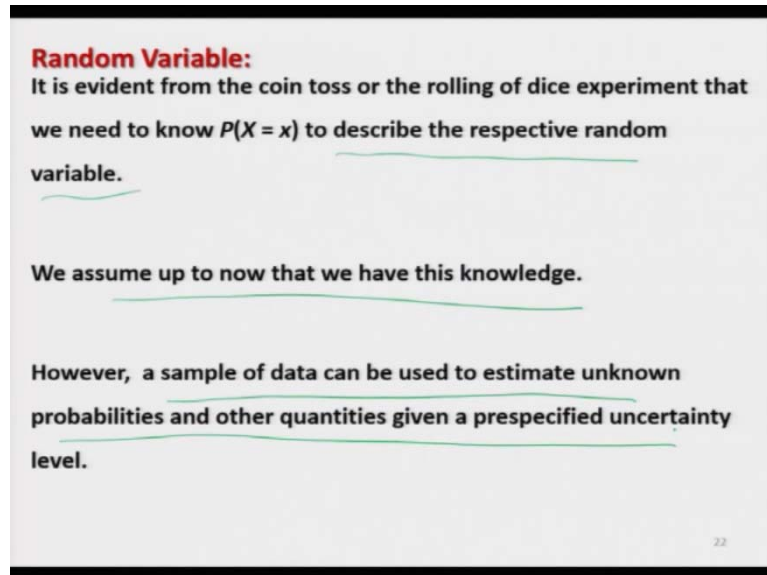
(Refer Slide Time: 34:30)

Continuous Random Variable:
There also exist random variables that take on a continuum of possible values.
These are known as *continuous* random variables.
One example is the random variable denoting the lifetime of a bulb, when the bulb's lifetime is assumed to take on any value in some interval (a, b) , $a > 0$, $b > 0$.

So, now, you can imagine that there also exists a random variable that take value on a continuum of possible values. These are known as continuous random variables. One example of this continuous random variable is by indicating the lifetime of a bulb. Now, if I say the lifetime of a bulb can be between this is 0 hours and say 1000 hours. That can be anything, we do not know. And then we have seen that in our homes also whenever we are trying to put a new tube light or a LED light or a bulb, anything, we do not know that how

long it is going to last. So, we can say that the bulb lifetime is assumed to take on any value in some interval say (a, b) where a and b are greater than 0.

(Refer Slide Time: 35:33)



So, now, once you have understood these two concepts of continuous and discrete random variables, can you recall that when we took the example of a coin toss or the rolling of dice experiment, that we had associated our probability with every value of the random variable to describe the respective random variable? So, up to now, we have assumed we actually assumed that this type of information, that is the information about the probability of such an event, that is known to us. So, we assume up to know that we have this type of knowledge.

However, if this knowledge is not there, one simple option is to draw a sample of data, consider a sample of data and use it to estimate the unknown probabilities and other quantities with a given pre specified uncertainty level. That is our basic objective, this is what we try to do. That whenever you are trying to work in data science, the process is so complicated that you just cannot compute the probability and you do not know the probability, you have to observe the data and then you have to compute such probabilities on the basis of a sample.

Once you obtain it on the basis of sample, there will always be some uncertainty. So that uncertainty should be known to you. So, that is what I am trying to say here that we try to estimate such probabilities.

(Refer Slide Time: 36:54)

Random Variable:
 A sample space is continuous if it contains an interval (either finite or infinite) of real numbers.

A sample space is discrete if it consists of a finite or countable infinite set of outcomes.

23

So, now, I can conclude this lecture by saying that our sample space is continuous, if it contains an interval, which can be either finite or infinite of real numbers. And a sample space is discrete, if it consists of finite or countable infinite set of outcomes. This is what you have to keep in mind.

(Refer Slide Time: 37:15)

Random Variable:
 More generally, we can say that it is mandatory to know $P(X \in A)$ for all possible A which are subsets of R .

If we choose $A = (-\infty, x]$, $x \in R$, we have

$$\begin{aligned}
 P(X \in A) &= P(X \in (-\infty, x]) \\
 &= P(-\infty < X \leq x) \\
 &= P(X \leq x).
 \end{aligned}$$

Handwritten notes: $(-\infty, x]$ above the first equation, and x is value below the last equation.

This consideration gives rise to the definition of the cumulative distribution function.

24

And now in case if I try to finally conclude it on this slide, I can say that in general, we can say that it is mandatory to know the probabilities. The probabilities for example, X belongs to a subset A , which is actually a subset of R . That is mandatory for us. So, now, in case if we try to choose here capital A to be like this, this interval, open interval minus infinity to close interval say here up to x , where x is belonging to a real number, then we have probability of

X belonging to A is given by probability that X belong to this interval and probability that X belongs to this interval from minus infinity to x . This can be written by here like this statement that probability that capital X is lying between minus infinity and x .

And this can be written as probability that capital X is smaller than x , that means the random variable and less than or equal to its value. So, now, if you try to see we can compute or converts such type of statement into probabilistic statement and this type of considerations gives rise to the definition of the cumulative distribution function. That we are going to discuss in the next lecture. So, now, we come to an end to this lecture.

But before leaving, let me tell you very honestly, I am dealing here with the data sciences, where you are going to deal with values. When we are trying to define the random variables or discrete random variables or continuous random variables, we have got a very strong mathematical theoretical definitions, which are coming from the measured theory. So, we have a measure theoretic approach for the probability. And actually, that is the most general definition. And this measure theory actually gives us all the fundamental for the development in the probability theory.

And whatever we are doing here in the probability theory, there are only a particular case of those definition which are defined in that measured theory. So, those who are from a statistics background, those who knows about that definition, they should not get confused. My idea here is very simple, how are we going to view these concepts in real data set? Well, in case if I have to teach a class of probability theory, or if I have to teach a class on measure theory, I will possibly express these things in a very different way. So, that is what you have to keep in mind and do not get confused.

So, you try to think about this phenomena, try to settle down inside your mind, that is most important. Because every time now whenever you have to do something, first you have to define a random variable and then you have to take a call whether it is continuous or discrete. And now, I will try showing you in the next coming lectures that the tools are different for continuous and discrete random variables. So, you try to think about it and I will see you in the next lecture, till then, goodbye.