

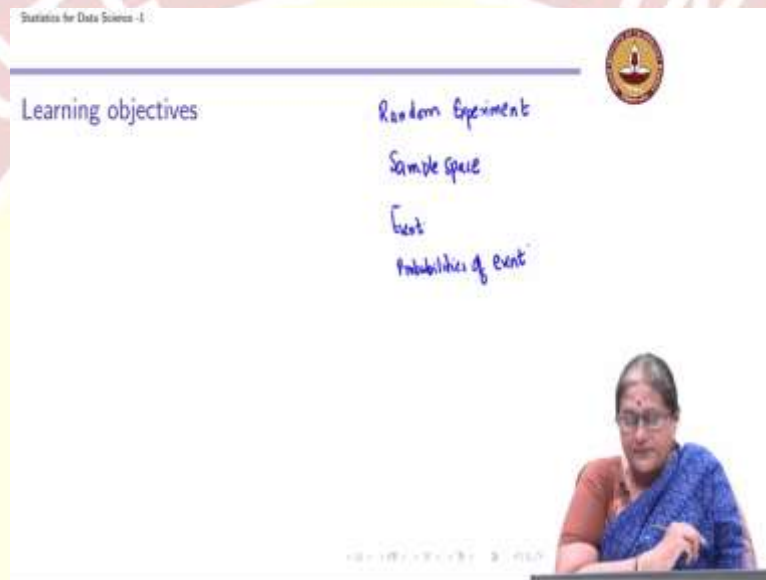
IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Professor. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras
Lecture No. 8.1
Discrete Random Variable

In this week, we are going to learn about the important concept of a Random Variable.

(Refer Slide Time: 00:22)



So, what we have learned so far is we have heard learned about what we call a Random Experiment. After a random experiment, we defined what was a sample space. A sample space is what we defined as a set of all outcomes of this random experiment. Then we define the notion of an event, which is a subset of a sample space. And then we define probabilities of the event.

When we define probabilities of an event, we approach that through the axiomatic approach to define probabilities. Now, what are we going to do after this then we introduce the notion of conditional probability we introduced what was based here. But basically what we have done is we have started introducing this entire framework of probability through the notion of random experiment and sample space.

So, by now, all of you should be really comfortable in computing probabilities of events. To help you compute the probabilities of events. We also introduced you to permutations and

combinations. So today, and the next 2 weeks, we are going to focus on the following. So, we are going to learn about random variables.

(Refer Slide Time: 02:00)

Statistics for Data Science - I



Learning objectives

1. Define what is a random variable.
2. Types of random variables: discrete and continuous.
3. Probability mass function, graph, and examples.
4. Cumulative distribution function, graphs, and examples.
5. Expectation and variance of a random variable.



In particular, after the end of these 2 weeks, you should be able to know what is a random variable. And then you are going you need to understand what is a discrete random variable, what is a continuous random variable. When you are talking about discrete random variable, which is going to be the focus of the first 2 weeks, we will be introducing notions of a probability mass function, how the graph of a probability mass function would look and then afterwards, we'll also introduce the notion of a cumulative distribution function.

Now, when we talk about both the probability mass functions and accumulator distribution functions, we are going to give a lot of examples, which might be very useful. Finally, we are going to introduce important concepts, namely the expectation and variance of random variables.

(Refer Slide Time: 03:01)



Random variable

Example: Rolling a dice twice

Example: Tossing a coin three times

Example: Application- life insurance



And we are going to see how we can apply this in our day to day lives. So, we begin with the notion of a random variable. So, what is the notion of a random variable? Let us start by revisiting a random experiment.

(Refer Slide Time: 03:19)



Random Variable

$$S = \{H, T\}$$

$$S = \{1, 2, 3, 4, 5, 6\}$$



So, recall when we talk about a random experiment, for example, when a toss a coin, I have a random experiment, the sample space is head and tail. When I roll a die once I can have any 1 of these outcomes, 1, 2, 3, 4, 5, and 6, any 1 of these 6 outcomes can happen. So, we say that every

time I can use the notion of a sample space to describe the outcomes of the experiment, but many time I might not be just interested in knowing about the outcome of a experiment.

(Refer Slide Time: 03:59)

Statistics for Data Science - I
1- Random variable

Random Variable

- ▶ When a probability experiment is performed, often we are not interested in all the details of the experimental result, but rather are interested in the value of some numerical quantity determined by the result.
- ▶ For example, in rolling a dice twice, often we care about only their sum of outcomes and are not concerned about the values on the individual dice.

Handwritten notes on the slide:

$$S = \{ (1,1), (1,2), \dots, (1,6), (2,1), (2,2), \dots, (2,6), \dots, (6,1), (6,2), \dots, (6,6) \}$$

The slide also features a small circular logo in the top right corner and a video inset in the bottom right corner showing a woman in a blue sari.

When a random experiment or a probability experiment is performed. We are not interested in the actual outcome. But we might be interested in the value of some numerical quantity determined by the result, I repeat, we might not be interested in the outcome itself, but we might be interested in the value of some numerical quantity that is determined by the result. What do we mean by this for example, when I roll a dice twice, I know there are 36 possible outcomes.

For example, I could have a 1 in my first toss, I could have 1 in my second toss. I could have 1 in my first toss a 2 in the second toss. I could have a 1 in a first toss and a 6 in the second toss. I notice that my sample space would have 36 such outcomes. This is something which we have already seen in our earlier lessons. But suppose I do not care about the individual outcomes, but I am interested only in the sum of the outcomes.

So, what is the numerical value or quantity I am associating and associating, and I am only bothered about the some of the outcomes.

(Refer Slide Time: 05:33)

Statistics for Data Science - I
1. Random variable



Random Variable

- ▶ When a probability experiment is performed, often we are not interested in all the details of the experimental result, but rather are interested in the value of some numerical quantity determined by the result.
- ▶ For example, in rolling a dice twice, often we care about only their sum of outcomes and are not concerned about the values on the individual dice.
 - ▶ That is, we may be interested in knowing that the sum is 7 and may not be concerned over whether the actual outcome was (1, 6), (2, 5), (3, 4), (4, 3), (5, 2), or (6, 1).
- ▶ These quantities of interest, or, more formally, these real-valued functions defined on the sample space, are known as **random variables**.
- ▶ Because the value of a random variable is determined by the outcome of the experiment, we may assign probabilities to the possible values of the random variable.



In this case, for example, I am not concerned, I am only bothered about whether the sum of outcome is 7. But I am really not concerned about whether that 7 has arisen because of outcome 1, 6, or 2, 5, or 3, 4, or 4, 3, or 5, 2, or 6, 1. I am not concerned about it, but I am interested in knowing that the sum is 7. Now, whenever I am talking about such a numerical quantity, these numerical quantities of interest, more formally, they are actually functions defined on the sample space, I am not going into the mathematical rigor of defining such functions.

In advanced courses, you will be subjected to the mathematical rigor. But what I want you to understand this with every outcome of the sample space, I am associating a quantity, a numerical quantity. So, this quantity of interest is what I refer to as a random variable. Since these random variables are, again, the values of the random variables are determined by the outcome of a random experiment.

And I can actually assign values to these outcomes of the random experiment, I can assign probabilities to the possible values of the random variable. So, let us look at an example to understand what we mean by numerical quantities.

(Refer Slide Time: 07:19)

Statistics for Data Science - I
Random variable
Example: Rolling a dice twice


Rolling a dice: Sample space

Experiment: Roll a dice twice

The sample space for this experiment is

$$S = \left\{ (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \right\}$$

(i, j)
i: First Roll
j: Second Roll



Let us revisit the experiment of rolling dice twice. So, when I roll a dice, twice, that is I have a 6 sided die. And I am rolling it twice. You already seen, I could have 36 possible outcomes. And those 36 possible outcomes are listed in the form of my sample space, which is given here I have s, I have 36 possible outcomes, it could be a (1, 1), (1, 2), up to (6, 6), by (i, j) I mean, i in the first toss, or the first roll of the dice, and j is outcome of the second roll of the dice. So, I am throwing a dice twice on my rolling a dice twice. So, this is something which you have already seen, we have defined events on this.

(Refer Slide Time: 08:21)

Statistics for Data Science - I
Random variable
Example: Rolling a dice twice

Rolling a dice: Sample space

Experiment: Roll a dice twice


The sample space for this experiment is

$$S = \left\{ (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,4), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6) \right\}$$

Consider the probabilities associated with the two questions

- Of the outcomes, how many outcomes will result in a sum of outcomes as 7?
- Of the outcomes, how many outcomes will have the smaller of the outcomes as 3?

$(1,5) \rightarrow 1$
 $(4,3) \rightarrow 3$
 $(3,3) \rightarrow 3$





Rolling a dice: Sample space

- ▶ Experiment: Roll a dice twice
- ▶ The sample space for this experiment is:

$$S = \left\{ \begin{array}{l} (1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), \\ (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), \\ (3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6), \\ (4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6), \\ (5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6), \\ (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6) \end{array} \right\}$$
- ▶ Consider the probabilities associated with the two questions
 1. Of the outcomes, how many outcomes will result in a sum of outcomes as 7?
 2. Of the outcomes, how many outcomes will have the smaller of the outcomes as 3?
- ▶ Notice, the experiment and sample space used to answer both the questions are the same.



Now suppose I am asking two questions. So, I am asking the questions out of out of these outcomes, how many outcomes do we have we have 36 outcomes? How many outcomes will result in a sum of 7? Notice I am just interested in the final value of 7. I am not interested in the individual outcomes as such. The other question I am interested in knowing is how many outcomes will have the smaller of the outcomes as 3.

What do I mean by smaller of the outcome, if I have an outcome (1, 5), the smaller of these outcomes is 1, whereas and (4, 3), the smaller of the outcome is 3. And for convenience, if it is (3, 3), I am going to take it as 3, both the outcomes are the same, we say the smaller of the outcome is itself. So, these two are the questions I am interested in answering. So, how do we answer these 2 questions?

So, these questions, I notice the experiment is the same, the sample space is the same. The questions are actually based on the outcomes and associating some numerical quantity. 1 is the sum of outcomes. And 1 is what is the smaller of outcomes. So, it is important for us to notice that the random experiment and the sample space that I am going to use to answer both these questions are the same. So, what is the random variable?

(Refer Slide Time: 10:06)

Statistics for Data Science - I
Random variable
Example: Rolling a dice twice

► Let X denote the sum of outcomes of the two rolls.
► Let Y denote the lesser of the two outcomes. If the outcomes are the same, the value of the outcome is taken as value of Y .

S: $\{(1,1), (1,2), \dots, (6,6)\}$

	X	Y
$(1,1)$	2	1
$(1,6)$	7	1
$(6,6)$	12	6
$(3,4)$	7	3

So, now let us start with the same example, let me denote the sum of the 2 rolls by X . Let me denote Y to denote the lesser of the 2 outcomes. So for example, I know this is my sample space, I have $(1, 1)$, $(1, 2)$, up to $(1, 6)$, and I have up to $(6, 6)$, this is what we have listed earlier. This was a sample space given here.

So, I am just writing the sample space again here. And I am telling X is denoting the sum of outcomes. So if I have $(1, 1)$ is my outcome, the value X will take is 2, which is $1 + 1$. If I have $(1, 6)$ is my outcome, the value X would take a 7, which is $1 + 6$, the value $(6, 6)$ is my outcome, the value X would take is 12, which is $6 + 6$. So, X is denoting the sum of outcomes. Similarly, Y is denoting the lesser of 2 outcomes.


So, what is the lesser of these 2 outcomes. I said, if the outcomes on both the rolls are the same, I am going to take it as itself, the lesser of $(1, 6)$ is again a 1, the lesser of $(6, 6)$ is a 6, suppose I had $(3, 4)$, the value of X would be a 7 and value of Y would be 3. So, this is how we are defining our X and Y here.

(Refer Slide Time: 11:52)

Statistics for Data Science - I
Random variable
Example: Rolling a dice twice

▶ Let X denote the sum of outcomes of the two rolls.
▶ Let Y denote the lesser of the two outcomes. If the outcomes are the same, the value of the outcome is taken as value of Y .

Outcome	X	Y	Outcome	X	Y	Outcome	X	Y
(1,1)	2	1	(3,1)	4	1	(5,1)	6	1
(1,2)	3	1	(3,2)	5	2	(5,2)	7	2
(1,3)	4	1	(3,3)	6	3	(5,3)	8	3
(1,4)	5	1	(3,4)	7	3	(5,4)	9	4
(1,5)	6	1	(3,5)	8	3	(5,5)	10	5
(1,6)	7	1	(3,6)	9	3	(5,6)	11	5
(2,1)	3	1	(4,1)	5	1	(6,1)	7	1
(2,2)	4	2	(4,2)	6	2	(6,2)	8	2
(2,3)	5	2	(4,3)	7	3	(6,3)	9	3
(2,4)	6	2	(4,4)	8	4	(6,4)	10	4
(2,5)	7	2	(4,5)	9	4	(6,5)	11	5
(2,6)	8	2	(4,6)	10	4	(6,6)	12	6



So, for the entire problem, we can see that these are I have 36 outcomes. For $(1, 1)$, the value is 2, $(1, 2)$ is 3, $1 + 3$ is 4, so 4, $2 + 6$ is 8, the minimum of $(1, 1)$ is 1, $(1, 6)$ is 1, $(2, 1)$ is again, 1, $(2, 2)$ is 2. So, you can see that $4 + 1$ is a 5, $4 + 2$ is a 6, $4 + 6$ is a 10. And here you can see the outcome, $5 + 1$ is a 6, minimum of $(5, 1)$ is a 1, so minimum of $(5, 6)$ is a 5, $5 + 6$ is 11. $6 + 6$ is at 12. And the minimum of $(6, 6)$ is a 6.

So, you can see that for each 1 of these 36 outcomes of my sample space, I have a value of X , and I have a value of Y , which is associated with it. Now why what, what, what next. Now, these outcomes are outcomes of a random experiment. And associated with each one of these outcomes is the value of X and the value of Y . So, when I have the outcomes, and these values, I can talk about associating, or I can talk about the following is, can I talk about probability of the values X taking the values?

Now let us go back here, X is taking a set of values, Y is taking set of values. On closer inspection, I can see X takes the value 2 takes the value again. Does it take the value 1, it does not take the value 1, it takes the value 2, it takes the value 3, I have an X taking a value here I take X again taking a value here, then I have yes, 4 yes, it takes a value 4 here it takes a value 4 here. 5 yes 5, 1. So, you can see that X is taking the value 5 yes 6, 7, 8, 9, 10, 11.

So, it is taking 7, 8, 9, 10, 11 and 12. So, if you look at the values this X is taking, it is taking the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12. These are the values this X is taking. I can see that X

is taking the value to when the outcome is (1, 1), it is taking the value 3 when the outcome is (1, 2), and it is taking the value when it is (2, 1), it is taking the value 4 when the outcomes are (1, 3), then it is also (2, 2), and it is taking when it is (3, 1). So, this way you can see that it is taking a particular value, where I have more than I could have only 1 outcome corresponding to the value as in this case, and in this case, or I could have more than 1 outcomes corresponding to the value. So, what are the values X is taking?


(Refer Slide Time: 15:32)

Statistics for Data Science - I
 Random variable
 Example: Rolling a dice twice

Sum of rolls of the dice:

- Let X denote the sum of outcomes of the two rolls.
- X takes the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12.

Value of X	Relevant event
2	$\{(1, 1)\}$
3	$\{(1, 2), (2, 1)\}$
4	$\{(1, 3), (2, 2), (3, 1)\}$
5	$\{(1, 4), (2, 3), (3, 2), (4, 1)\}$
...	...
9	$\{(3, 6), (4, 5), (5, 4), (6, 3)\}$
10	$\{(4, 6), (5, 5), (6, 4)\}$
11	$\{(5, 6), (6, 5)\}$
12	$\{(6, 6)\}$



X is taking the values, which are which is denoted as the sum, it is taking the value 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12. These are the values X is taking. And however defined X , I have defined X to be the sum of the outcomes of both the throws or both the roles of the dice. Now given this, now, let us do the following. I see what is the value of X . Now, let me take the value of X to be 2. Now, what is the relevant event, I know X will take the value too, when the outcome is (1, 1). So, the relevant event that is corresponding to 2 is just the event (1, 1), there is no other way X can take the value 2.

Now X can take the value 3, if my first outcome is a 1 and second outcome is a 2, first outcome is a 2 and second outcome is a 1. So, these is again the relevant event, which will give me the value of X equal to 3. So, if I am going to look at each one of them, I get 2 with the event (1, 1), 3 I have the event which is (1, 2) and (2, 1), 4 I have the event (1, 3), (2, 2) and (3, 1), this the

sum of these outcomes, which I can define as a relevant event will give me the sum of 4, 5, it is going to be (1,4), (2, 3), (3, 2) and (4, 1), and so forth, they keep going.

And then you can see 9, it is going to be (3, 6), (4, 5), (5, 4) and (6, 3), all of them add up to 9. For 10, I see the events are going to be (4,6), (5, 5) and (6, 4), for 11, it is going to be (5, 6) and (6, 5), and 12 I see the relevant event is again, just (6, 6). So, what we have done is first we have associated a value with each of the outcomes. And then I have just nabbed I have seen that the X takes the value 2 through 12, X takes these values, I have mapped the relevant event for each value X takes.

Now what is the point of mapping this relevant event? We recall what we said is because these random variables are defined on my probability sample spaces, what I can actually see and what we have defined earlier is the following is because the value of a random variable is determined by the outcome of an experiment; we may assign probabilities to the possible values. Now come back to the example what are the possible values of the random variable? I see that the possible values of the random variable are 2 through 12.

(Refer Slide Time: 18:58)

Statistics for Data Science - I
 1. Random variable
 1. Example: Rolling a dice twice

Sum of rolls of the dice

- Let X denote the sum of outcomes of the two rolls.
- X takes the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12.

Value of X	Relevant event
2	$\{(1, 1)\}$
3	$\{(1, 2), (2, 1)\}$
4	$\{(1, 3), (2, 2), (3, 1)\}$
5	$\{(1, 4), (2, 3), (3, 2), (4, 1)\}$
...	...
9	$\{(3, 6), (4, 5), (5, 4), (6, 3)\}$
10	$\{(4, 6), (5, 5), (6, 4)\}$
11	

Handwritten notes on the slide:

- $P(X=2) ?$
- $P(X=3) ?$
- $P(X=12) = ?$

A video inset shows a woman in a blue sari speaking.

These are the values it can take the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12. So, the question is, can I assign a probability of X taking the value 2. Can I assign the probability of X taking a value 3? Can I do this for all the values X is taking? That is the question we are asking. To answer this

question we are finding, and we are mapping to each value that X is taking what is the relevant event? We have already seen how to come up with probabilities of events.

(Refer Slide Time: 19:42)

Statistics for Data Science - I
 Random variable
 Example: Rolling a dice twice

We say X is a random variable taking on one of the values 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 with respective probabilities

Probability of X	Probability of relevant event	Probability
$P\{X=2\}$	$P\{(1,1)\}$	$\frac{1}{36}$
$P\{X=3\}$	$P\{(1,2), (2,1)\}$	$\frac{2}{36}$
$P\{X=4\}$	$P\{(1,3), (2,2), (3,1)\}$	$\frac{3}{36}$
$P\{X=5\}$	$P\{(1,4), (2,3), (3,2), (4,1)\}$	$\frac{4}{36}$
$P\{X=6\}$	$P\{(2,4), (4,2), (3,3), (5,1), (1,5), (6,1), (1,6)\}$	$\frac{6}{36}$
$P\{X=7\}$	$P\{(3,4), (4,3), (5,2), (2,5), (6,1), (1,6)\}$	$\frac{6}{36}$
$P\{X=8\}$	$P\{(4,4), (5,3), (3,5), (6,2), (2,6), (7,1), (1,7)\}$	$\frac{6}{36}$
$P\{X=9\}$	$P\{(3,6), (4,5), (5,4), (6,3)\}$	$\frac{4}{36}$
$P\{X=10\}$	$P\{(4,6), (5,5), (6,4)\}$	$\frac{3}{36}$
$P\{X=11\}$	$P\{(5,6), (6,5)\}$	$\frac{2}{36}$
$P\{X=12\}$	$P\{(6,6)\}$	$\frac{1}{36}$

Once we do this mapping, we can see that the probability of X is the same as a probability of the relevant event. So, what is the X value of X ? I am looking at value of X , X takes the value 2. I am interested in knowing what is probability $X = 2$, I know $X = 2$, is the relevant event test, just this set, I need to know what is the probability of this set happening. And we know the probability of this happening is $1/36$. This is something which we have already seen in an earlier discussions.

Similarly, if X takes the value 3, I am interested in knowing what is the probability of $X = 3$? So, I know what are the outcomes that give me $X = 3$. So again, I go back, and I see it is $(1, 2)$, and $(2, 1)$, these are the, this is the irrelevant event. And I know the probability of this event is $2/36$. So, if I continue in this way, I get probability $X = 2$ is $1/36$. Probability of $(X = 3) = 2/36$, $X = 4$ is $3/36$, probability $X = 5$ is $4/36$.

I can keep continuing it this way. And I can verify that probability $X = 9$ is $4/36$, $X = 10$ is $3/36$. Probability of $X = 11$ is $2/36$. And probability $X = 12$ is $1/36$. So, what we have established is, I know that this X , I say X is a random variable, which takes the values 2 3 4 5 6 7 8 9 10 11, and 12. And I can also assign a probability to the value X , taking a particular value, which I get from

recognizing the probability X taking a particular value, I will find out what is the probability of the relevant event and I have assigned probabilities to each of the values X takes.

Now let us look at the other. So, I asked two questions. The first question I asked is, what is the probability of the sum equal to 7? That was a question. So, I am not interested in anything else, I just need to check whether X takes a value 7, I see that X takes a value 7, I need to know what is the probability X would take the value 7, I know probability X would take a value 7 is same as probability (1, 6), (2, 5), (3, 4), (4, 3), (5, 2) and (6, 1), which is same, which would be $6/36$.

So, I am not interested in the individual outcomes, but the probability that the sum is equal to 7. And I know that that happens with a probability of $6/36$. So, what we have done is we have defined what is a random variable, and we have assigned probabilities to that random variable.

(Refer Slide Time: 23:17)

Statistics for Data Science - I
Random variable
Example: Rolling a dice twice

Lesser of the two values

(i, j) (j, i) $Y = i$ $Y = j$

► Let Y denote the lesser of the two outcomes. If the outcomes are the same, the value of the outcome is taken as value of Y .



Lesser of the two values

- Let Y denote the lesser of the two outcomes. If the outcomes are the same, the value of the outcome is taken as value of Y .

$$\begin{aligned}(1,1) &= 1 \\ (1,2) &= 1 \\ (2,1) &= 1 \\ (2,2) &= 2 \\ (2,3) &= 2\end{aligned}$$



Now on the same sample space, what do we mean by same sample space, I have again, roll 2 die, I have the same sample space. Now I am going to define Y to be the lesser of the 2 outcomes? What do we mean by it? If (i, j) is an outcome, if $i < j$, then Y will take the value i . If $i = j$ Y will again take the value i , I saw, I can define that Y takes the value i if $i \leq j$. So, what are the values this Y would take?

So for example, if I have $(1, 1)$, the value is 1, Y is equal to 1, $(1, 2)$ the values again equal to 1. $(2, 1)$, the values again equal to 1, $(2, 3)$ the value would be equal to 2. $(2, 2)$, the value is equal to 2. So, these are the values that I am associating with each of the outcomes. So, I know that for each of these 36 outcomes, I can see that again, I can go back here.

So if you go, so we can go back here, you can see here, the value is 1. So Y takes the value 1, the second value Y takes is 2. So this is where it takes the value 1, 2, it takes the value 3 takes the value 4, takes the value, 5, and 6. So you can see that on the same sample space, now I am interested in the value Y takes. So Y takes the values 1, 2, 3, 4, 5, and 6, these are the values this variable, or this Y takes.

(Refer Slide Time: 25:21)

Statistics for Data Science-I
 Random variable
 Example: Rolling a dice twice

Lesser of the two values

$Y = 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12$
 $P(Y=1)$

Let Y denote the lesser of the two outcomes. If the outcomes are the same, the value of the outcome is taken as value of Y .

Y takes the values 1, 2, 3, 4, 5, and 6.

Y	Relevant event
1	$\{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1)\}$
2	$\{(2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (4, 2), (5, 2), (6, 2)\}$
3	$\{(3, 3), (3, 4), (3, 5), (3, 6), (4, 3), (5, 3), (6, 3)\}$
4	$\{(4, 4), (4, 5), (4, 6), (5, 4), (6, 4)\}$
5	$\{(5, 5), (5, 6), (6, 5)\}$
6	$\{(6, 6)\}$

$1(Y=1), P(Y=1), \dots, P(Y=6)$

So, you can see why it takes the values 1 2 3 4 5, and 6. So, I can again repeat what I did for the earlier case. So, I see Y , with Y takes the value 1, the relevant event is when we would Y take the value 1, it takes the value 1 when my outcome is either a (1, 2), (1, 3), (1, 4) or all outcomes in which 1 appears either in the first toss, or in the second toss. So, what are the outcomes in which 1 appears either in the first toss or second toss. So, I have (2, 1), (3, 1), (4, 1), (5, 1), and (6, 1). So, these are the outcomes in which 1 appears.

And for these 2 outcomes, Y takes the value 1. So, I can say that Y is taking the value 1, the relevant event is this following, which is listed all the possible outcomes I have just discussed. Now, when we take the value 2, Y would take the value 2, when 2 is equal to (2, 2), (3, 2), (4, 2), (5, 2), (6, 2), (3, 2), (4, 2), (5, 2), and (6, 2). So, you can see that the outcomes were the lesser of the 2 outcomes is 2, and that you can see is (2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (4, 2), (5, 2) and (6, 2).

The third thing, when would Y take the value 3, again, I have a (3, 3), I will have a (3, 4), I will have a (4, 3), I will have a (3, 5), (5, 3), (3, 6), and (6, 3), which I can list in the same way, I have all these outcomes, where Y takes the value 3. Similarly, for 4, Y takes the value when it is (4, 4), (4, 5), (4, 6), (5, 4) and (6, 4). 5 and it is (5, 5), (5, 6) and (6, 5), and Y takes the value 6, only when the outcome is (6, 6).

So, you can see immediately what you notice is the value Y is taking whereas X was the sum which took the value 2 3 4 5 6 7 8 9 10 11 and 12. And I obtained what was the probability $X = 2$, $X = 3$ up to $X = 12$. So similarly, here, I have Y , which is taking value 1, 2, 3, 4, 5, 6. And I can find out based on what is the relevant event, I can also tell what is probability $Y = 1$, $Y = 2$ up to probability $Y = 6$. I can tell this similarly, like the way we did earlier. So, let us see what are the probabilities.

(Refer Slide Time: 28:26)

Statistics for Data Science - I
 Random variable
 Example: Rolling a dice twice

► We say Y is a random variable taking on one of the values 1, 2, 3, 4, 5, and 6 with respective probabilities

Y	Relevant event	Probability
1	$\{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1)\}$	$\frac{11}{36}$
2	$\{(2, 2), (2, 3), (2, 4), (2, 5), (2, 6), (3, 2), (4, 2), (5, 2), (6, 2)\}$	$\frac{9}{36}$
3	$\{(3, 3), (3, 4), (3, 5), (3, 6), (4, 3), (5, 3), (6, 3)\}$	$\frac{7}{36}$
4	$\{(4, 4), (4, 5), (4, 6), (5, 4), (6, 4)\}$	$\frac{5}{36}$
5	$\{(5, 5), (5, 6), (6, 5)\}$	$\frac{3}{36}$
6	$\{(6, 6)\}$	$\frac{1}{36}$

So, I have $Y = 1$, Y equal to 1, I have this is my relevant event. And I can see there are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 in the event. And the probability of this event happening is $11/36$ so probability $Y = 1$ is $11/36$. Similarly, the probability with which Y takes the value 2, I have 1, 2, 3, 4, 5, 6, 7, 8, 9. And all of them are equally likely. So, it is $9/36$, probability $Y = 3$, is 1, 2, 3, 4, 5, 6, 7, which will give me a $7/36$. Probability $Y = 4$ is $5/36$, $Y = 5$ is $3/36$. And finally, probability $Y = 6$ is $1/36$.

So, this was an example where we have defined 2 random variables, which are capturing different numerical quantities on the same sample space, which comes from the same experiment.

(Refer Slide Time: 29:45)

Statistics for Data Science - I
Random variable
Example: Tossing a coin three times

Tossing a coin three times: Sample space

First Second Third

$S = \{ H H H, H H T, H T H, H T T, T H H, T H T, T T H, T T T \}$

$2 \times 2 \times 2 = 8$

H H H
H H T
H T H
H T T
T H H
T H T
T T H
T T T



Now let us look at another problem. Let us look at another example. Again, this is an example which we have seen earlier, and I am tossing a coin 3 times. So, when I toss a coin 3 times I am observing what is the outcome of each of the tosses, I know the sample space in this is my first toss I record, what is my first toss, my second toss and my third toss, this is what I am recording.

So, if I look at the outcome, I could have a head, head, head, I could have a head in my first toss, head in the second toss or tail in my third toss, head in my first toss, tail in my second toss, head in my third toss, I could also have a head in my first toss, a tail and my second toss, a tail in the third toss, I could have a tail in my first toss, head in a second toss, tail in my third toss, tail, head, head; tail, tail, head and tail, tail, tail.

So, you can see that this 1, 2, 3, 4, 5, 6, 7, 8 are the set of possible outcomes that is I can have a head in my first head in the second. So, it is basically head appearing in there. So, I have 3 tosses. The first toss could be head or tail. So, there is 2 ways of happening it. Second toss also could be head or tail. Another 2 ways of happening third toss could also be head and tail, another 2, so $2 \times 2 \times 2$, which is 2^3 , which is 8 possible outcomes for this experiment.

(Refer Slide Time: 31:28)

Statistics for Data Science - I
Random variable
Example: Tossing a coin three times



Tossing a coin three times: Sample space

- ▶ Experiment: Toss a coin three times.
- ▶ The sample space for this experiment is
$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$
- ▶ Consider the probabilities associated with the two questions:
 1. Of the three tosses, how many tosses will be heads?
 2. Of the three tosses, which toss results in a heads first?, i.e first, second or third toss is a head?

Navigation icons



Statistics for Data Science - I
Random variable
Example: Tossing a coin three times



Tossing a coin three times: Sample space

- ▶ Experiment: Toss a coin three times.
- ▶ The sample space for this experiment is
$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$
- ▶ Consider the probabilities associated with the two questions:
 1. Of the three tosses, how many tosses will be heads?
 2. Of the three tosses, which toss results in a heads first?, i.e first, second or third toss is a head?
- ▶ Notice the experiment and sample space used to answer both the questions are the same.

Navigation icons



And the sample space for this experiment is hence given by these 8 possible outcomes. Now, suppose for this experiment, again, I am asking two questions. The first question I am asking is, I want to know how all of this 3 tosses how many tosses will be heads? In other words, I am counting the number of heads in each of the tosses. The second question I am asking is, of the 3 tosses, which toss results in a head first, what do I mean by it, is my first toss a head, or the second toss a head, or the third toss a head.

For example, in this outcome, my first toss is a head. In this outcome, my first toss is a head. In this outcome, again, the first toss is a head. In this, my first toss is a head. In this, my second, the

head appears for the first time in the second toss, here also the head appears in the for the first time in the second toss, here, it appears for the first time in the third toss, here, it does not appear at all.

Now, when I am counting the number of heads, I know I have 3 heads here, I have 2 heads here, I have 2 heads here, I have 1 head here, I have 2 heads here, I have 1 head here, I have 1 head here, I have no head here. With a caution and not put a number here because here, I am not telling it is a zeroth toss because I do not know whether what is a zeroth toss. So I just left for now I am calling this nil.

So, you can see that on the same sample space and with the same outcomes we have defined 2 random variables or 2 numerical quantities which we seek to answer. Again, observe that the experiment and the sample space are the same.

(Refer Slide Time: 33:33)

Statistics for Data Science - I
 Random variable
 Example: Tossing a coin three times

► Let X denote the number of heads that appear. Let Y denote the toss in which a head appears first.

Outcome	X	Y
HHH	3	1
HHT	2	1
HTH	2	1
HTT	1	1
THH	2	2
THT	1	2
TTH	1	3
TTT	0	Nil

So, now let us define X to be the number of heads that appear. And Y to be that toss in which a head appears first that is the order of the toss, whether is the first toss, or the second, or the third toss. Again, as in the previous example, for each outcome, let us find out what is the value of X and what is the value of Y . So, what do I have here? What are the outcomes I have these 8 outcomes.

In this outcome I know the value of head there are 3 heads, so value of X is 3, and the head is appearing in the first store. So, the value of Y is 1. Here I have 2 heads, head appears in the first toss so value of Y is 1, X is 2. Again access to a head appears in the first toss to and 1 head appears in the first toss and there is only 1 head so both X and Y take the value 1, here head appears in the second toss, Y takes the value 2 and there are 2 heads.

Here again head appears in the second toss Y value is 2 but there are only 1 head, head appears in the third toss and there is only 1 head here there are no heads so X takes the value 0, whereas here, I am writing nil or none because Y does not appear does not appear in any of the tosses so I am just assigning a value nil. Now this value could be a very high value or not. But what I want you to see is the Y takes a value corresponding to this, what is the value I need to give to this nil is something it could be any real value. For now, I want you to understand that these are the values that Y takes. So, Y and X are defined on the same sample space.

(Refer Slide Time: 35:33)


Statistics for Data Science - I
 Random variable
 Example: Tossing a coin three times

Number of tosses that will be heads

- Let X denote the number of heads that appear.
- X takes the values 0, 1, 2, 3

Value of X	Relevant event
0	$\{(TTT)\}$
1	$\{(HTT), (THT), (TTH)\}$
2	$\{(HHT), (HTH), (THH)\}$
3	$\{(HHH)\}$

$P(X=0), P(X=1), P(X=2), P(X=3)$





Number of tosses that will be heads

- ▶ Let X denote the number of heads that appear.

- ▶ X takes the values 0, 1, 2, 3

Value of X	Relevant event
0	$\{(TTT)\}$
1	$\{(HTT), (THT), (TTH)\}$
2	$\{(HHT), (HTH), (THH)\}$
3	$\{(HHH)\}$

- ▶ We say X is a random variable taking on one of the values 0, 1, 2, and 3 with respective probabilities

- ▶ $P\{X = 0\} = P\{(TTT)\} = \frac{1}{8}$
- ▶ $P\{X = 1\} = P\{(HTT), (THT), (TTH)\} = \frac{3}{8}$
- ▶ $P\{X = 2\} = P\{(HHT), (HTH), (THH)\} = \frac{3}{8}$
- ▶ $P\{X = 3\} = P\{(HHH)\} = \frac{1}{8}$



And for each outcome, I have a value. So, now let us look at the X that is number of heads in the outcome. So what are the values this X is taking? So again, go back here, you can see that X takes the value 0 1 2, and 3, so X takes 4 values, and what are the values X is taking? X is taking the value 0 1 2 and 3. So, now let us look at the relevant events where X takes the value 0, I know X takes the value 0 means that all the 3 tosses result in a tail. So, the relevant event is TTT. X takes the value 1 when 2 tosses, so 1 of the tosses is a head. So, I have the first toss, second toss, third toss. So, the head in the first toss or head in the second toss or head in the third toss.

So, the possible ways it can happen are the following. So, what are the possible events, the relevant event of X taking the value 1 is HTT, THT or TTH. Similarly, X is taking the value 2. So I can have my first second and third toss, H can be the first 2 tosses, or the first and the third toss or the second and the third toss. So, the relevant event is all these 3 outcomes put together, which is HHT, HTH, and THH.

Similarly, X takes the value 3, if all the 3 tosses result in head, and that can happen only in 1 way. So, the relevant event is again, HHH. So what we have done now is I know x , which is the number of heads in an outcome that appear is taking the value 0, 1, 2, 3 X takes the value 0, the relevant event is all my 3 tosses are tails. And for $X = 3$, the relevant event is all the 3 tosses are heads, and for 1 and 2, I have listed what are my relevant events?

So, as in the earlier case, I can find out what is the probability of $X = 0$, $X = 1$, probability of $X = 2$ and probability of $X = 3$, which are nothing but the probability of the relevant events. So, what

is probability of $X = 0$? Probability of $X = 0$ is same as probability of this event TTT happening and I know that that is equal to $1/8$, because my sample space has 8 equally likely outcomes. Similarly, probability of $X = 1$ is probability of this happening, which is equal to $3/8$.

Again, in the assumption of equal likelihood probability of $X = 1$ is $3/8$. What is probability of $X = 2$? Again, there are 3 outcomes probability of $X = 2$ is also $3/8$, whereas probability of $X = 3$ is just $1/8$ because there is only 1 outcome that satisfies the condition that $X = 3$.

So, what we have done here to see that from a sample space of tossing a coin, or an experiment of tossing a coin 3 times we list down the sample space define the random variable to be the number of heads we have got, what are the values his head can take, and what is the probability with which X takes those values. Now for the same sample space, we are interested in knowing which toss results in a head first. To understand this, let us go back to our earlier table.

So, if you look at this table, you can see that Y takes the values what are the Y values, Y takes the value 1 2 3 and nil these are the values Y is taking. As again now I am not assigning a numerical value here, but I could assign a numerical value, I could assign a value 0, but for now I am just writing nil because Y equal to 0, what physically I can interpret it as the zeroth toss, this does not make meaning. So, I am just writing that there is no toss corresponding to this outcome where the first (toss) head appears for the first time. So, you can see that Y takes these 4 values that is what I can see that Y takes these values,

(Refer Slide Time: 40:56)


Statistics for Data Science - I
Random variable
Example: Tossing a coin three times

Which toss results in a heads first

Let Y denote the toss in which a head appears first.
 Y takes the values 1, 2, 3, and NIL

Value of Y	Relevant event
1	(H, H, H), (H, H, T), (H, T, H), (H, T, T)
2	(T, H, H), (T, H, T), (T, T, H)
3	(T, T, T)

Handwritten notes: $Y=1$ H H H H T H T H T T



And now, again, let us see what when Y takes the value 1, when we take the value 1 again, Y would take the value 1 is same as first toss is head that is in which toss the head appears first So, again, I have my first toss, my second toss, my third toss that is 1 2 3 my first toss is a head, then my second toss could be a head or a tail, tail or a head, head or a head or tail or tail. Is there any other possibility that can happen other than this?

(Refer Slide Time: 41:41)


Statistics for Data Science - I
Random variable
Example: Tossing a coin three times

Which toss results in a heads first

Let Y denote the toss in which a head appears first.
 Y takes the values 1, 2, 3, and NIL

Value of Y	Relevant event
1	(H, H, H), (H, H, T), (H, T, H), (H, T, T)
2	(T, H, H), (T, H, T), (T, T, H)
3	(T, T, T)

Handwritten notes: H H H H H T H T T T T T T T T T





Which toss results in a heads first

► Let Y denote the toss in which a head appears first.

► Y takes the values 1, 2, 3, and NIL.

Value of Y	Relevant event
1	$\{(HHH), (HHT), (HTH), (HTT)\}$
2	$\{(THH), (THT)\}$
3	$\{(TTH)\}$
NIL	$\{(TTT)\}$

► We say Y is a random variable taking on one of the values 1, 2, 3, and NIL with respective probabilities

$4/8$
 $2/8$
 $1/8$
 $1/8$



So, you can see that if my first toss is a head, the relevant events that could happen as head with both the tosses falling as a head or head tail head or head tail, tail head or head TT. So, these are the relevant events. So, now, let us look at Y taking the value 2. So, head Y is denoting the toss in which head appears first. So, again, I look at this HHH, HHT, HTH, HTT, TTH, THT, THT, THH and TTT. These are my possible outcomes. So, here I have head appearing for the first time in the first 4 tosses. Now, in this among these 2 tosses, I have head appearing first in the second toss for these 2,

So, the outcomes that satisfy that head appearing first for the first time and the second toss is THH and THT. For the third time again, we can see for the third time, it would be only this outcome, which is TTH, and that outcome, X head appearing for the first time in the third toss is given by TTH. And then of course, in the nil, which corresponds to this outcome, because head does not appear at all, and the way head does not appear. So, head appears first does not happen in this outcome that the head appears first.

So, these are the values that Y takes Y takes values 1, 2, 3 and nil. As earlier, I can also associate probabilities with the values Y takes, and what would be these probabilities. Again, you can see that the probability of Y taking the value 1 is equivalent to the probability of this event happening. Remember, there were 8 outcomes in my sample space, I have a 4/8. Probability of $Y = 2$ items, or 2/8, $Y = 3$ is just 1/8, probability Y equal to nil again, will come back to what is this

nil is again, $1/8$. I can represent this nil with any real valued number, which makes sense but for now, I am just telling that Y takes the value nil.

(Refer Slide Time: 44:25)


Statistics for Data Science - I
 Random variable
 Example: Tossing a coin three times

Which toss results in a heads first

- Let Y denote the toss in which a head appears first.
- Y takes the values 1, 2, 3, and NIL.

Value of Y	Relevant event
1	$\{(HHH), (HHT), (HTH), (HTT)\}$
2	$\{(THH), (THT)\}$
3	$\{(TTH)\}$
NIL	$\{(TTT)\}$

- We say Y is a random variable taking on one of the values 1, 2, 3, and NIL with respective probabilities
 - $P\{Y = 1\} = P\{(HHH), (HHT), (HTH), (HTT)\} = \frac{4}{8}$
 - $P\{Y = 2\} = P\{(THH), (THT)\} = \frac{2}{8}$
 - $P\{Y = 3\} = P\{(TTH)\} = \frac{1}{8}$
 - $P\{Y = \text{NIL}\} = P\{(TTT)\} = \frac{1}{8}$



Or if you say test the value 0, you should qualify by saying that what do you mean by Y taking the value 0, by taking the value 0 means that my outcome, the head never appears first in any of the tosses. So, probability $Y = 1$ is $4/8$, $Y = 2$ is $2/8$, $Y = 3$ is $1/8$ and Y takes the value in nil is $1/8$. So, what we have seen in the earlier two examples is given a random experiment and sample space I am associating some numerical quantity to each outcome and, and using this concept of a numerical quantity to answer some things about the experiment.

For example, it could be some of the dice or the lesser of the 2 dice or the number of heads that appear in each outcome or what is a count the number or the order of the outcome for which the head appears for the first time.