Essentials of Data Science with R Software- 1

Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture 43

Bivariate Probability Distribution for Discrete Random Variable

Hello friends, welcome to the course Essential of Data Science with R Software- 1, in which we are trying to understand the basic concepts of probability theory and statistical inference. So, up to now from the beginning of the course you can see what we have done, we understood some basic commands of R software that are useful in this lecture, that now you can realize that why I have chosen some specific commands.

Then we had considered the probability theory and then we associated the probability with the univariate random variable and then we considered the discrete and continuous random variable and then we consider the probability mass functions and probability density function. And when we introduced the univariate random variable then we had discussed different types of properties like as probability function and conditional probability distribution and so on.

So, now we would like to extend it to more than one random variable and basically we will be concentrating on bivariate random variable and once we understand the setup and concepts of bivariate random variable it is not difficult to extend it to a multivariate setup. So, now the question comes what are these bivariate or multivariate random variables, how they are useful and what are their properties?

So, definitely if you try to see whenever you are trying to use the probability or random variable in a real-life data situation, you are essentially trying to model some phenomena, you want to write down the probability model, how the probabilities are distributed over the range of values of the variable.

Now, do you think that all the phenomena process they are dependent only on one variable, but they depend on more than one variables also. In many situations, you will see that the outcome is dependent not only on one variable but more than one variable. So, they can be two variables, three variables and so on. So, now the question is that how to handle these types of situation? Well now I am sure that you have understood that how the theory of probability, theory of random variables etc., they enter to help us in modeling the phenomena.

So, now whatever we have done in the univariate case we would like to extend it to a multivariate case, well I will be considering here mainly the two variables that is bivariate case, because you have to first understand the basic fundamental that how the things are happening, whether the things are extended or things are defined or fresh and once you know that how are you going to extend a concept from a univariate to a bivariate do you think that is it difficult to extend to trivariate or four variate or any multivariate thing.

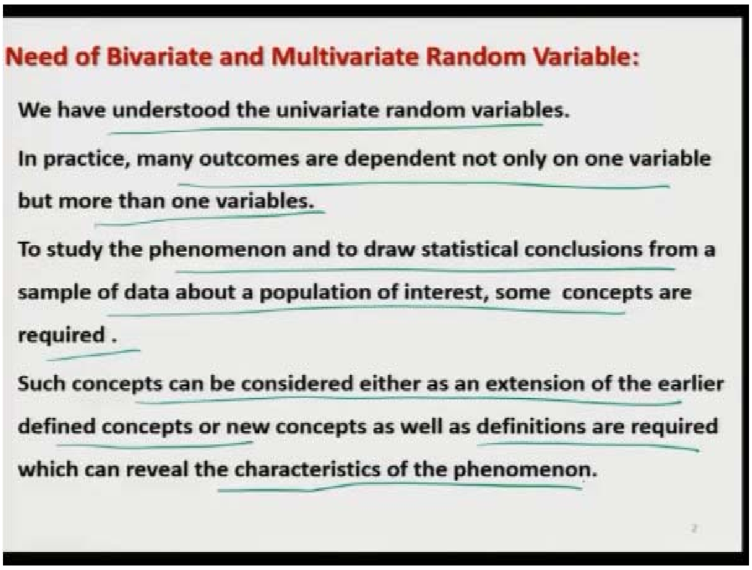
That is the same story in our childhood, my mom explained me how to sum two and three and she just gave me an exercise to add two, three and five and I was able to do it, then she asked me to do two plus three plus four plus five, then again I was able to do it because I understood the basic fundamental of addition, then even if you give me 100 terms I can do it easily.

So, that is the same story here in the multivariate setup also, so my request to you all is that please try to concentrate on the basic fundamentals, there will be some new concept, new terminologies you have to understand, what are they trying to indicate, under what type of condition you are going to use them and I am promising you that this is going to be a very useful topic for you when you are trying to deal with real data in data science.

So, as we discuss in the case of univariate, we have two types of random variable discrete and continuous, so here also we will have a similar situation that when we have more than one variable all are discrete or all are continuous or there can be a combination of them, but to make the things simpler and my objective is to explain you the basic fundamentals, so I will be taking here only two cases, when all the variables are discrete and when all the variables are continuous.

So, in this lecture I am going to talk about the bivariate probability distribution and related concepts when we have discrete random variables and after that I will show you that how you can implement it in the R software and then I will be showing you that how you can translate these things in a continuous random variables' setup. So, let us begin our lecture and let us try to understand this bivariate probability distribution for the discrete random variable.

(Refer Slide Time: 5:55)



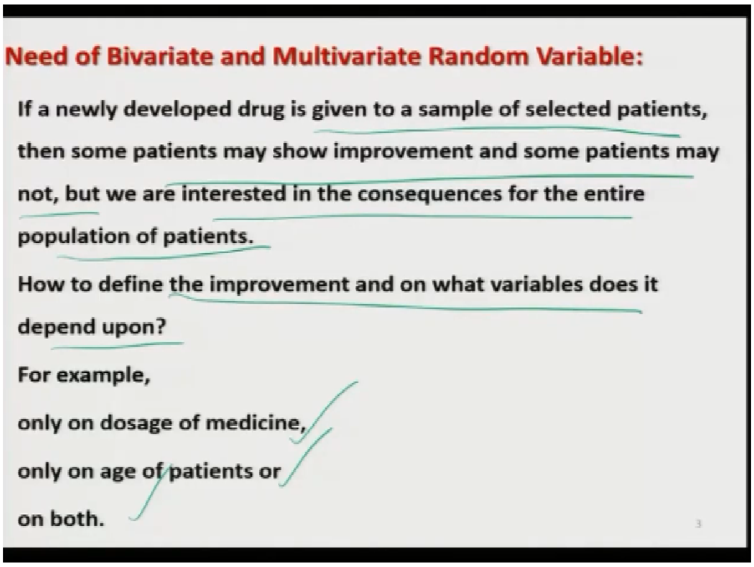
So, now the first question comes what is the need of this bivariate or multivariate random variable? So, now we already have understood the univariate random variable and in practice, many outcomes are dependent not only on one variable but more than one variables and to study the phenomena and to draw statistical conclusion from a sample of data about a population of interest some concepts are needed, some concepts are required.

Now, these concepts can be considered as an extension of the earlier defined concepts. For example, the way you have defined the PDF, PMF you have given some conditions, those condition can be simply extended to a bivariate, trivariate or a multivariate setup or the second option is this that we need some new definitions which can reveal the characteristic of the phenomena.

One thing I would like to make it here clear that when we talked about the univariate distribution, then only the values of the variable that was going to affect the outcome but think about a bivariate, trivariate or a multivariate situation there will be one difference. Suppose, if I say there are two random variables x and y, so x and y both are going to impact the outcome and that that impact has to be studied on the basis of sample of data but do you think that there is going to be one more aspect that will automatically be entering into the data, that will be the joint variation of x and y, means how x varies separately, how y varies separately and how x and y behave jointly and that is the advantage of having a bivariate or a multivariate setup.

For example, if you try to see the yield of a crop, do you think does it depend only on the quantity of fertilizer, but it also depends on the quantity of fertilizer, irrigation, temperature etc. Similarly, what do you think the mileage of a car, does it depends only on the quantity of petrol? No, it also depends on the type of engine, the driving conditions etc. So, these types of features can be captured in the statistical analysis by the consideration of bivariate or multivariate random variables and their related concepts.

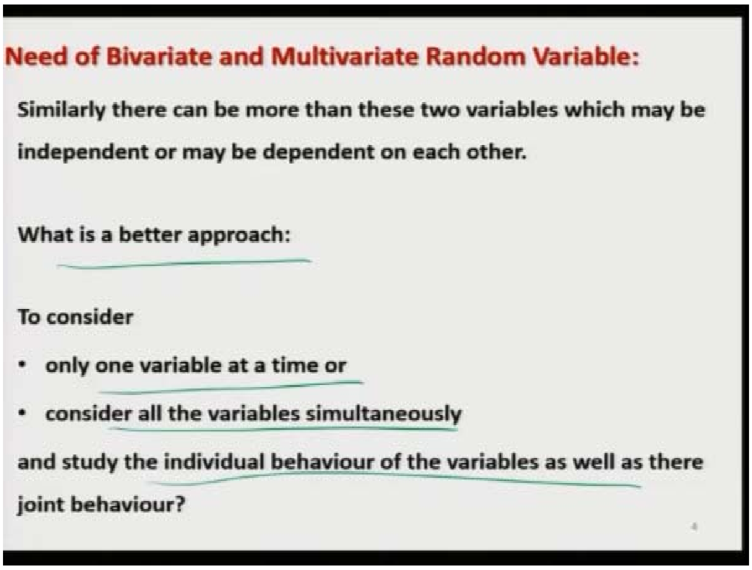
(Refer Slide Time: 8:56)



So, now for example, if you try to see here suppose, a new drug is developed and you want to study whether this drug is effective or not, so what are you going to do? You are going to give the drug or you are going to administer the drug to a sample of selected patients. Now, do you think that all the patients are going to show the change or improvement? Not really, but what will happen, some changes will be less, some changes will be more and why? Because they are dependent on the patients also, their age, their height, their weight, their health condition etc.

So, what will happen that some patient may show improvement and some patients may not but in case if you try to see this behavior only from a sample do you think that this is your objective, you want to stop here? No, we are interested in the consequences for the entire population of the patients. So, now the question is how to define the improvement and on what variable does it depend upon. For example, does it depend only on the dosage of medicine age or patient or both? So, these types of questions they can be answered by the multivariate random variables.

(Refer Slide Time: 10:30)

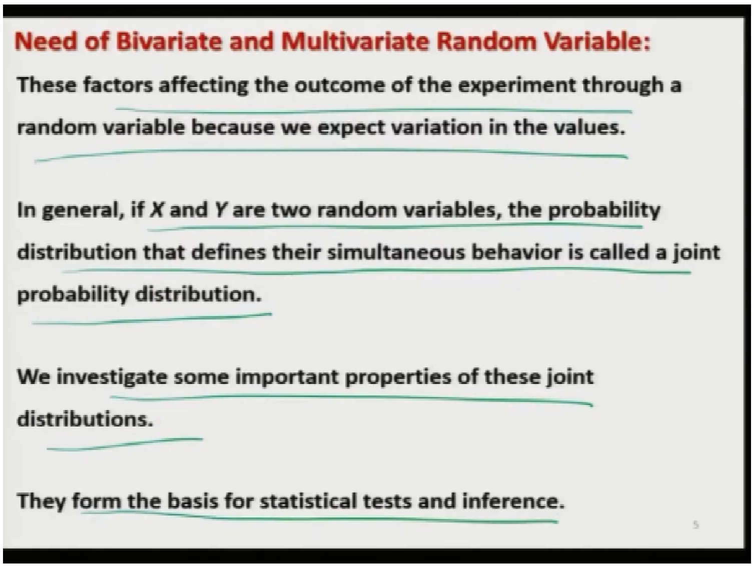


Similarly, if I try to say if there are two variables and if we want to know whether those two variables are going to impact the outcome or not and is there any joint effect present or not. For example, if I say the marks in the examination are going to depend on the number of hours of study and the price of petrol, do you think that marks are going to dependent on the price of petrol?

Well, if you just want to have a discussion for time pass, well, you can say that they are dependent on the petrol price because the student has to come to the college on the car, scooter or bike and the price of petrol increases, the person may feel bad and then the marks get bad, but even if you try to consider these types of irrelevant factors their impact is going to be extremely less but you know that the numbers of study they are really going to impact the marks in the examination.

So, with such type of analysis you can also find whether the variables are independent, whether do they have any joint effect or not and once you can have this type of information, then you can decide for what is a better approach, whether you want to consider only one variable at a time or you want to consider all the variables simultaneously and you want to study the individual behavior of the variable as well as their joint behavior. So, it all depends on you, on your objective.

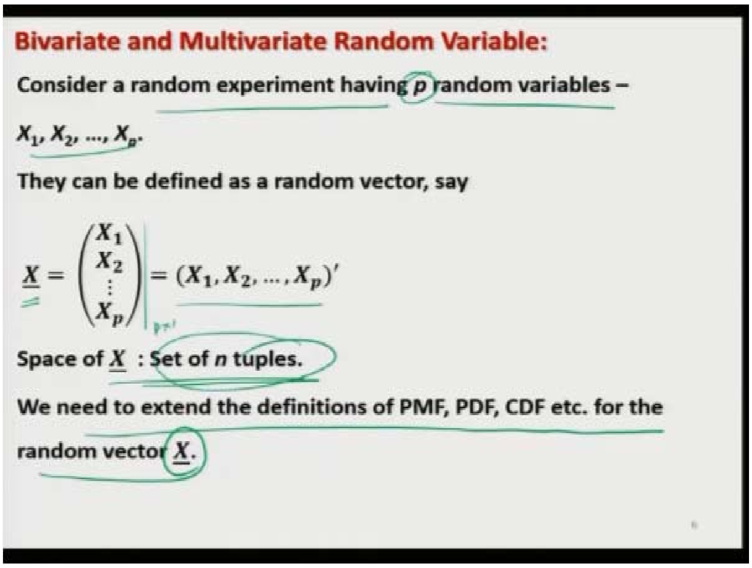
(Refer Slide Time: 12:05)



So, these factors which are affecting the outcome of the experiment they can be expressed through a random variable, why they are expressed through a random variable? Because we expect the random variations in the values of the observations. Once there is a random variation and if the variables are dependent on each other then definitely they will have a joint variation and then joint variation is also going to be random, so obviously, the concept of probability distribution enters because the probability distributions can only describe the joint behavior which is also random.

So, in general if we say that there are two random variables X and Y, then the probability distribution that define the simultaneous behavior is called a joint probability distribution and we would like to investigate some important properties of this joint distribution and these are the properties which are going to form the basis for the statistical test and inferences that we are going to learn in some time.

(Refer Slide Time: 13:19)

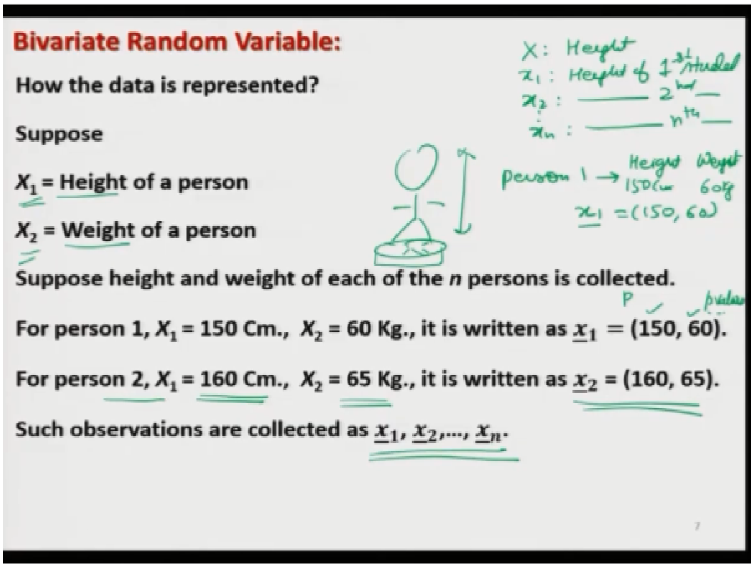


So, let me try to give you a very general structure and then I will try to show you that how these things are translated in terms of only bivariate setup. My objective is that I want to give you the idea of the univariate distribution but I want you to extend it to the multivariate setup yourself and for multivariate setup you can very easily use the concept of vectors and matrices that you have used in your mathematics.

So, now suppose, there is a random experiment which is suppose depending on p random variables, X_1, X_2,…, X_p all these variables can be defined in a random vector. So, we try to denote here a random vector by writing here X and then underscore that is the standard notations in mathematics to express a vector. This vector is from the concept of vectors and matrix, so this can be written as a column vector of order p cross 1 or this can be expressed as a row vector and a transpose.

So, the space of X vector is going to be set of n tuples, what is the meaning? I will try to explain you and we need to extend the definition of probability mass function, probability density function, cumulative distribution function etc., for the random vector X. So, what is the meaning of this set of n tuples, what does this actually mean?

(Refer Slide Time: 14:56)



Suppose, if I try to take here an example of bivariate setup that we have only here, two variables which are going to impact the outcome. So, we also have to understand that how the data will look like when we are trying to understand the bivariate setup, do you remember that we have considered this example that if X is the height this is my random variable, so x_1 is going to represent the height of first student and x_2 is going to represent the height of second student and similarly here x_n is going to represent the height of nth student.

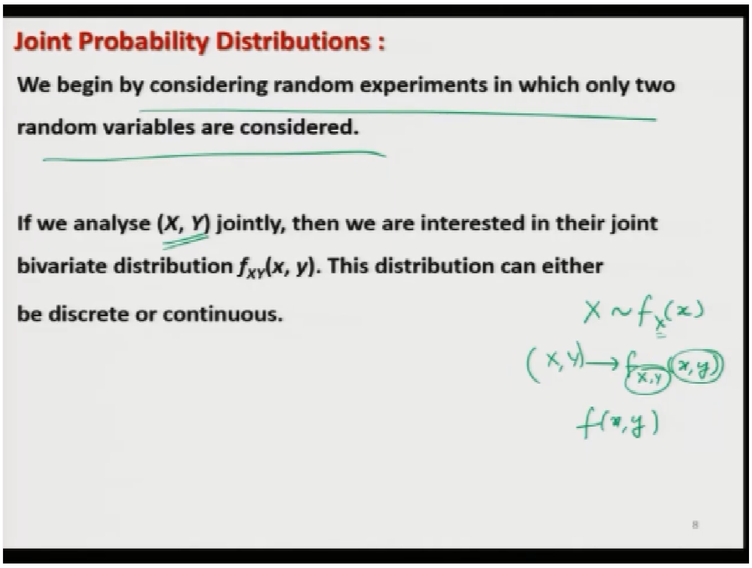
So, now this concept has to be extended for a bivariate setup, so suppose there are two random variables here height and weight. So, we denote say height by X_1 and weight by X_2. So, now there is one human being which comes and we try to measure the height and we try to measure the weight on a weighing scale like this. so, now for person number 1 we will have here two observations, one on the height and another on the weight.

So, suppose person number 1 comes and the value of the height comes out to be suppose 150 centimeter and the weight comes out to be 60 kilogram, so now we try to express this thing as the first set of observation and instead of writing here x_1 we try to write down here x1 vector.

So, this is going to be something like here 150 and 60 and similarly the person number 2 comes and the height of the person number 2 is 160 centimeter and weight is 65 kilogram, then it is written as here x_2 vector 160, 65. So, you can see here these are the observation on the random vector, so if there are two variables each pair or every observation is consisting of a pair of observation and if there are more than two variables then definitely instead of only two values there will be more value.

Suppose, if I say if there are here p variables then we will have the value like 150, 160… up to here p values and then we are going to repeat this experiment or we are going to get a small n number of observations. So, every observation will be in this case like a paired observation or like this one x_1 vector, x_2 vector, x_n vector. So, that is the understanding of the symbols and notation about the representation of the values of the random variable.

(Refer Slide Time: 18:05)

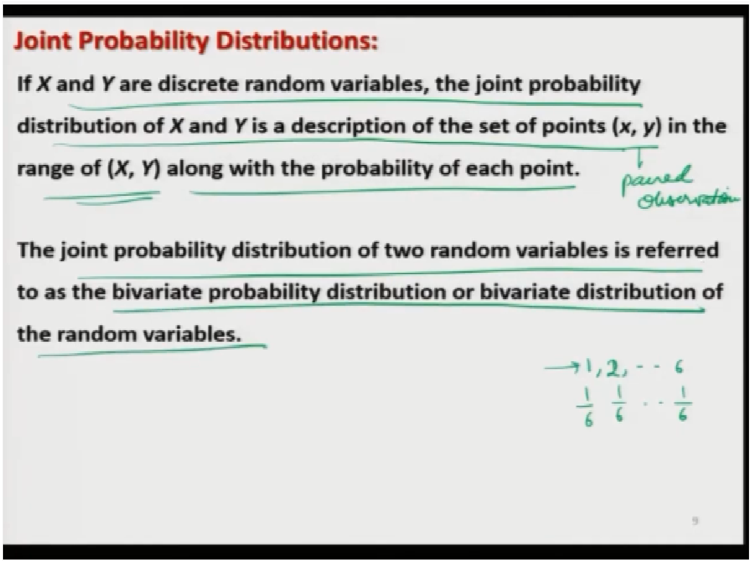


So, now we begin by considering random experiment in which there are only two random variables and we try to analyze the joint behavior of those two random variables. Let us call them as X and Y, so capital X and capital Y are going to indicate the two random variables. So, if you remember, earlier we had only one random variable X and we said the probability function either it is PMF or PDF that was going to be denoted by like as f of a small x and in the subscript there will be capital X which is indicating the corresponding random variable.

So, now in case if I have a two random variable, I can extend this definition and I can write down this symbol as here X comma Y and x comma y like this one. So, this is going to indicate the corresponding random variable and this values inside the parenthesis they will be indicating the values but just for the sake of simplicity we will simply write down here f x, y as we did in the earlier case.

So, this definition of PDF, PMF, CDF etc., I am going to simply extend from the univariate case, so it is very important for you that if you want to understand this bivariate setup you have to revise the univariate setup or at least those things should be there in your mind, otherwise it will be very difficult for you to understand this new concept and once I define this joint probability function after that I will try to introduce some new concepts.

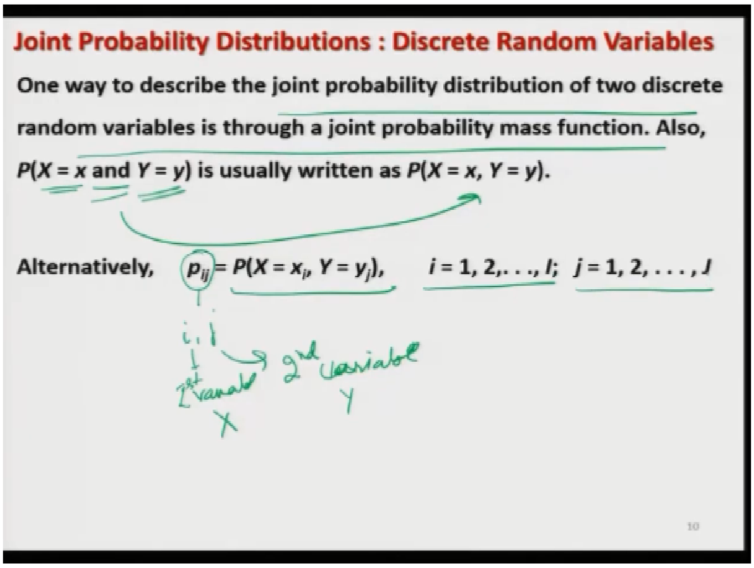
(Refer Slide Time: 19:42)



So, now this random variable X and Y they can be actually both either they can be discrete or continuous but now in this lecture first we try to concentrate on the discrete random variable. So, if X and Y are the discrete random variable then the joint probability distribution of X and Y is a description of the set of points (x, y) which are the paired observations. Paired observation means if a person number 1 comes you are going to measure the height and weight of the same person, you are not going to height or the person number 1 and weight or the person number 2.

So, these points are observed in the range of capital (X, Y) along with the probability of each point, just like if you remember we had considered the rolling of a dice we had the point 1, 2, 3, 4, 5 up to 6 and then we had indicated the probability of observing 1, 2, 3, 4, up to 6, this is 1 by 6. So, now instead of having only this univariate point we will have bivariate points or the paired observations and they are going to characterize the probability function. So, the joint probability distribution of two random variable is referred to as the bivariate probability distribution or bivariate distribution of the random variables.

(Refer Slide Time: 21:10)

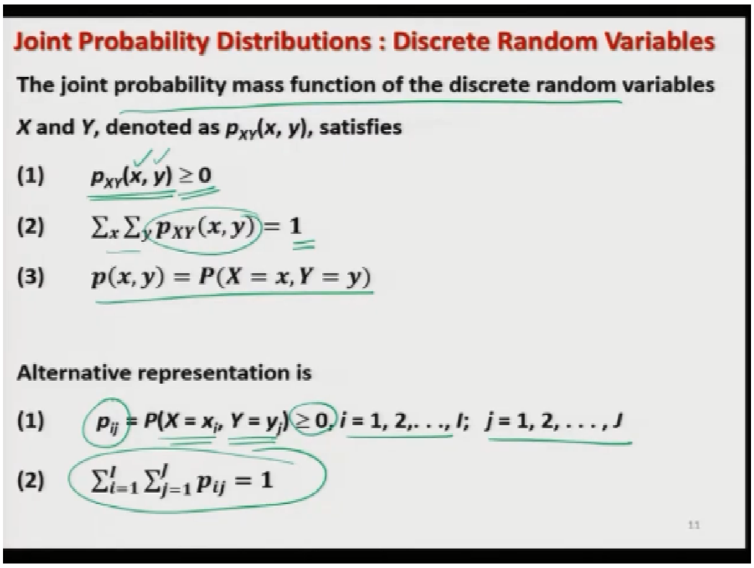


Now, we try to define this probability function or the joint probability distribution of two discrete random variables X and Y through the concept of joint probability mass function. If you remember you had defined the probability function as probability mass function when x are discrete so similarly, I can call is here as a joint probability mass function and means symbolic notation will be that if I want to find out the probability that X is equal to x and Y equal to y to write down here like this probability X equal to x comma Y equal to y.

And then you are going to observe different data point, different values of (x_i, y_i) so what we try to denote here the probability that X equal to x_i and Y equal to y_j this is going to be indicated by small p_{ij}. So, that is our symbol that if it is written here like as here ij so that means i will correspond to the first variable and this will correspond to the, j will correspond to the second variable.

So, first variable here is X and second variable here is Y and i and j they are going to take different values so just for the sake of understanding or simplicity in understanding I have taken the values of i goes from 1 to capital I and j goes from 1 to capital J so that it is not difficult for you to recall that this capital I is going to indicate the range of i and capital J is going to indicate the range of j.

(Refer Slide Time: 22:49)

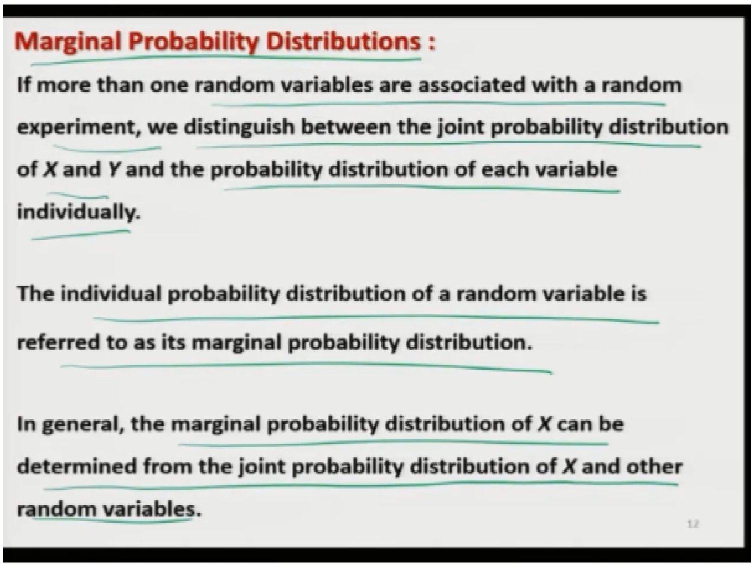


So, now do you remember the conditions that we have defined for a function to be a probability mass function, we have defined that the probability at every point should be greater than 0 and the sum of all the probabilities over the range of random variables should be equal to 1. The same definition now can be extended to a bivariate setup, how?

So, we try to define here the joint probability mass function of X and Y, so now we know what the meaning of this symbol p_{XY} (x, y)? These are the probability at the points which are observed as a small x and a small y, so this is going to be greater than 0 or equal to 0 at all the points and the sum over the values of or the range of x and y for this probability that is always going to equal to 1 that is double summation over x and summation over y p_{XY} is going to be 1 and the meaning of p_{XY} is probability that X equal to x and Y equal to y.

Another representation if you want to use the representation of p_{ij} that is x_i and y takes the value y_j, i goes from 1 to i, j goes from 1 to j then the first condition will come out to be that p_{ij} are greater than or equal to 0 and the summation over i and summation over j p_{ij} should be equal to 1. So, both this notations are possible and you can see that if you have understood the basic condition for a function to be a probability mass function then understanding this bivariate probability mass function is not difficult and this condition have been extended from univariate to bivariate.

(Refer Slide Time: 24:35)



Now, once I have defined the joint probability density function, then we are also interested in the marginal probability distribution, because you see when there are two variables X and Y they are trying to give the joint effect of X and Y on the outcome but you are also interested then in knowing that what is the individual effect of those variables.

For example, if I try to take a very simple example suppose the marks of a student, they depend on the number of hours the student studies and the numbers of hours the student plays, that we know that playing is also important for a good health and once you have a good health you can study better and then you can get better marks. So, yes, so now I have here three types of impacts, the first effect will be the effect of number of hours of study on the marks, the second effect will be the number of hours of playing on the marks and the third effect will be their joint variation.

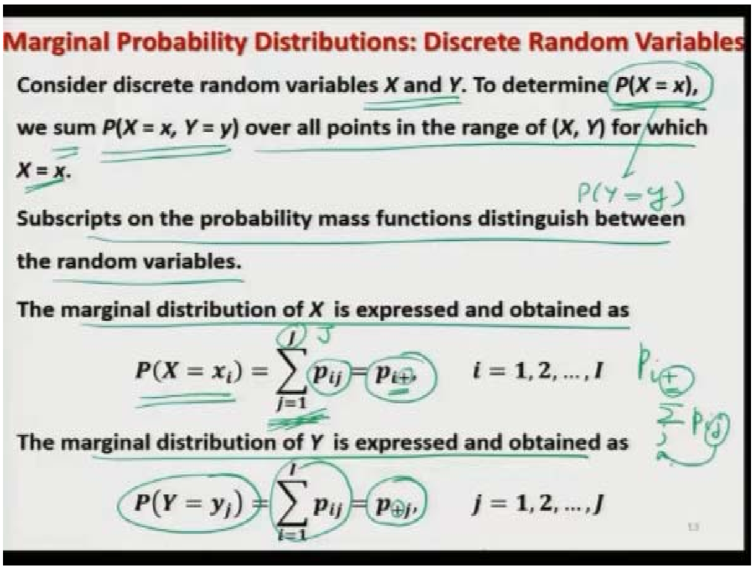
So, now you get only a numerical value in the form of paired observations, that if a student has played for 5 hours a week and then studied 10 hours a week then the marks are like this, so these marks are going to dependent on both this variable. So, now my interest will be suppose, I want to know from those marks that what is the effect only of number of hours of playing or what is the effect of number of hours of studies only, then this concept is actually called as marginal distribution.

And second concept can be what will be the marks if the student has played only 5 hours a week or what will be the marks in case if somebody has studied only for 7 hours a week, and you ask these types of question many times to your parents, teachers, friends that means if I study say 3 hours a week then what do you expect my performance will increase or not or you try to say means if I try to study 7 hours a week what do you think, that how many marks can be increased.

So, these types of question they can be answered by the concepts of marginal probability distribution and conditional probability distribution, so we try to discuss them here. So, now first we try to understand the concept of marginal probability distribution. So, if there are more than one random variable which are associated with a random experiment, we distinguish between the joint probability distribution of X and Y and the probability distribution of each of the variable individually.

So, the individual probability distribution of a random variable is referred to as its marginal probability distribution and in general, the marginal probability distribution of X can be determined from the joint probability distribution of X and other random variable Y or in the case of multi variable case there can be more than two random variables. So, the moral of the story is this that the marginal effect of the random variable has to be determined only from the jointed distribution, you are not going to get any distribution from outside.

(Refer Slide Time: 28:22)



So, let us try to first understand that how you can obtain such a marginal distribution from the joint probability mass function. So, now we have here two random variables X and Y both are discrete and suppose we want to find out the probability of X equal to x, so you can see here the joint probability, the distribution contains the observations on X and Y but we want to compute only the probability which is related only to the one variable X.

So, what we try to do here, we try to consider the joint probability X equal to x and Y equal to y and we sum this probability over all points in the range of X, Y for which X is equal to x, that is a very simple thing if I try to show you mathematically you will understand it very quickly. So, the subscript on the probability mass functions distinguishes between the random variable because you have found here probability of X equal to x and then you can also find out here probability of Y equal to y.

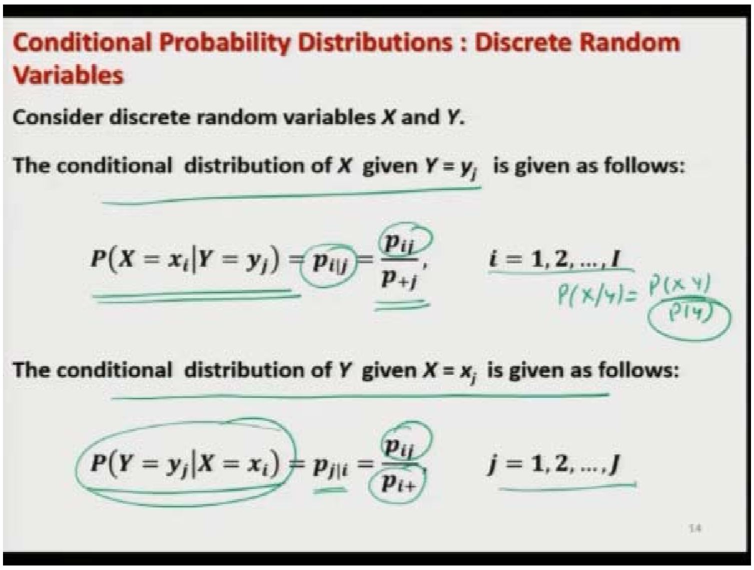
So, there are going to be two types of marginal probability distribution that are going to be associated with the joint probability function of X and Y. So, now if you try to see the marginal distribution of x can be expressed and obtained as probability that x equal to x_i where I am trying to sum this all the probabilities p_{ij} but my range of sum is over j, j goes from 1 to capital J. Remember one thing so when you are trying to find out the probability of x then you have to sum over the range of y and vice versa.

So, the marginal distribution of y can be expressed and obtained as probability y equal to y_j but now you have to sum it over the range of another variable x, so this will become here $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$ and we try to indicate this marginal distribution as probability of x equal to x_i can be indicated by p_{i+} that is the in the subscript you have to write i plus and similarly for the marginal distribution of y we are indicating it by p_{+j}, plus j is going to enter in the subscript.

So, you can see here when you are trying to indicate over the value of second subscript we try to replace it by here a sign plus and when we are trying to sum it over the first variable then we try to once again replace it by the sign plus and the location or the position of this plus sign that is going to determine that at what place you have summed up.

For example, if you tell me only like here p_{i+}, so I know that the summation has been done on this second subscript, and this is what you have done, means you have obtained here the summation over j and summation was j over what p_{ij} over this j. So, you can see here it is very easy to remember it also.

(Refer Slide Time: 31:33)

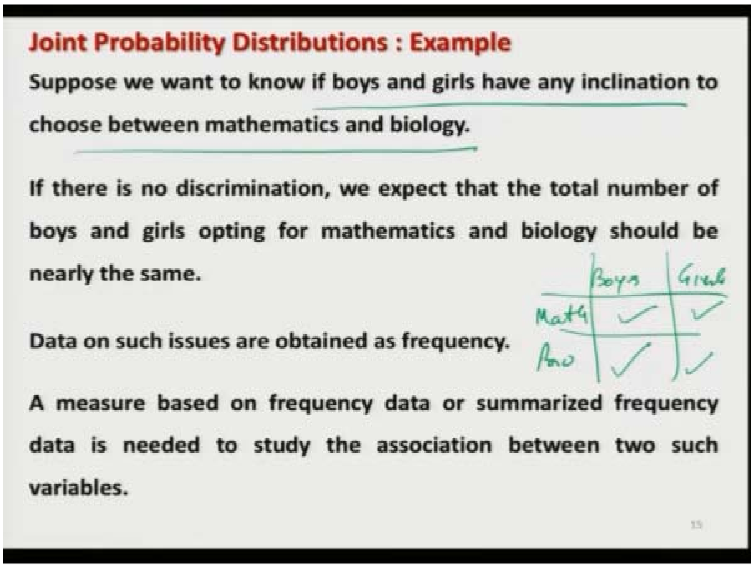


And similarly, if you want to find out the conditional probability distribution in case of discrete random variable then the conditional distribution of X given Y equal to yj is obtained as like this, that we try to write down this probability, a probability that X equal to xi given Y equal to y_j this is simply here p_{ij} upon the marginal probability distribution p_{+j}.

And in case if you want to find out the conditional the distribution of Y given X equal to x_j then it is indicated here as probability that Y equal to y_j given X equal to x_i and this is indicated by p_{ij}/p_{i+}. So, you can see here this is very similar to the conditional probability that when you wrote the probability of X given Y was equal to probability of x, y divided by here probability of here y. So, this is if you try to see this probability of y is simply your the marginal probability distribution that is what you are going to write down here.

Now, you know how to compute this probability, in the case of discrete you are going to find out them by summation and in the case of continuous you are going to take the integral, that is all, as simple as that, so it is not difficult to remember also and the probability that X given Y that is X equal to x_i, Y equal to y_j is indicated by this symbol p and in the subscript i given j and for the probability of Y equal to y_j given X equal to x_i this is indicated by $p_{j|i}$ in the subscript, and obviously it means I will go from 1 to i and j will go from 1 to j. So, now I hope that it is not difficult for you to understand these three-concept joint probability marginal probability and conditional probability.

(Refer Slide Time: 33:28)



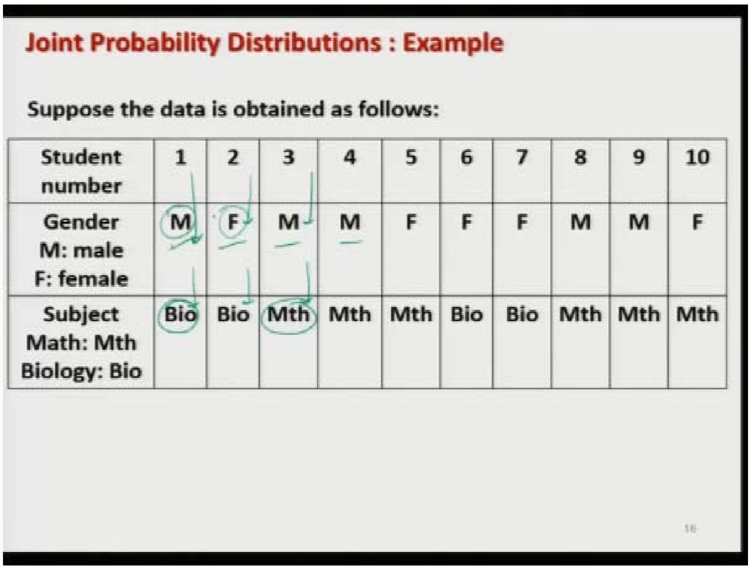
So, now I am going to take here a very simple example to illustrate this concept and in the next turn I will try to show them that how you can compute such things in the R software. So, suppose, we want to know if boys and girls have any inclination to the choice of the subject between mathematics and biology, yes, you have heard many times we say that girls prefer biology more or boys prefer mathematics more or vice versa.

So, now if you want to know this answer, what you have to do, you have to collect the data on the subjects taken by boys and girls and we expect in that sample if the number of boys and number of girls who are trying to take or opt the mathematics and biology if they are nearly the same that you cannot say that there is any discrimination between the two but suppose if you try to take a sample of 100 student and 90 percent of the girls are choosing the biology, possibility you can conclude that girls are indicating the or they are preferring the biology subject.

So, now how to conduct it such an experiment, how to get the data and how we have to compile it and how we have to infer these types of conclusion from there that is my objective in this example to explain you. So, we try to collect the data on four aspect on boys, girls and then on the subject here maths and here biology, and now these observations can be compiled in a sort of table that how many boys are going to take maths, how many girls are going to take maths, how many boys are going to take biology and how many girls are going to take biology.

So, such type of frequencies can be obtained whenever we try to work on the real life of data. So, now we need to understand this association and we try to compute different types of marginal and conditional probabilities.

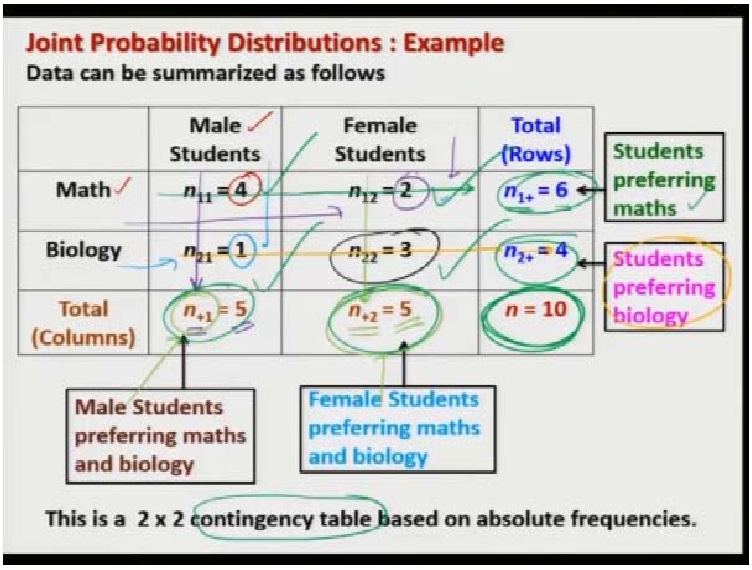
(Refer Slide Time: 35:41)

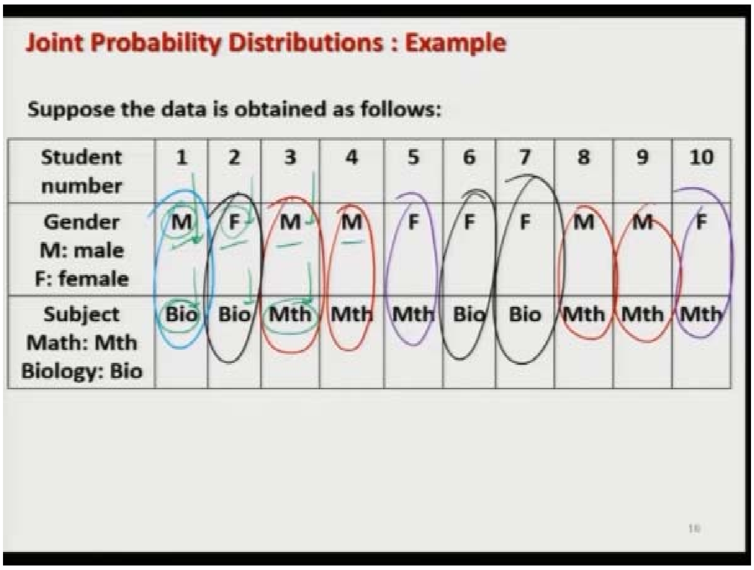


Student number	1	2	3	4	5	6	7	8	9	10
Gender M: male F: female	M	F	M	M	F	F	F	M	M	F
Subject Math: Mth Biology: Bio	Bio	Bio	Mth	Mth	Mth	Bio	Bio	Mth	Mth	Mth

So, suppose just for the sake of understanding I have taken this example, I have taken here only 10 students which are indicated by male, female, male, male etc. and then I have collected the data on their choice of subject. For example, issue number 1 this student is a male and the student has taken biology, second student is a female and she has taken biology, third is student is male and he has taken mathematics and so on we have collected the data on 10 students. So, this biology is indicated by Bio, mathematics is indicated by Math and gender is indicated by M and F for male and female respectively. Well, I have taken here only 10 observation otherwise my slide will become too clumsy.

(Refer Slide Time: 36:31)





Now, I try to simply count the numbers from here, you can see here number of male students who are taking maths, you can see here what is that number, let me try to use different color pen. What we want male and math, so male and math is here 1, male and math here is 2, male and math here is 3 and this is here 4. So, there are four students who are male and they have taken here the subject mathematics. So, I try to write down here male and math which is here four and similarly, I try to find out here what are the number of male student who have taken the biology.

So, now I try to find out here male student biology 1, then there is no other student who is male and has taken the biology. So, I will try to run down here, this male student who are taking the biology as a subject this is here 1 and similarly, we try to do the same exercise for female students. So, female student who are trying to take mathematics they are here, here you can see here 1 and here 2.

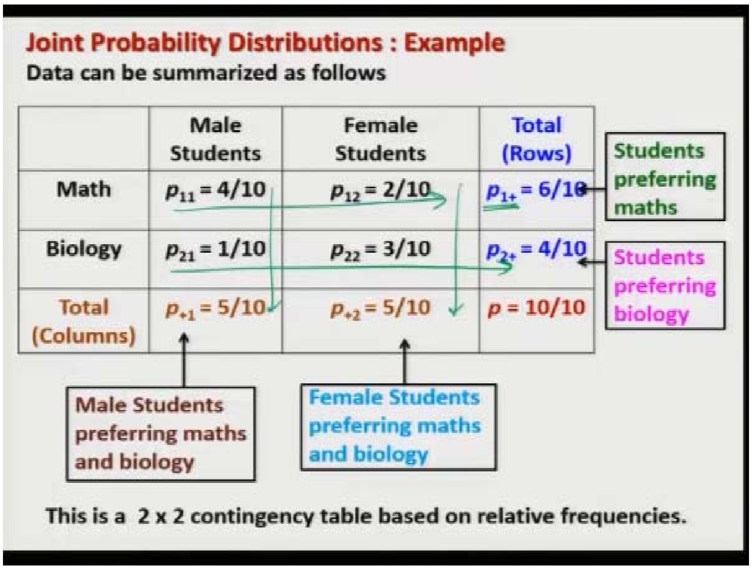
So, I try to write down here female student taking mathematics here this is here 2 and similarly here female students taking biology this is your 1, 2 and 3 and this is your given by here this 3. So, now this is how I have compiled this data and this data is called here as say 2 by 2 contingency table, that is the name and what we have done, we have actually obtained the absolute frequencies.

So, now if you try to see here I try to add these numbers row wise and column wise. So, if I come to add here row wise you can see here this is 4 plus 2 that is n_{11} plus n_{12}. So, I am trying to sum it over the second subscript, so I am indicating here is n_{1+} and which is equal to here 6 and if you try to see what is this indicating, this is trying to indicate the students who are preferring the mathematics and similarly, if I try to take here the second row, we try to add here n_{21} and n_{22} which is giving here n_{2+} which is equal to here 1 plus 3, 4. So, this is indicating the students who are preferring the biology.

And similarly, if you try to add them column wise, so if I try to add the values in the first column n_{11} plus n_{21} this is indicated by n plus 1 because we are trying to add them on the first subscript and this value is here 4 plus 1 that is 5 and similarly for the second column, we are trying to add here n_{12} and n_{22} that is indicated by here n_{+2} in the subscript and this will become here 2 plus 3 equals to 5.

So, this n_{+1} this is going to indicate the number of male students who are preferring the mathematics and biology and this number here n_{+2} that is going to indicate the female students who are trying to prefer mathematics and biology. So, this is how you can see that you can obtain different type of information from this thing and if you try to see what you have obtained here this n_{1+}, $n2_{+}$, n_{+1} and plus 2, what are these things they are the marginal frequencies and you know that if you try to divide the frequency by the total number of observations that will give you relative frequencies, so here the total number of observations here are 10.

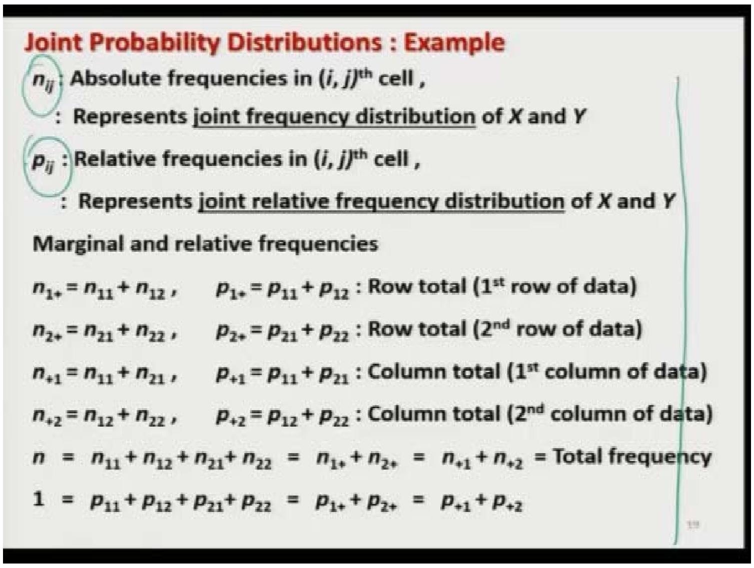
(Refer Slide Time: 40:33)



So, if I try to divide all these values this here 4, 2, 1 and here 3 by here 10 then we are getting here these types of values and these are the relative frequencies and you know that relative frequencies are trying to measure the probabilities. So, now you know that these things are very simple and if you try to add here the values in the first row that will give you here the marginal probability of the variable in the first row and similarly for the second row this will give you the marginal probability when you are trying to obtain the marginal distribution for the variable in the second row and the same thing will hold for the first and second column also.

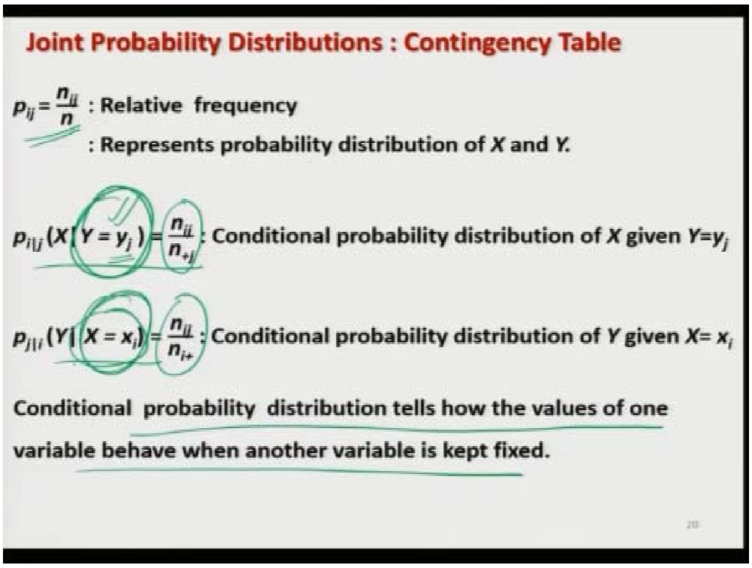
So, you can see here you can very easily compute this marginal probability distribution by arranging the data into a 2 by 2 contingency table. Now if you have more number of variable even there can be 100 variable in the row and 200 variable in the column, exactly the same concept has to be extended without any problem.

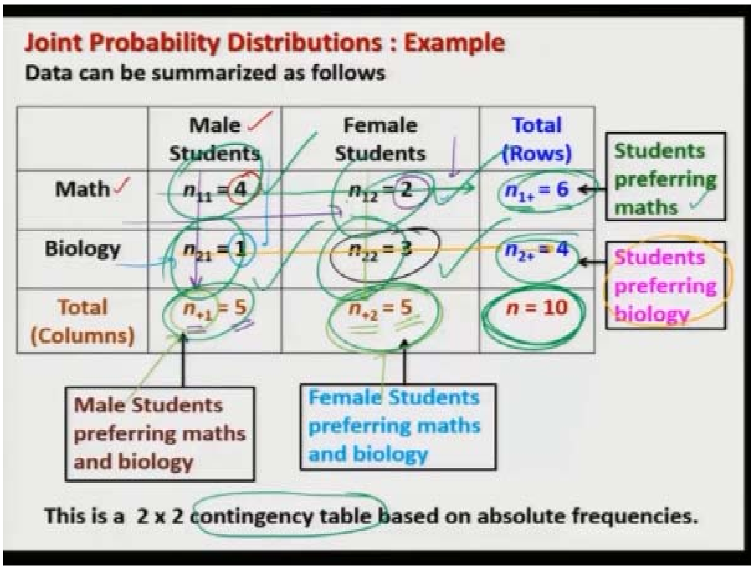
(Refer Slide Time: 41:32)



And yes, in this slide I have just written all the symbols and notation which I have just explained you so you can have a look from your slides, they are the same thing that I just explained you in this example. So, for example, you can see here this is n_{ij} is the absolute frequency, p_{ij} is the relative frequencies in the ij-th cell and so on. And 1 plus is $n_{11} + n_{12}$, so these are very simple thing.

(Refer Slide Time: 42:01)



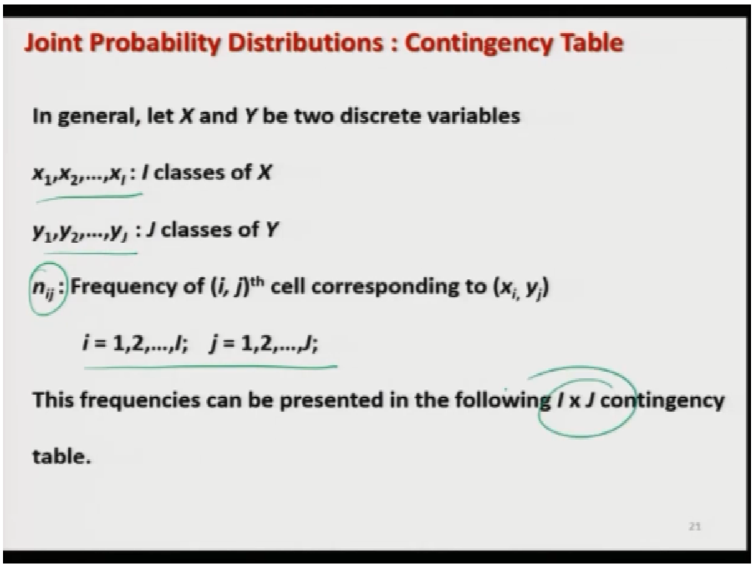


Now, in case if you want to compute here the relative frequency, you can just obtain it by here p_{ij} which is equal to here n_{ij}/n and that is actually represent the probability distribution of X and Y, so that you can also obtain from this table, you can see here that in this example these are the values which are the values of n_{11}, n_{12}, n_{21}, n_{22} and from there you can compute these relative frequencies.

Once you obtain these relative frequencies, you can very easily obtain the conditional probability distribution of X given Y or Y given X. For example, if you want to know the probability distribution of X given Y equal to y_j that is the conditional probability distribution of X for a given value of Y equal to y_j that is the y equal to y_j is fixed, this can be obtained by simply here n_{ij}/n_{+j} and similarly the conditional probability of Y given X equal to x_i that means we are trying to fix the value of Y equal to x_i that can be obtained from n_{ij}/n_{i+}.

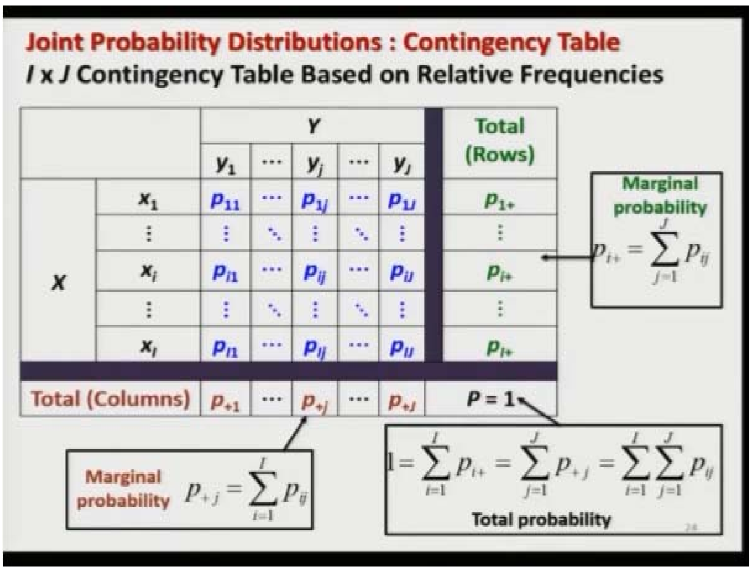
So, you can see here that this probability is not difficult to obtain from the given set of data and this conditional probability distribution tells us how the values of one variable behave when other variable is kept fixed. So, you can see here, here in this case Y equal to yj is kept fixed and in the second case X equal to xi is kept fixed.

(Refer Slide Time: 43:37)



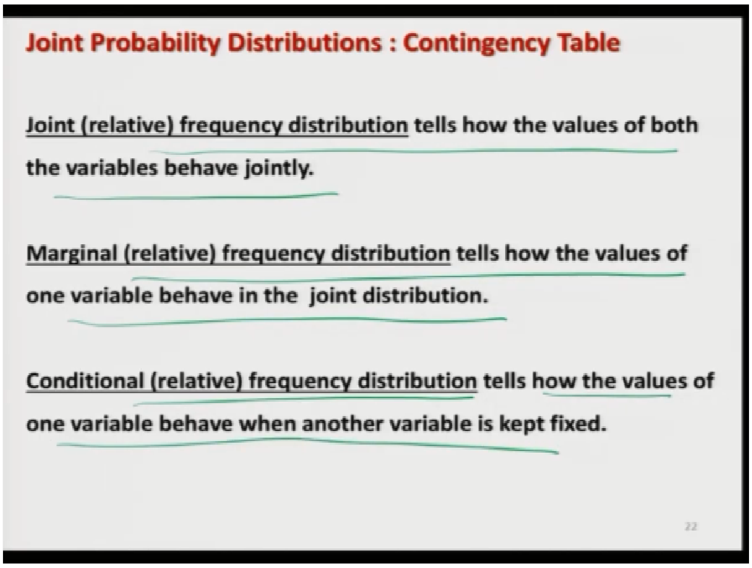
So, now in case if you simply want to extend it to a general thing, well I will not explain you in detail but I can just give you a quick idea that if you have instead of two classes you have here I classes x_1, x_2, x_i for X and y_1, y_2, y_j for J. so, there are I classes for X and J classes for Y and then n_{ij} definition will remain the same, only the range of i and j will change and these values can be represented now instead of 2 cross 2 contingency table but in i cross j contingency table.

(Refer Slide Time: 44:12)



And how they are going to be done means, I can explain you here, you can see here that these are values you can see exactly here n_{1+} , n_{2+} up to here n_{i+} n n_{+1} n_{+2} n_{+j} etc. and then you can find out their marginal probabilities by finding out the row totals, column totals and the total and then you can just divide them by the number of observation and you can obtain the same table, same I cross J contingency table based on the relative frequency exactly in the same way as I said earlier and you can obtain different types of probabilities over here.

(Refer Slide Time: 45:02)



So, now if you try to see here that the joint relative frequency distribution tells us how the values of both the variables behave jointly, the marginal relative frequency distribution tells us how the values of one variable behave in the joint distribution and conditional relative frequency distribution tells how the values of the one variable behave when another value variable is kept fixed.

So, now we come to an end to this lecture, I am sorry it was a pretty long lecture but you can take a break between the two, means you can just take a break and can watch in however many parts you want to wish, but yes, but I wanted to complete these things in a single shot, otherwise it will become difficult for you to understand and now if you try to see what we have done, we have got the data that was based on two random variables and from there we have obtained the marginal distribution that are going to help us in knowing the individual effect of those random variables, we have obtained the joint effect, we have obtained the conditional effects etc.

So, now you can see that in real life these things are going to occur very often for example, if I try to consider the same example of the shopping website, the shopping pattern is dependent on the age, income and the gender, occasion etc. So, ultimately the shopping website is interested in the sales and sales is going to dependent on all these variables like for example, during the festival time the sale is more, so that will give an idea to the shopping websites to offer some discounts, some sale etc. and similarly if somebody wants to know what is the pattern of sale only for the males or only for the females.

Or somebody wants to know what is the sale given the salary is more than 5000 rupees a month, how are you going to answer these questions? You are going to answer them by using this joint probability distribution using the concept of joint probability, marginal probability and conditional probabilities and you can take very interesting outcome which are going to be very useful for you.

So, in the next class I will try to show you that how you can compute these things automatically in the R software, so we try to think, try to have a look in the various example in the books, try to attempt them, try to solve them and I will see you in the next lecture till then good bye.