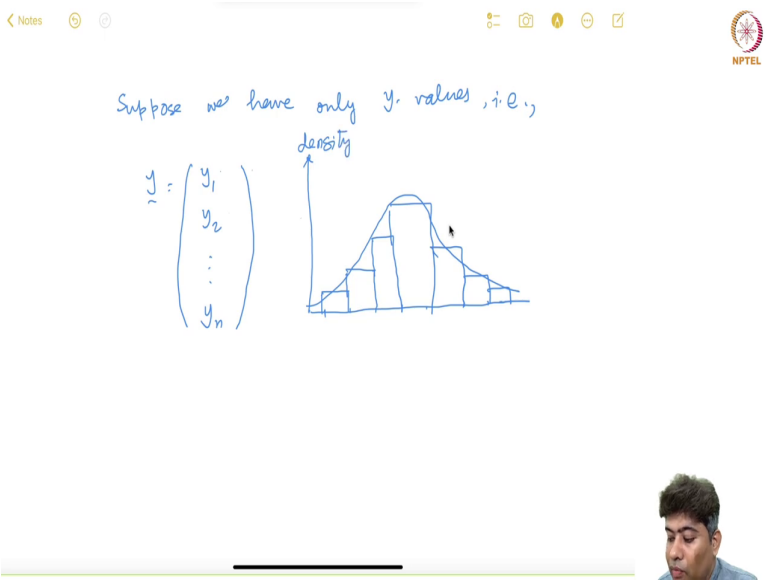
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 08
Understanding the joint probability from data perspective

Hi, now we are going to discuss a regression line as a regression function and some probabilistic aspect of regression line.

(Refer Slide Time: 00:32)

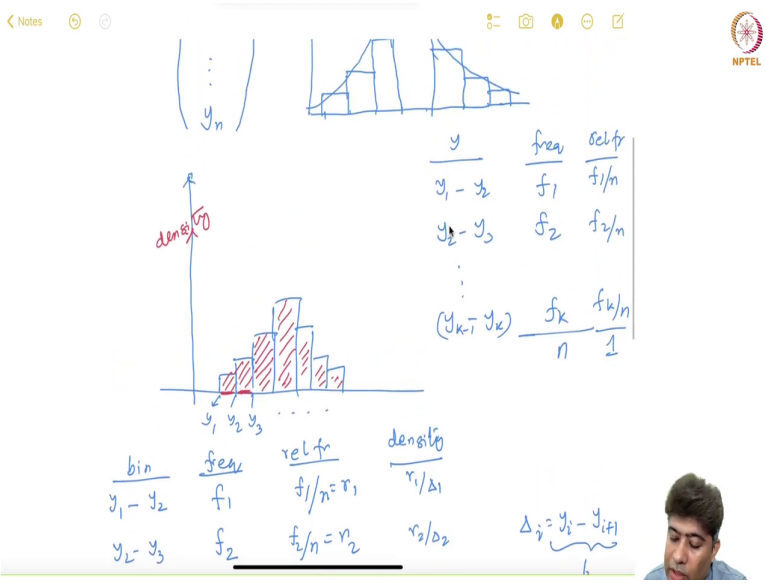


Suppose we have a value only values y, suppose we have say we have only y values that is y 1, y 2, y n. And if we want to understand how the distribution of y behaves; so, probably what we will do? We will try to figure out the frequency distribution of y. So, what we will do? We

will create some bins like this and then each bin how many samples we have we see and that will be our frequency of that bin and we generally draw a histogram out of that bin.

And this histogram gives us some sense of the distribution of the y, but if you look into the carefully, generally when we draw the histogram we generally do not draw frequency on the y axis rather we draw density.

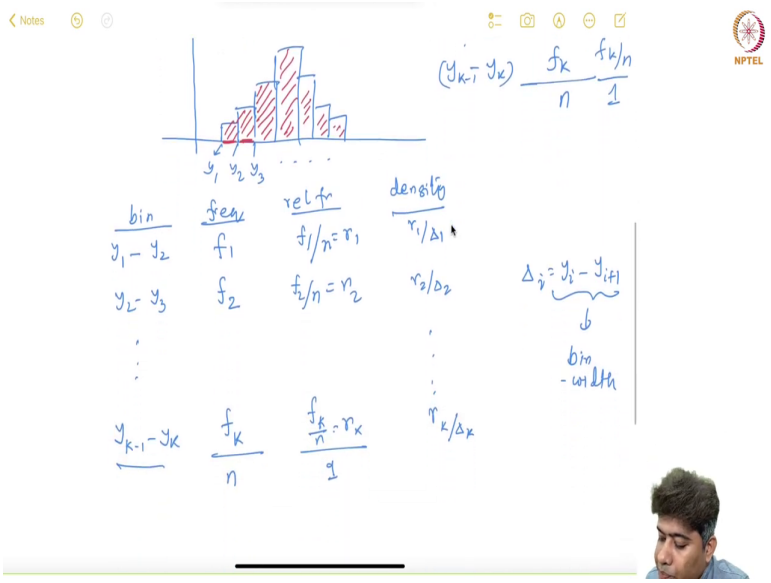
(Refer Slide Time: 02:26)



How we draw the density? Generally what happens, we first draw the you know relative frequency; that means, we calculate the how many first we can in each beams we have the frequencies. So, we have the data say y from y 1 to y 2, how many in this bin how many y values are there. So, frequency you can compute frequency 1 from y 1 to y 3, you can have how many frequencies are there in this way you can have y k minus 1 to y k in this group you have f k and total number of samples you have n.

Now, what you can do is you is the we generally instead of drawing the frequency you can always draw the relative frequency which is essentially f 1 by n f 2 by n f k by n. And if you sum them up it should be exactly equal to 1. Now, what we do generally instead of; so, we can call.

(Refer Slide Time: 04:05)



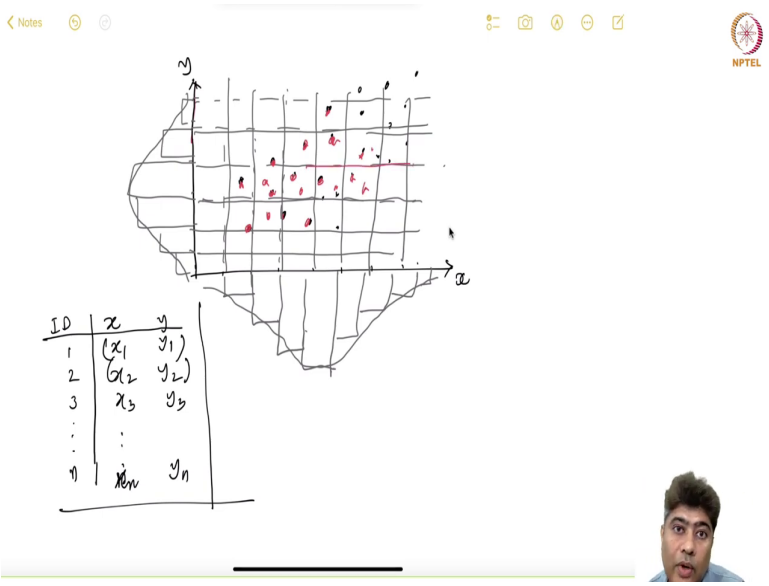
So, let us write it a little bit carefully; so, we have the bins here bins are like from say this could be y 1, this could be y 2, this could be y 3 in this way. So, y 1 to y 2 how many frequency, how many samples are there? And relative frequency is essentially f 1 by n which is r 1. Then y 2 to y 3 will be f 2 and relative frequency of this will be f r 2 which is f 2 by n.

And then y k minus 1 to y k which will have k many samples, total n samples are there in the data set and we have f k by n which is r k. So, if I sum them up all the relative frequency it will give me 1. Now, we will define density essentially densities r 1 by delta 1 r 2 by delta 2

in this way r k by delta k. Now, what is delta i? Delta i is basically y i minus y i plus 1; so, that if this bin width; so, delta is the bin width; so, delta is the bin width.

So, if I have this let me take a different color; so, this is the bin width. So, what we are doing? We are if we divide the relative frequency by the bin width then we get density and that is what we plot here on the y axis. So, the advantage is in this case we get the area of the box is the relative frequency itself. So, the total area under the all boxes will be essentially 1; so, the in the histogram the total area represents the probability idea of probability.

(Refer Slide Time: 06:58)



Now, this is for the univariate case, what happens if I have two continuous variable say x and y. Well, what generally we do is instead of a one generally we plot x on the x axis and y on the y axis and how the data will look like. So, x and y and maybe some ID variable will be

there. So, 1 2 3 up to n; so, suppose I have n variables; so, I have x 1 y 1. x 2 y 2, x 3 y 3 in this way x n y n.

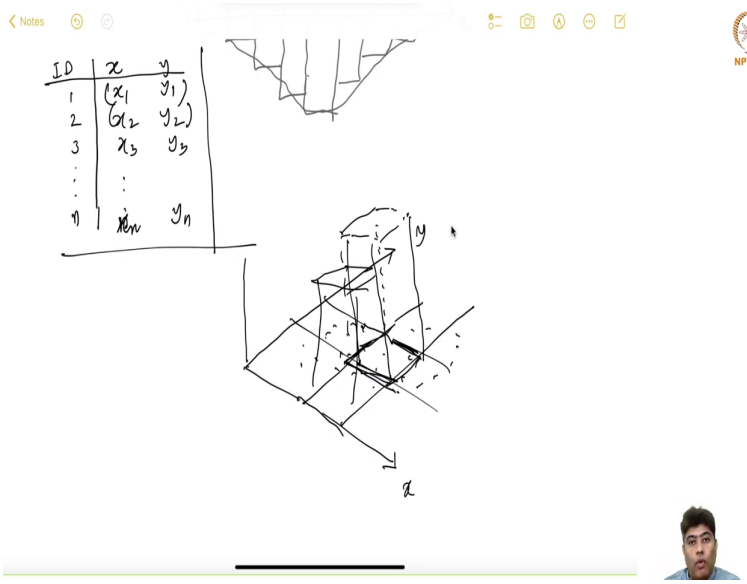
Now, this x 1 y 1, I will take this tuple and I will try to plot them somewhere here. Then x 2 y 2, I will take them and I will try to plot them somewhere here, then x 3 y 3 I will try to plot them here maybe x 4 y 4 here; so, in this way we will try to plot them on the Cartesian product; suppose this is the data that we have right. Now, the question is we can if we just want to get the distribution of x, then what we will do?

We will just ignore the y values and we will just create the bins and see how many points are there in each of these bins. So, each of these bins how many points are there and based on that we will create a histogram. Similarly, if I am interested in; so, this will give me some idea about the how the distribution of x behaves. Similarly, if I am interested in the x the y axis, and the how the distribution of y behaves what I will do? I will create bunch of bins.

And I will see how many points falls in these bins and then I will just create the frequency density relative frequency density for these each of these bins and I will draw the histogram of the y and that will give me how the distribution of y behaves. Now, the question is what is the joint distribution of x and y? How x and y behaves together? So, what we can do?

We can just create this bunch of joint bins. And now, these joint bins are not anymore bins they are almost like a boxes; so, I am creating now bunch of boxes. And each of these what we will look for in each of these boxes which points how many points are there. So, these points these are there and basically I will see we will count in each of these boxes how the points are there.

(Refer Slide Time: 11:27)



Now, what we can do; so, essentially how it will look; so, we will try to draw some kind of a 3D picture; so, our points are this is my x this is my y and suppose that is how the points are. So, now, for x for x we have bins for y we have a bins and in this bin how many points are there, based on that we will draw based on that we will draw a 3D bar on third axis.

Now, so, three dimension histogram will be created, a three dimension histogram will give us the idea how distribution of joint distribution of x and y. And; that means, that here in this case; so, when we were doing say look into this one. So, area of the box under the curve was; so, area probability was area under the curve, but in the joint distribution.

Now, the volume of the building we can think of as a building the volume of the building is our joint probability of x and y to be in that box; so, we will continue on this and in the next video.