

Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering,
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology Guwahati
Lecture 21
Histogram & Boxplot

Hello, welcome back. In the last lecture, we learned about scatter plot and bar plot. This lecture will be a continuation of that bar plot to create a boxplot, and histograms. I will start first with histogram, with an example I will explain why do you need a histogram in place of bar plot and then I will move into box plot.

(Refer Slide Time: 0:59)

How to plot marks in an exam?

Roll No	Marks
1	45
2	30
3	10
:	:
:	:
:	:
40	12

How to plot marks in an exam?



Roll No	Marks
1	45
2	30
3	10
:	:
:	:
:	:
40	12

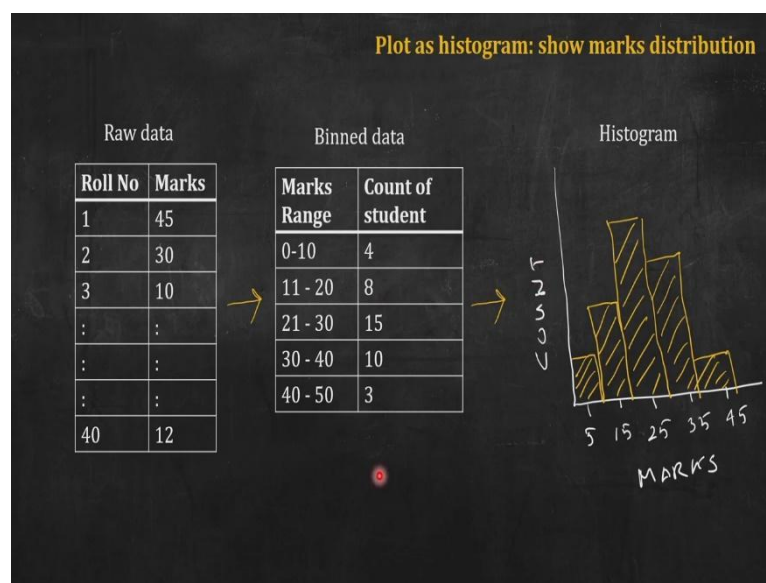
Suppose, I have a data from an examination, I have 40 students, their roll numbers are 1 to 40. And they have got different marks for out of 50. The first guy has got 45, and the second student has got 30, and so on. Now, how can I represent this data? We have learned a bar plot earlier. So, I can consider, each of this student is a category because he or she is a individual.

And then I can plot those students in the horizontal axis and their marks in the vertical axis and show them by a bar plot that will look something like this. I have just roughly drawn by hand. Now looking at this diagram, where I have marks in the vertical axis, and the roll numbers are the categories or the classes of the students in the horizontal axis.

These graph represents the data, but it actually does not give any story out of it. It does not tell me any story. What I know here? The first one has got around 45, second one has got around 30, something like that. That is, the information I get only. Now, what you, if you are looking at this data, you may be interested to know, I am not bothered about how individual student has got, that individual student will be interested to know about.

But rather from a outside, as a teacher, or somebody who is looking at the data, you will be interested to know the distribution pattern of the data. By distribution pattern, I mean, how many people have got above 40? How many people have got below 10? Something like that. And that is where histogram or frequency histogram comes into work.

(Refer Slide Time: 2:39)



So, what I have to do to create a histogram? I have to take the raw data, and then I have to bin the data. What do I mean by bin? Bin means you create some bag, buckets, for example, I

take the first bucket as the range from 0 to 10. So, anybody who has got a mark between 0 to 10, belongs to this first bucket or first bin, and then so on, I can keep on going 11 to 21, bucket, another bucket, third bucket is 21 to 30, something like that, up to 40 to 50, my last bucket.

So, these are my bins, I have created five bins. And then I look into my raw data and I count how many students are in the bin for the first one? Where, what do I mean by that? I count how many student has got marks from 0 to 10? So, the result says here in the raw data, if I look into I see that, maybe 4 student has got marks from 0 to 10. So, I put that corresponding to 0 to 10 bin. Similarly, there will be 10 student who has got marks between 30 to 40.

So, the count for the bin 30 to 40 is 10. Now I have these two variable marks range or bins, and the count of the student in each of these bins. Now I will create a bar plot for these pairwise data. What I will do? I will take the middle point of each bin, midpoint of each bin, for example, 0 to 10, the midpoint will be 5.

So for 5 I will put a bar of height 4 and that is what I have done in this histogram, then 15 is the midpoint for the second bin marks 11 to 20. And I have created a bar there of height equal to 8. In this way, I have created all the bars. And this is my histogram where I have the marks in the horizontal axis and the count number of student in each category; each bin is on the vertical axis.

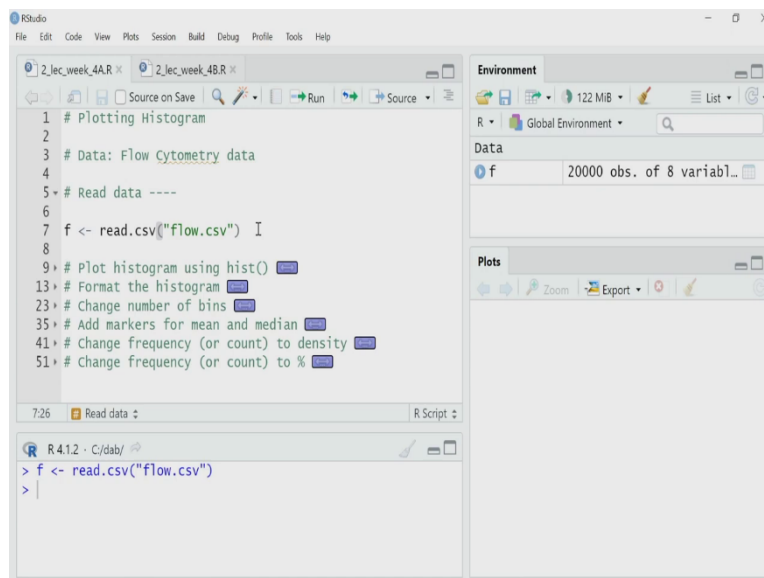
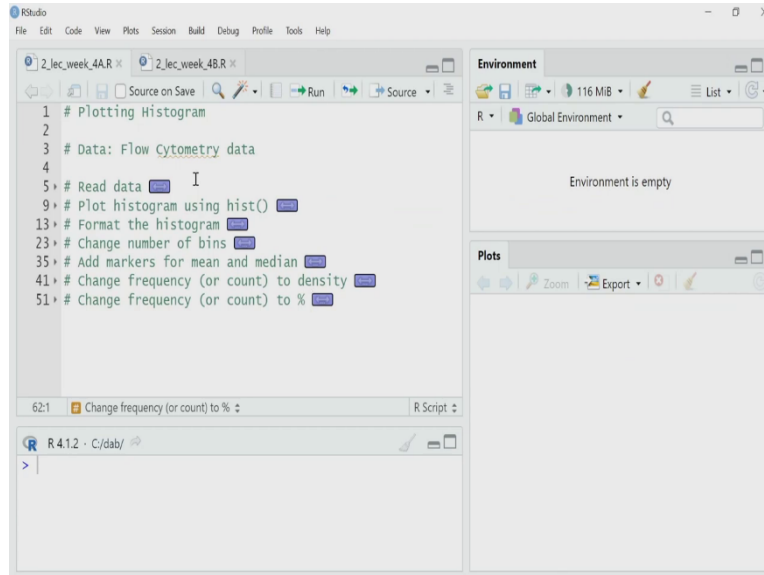
Looking at these you can easily see they are quite a good distribution of mark there is not much skew in it. Most of the people have got marks in the third bin which is 21 to 30. So, that is what you expect in a on an average class. And there are some people who are a bit outlier who have got very high marks so they are in the fifth bin and there are some people, students who are performing a bit low so they are in the lowest bin.

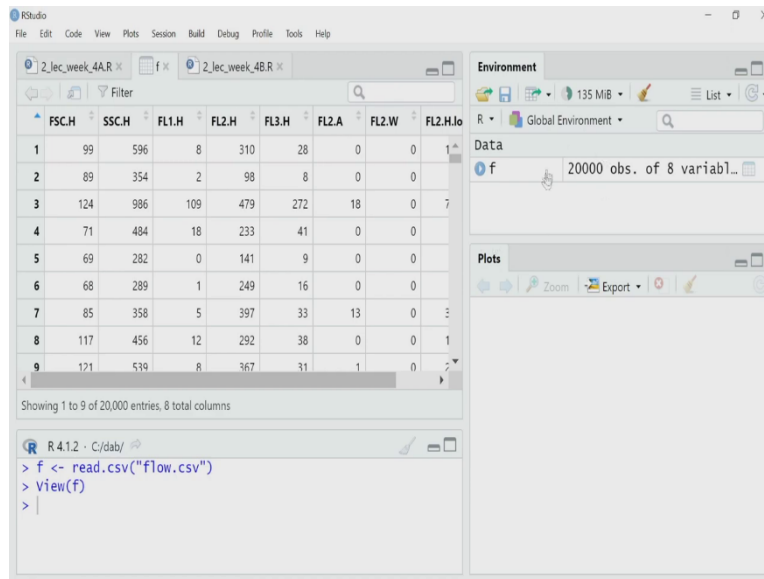
Now looking at this histogram as a whole, I have got a comprehensive information about the distribution. So, this is histogram, sometimes we may call it frequency histogram. And here I have plotted the count, raw count in the vertical axis; you can convert it into fraction also. For example, I have 40 students. So, in the first bin, 0 to 10, I have 4 students. So 4 divided by 40 is 1 by 10. So, you can plot the fraction values here on the vertical axis.

So, the height of this first bar will be 0.1. You can convert these into percentage also, if you want to say, I want to know the percentage of students in different groups. So, in that case, 0.1 is 10 percent, you can calculate the percentage from the original count data also. So, this

is histogram. So, now I will move to R studio and I will analyze the data coming from flow cytometry, and I will try to plot a histogram using R.

(Refer Slide Time: 5:52)



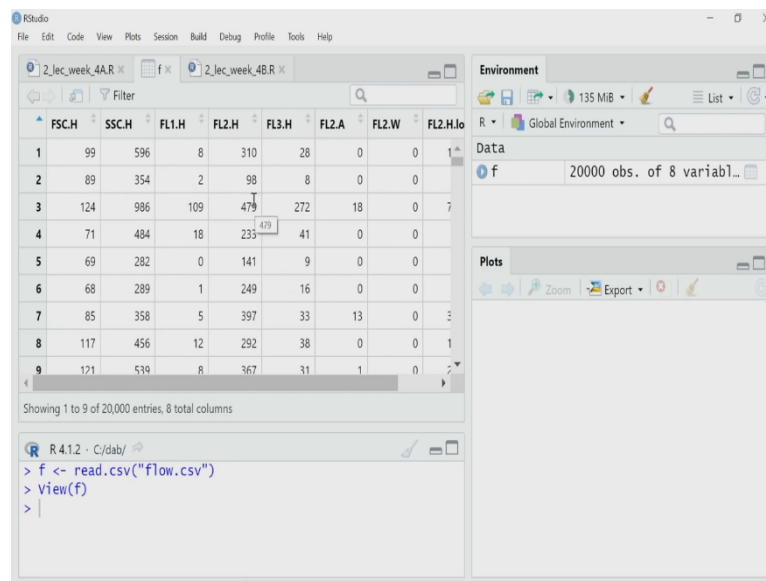


`f <- read.csv("flow.csv")`

So, I have a data set coming from flow cytometry, what I have done? I have extracted it as a csv file. So let me first read that file, load it into my space and then I will discuss what is what is there in this data. So, I am reading this flow dot csv file and storing the data in f let me open f. So, if you can see here, we have 20,000 observation, written for 8 variable. So, this 20,000 is actually 20,000 individual cells, if you remember, in flow cytometry, you assay individual cells, one at a time.

So, we have pass 20,000 cells. So, 20,000 observations are there and for each cell, we have a major 8 variable, what are those FSH, H stands for height, I will not go in detail of what do what does that mean? You must be knowing our flow cytometry.

(Refer Slide Time: 6:44)



`f <- read.csv("flow.csv")`

So, for our scatter SSC side scatter FL1, FL2, FL3, these are the fluorescence channels. So, for each of the cell, I have these data, for example, if I look into the FL2 h, 310 is the value of fluorescence in the second channel FL2 channel for the cell 1 and the unit is arbitrary. Similarly, 98 is the value of the flows and intensity for FL2 channel for the second cell something like this.

So, in this way, I have 20,000 cells. Now, if you just create a bar plot for these, it will be meaningless. Rather in this case, as you have seen in any flow cytometry data presentation, what we do? We try to create a histogram distribution of Rosen's intensity right. So, that is what I will do.

(Refer Slide Time: 7:30)

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

2 Lec_week_4A.R * f * 2 Lec_week_4B.R *

Filter

	FSC.H	SSC.H	FL1.H	FL2.H	FL3.H	FL2.A	FL2.W	FL2.H.lo
1	99	596	8	310	28	0	0	1
2	89	354	2	98	8	0	0	
3	124	986	109	479	272	18	0	7
4	71	484	18	233	41	0	0	
5	69	282	0	141	9	0	0	
6	68	289	1	249	16	0	0	
7	85	358	5	397	33	13	0	3
8	117	456	12	292	38	0	0	1
9	121	539	8	367	31	1	0	

Showing 1 to 9 of 20,000 entries, 8 total columns

```
R 4.1.2 · C:/dab/
> f <- read.csv("flow.csv")
> view(f)
> |
```

Environment

R · Global Environment

Data

f 20000 obs. of 8 variabl...

Plots

Zoom Export

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

2 Lec_week_4A.R * f * 2 Lec_week_4B.R *

Filter

	FSC.H	SSC.H	FL1.H	FL2.H	FL3.H	FL2.A	FL2.W	FL2.H
239	67	27	0	0	0	0	0	0
346	82	56	17	0	0	0	0	0
669	65	13	0	0	0	0	0	0
729	62	24	0	0	0	0	0	0
992	67	30	0	0	0	0	0	0
1366	79	151	0	0	2	0	0	0
1990	81	66	0	0	0	0	0	0
2339	83	131	0	0	0	0	0	0
2469	66	140	0	0	0	0	0	0

Showing 1 to 9 of 20,000 entries, 8 total columns

```
R 4.1.2 · C:/dab/
> f <- read.csv("flow.csv")
> view(f)
> |
```

Environment

R · Global Environment

Data

f 20000 obs. of 8 variabl...

Plots

Zoom Export

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

2 Lec_week_4A.R * f * 2 Lec_week_4B.R *

Filter

	FSC.H	SSC.H	FL1.H	FL2.H	FL3.H	FL2.A	FL2.W	FL2.H
13693	115	1023	48	867	1023	1023	555	
3004	109	1023	42	863	1023	1023	558	
17457	106	1023	44	861	1023	995	563	
14387	145	1023	88	846	1023	279	326	
4948	128	1023	167	845	1023	629	352	
210	113	687	63	841	1023	467	180	
5401	115	1023	36	834	938	736	550	
1334	108	788	69	828	1023	575	404	
19150	91	1023	37	825	889	745	584	

Showing 1 to 9 of 20,000 entries, 8 total columns

```
R 4.1.2 · C:/dab/
> f <- read.csv("flow.csv")
> view(f)
> |
```

Environment

R · Global Environment

Data

f 20000 obs. of 8 variabl...

Plots

Zoom Export

```
1 # Plotting Histogram
2
3 # Data: Flow Cytometry data
4
5 # Read data ----
6
7 f <- read.csv("flow.csv")
8
9 # Plot histogram using hist()
13 # Format the histogram
23 # Change number of bins
35 # Add markers for mean and median
41 # Change frequency (or count) to density
51 # Change frequency (or count) to %
```

Environment

R Global Environment

Data

f 20000 obs. of 8 variabl...

Plots

```
R 4.1.2 · C:/dab/
> f <- read.csv("flow.csv")
> view(f)
> |
```

```
1 # Plotting Histogram
2
3 # Data: Flow Cytometry data
4
5 # Read data ----
6
7 f <- read.csv("flow.csv")
8
9 # Plot histogram using hist() ----
10
11 hist(f$FL2.H)
12
13 # Format the histogram
23 # Change number of bins
35 # Add markers for mean and median
41 # Change frequency (or count) to density
```

Environment

R Global Environment

Data

f 20000 obs. of 8 variabl...

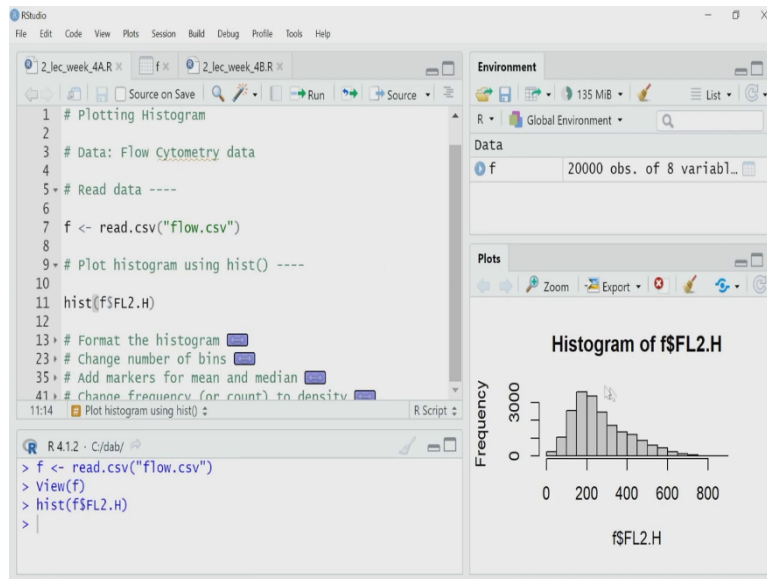
Plots

```
R 4.1.2 · C:/dab/
> f <- read.csv("flow.csv")
> view(f)
> |
```

`hist(f$FL2.H)`

I will try to plot a histogram for all cell for this FL2 H. So, for this column FL2 H I will try to plot a histogram. So, let me see, to do the create the histogram R has an inbuilt function called hist, I will use that, so, I will use f dollar sign FL2 H, FL2 H is the variable of the F object. So, I want that column data and I want to create a histogram. So, I am using just one argument for hist function and that will create me a plot. So, let us first create that and then I will cross it.

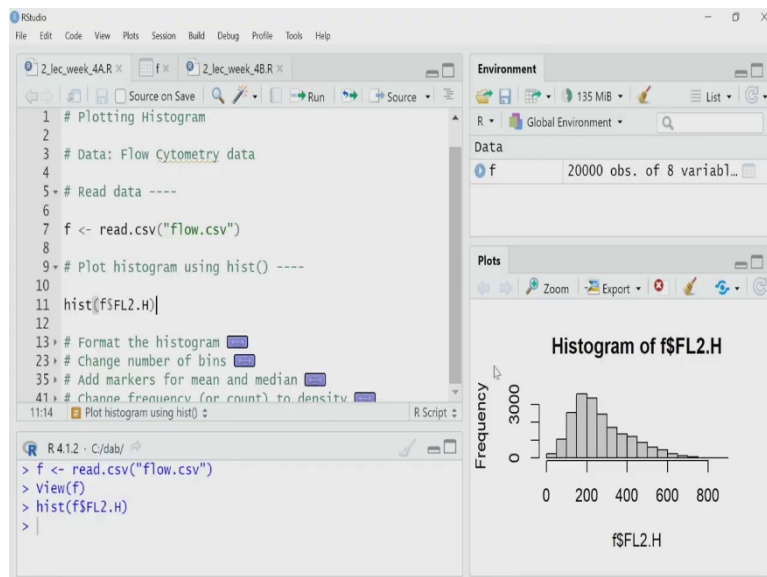
(Refer Slide Time: 8:04)

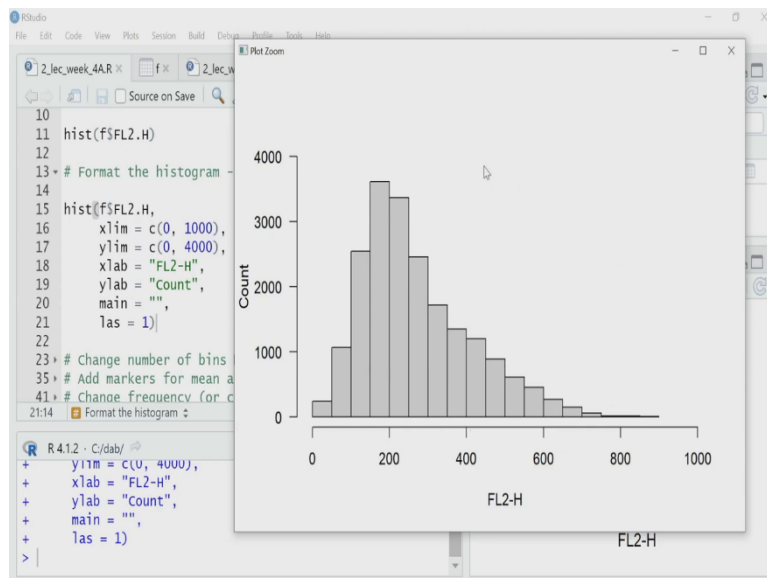
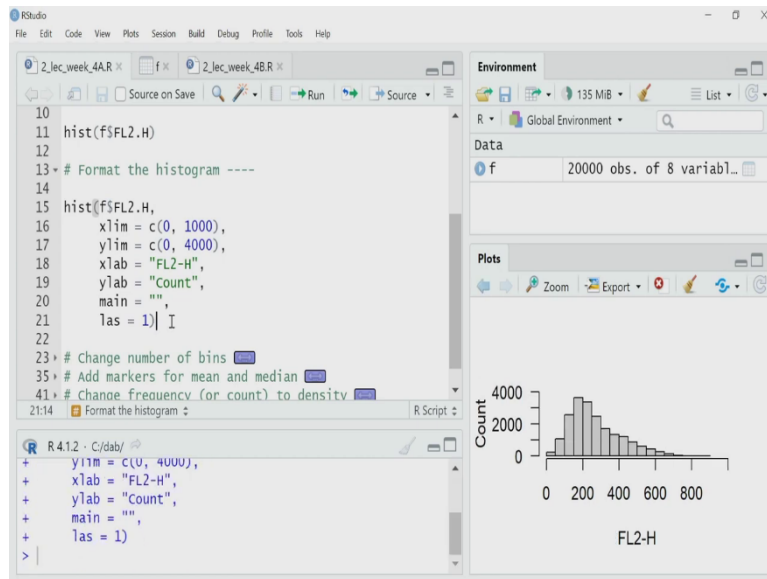


`hist(f$FL2.H)`

So, this is my histogram, I will not zoom into it, but you can easily see what we have. It has frequency written on the vertical axis, in a sense, R by frequency they mean count. And in the vertical axis, you have the readings for FL2 H, and it has lots of multiple binning has been done, and for each bin, it has plotted the bar. So, this is what the distribution frequency distribution of FL2 H reading in my 20,000 cell in this particular experiment.

(Refer Slide Time: 8:42)





hist(f\$FL2.H,

xlim = c(0, 1000),

ylim = c(0, 4000),

xlab = "FL2-H",

ylab = "Count",

main = "",

las = 1)

Now, what I obviously will prefer to do I want to do some formatting of these plots so that it looks more aesthetically better and gives more information. So, let me first do the formatting

and create a new plot and then I will explain what I am doing, what are the argument I am using here, So, let me zoom, now it looks more professional.

So, what I have done I have written `FL2 H` the way you write in the horizontal axis and on the vertical axis, I have count written. So, if you see I have said `x Lab x label equal to FL2 H`, and I have given the second argument to `hist` that `y lab equal to count`, so that is how it has written these two texts.

And I have done some rearrangement or rescaling of the vertical horizontal axis to accommodate all the data all the histogram bars within a plot nicely. So what I have done, I have set the horizontal axis `x limit equal to 0 to 1000`. Whereas for the vertical axis, I have chosen `y limit equal to c 0 to 4000`. So, I am scaling from 0 to 4000 and it looks much better.

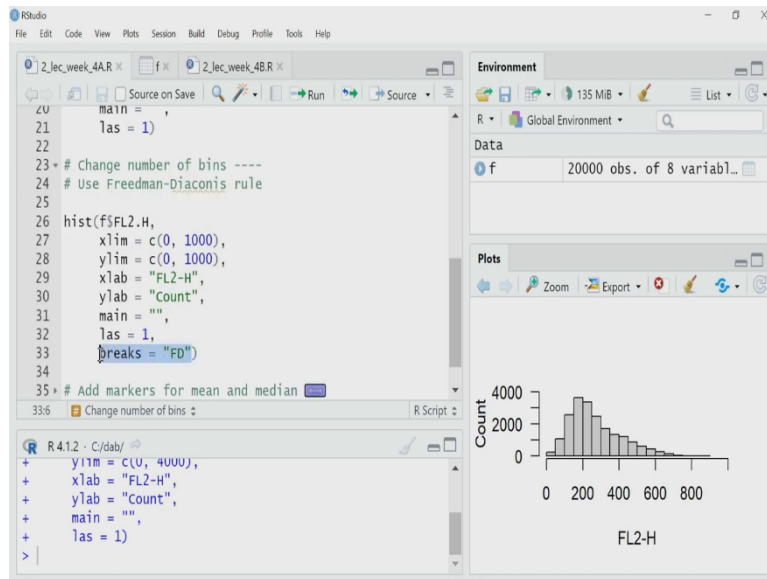
And another thing I have removed the title of this graph, which is usually not required, so `main` is blank I have kept here as a argument and `las` it actually written as `1` because I want now the numbers, this 4000 3000 there is a tick labels for the vertical axis to be horizontal. By default they are vertical, I want to convert them into horizontal the way it is there in this plot. So, I have set `las` is equal to `1` and that has given me this nice, diagram nice histogram.

Now one question you must be thinking that, how many bins should be there in my data? In the example of student marks, it is very easy to decide because I have 50 marks 0 to 50, you can say, make a bin of 10 marks each, 0 to 10, 11 to 20, something like that. So, it is very easy to decide. But what should I do in this case? Where I have a large variation in data.

If you can see from starting from 0 to even 900 something it is going the intensity is going. So, how should I decide what should be the number of bins? Now one point that we have to remember is that number of bins decide the shape of this histogram, if you reduce the number of bins the shape will change, if you increase the number of bins the shape will change. So, what should be the optimum number of bins?

There are some thumb rules, here what we have done, I have used the default option of `R`, which has its own function by default, it will calculate the optimum number of bins and plot it. But now, suppose I want to change the bin numbers, I have idea but or suppose I want to use some other function or other rule to decide how what should be the number of bins optimal number of bins for my histogram?

(Refer Slide Time: 11:37)



```

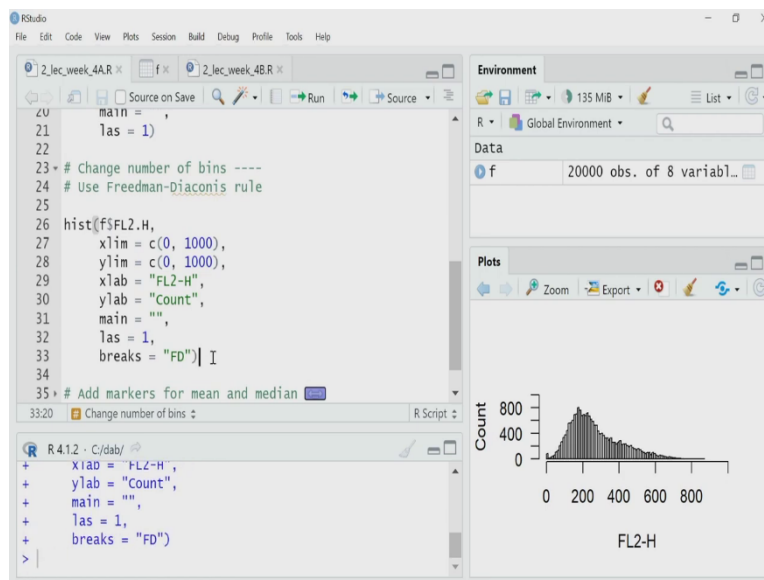
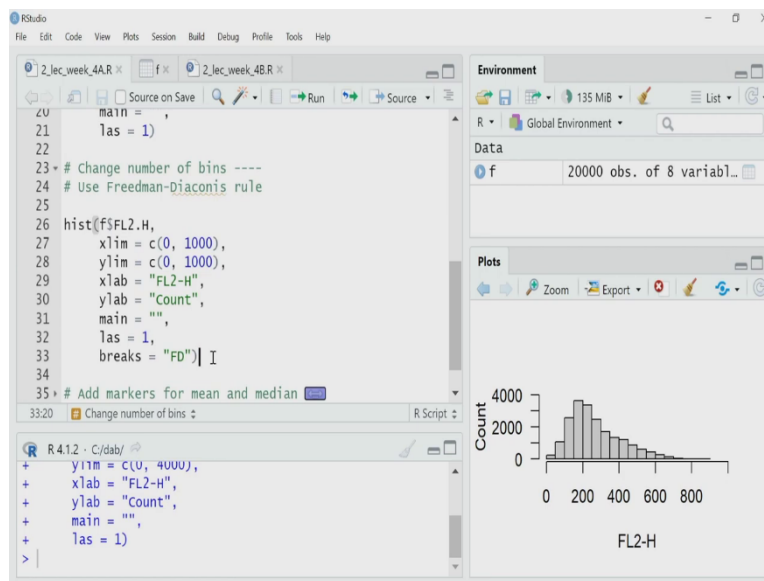
hist(f$FL2.H,
      xlim = c(0, 1000),
      ylim = c(0, 1000),
      xlab = "FL2-H",
      ylab = "Count",
      main = "",
      las = 1,
      breaks = "FD")

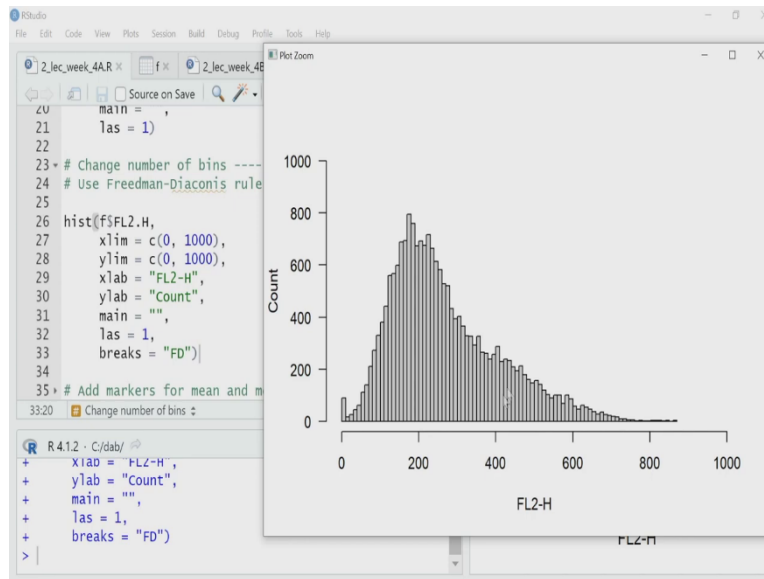
```

So, that option is also there. So, I will do that now. So, what I will do? I will use the there is a rule called Freedman Diaconis rule. And I will use that rule to create a histogram in R, so I will force R hist function to use that rule. What I am doing I am just changing one argument here, I am adding one la argument at the end I have written breaks equal to FD.

So, FD stands for Freedman Diaconis rule. So, I could have placed any value there also I could have said makes bricks equal to 50. So, then it will create 50 bins for my data set, here I am leaving it to this rule, FD rule.

(Refer Slide Time: 12:15)



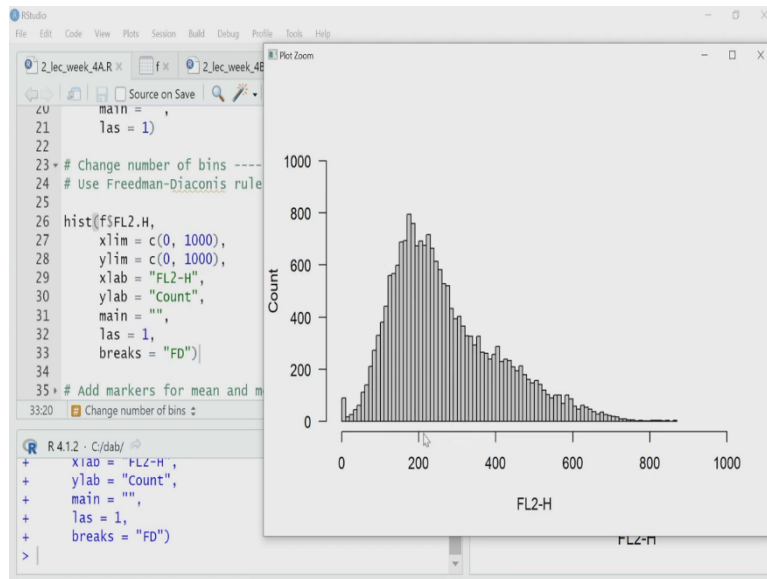


```
hist(f$FL2.H,
     xlim = c(0, 1000),
     ylim = c(0, 1000),
     xlab = "FL2-H",
     ylab = "Count",
     main = "",
     las = 1,
     breaks = "FD")
```

So, let us create a histogram using this rule. Now, you can easily see the pattern of the diagram the histogram has slightly changed, it is now much smother. And I can see as if there is two hump one is this is the main hump near 200. And then maybe I have another subpopulation something here, which is coming around 400, to say near 400.

So, as if there are slightly another peak, very low one, which was not visible when I was using the default bin numbers. So there, there is no hard and fast rule how you will decide the bin. Once you understand the data, and you understand the physical meaning of it, and you know what you want to see in that, observe in the data, you can actually decide the optimum number of bins.

(Refer Slide Time: 13:18)



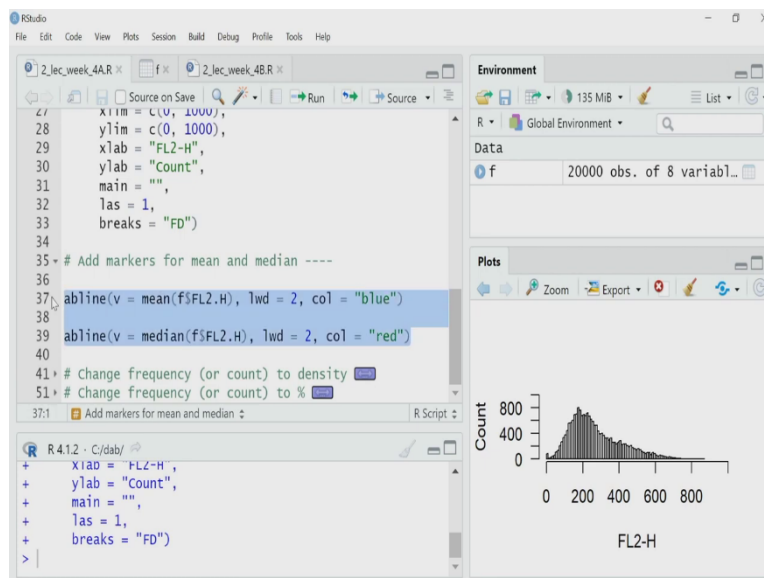
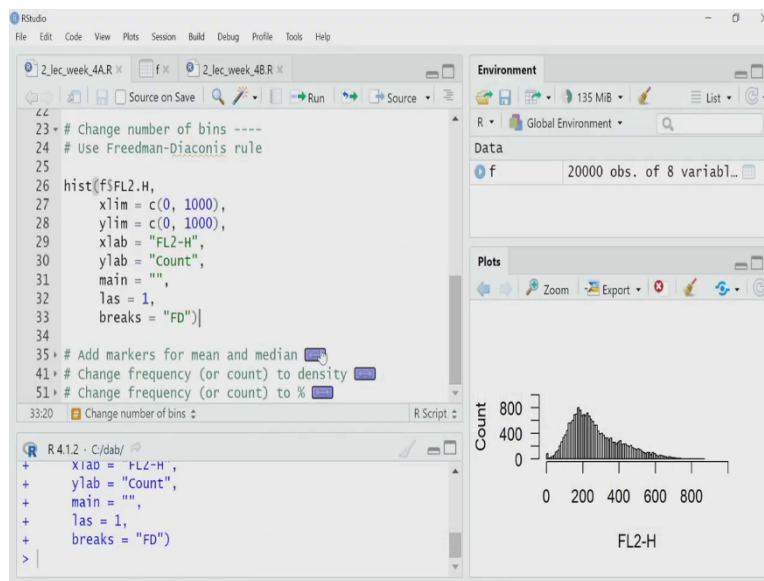
```

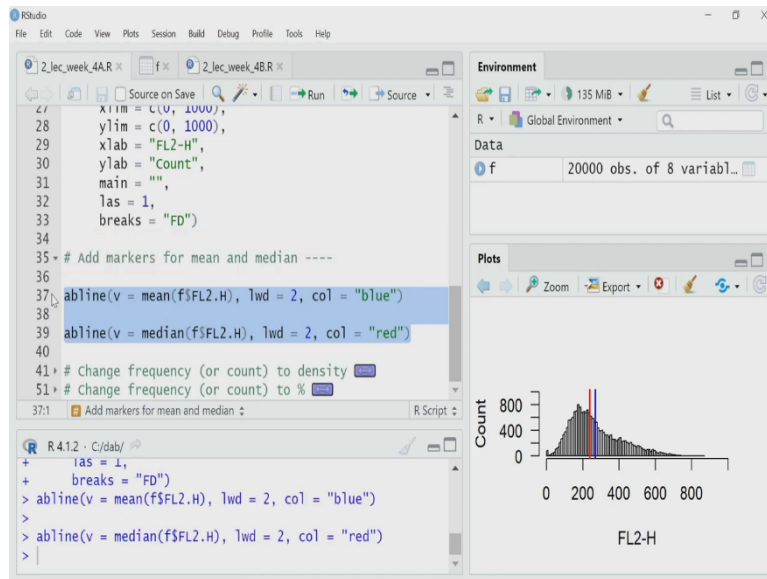
hist(f$FL2.H,
     xlim = c(0, 1000),
     ylim = c(0, 1000),
     xlab = "FL2-H",
     ylab = "Count",
     main = "",
     las = 1,
     breaks = "FD")

```

Now, sometimes this histogram, this is giving me a distribution. But sometime I want to add some more statistics to this data. For example, I may want to add a marker for the mean of this data, or the median of this data. Mean and median will be same if I have a very symmetric data, but it is a skewed data in this case. So, I want to mark this graph for mean of this data, and also the median of this data. So how can I do that?

(Refer Slide Time: 13:37)





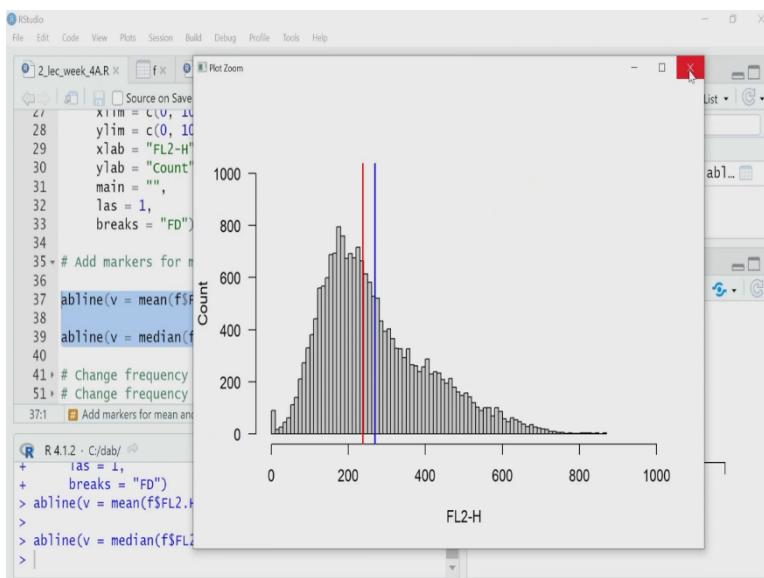
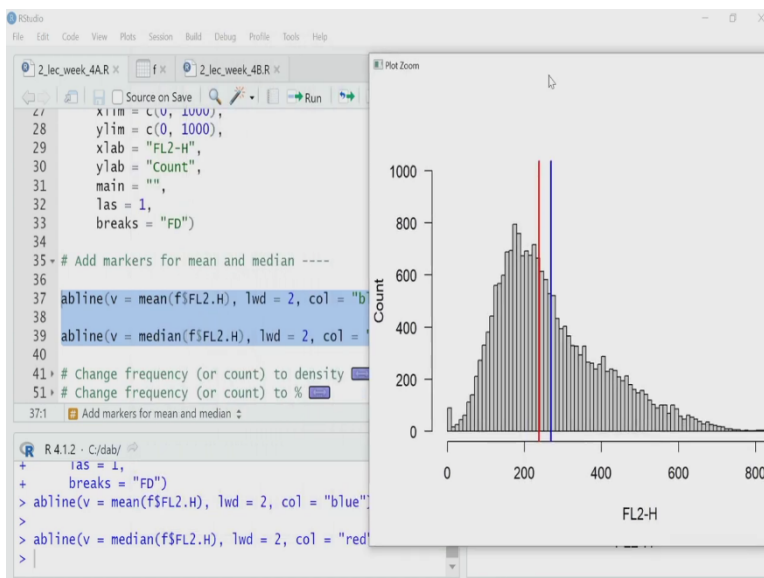
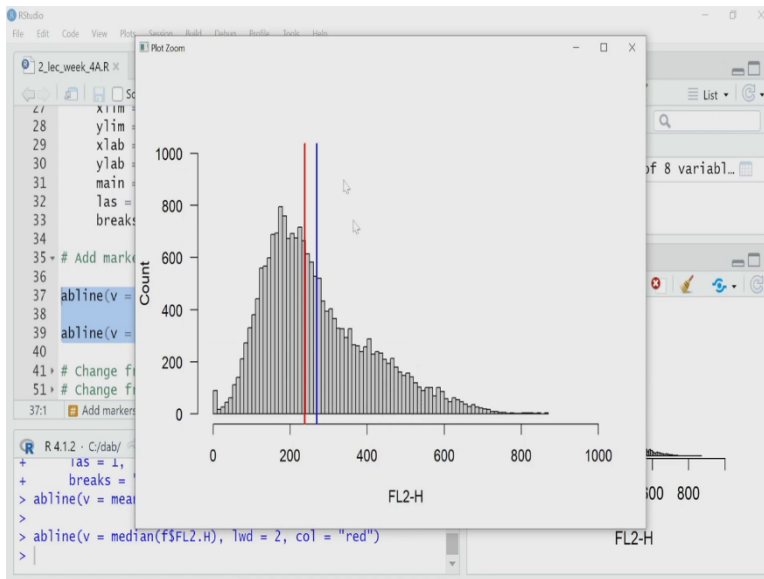
`abline(v = mean(f$FL2.H), lwd = 2, col = "blue")`

`abline(v = median(f$FL2.H), lwd = 2, col = "red")`

To achieve that, what I will do? I will use the `Abline` function. `Abline` function adds a straight line either vertical or horizontal. So, I am saying `Abline` in the functions argument, I am saying `v` that is vertical equal to mean of `FL2 H` of these data, So, it will calculate the mean and use that to draw a vertical line, I want the thickness or width of the line as 2 and I want the color equal to blue.

And the second line is for the median, the same `Abline` function I am using but I am using the function `median` to calculate the median of the same data, I am using a different color for that line, I am using it as a red line and both of them will be vertical. So `v` equal to I have written as the first argument.

(Refer Slide Time: 14:24)



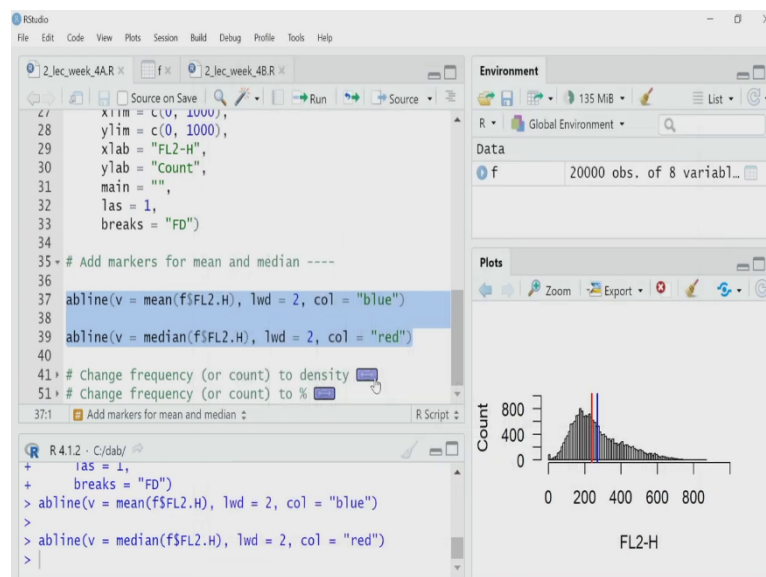
```
abline(v = mean(f$FL2.H), lwd = 2, col = "blue")
```

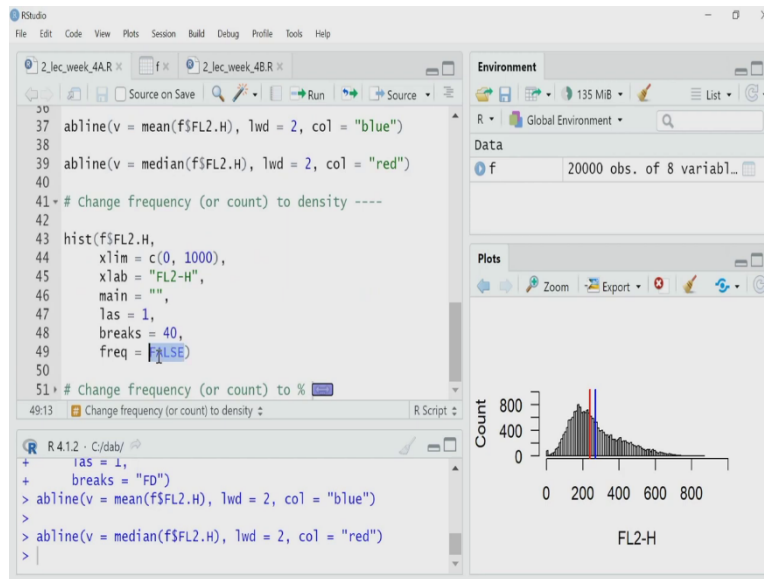
```
abline(v = median(f$FL2.H), lwd = 2, col = "red")
```

So, if I execute these two lines, now, let me zoom you can now see nicely I have marked the mean and the median. So, my mean is blue and the median is red. So, how what I have plotted here if you see I am plotting by default R is plotting the frequency what they call frequency. Actually we are plotting the count on the vertical axis.

But sometimes you do not want count, you want the frequency to be plotted the frequency means in the fraction values on the vertical axis. R has that option. I will go to the default option, and then I will tweak it a bit to create something else.

(Refer Slide Time: 15:03)





`hist(f$FL2.H,`

`xlim = c(0, 1000),`

`xlab = "FL2-H",`

`main = "",`

`las = 1,`

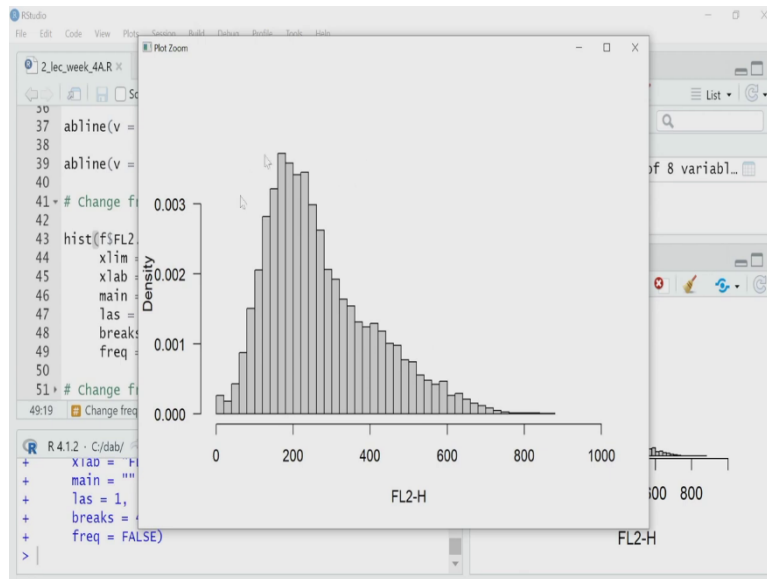
`breaks = "FD",`

`freq = FALSE)`

So, what I have to do? If I add another argument frequency, `freq` equal to false, then what will happen? R will not plot the count data; rather it will plot the density data. What do I mean by density? If you remember our discussion on probability distribution function, we know, I can draw a probability distribution function PDF, where the vertical axis is the density and they have the PDF.

And the area under the curve of that is the probability and the total area under the curve will be equal to 1. So, the similar thing it will do here, it will calculate the density and put it in the vertical axis and draw the bar plot. And the area under this whole curve will be equal to 1.

(Refer Slide Time: 15:57)

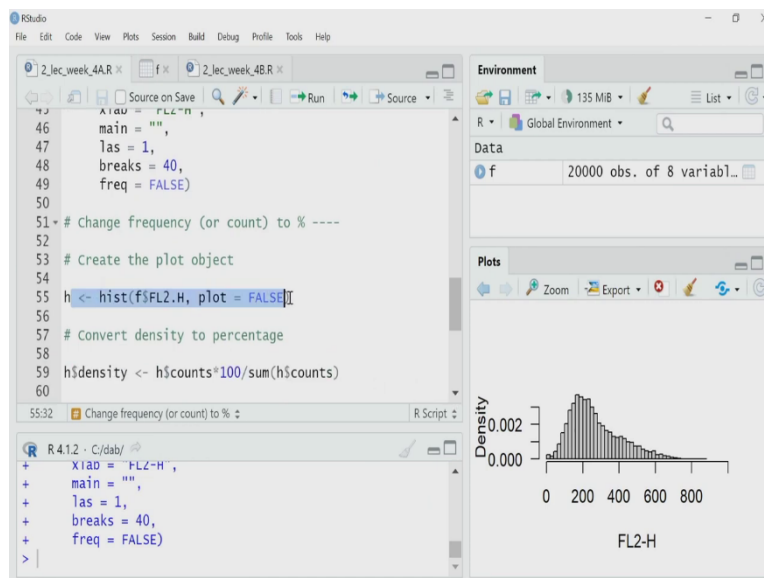
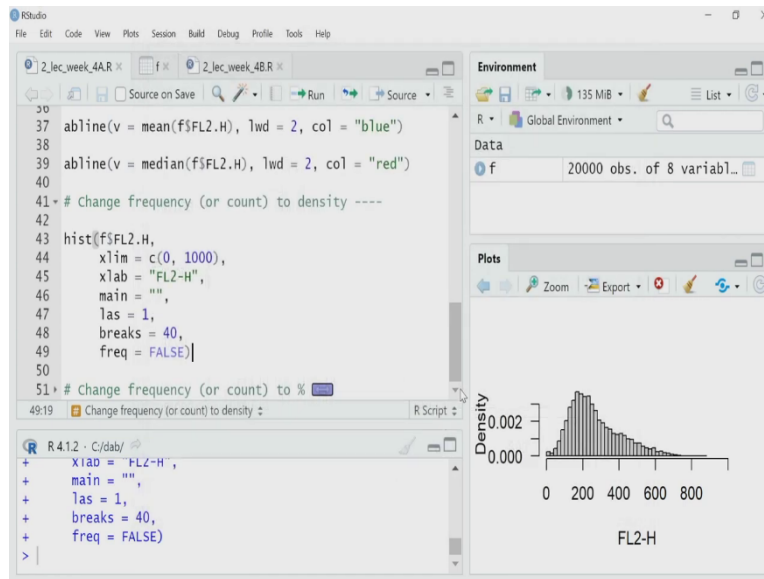


So, let me execute that and see how does it looks like? Here I am. So, what we have here, this vertical axis is actually now the density. And that is why it is density, I could change the axis limits from 0 to I can increase it so that I can accommodate the all the bars inside that, I love doing that.

Because if you are looking to this way of representation, usually in day to day life, you do not actually represent data in this way. Yes, you represent fractions as a frequency on the vertical axis or sometime percentage. But that is not the density the way we understand PDF, what do you want usually?

Suppose that I will not put the block count on the vertical axis on my histogram. Rather, I will put the percentage. So, I want to know how many percentage cell are there in a particular bin? That is what I said for the marks data also that we explained few minutes back. So, how can I plot the percentage data in the vertical axis? Let us try that.

(Refer Slide Time: 17:00)



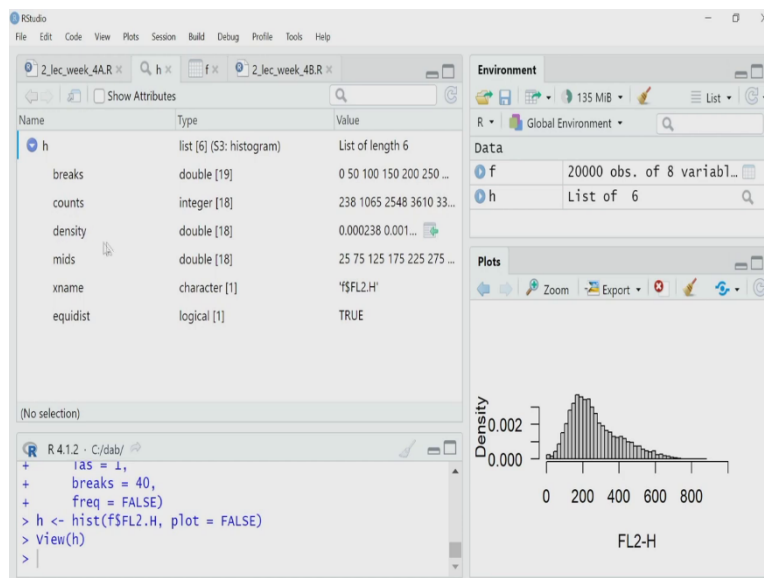
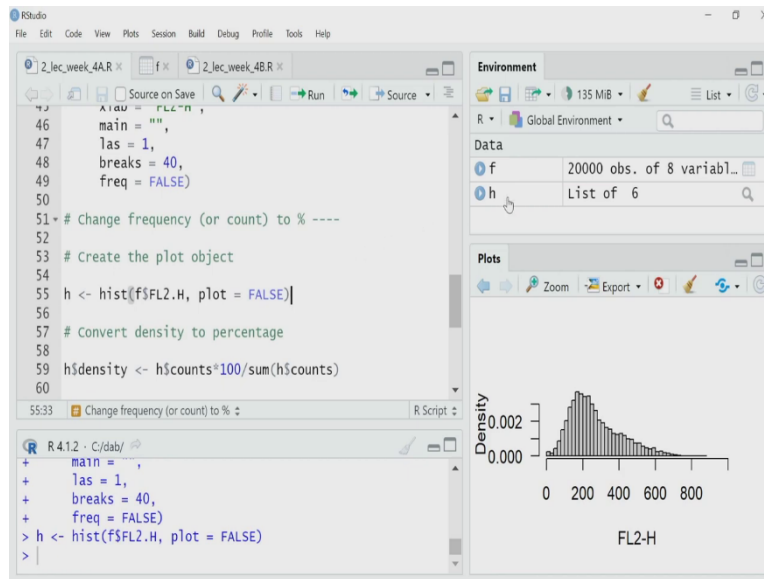
`h <- hist(f$FL2.H, plot= FALSE)`

`h$density <- h$counts*100/sum(h$counts)`

I have to play a bit with the options that they are for histogram function. What I am doing here in this line is that I am calling hist function and giving the data FL2 H. And I am saying plot equal to false. That means I am saying do not plot it. But what I am saying? I am doing these calling this function and assigning the output of that function to an object h.

So, now, in this line of the script, I will not plot the graph the histogram, but I will store the data of this histogram into a variable h and then I will look into these variable h and see each different elements of that.

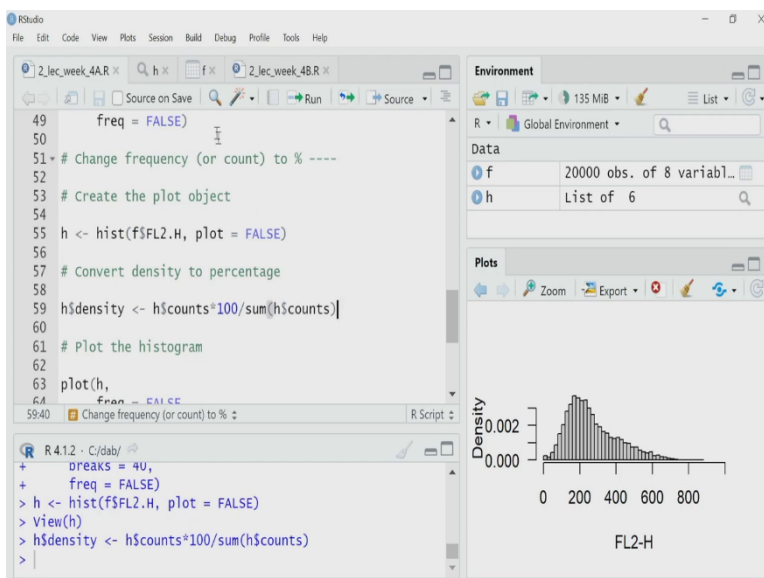
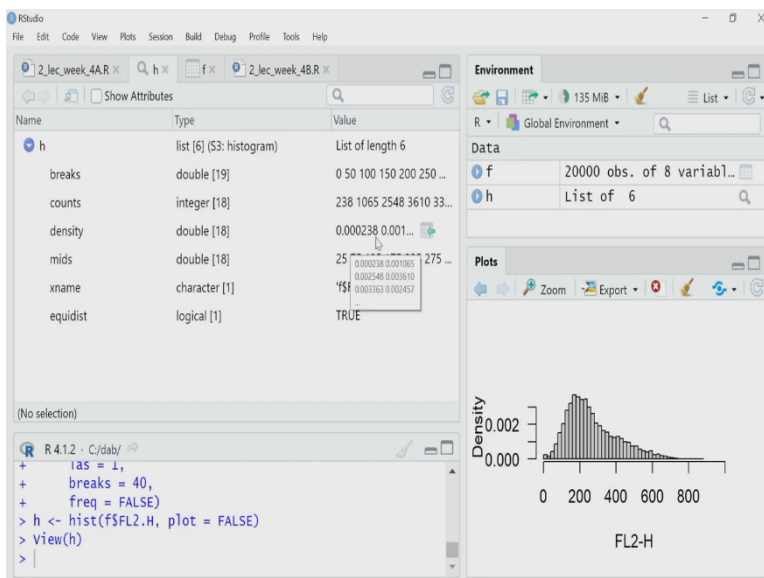
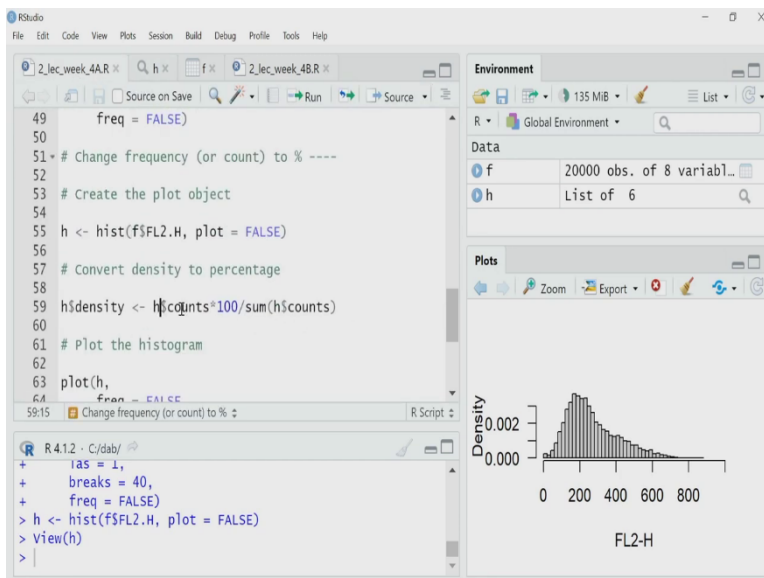
(Refer Slide Time: 17:43)



So, let us first do that. So they see there is no new plot created, but h variable is created here. Let me open that. If you look into that h has multiple things, I will request you to look into do two things this counts and density. So, the counts means it has divided data in different bins, and for each bin, it has counted, how many cells are there.

So, for the first bin, it has 238 cells, second bin, it has one 1065 and so on. And the density is the density that it calculated and plotted just now that we plot it earlier when I said, frequency equal to false. So, what I will do? I will change this density data in this h object. I will intentionally tamper it, so that now density will not store this density data rather it will store the percentage data, how should I do that?

(Refer Slide Time: 18:39)

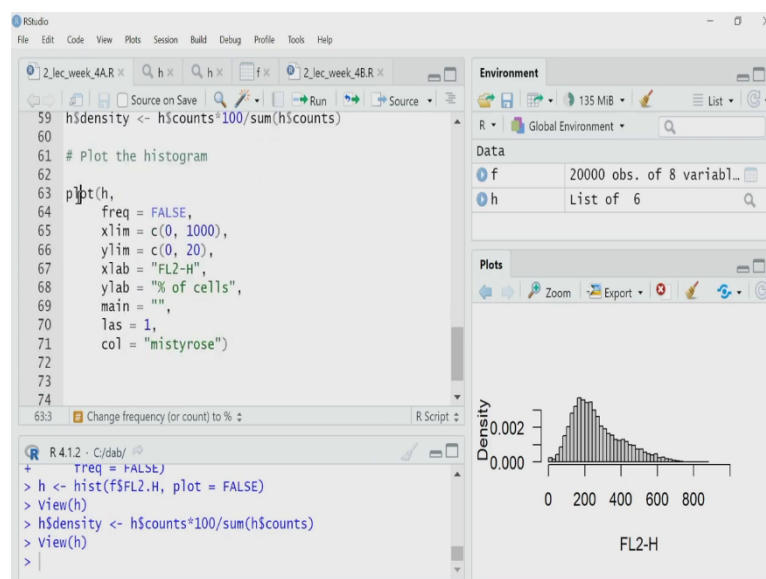


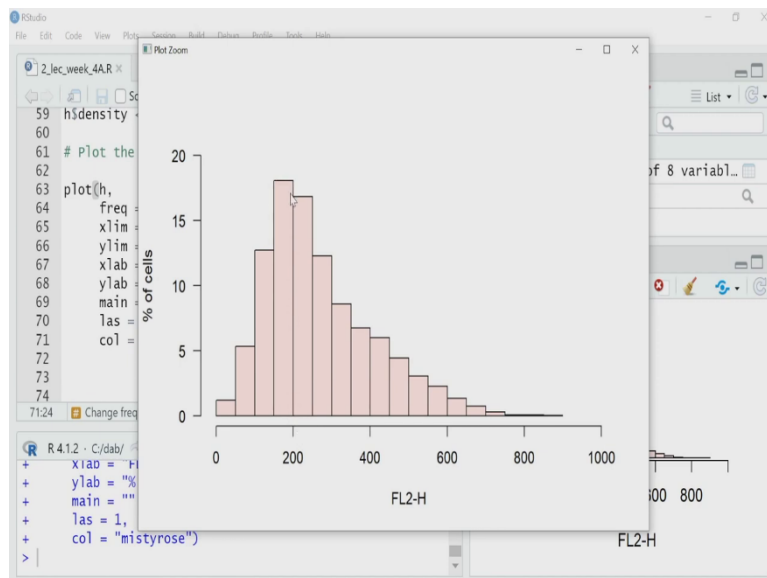
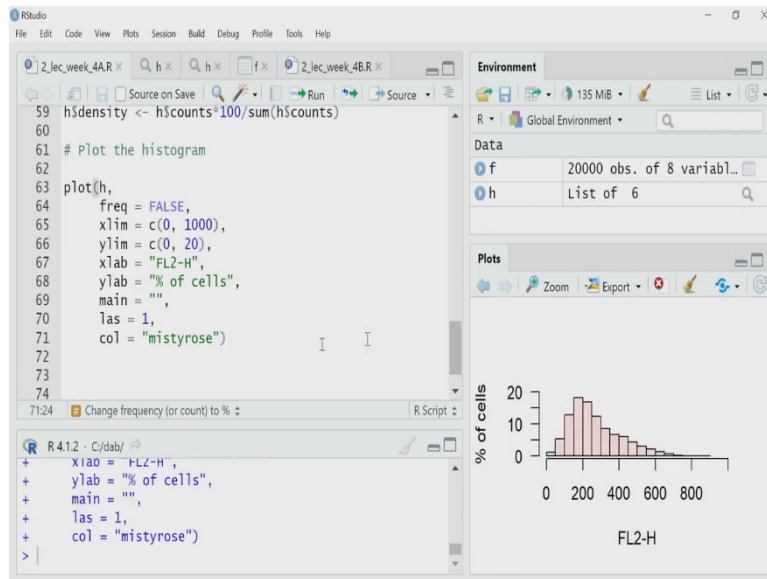
In the next line I am doing it, what I am doing? I am taking the count data from h so, h dollar counts and I am multiplying by 100 because I want to calculate percentage and then I am dividing it by the sum of all the counts. So, for each bin, there is a count you sum them and you put that in the denominator, sum of all the counts and the numerator you have individual counts for each bin and you multiply that by 100.

So, I get the percentage and now I assign that data to a not a new variable, but to the existing variable h dollar density in the density variable, density object in the h variable. So, that is how it will get replaced. So, just to remind you, remember these values, I have a density first value is 0.30238 and it should change it should become percentage data.

So, if I execute it something has done. So, let me go back to the h data. And now you see the density is something is there, density is changed to 1.19. So, it is no more than density it calculate like frequency distribution density. It is now the density in terms of percentage of cell in that bin.

(Refer Slide Time: 20:05)





plot(h,

freq = FALSE

xlim = c(0, 1000),

ylim = c(0, 20),

xlab = "FL2-H",

ylab = "% of cells",

main = "",

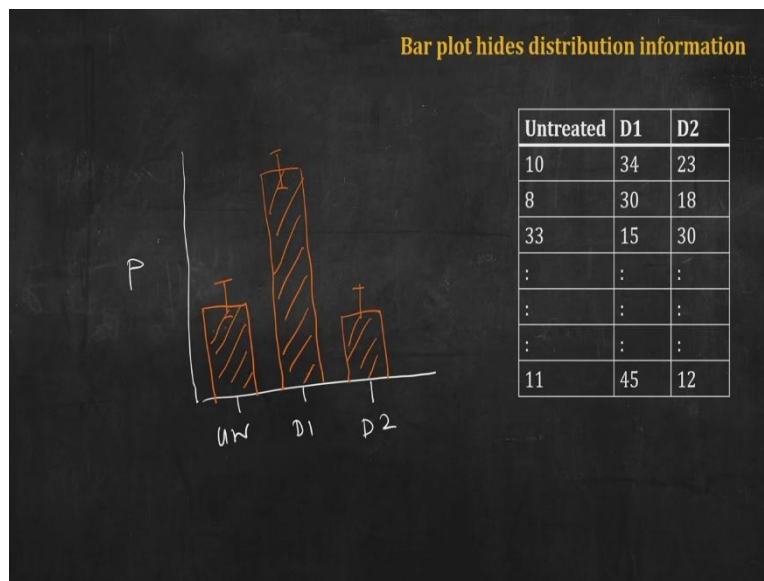
las = 1,

```
col = "mistyrose")
```

And now I want to plot this h object which has all these information to draw the histogram. So, now, what I will do? I will call the plot function, I will not call anymore the hist function because already I have created the histogram, I will call the plot function give the histogram object h as an argument, the first argument I will set the frequency equal to false because I want to plot the density not the count, and then I have all these other argument for the graphical thing issues like the color, limit of the x and y axis, let me zoom.

Now, you can see I have a nice plot where I do not have a raw count in the vertical axis rather I have percentage of cell. So, I can say in this bin near 200 almost maybe 17 percent of the cells are there. So, now, it is much more meaningful rather than looking into the count. So, this is how you can create histogram you have to know what you want to understand out of these data and based upon that you have to choose, how will you represent the data in a histogram.

(Refer Slide Time: 21:14)



Now, let me move to the second topic of today's discussion that is a boxplot. Many times we represent data either as scatterplot or bar plot, but plotting data for as a bar plot is not a good idea every time, let me explain with an example. So, here I have suppose a data set, a hypothetical data set from an experiment where we have suppose 30 or 40 animals and I have three groups of animal in each group I have suppose for 40 animals.

And I am testing some effect of one or two drugs. So, I have one group which we call untreated suppose, that has 40 animals, and D1 is another group where I have treated with a particular drug or a dose of a drug and D2 is another group of animals both having 40 animals each. Now, we have treated these animals with the drug and we have also untreated population and then we have measured some parameter, suppose some blood parameter we will not go into that now, I can represent this data in tabular form in this way.

Now, to represent it visually graphically, what we may attempt to do in this case is to create a bar plot for bar plot what I will do, I will take the mean of each of these category, right and then I will plot those mean the height of those bar are these bars are representing those means, and then I know to calculate the error, either the standard error or the standard deviation I will calculate from the data and I will draw the error bar representing those errors.

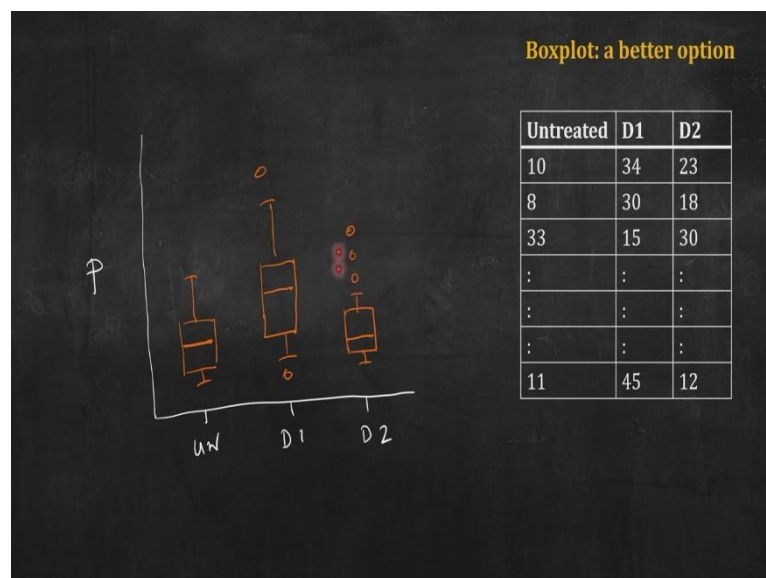
That is fair enough that is how we usually many times represent these type of data. And as you can see, here, D1 has the higher effect much higher effect than untreated, whereas D2 has almost similar effect with respect to untreated one. Now, these type of plot when I have a large data set in each category we have large number of samples, is actually a bit confusing.

So, this type of plot actually many times particularly for large data set, when I have large number of samples in each category are actually not recommended one because they hides lots of information, let me explain what do I mean by they hides lots of information. For example, here if I have these 40 animals in these each category, I have calculated the mean and the, suppose the standard deviation or standard error, but I know as they are animals their responses must be quite varying.

So, there must be quite a large variations or heterogeneity in the responses in each category also, the same phenomena will happen if you do some study on human being or even if you are doing assay on cells, where you have taken some large samples from large number of cells, then you will see a get a heterogeneous population.

Now, these heterogeneous population, some cell will be near the mean we have the mean behavior, but at the same time, many of them may be outliers also. So, bar plot hide that heterogenetic information, it only tells me the mean and the standard deviation or the standard error, it does not talk about the outlier or the dispersion of the data as a whole. And that is where box plot comes into work.

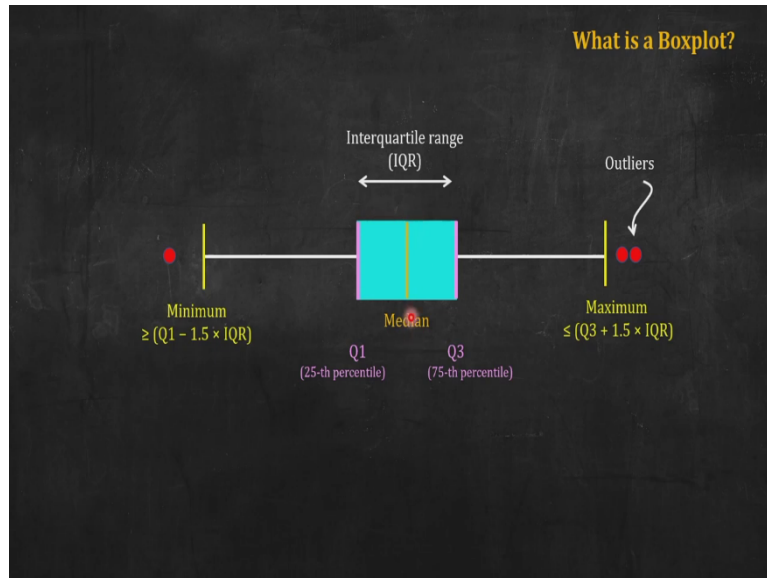
(Refer Slide Time: 24:23)



Let me show how we will represent the same data roughly as a box plot. A box plot will have box, so I have three boxes for each three category just in place a bars I have the boxes, and I will explain what does a box represent. And then each from each box, you have something thing like thing going out from each box you can see those are called whiskers.

And so a box plot is made up of boxes and whiskers, apart from then, you can see here for example, D2 I have three circles I have marked, those are called outlets. Now, I will explain what these box is? What a box is? What a whisker and what are these outlets?

(Refer Slide Time: 25:00)



So, here I have drawn a box plot one single box, I have drawn it horizontally to accommodate in space, but the formalism will be the same when you will draw vertically also. So, box plot will have a box and the in the middle of the box, there will be a line, that line represent the median, how do I get the median of the data? So, suppose I have these 40 untreated animals, so, I have measured their blood parameter.

So, what I will do? I will arrange these data for this population in ascending order, suppose, lowest to highest. And then I will find the middle point, the middle data, that is the median, and these bar, these vertical bar represent here, that median, so the bar in the middle of the box represent the median of the data. Now, let me go to the ages of these box on the lower side, on the left hand side of this box, I have marked it by this pink, this is the quartile 1 of the data, how do I get quartile 1?

So, take the data less than the median. So, these are median, take the data below the median and find the middle of that. So, that is your quartile 1. Similarly, the pink line on the other edge, the right hand side edge represent quartile 3, third quartile, what is that? I have the median of the data, you go up on the highest side, take the median of that, the middle point of that, that is my third quarter.

So, the box, the lower part, one edge of the box is quartile 1, then the middle line is median or it is quartile 2, and the upper edge, the last line in the box is the quartile 3 and you can easily understand the distance between these quartile 1 and quartile 3 will be called interquartile range or IQR. Now, let me go to the whiskers from this box we have extended whiskers. What does a whisker represent?

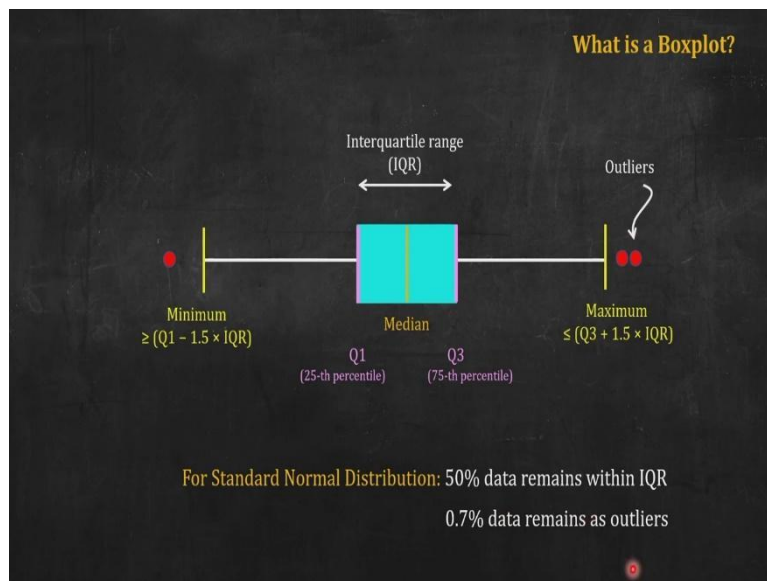
Here, I have written this yellow thing here as the bar where the whisker has ended, I am calling it maximum, whereas on the other side, I am calling the end of the whiskers minimum, but please do not get confused that with the minimum and maximum in the data, no, they do not represent minimum or maximum in the data they represent minimum or maximum in some other sense, what is that? Let me go to the higher side of Q3.

So, this whisker represent a range which is representing data from Q3 plus 1.5 into the interquartile range. So, what is the end of this whisker representing? That end is representing the highest value in my data, which is within the range of Q3 quartile 3 plus 1.5 of IQR. So, that is the maximum on that side, what is represented by these on the lower side? So, the whisker in the lower side ends at the value of the data, which is the minimum value within the range of Q1 minus 1.5 into IQR.

So, these whisker on the higher side is going up to the data point, which is within Q3 plus 1.5 into IQR. And this lower whisker going up to the data point, which is within the range of Q1 minus 1.5 into IQR. And then what does this red point represent? As I said, they represent, these red circles represent the outliers. So, these outliers are outside this range of 1.5 into IQR, the interquartile range on both side, after Q1 and after Q3.

Now, why do we choose 1.5 IQR and represent boxplot in this way, to understand that we have to go to standard normal distribution, I will not go into the mathematics of that, but it can be shown that if your data follows standard normal distribution, that means it is a normal distribution with a mean 0 and variance equal to 1.

(Refer Slide Time: 29:34)



Then 50 percent of the data will remain inside this box that is within the interquartile range. Whereas this outlier on both sides they will constitute only 0.7 percent of the data. So, 0.7 percent of the data will represent for a standard normal distribution will represent the outliers and that is why we have designed the boxplot in this particular fashion. So now I will show how to use R to create boxplot for a particular data set.

(Refer Slide Time: 30:06)

```
1 # Boxplot
2
3 # Data: Default "insectSprays"
4
5 # Read data ----
6
7 d <- InsectSprays
8
9 # Create boxplot using boxplot() function ----
10
11 boxplot(count ~ spray, d)
12
13 # Format the boxplot ----
14
15
```

RStudio

```

1 # Boxplot
2
3 # Data: Default "InsectSprays"
4
5 # Read data ----
6
7 d <- InsectSprays
8
9 # Create boxplot using boxplot() function ----
10
11 boxplot(count ~ spray, d)
12
13 # Format the boxplot ----
14
15

```

7:18 Read data R Script

R 4.1.2 · C:/dabold/

```

> d <- InsectSprays
>

```

Environment History

R · Global Environment

Data

d 72 obs. of 2 variables

Files Plots Packages Help Viewer

RStudio

2 lec_week_4BR d

	count	spray
1	10	A
2	7	A
3	20	A
4	14	A
5	14	A
6	12	A
7	10	A
8	23	A
9	17	A

Showing 1 to 9 of 72 entries, 2 total columns

R 4.1.2 · C:/dabold/

```

> d <- InsectSprays
> View(d)
>

```

Environment History

R · Global Environment

Data

d 72 obs. of 2 variables

Files Plots Packages Help Viewer

RStudio

2 lec_week_4BR d

	count	spray
17	10	B
18	14	B
19	17	B
20	17	B
21	19	B
22	21	B
23	7	B
24	13	B
25	0	C

Showing 17 to 26 of 72 entries, 2 total columns

R 4.1.2 · C:/dabold/

```

> d <- InsectSprays
> View(d)
>

```

Environment History

R · Global Environment

Data

d 72 obs. of 2 variables

Files Plots Packages Help Viewer

RStudio interface showing a data table with columns 'count' and 'spray'. The table displays rows 48 through 56. The 'spray' column contains categorical values 'D', 'E', and 'E'. The console shows the R commands: `d <- InsectSprays` and `View(d)`.

count	spray
48	4 D
49	3 E
50	5 E
51	3 E
52	5 E
53	3 E
54	6 E
55	1 E
56	1 E

```
R 4.1.2 · C:/dabold/  
> d <- InsectSprays  
> View(d)  
>
```

RStudio interface showing a data table with columns 'count' and 'spray'. The table displays rows 60 through 68. The 'spray' column contains categorical values 'E' and 'F'. The console shows the R commands: `d <- InsectSprays` and `View(d)`.

count	spray
60	4 E
61	11 F
62	9 F
63	15 F
64	22 F
65	15 F
66	16 F
67	13 F
68	10 F

```
R 4.1.2 · C:/dabold/  
> d <- InsectSprays  
> View(d)  
>
```

RStudio interface showing a data table with columns 'count' and 'spray'. The table displays rows 1 through 9. The 'spray' column contains categorical values 'A'. The console shows the R commands: `d <- InsectSprays` and `View(d)`.

count	spray
1	10 A
2	7 A
3	20 A
4	14 A
5	14 A
6	12 A
7	10 A
8	23 A
9	17 A

```
R 4.1.2 · C:/dabold/  
> d <- InsectSprays  
> View(d)  
>
```

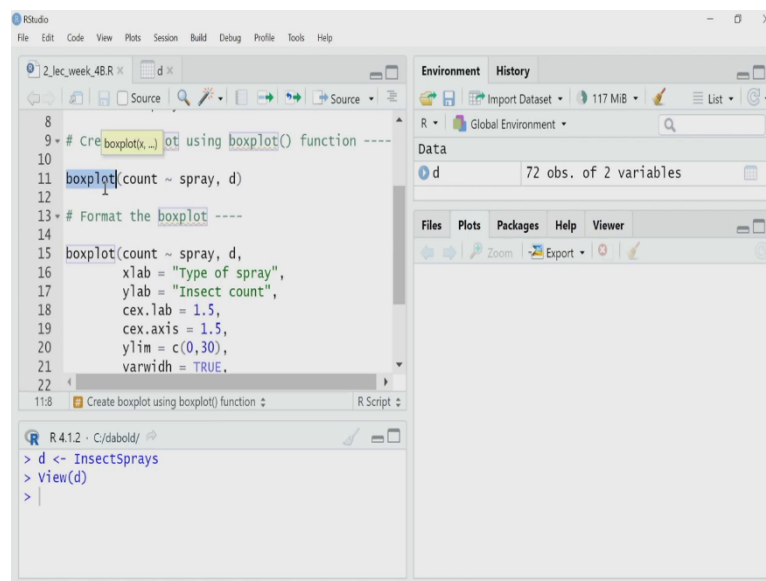
```
d ← InsectSprays
```

```
boxplot(count ~ spray, d)
```

So, here I have a script to create box plot, the data that I am using is a default inbuilt data set from R insectspray. So, what you have here, here, they have tried different insecticides on insects, and then made some counts of insects. So let me read that data and load that in a variable d. Let us check the data, is a two column data, two variables count and spray.

The spray, as you can see is a category data, A, B, C, these are the label E, up to f these are different types of spray A B C D E up to F. And for each case, you have the count of insects, when they have applied those insecticides the spray is on the insect and they have made some count. So, now I want to represent these data not like a bar plot but as a box plot.

(Refer Slide Time: 31:03)



The screenshot shows the RStudio interface. The script editor contains the following code:

```
8  
9 # Create boxplot using boxplot() function ----  
10  
11 boxplot(count ~ spray, d)  
12  
13 # Format the boxplot ----  
14  
15 boxplot(count ~ spray, d,  
16         xlab = "Type of spray",  
17         ylab = "Insect count",  
18         cex.lab = 1.5,  
19         cex.axis = 1.5,  
20         ylim = c(0,30),  
21         varwidth = TRUE,  
22  
11:8 Create boxplot using boxplot() function R Script
```

The Environment pane on the right shows the following data:

Object	Size
d	72 obs. of 2 variables

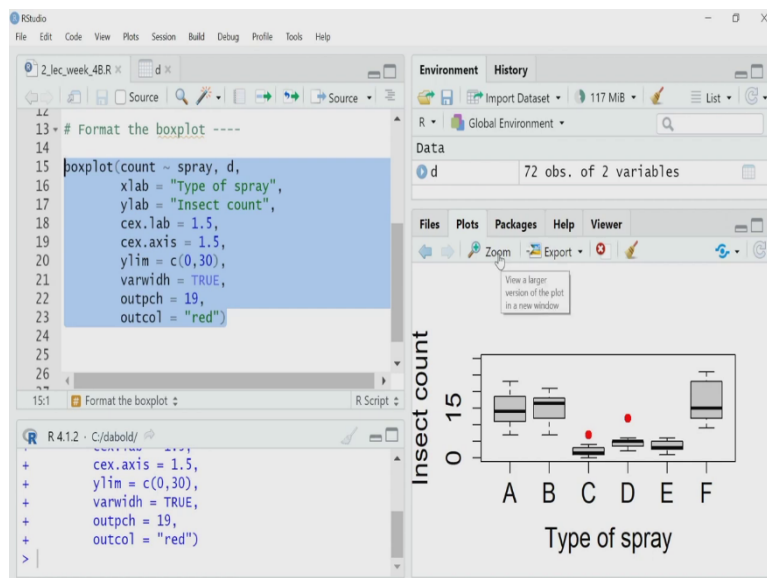
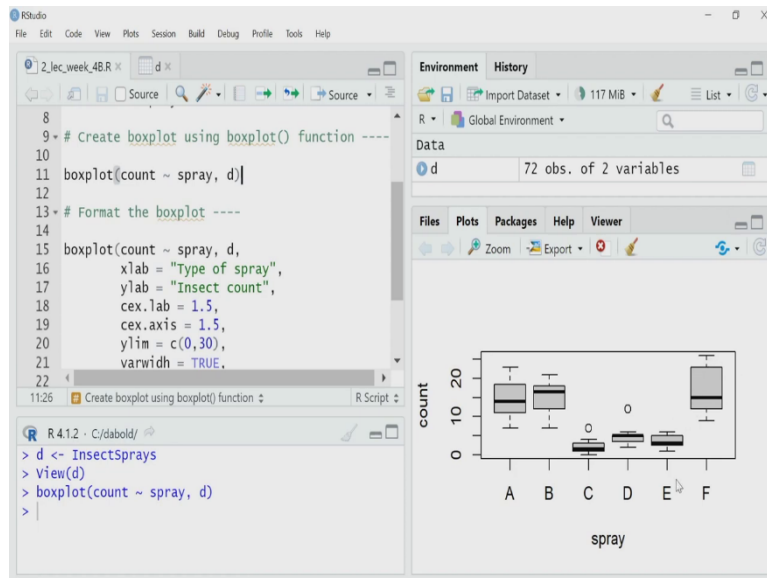
```
boxplot(count ~ spray, d)
```

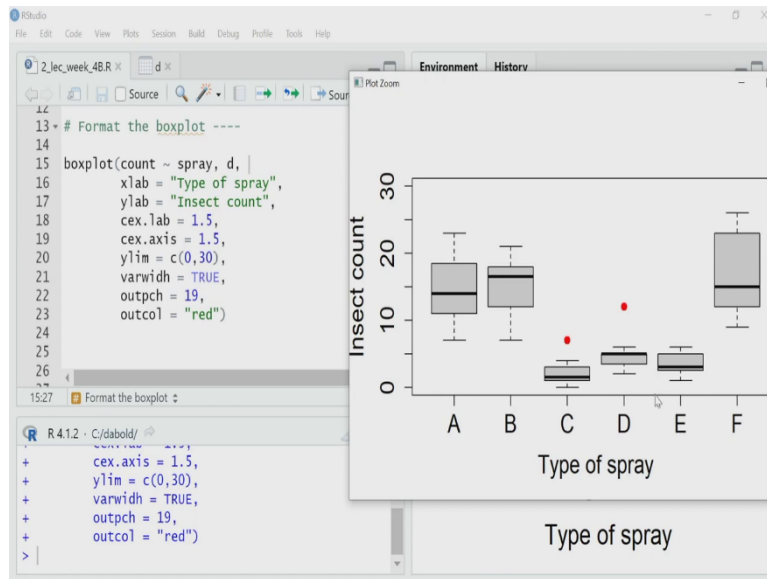
So, what I will do? R has a function called box plot, and I will use that. In the box plot, obviously, you have to give the data as the argument that is the second argument. But the first argument is the formula. What does formula represent here? Here I have written count tilde spray. That means I am telling the boxplot function that see my group is based in sprayed variable.

So, there are groups in spray variable A B C D E, you have to consider that as a group variable. And the count is a vector of data where the numerical values are there. So, take the

values in the count variable, arrange them as per the group information given in spray variable, and give me the boxplot, it will do that.

(Refer Slide Time: 31:53)





`boxplot(count ~ spray, d,`

`xlab = "Type of spray",`

`ylab = "Insect count",`

`cex.lab = 1.5,`

`cex.axis = 1.5,`

`ylim = c(0, 30),`

`varwidth= TRUE,`

`outpch = 19,`

`outcol = "red")`

So, here I have got the box plot. So you can see each of these categories the spray A to F are there in the horizontal axis, and the vertical axis I have the counts and the data is represented at boxes and whiskers along with the outlier. I will do some makeover of that, edit some takes and all these things, and then I will explain what these box plot is telling us.

So, let me first run the modified code for this box plot then I will explain each of the argument. Now I have got the box plot which is looking a bit better than the previous one. Let me zoom it and then I will explain each of these argument that I am using. Obviously, the first argument in box plot is now still the formula count tilde spray.

Then I am declaring the data d, then I am writing x lab label of x, I am saying this is type of spray and that is in apostrophe. That is why type of spray is written here in the horizontal

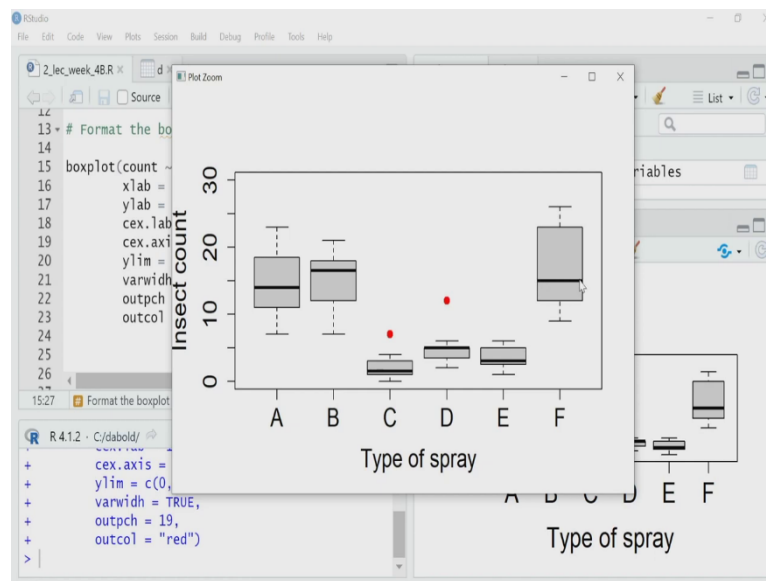
axis, then I am labeling the y axis the vertical axis y lab equal to insect count again within apostrophe.

So, insect count has been used as label of the vertical axis in the plot, I have changed the font size. So, what I am saying, that for the both the label and the axis tick this 0 10 this are the axis mark A B C are also axis marks, and a type of spray and insect count are the labels, so I want to increase the font size of both of them, and I want to make them 1.5 times off the default one. So I am written `cex dot lab equal to 1.5 cex dot axis equal to 1.5`.

Then I have the change the limit of this vertical axis, I have made it from 0 to 30. So, I have written `y limit equal to c 0 30`, and I have chosen variable with `varwidth equal to true`, although you do not see much effect of that here. See many times what can happen, in each category may not have the same number of samples like, maybe in one case you have 40 cases and samples whereas the in the other category the sample size is 30 or something like that, and you want to represent that data also in the same box plot then what you can write?

You can make this variable `varwidth equal to true`, then these, size of these boxes that width of these boxes will vary. Now in this case, actually all the sample has the same number of samples in that each category has the same sample size, that is why the variable it does not have much effect to see here. Then what I have done, I have marked these outlier by red color solid fill. So, I have set `outpch equal to 19` so it is a filled circle and `out color equal to red`. So, I am using red color fill circle as outlier.

(Refer Slide Time: 34:46)



`boxplot(count ~ spray, d,`

`xlab = "Type of spray",`

`ylab = "Insect count",`

`cex.lab = 1.5,`

`cex.axis = 1.5,`

`ylim = c(0, 30),`

`varwidth = TRUE,`

`outpch = 19,`

`outcol = "red")`

Now, let me look into a bit details in the diagram. So, as I explained these if you take the F category of spray, these thick line inside the box represents the median of the data in that category. So, that is quartile 2, the box lower end or lower edge of the box where the box end is my Q1, whereas the upper edge of the box is Q3, quartile 3, and where the whisker ends.

For the lower side of the whisker, lower side whisker start from Q1, you go down up to the length of 1.5 into IQR, 1 IQR is the length of this box, you go up down up to the length of 1.5 IQR, and find where you have the minimum data. So, the minimum data is somewhere here. So, the whisker ends there.

For the upper whisker, you start from Q3 and go keep going up to 1.5 into IQR, so, you add 1.5 into IQR, above Q3, and you stop at the maximum value in your data, and that is where it

has stopped. So, this is the upper whisker ends at the data point, which is within the, it is the maximum value within the range of $Q3 + 1.5 \text{ IQR}$. Whereas, the lower whisker has ended at a minimum value, which is within the range of $Q1 - 1.5 \text{ IQR}$.

Now, let me look into this D and C, in this case, you can see the median are either close to $Q1$ or $Q3$, that type of data represent actually that the data is very skewed in one side. So, in this case, for example, D, the data is skewed in one direction, whereas for C, the data is skewed in the other direction.

And for both C and D, and there are two outliers, and those are represented by this red dot. And as I said, they are beyond that limit of 1.5 IQR . So, that is all for this video. So, in this lecture, we have learned how to create a boxplot and histogram. See you in the next lecture. And we will learn more about making data visualization using R. Thank you.