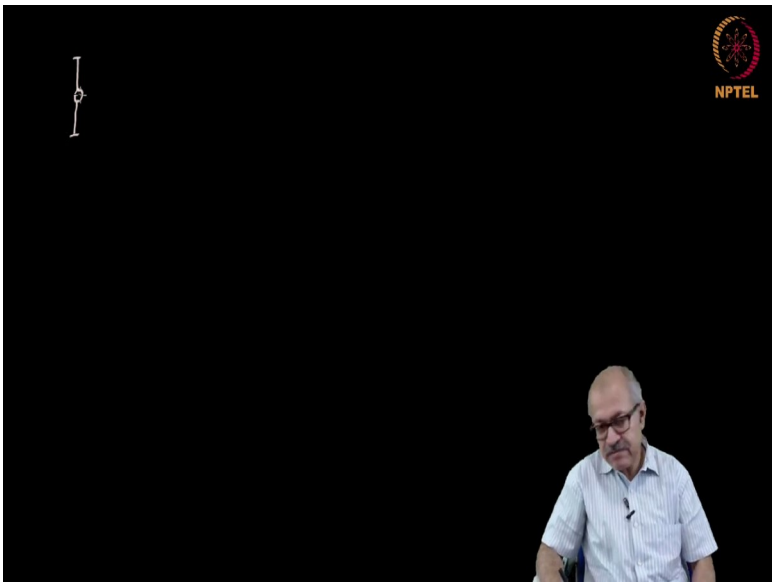
Research Methodology
Prof. Soumitro Banerjee
Department of Physical Sciences
Indian Institute of Science Education and Research, Kolkata

Lecture - 38
Box and Whisker Plot

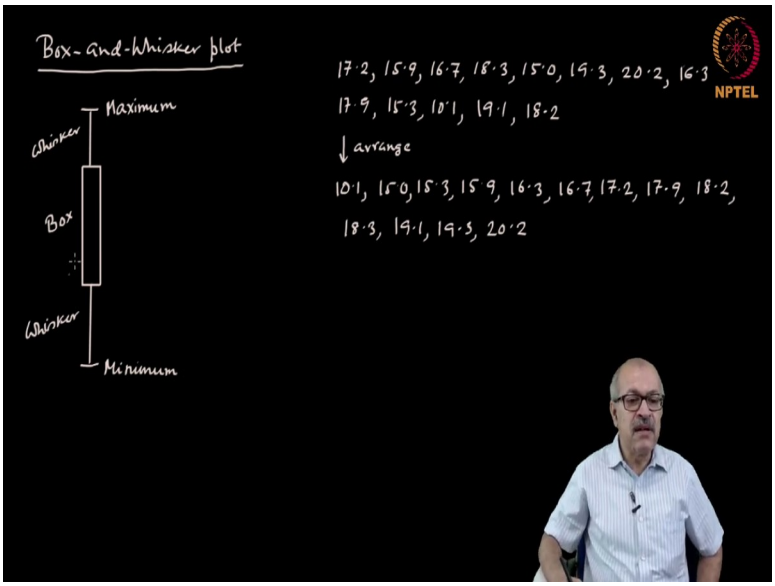
In whatever we have done, be it the measurement of a value, or the measurement of a proportion, we are basically obtaining a mean and an error bar around it.

(Refer Slide Time: 00:32)



Representation of the data will be in this form, the mean and the error bar. But in this kind of representation, one does not get a feel for the spread of the data. How big was the spread? What was the minimum value? What was the maximum value? All these information are lost in this kind of representation. And there are certain applications in which that information becomes important.

(Refer Slide Time: 01:22)



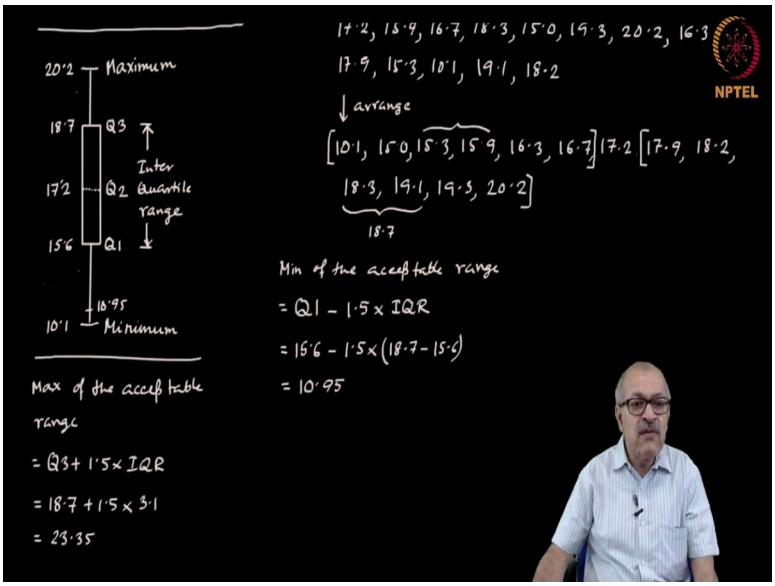
In those cases another type of data representation is used, which is called the Box and Whisker Plot. In a box and whisker plot, there is a box and there are whiskers. For example, you can have this as the box and these are the whiskers.

Now, how do we represent a data-set like this? A data-set will be a number of numbers. Now, whenever such a number of numbers are obtained or available, we first represent that in a ascending order so that the minimum value is easily identified. The minimum value is plotted here and the maximum value is plotted here.

So, these are easy to identify. While I work, let me also give an example so that we can work with that example. Suppose I have a data-set which is in whatever unit. You may put 17.2, 15.9, 16.7, 18.3, 15.0, 19.3, 20.2, 16.3, 16.3, 17.9, 15.3, 10.1, 19.1 and 18.2.

So, how many data points are there? 13 data points, and with that we are trying to construct this kind of a box and whisker plot. The first step will be to arrange in order. So, if you arrange in an ascending order, I can see that 10.1 is the minimum. So, 10.1, 15.0, 15.3, 15.9, 16.3, 16.7, 17.2, 17.9, 18.2, 18.3, 19.1, 19.3 and 20.2 that is ascending order, which immediately identifies the minimum.

(Refer Slide Time: 06:04)



So, the minimum is 10.1 and the maximum is 20.2. Now, in between the minimum and the maximum there would be a particular value, which is the median value. Not the mean, but the median value. The median value is the middle value of the data set. Now, there are 13 data points. Therefore, the middle value is the 7th data point, which is 17.2. So, 17.2 is the median value.

Now, between the minimum and the median, there would be half the data points and we take the median of that, which means the median of the values from here to here, because 17.2 is the median. Now, there are 6 data points here. 6 is an even number. So, in order to find the median of that, we have to take the average of these two values: 15.3 and 15.9, whose mid value is 15.6. So, this point would be 15.6.

The rest, which means that these 6 data points from here to here, their median is the mid value of these two. The mean of these two comes to be 18.7. So this point is 18.7. Let me put it the way: they actually carry names.

Here we have a position which is called the Q1, and this point, the median of the whole data set, is called the Q2, and this is called Q3. So, this is a quartile, this is a quartile, this is a quartile and this is a quartile. It is immediately clear that one fourth of the data set is in each of these quartiles. And by looking at it you can easily figure out what is the spread of the data.

Therefore it is easy to plot the box and whisker plot. It is basically the problem of finding the median values. This range is called the inter-quartile range. Now, it is easy to plot such box and whisker plots. One major applications of the box and whisker plots is to identify outliers. Sometimes in an experiment we get outliers—some value which is not normal, which is beyond an acceptable range.

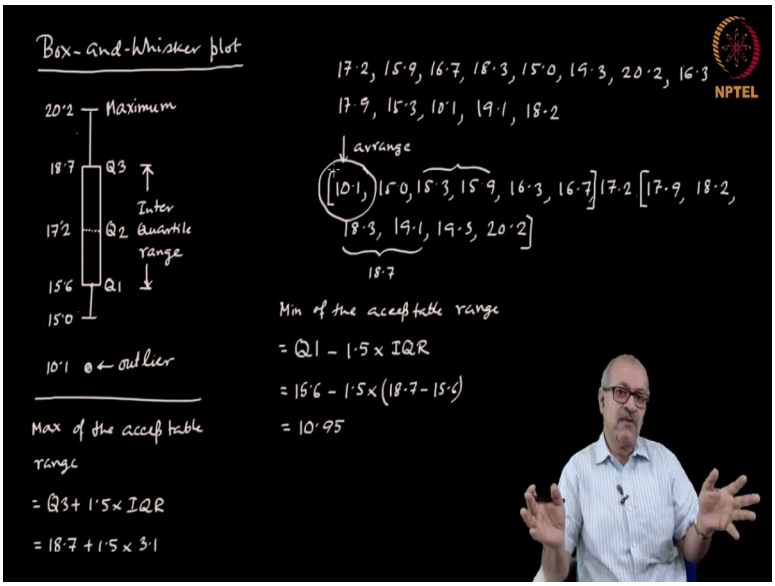
But what do you do then? The accepted practice is that we always present that outlier in the data representation, but we identify that as an outlier. How do we identify that? The standard procedure is that, the minimum of the acceptable range is, Q1 (the minimum of the box) minus 1.5 times the inter-quartile range. The inter-quartile range is the distance between Q1 and Q3. For example, in this particular problem the inter-quartile range is 15.6 minus 18.7 or 18.7 minus 15.6. This turns out to be the inter quartile range, 3.1. Now, Q1 minus 1.5 times inter-quartile range turns out to be 10.95.

Now, notice that 10.95 is above 10.1. If the minimum of the acceptable range ends at 10.95, then 10.1 is an outlier. 10.95 is somewhere here. So this is then identified as an outlier.

Similarly, the max of the acceptable range is the topmost point of the box, i.e., Q3, plus 1.5 times inter quartile range. In this particular problem, it is the 18.7 plus 1.5 times inter quartile range was 3.1, that comes out to be 23.35. And you can see that the largest point in the data set is below that: 20.2.

Therefore, we conclude that this whisker is ok, but this whisker is not. This whisker needs to be modified identifying the outlier as well as by plotting the whisker properly. Here we will plot a point and identify that as an outlier.

(Refer Slide Time: 14:40)



But the extremity of the whisker will be the last point within the minimum. That means there is a range from the minimum to the maximum and within that the extreme points will be the extremity of the whisker.

So, the whisker should end at this value which is above this. So, it will end at 15.0 and this point will have to be plotted, will have to be shown as an outlier. So, this is the outlier, this is how we will present the result. Remember we cannot ignore the outlier. We cannot simply refrain from reporting the data, the outlier. We have to put it in the box and whisker plot, so that the reader knows that one data point was obtained like this. We do not know why this data point was there. For the purpose of general calculation we can go ahead with the rest of the data, ignoring this, but this data must be presented in the paper, because there is a possibility that this data was actually not an error, but due to some physical process we got this data. This might provide information for further development of science. Therefore, this should not be missed. This should be presented.

But for the rest of the calculation, we can go ahead with this. That means, by actually doing the box and whisker plot, we identify the data and thereby we drop this particular point, and we can then continue with the rest of the data to calculate the mean, the standard deviation, the value and the error bar—all that can be calculated by taking the rest of the data.

But this number has to be presented using a box and whisker plot, so that the reader identifies that as a obtained outlier. We do not yet know why this outlier was there. But then that has to be presented. That cannot be missed, that cannot be hidden, you have to present it.

So, with that I will end today's class. I will continue with that in the next. But the main point that I wanted to make today: there are two important points, take home messages, is that there are certain ethics of science, details of which I will come to later. Whatever result you get, you have to present. You might know that certain results are erroneous. Still you have to present, and state that it is an outlier: I believe that it is erroneous.

Therefore, I am calculating the rest, going ahead with the rest of the calculation with the other points, not with the outlier. And whenever you are doing a sampling, the sample that you got that gives you a result, which may not tally with your expectation, your belief, and then you cannot say that, no, this particular data set is bad, and therefore, I will take another data set which will conform. No, that is not the proper way of science.

In doing science, when you do an experiment, you would have some expectation alright, but you have to do the experiment in such a way that your subjective beliefs do not interfere with your judgment. Your subjective beliefs do not interfere with the outcome of the experiment. This is very important. And we will have more to deal with this, but I just wanted to mention before we end this class.