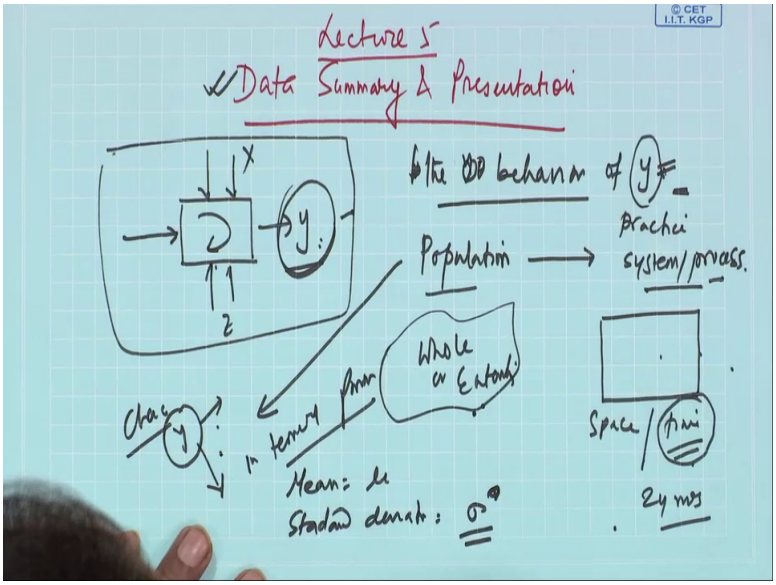
Design and Analysis of Experiments
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur
Introduction and Mathematical Preliminaries

Lecture – 05
Data Summary and Presentation

Student: Yes, sir.

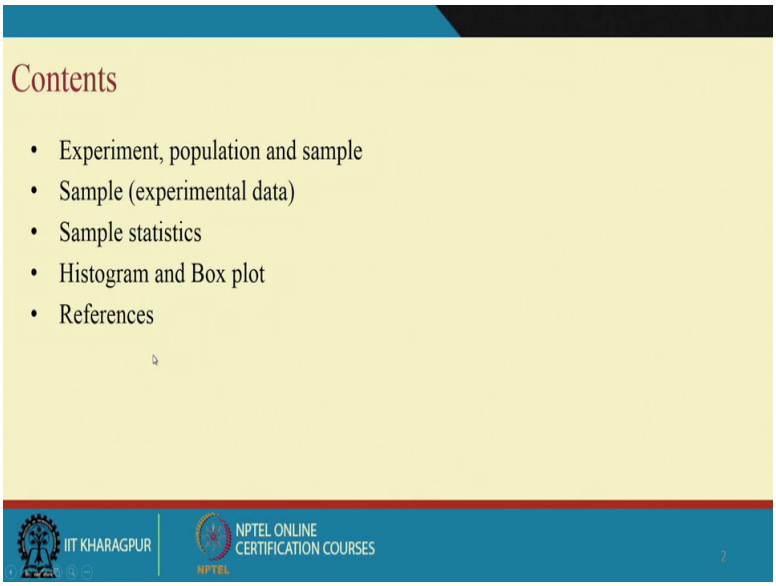
Welcome to the fifth lecture of design and analysis of experiment.

(Refer Slide Time: 00:29)



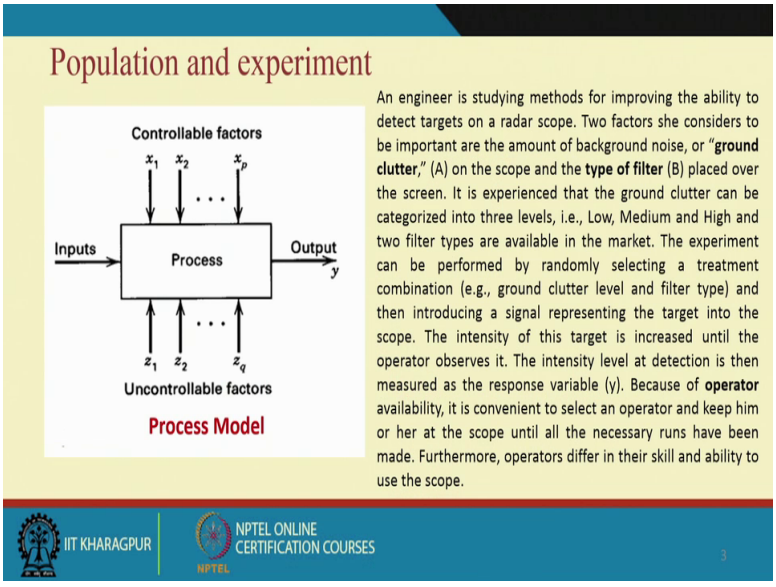
Today, we will discuss data summary and presentation.

(Refer Slide Time: 00:46)



The content of today's presentation is experiment, population and sample, sample and sample statistics, histogram box plot and some other measures like median, mode, quartile, inter quartile range all those things. So, let us start from where we ended in last class.

(Refer Slide Time: 01:01)



Last class, I explained that, there is our process model. So, in the process there will be input, there will be controllable factors, there will be uncontrollable factors and then there is a response variable y and we are interested to know the behavior of y.

So, the process can be a physical process, can be a virtual process, can be a manufacturing system and can be a service system whatever maybe. So, as you know that the device objectives are to find out the behavior of y. This is the first and important objective that behavior of y means the response variable. So, when we conduct the experiment we get data on y, apart from the information related to x, z and inputs. For example, in this study that engineer is studying the method for improving the ability to detect targets on a radar scope; what I have explained in the last class.

So, you see that here the response variable y is the intensity level at detection. Suppose, you are interested to see the behavior of intensity level based on experimental data; the sole purpose of doing this experiment is to analyze it and based on analysis we have to infer about the radar scope, radar scope behavior, in general; in the sense at the system at the process level.

So, in statistics there is important concept called population. So, and it is equivalence in practice, we can say that system or process system oblique process. Population means the entirety or whole or entirety.

Student: (Refer Time: 03:32).

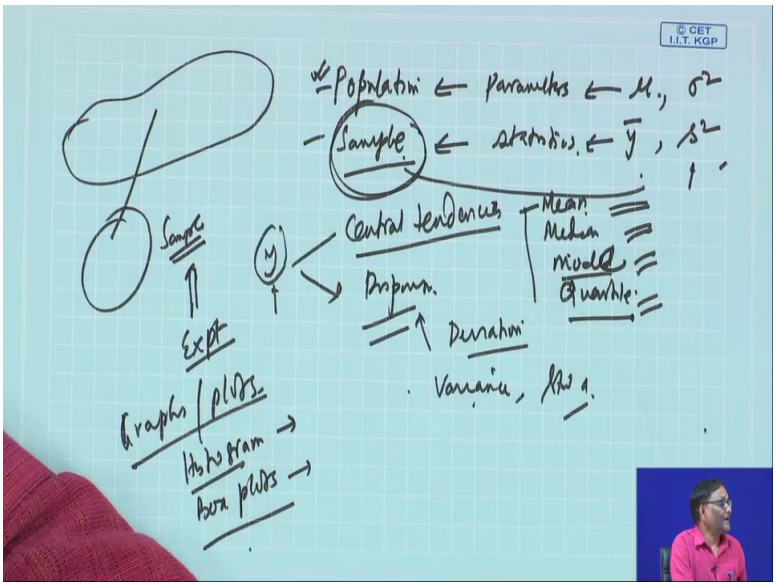
And in system means, the system within a particular boundary, that boundary will be space boundary and or time boundary. So, that means, if we consider the radar scope example; here what happened, the space is that we are interested in the behavior of the or the performance of the radar scope, given ground color clutter and type of filter used and also in the presence of difference level of operators. And what is the time? Time is that maybe we will conduct certain experiment, like that 24 numbers of experiments we have conducted.

In that case, when the time is such a sort, it will not represent the population. The system will not represent the population. Suppose, if I consider the entire life span of the system or the process, then we will get or we can do that lacks of millions of such experiments or in real life what happened, that the entire life cycle of the process or the system that there will be a billions of uses or millions of items will be produced. Now, all those taken together will represent the population.

So, when we talk about population another important thing relates to population is that population with respect to certain characteristics. These characteristics usually, the y, the

response variable and we have seen here, in this case the intensity level at detection. These characteristics can be summarized in terms of parameters. So, what are those parameters? Parameter can be the mean and standard deviation, sigma. But, then may be many more parameters, but as we know that we cannot wait for observations of the entire span of the system or infinite counter will be infinite number of items to be produced, what we basically do? We do experiment or we collect data in our case it is the experimental data.

(Refer Slide Time: 06:27)



So, that means, we have such population and from here we get certain represented number observation that is the sample and this sample, here, in our case what the experiment we conducted that is this experiment gives us some data. This data represent or constitute the sample. So, population followed by sample; this is another important concept. Population is characterized by parameters and sample is characterized by statistics. So, if I say the population parameter is mu then the sample statistics will be the sample average y bar, because y is the variable of interest. If I say population parameter is sigma square, then sample statistics, corresponding sample statistics s square will be sample variants.

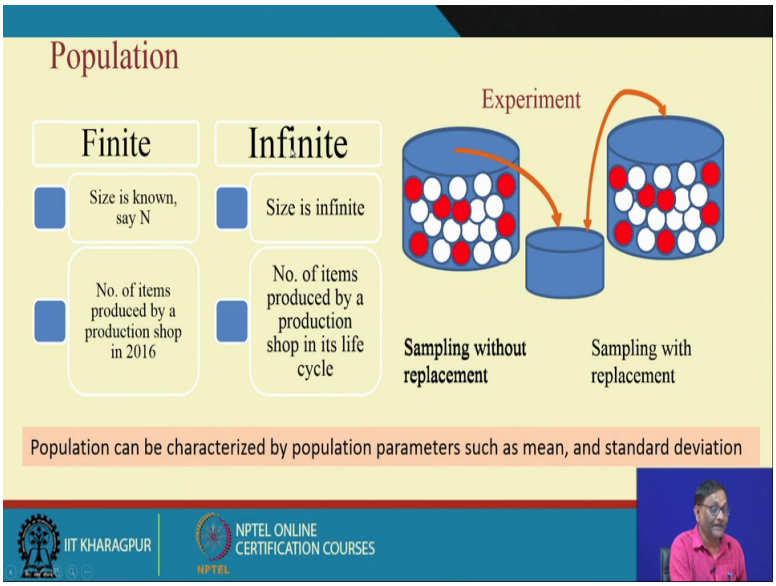
So, the population variable y, it has central tendencies, it has dispersions also. By central tendencies we mean; mean, median, mode, then your quartile usually. There may be percentile quartile others. By dispersion, we mean that what the deviation from the target is basically deviation from target. If it is deviation from the mean, then it can be it is known as variance; square root of variance is standard deviation. So, what we will see with the today,

that can we get some measure of the central tendencies or the deviation of the variable or response variable y or the random variable of interest y.

So, in order to know all those things you can have conducted experiment and you got a sample. Now, you compute corresponding that sample estimates; like for mean y bar, for standard variance s square, similarly, that the data of the median value, the mode value, the quartile value and all those things. So, another way of presenting or summarizing the data is through graphs, plots. So, there are many ways, but for most important one is histogram, because it will give us the idea that what will be the tentative distribution of this response variable y. Another one is box plot, this also this gives us that what is the central tendency as well as the dispersion in the data from 1 plot.

So, these few things we will be describing within half an hour of time.

(Refer Slide Time: 09:59)



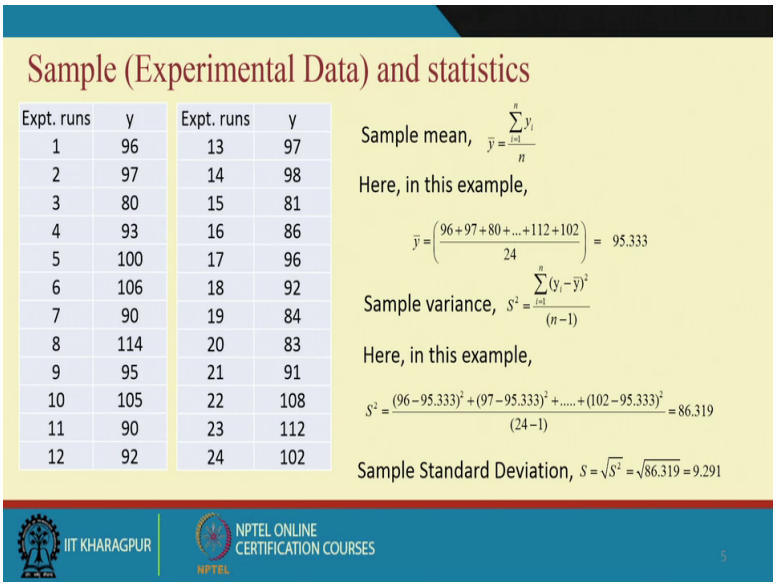
So, population can be finite and it can be infinite. Finite mean it is accountable with size n, infinite means size infinite. So, if you want to get a finite population you do sampling without replacement and if you want to get infinite population you do sampling with replacement; by without replacement we mean that suppose, you consider a container with red and white balls and you have certain parameters to measure like this that what is the proportion of white ball in this container. So, then what you do? You pick up 1 ball and keep in this container, then again, see that what color it is, again you pick up another 1, see color and if you do for

substantial number of times what will happen you will get some white and black ball and then you find out the proportion.

So, this is without replacement and as a result what happened number of balls in the main container will be reduced and if you continue doing this for a sufficiently large amount of time, there will be no ball remaining in this container. That is simple finite one, because there is no other observation or item to be measured. In the second case, you are doing the same thing, but you are replacing this. As a result, what is happening; the number of balls in the container remains constant all through. So, you can get infinite number of observations.

So, as I told you population is characterized by mean, standard deviation; these are the parameters of the population and we basically infer these population parameters from sample and in our case this sample is coming through experimental data.

(Refer Slide Time: 11:52)

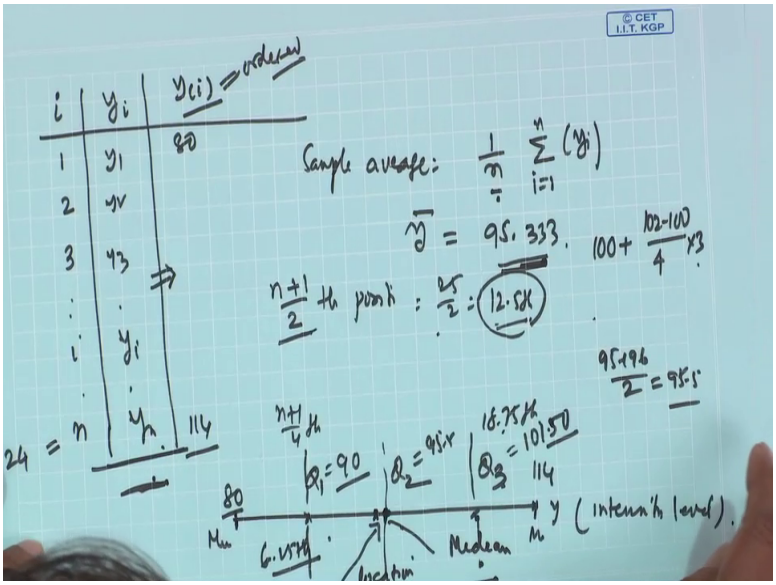


The experimental data is constituting the sample. Here, I want to tell you one more thing that here there although the controllable factors are there and we have seen that ground clutter as well as the type of filter contributes or effect in y, the response variable, but in this case we are not considering all those ground clutter and type of filter x variables, just for the sake of presentation here, that is we are only considering y values. We are assuming that is what we object from the experimental data, because if we bring those x that control variable. So, we have to have some additional information and accordingly we require some different way of

analyzing, maybe computing statistics for different levels of ground clutter, different levels of filter.

The sense of presentation, now in that suppose given a data, you got an experimental data set; now, how to compute mean, standard deviation variance and how to go for different kind of plots? So, let us say that there are 24 observations and if as you have 24 observations; in general the data structure will be like this, sample data structure will be like this i or j. Let us start from i 1, 2, 3 like I suppose there will be n number of observation in this experiment n equal to 24.

(Refer Slide Time: 13:35)



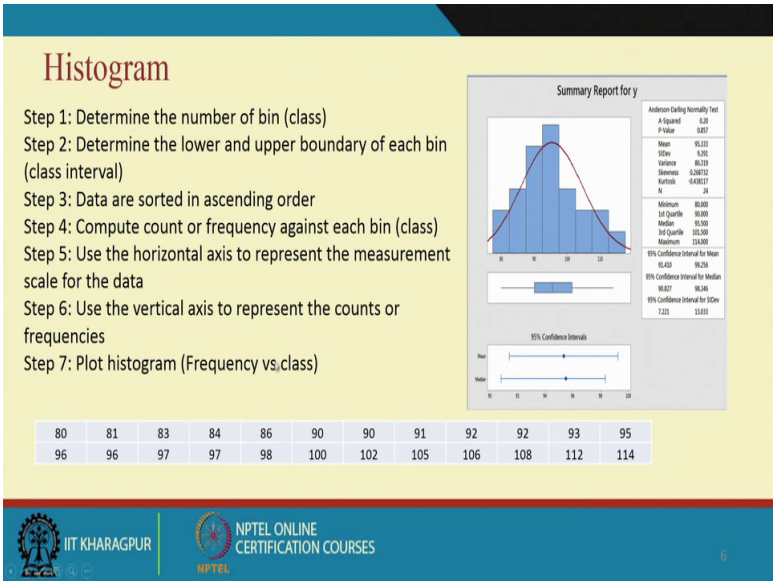
Then, you have y values which are basically y i. So, y 1, y 2, y 3, like this y i then y n, these number of observation is there. This is what your sample here is.

Now, you want to compute the sample average. Sample average you can compute like this using this formula 1 by n i equal to 1 to n, y i; that means, you take sum of all those values and sum of all those values for this given example and when you divide by 24, that is the number of observations, then you will get sample average y bar equal to 95.333.

So, what does it mean? We say that if we plot, suppose, this is my y, which is the intensity level and this is the minimum value and this may be the maximum value then average minimum value for this example is 80, maximum value 114. So, 95.333 it will be somewhere 19 and 15 maybe somewhere here.

So, this is the location, this represent the location on the data stream. So, this is average. Now, let us considered that what will be the median of the data sheet? That mean, I have y 1 to y n data, what you do? You find out y within bracket i, this is the ordered data; ordered data means from minimum to maximum. So, minimum is 80, maximum is 114 and in between the data sheet will be there, just I want to show you this data.

(Refer Slide Time: 15:31)



Here, you see this is the ordered data 80, 81, 83, 84, 86 like this 100 and this.

So, what you want to know? You want to divide it into 2 halves; 50 percent left and 50 percent right. What you do? You first you find out n, then you find out n plus 1 by 2th position. As n equal to 24, so 25 by 2 it is 12.5th position. So, if I say this is my first position and it is my 24 position, so somewhere here it will be 20, this is 25.33. So, somewhere here what will happen? Somewhere here maybe this is the position of median because this is the 12 point fifth position, 12.5 position median.

So, how do compute this? What is the value? So, what you require you find out that 12.5 these position is not it is a count type of data, you are not getting that 12.5. You are getting another 12th position, data value is y value is 95 and 13th position 96; you take the average of the 2. So, 95 plus 96 by 2, it is 95.5. So, the median value is 95.5. Now, you may be interested to know the quartile, that mean, by doing median you are dividing the data stream

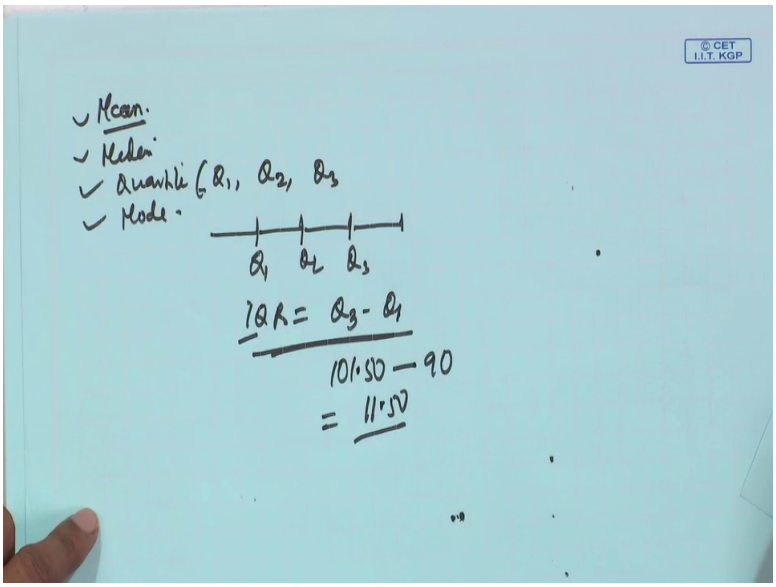
into 2 half's, by making quartile you are basically finding out the 4 parts. This one is known as first quartile, this will be known as third quartile and median is your second quartile.

So, if median is n by 2th position data, then first quartile will be n plus 1 by fourth position data. So, this is 12.5 by 2, it will be 6.25, this position and similarly, this will be that 12.5 plus 6, 18.75 this fifth position. What do you require? You have to go to the data and see that what is the 6.25th position?

So, 1, 2, 3, 4, 5, 6; 6th position is 90 and 7th position is 90, in between what about data either that will be 90. So, first quartile is 90, second quartile we found 95.5. Now, what will be the third quartile? Third quartile will be the 18. So, 12, 13, 14, 15, 16, 17, 18 this is 100 and 19th position is 102. So, that means what happened, your value will be 100 plus 102, minus 100 by what happens, 4 into 3, where is in it is 0.75, 3 by 4 this position. So, these value the third quartile value this value will come around 101.50.

Another one is mode. Mode is basically, now suppose you just see that the data, which data say is occurring the maximum amount of time. So, if we see this data 90 occurring twice, 92 occurring twice, then 96 twice, then these 3 occurring twice, all other once. So, then it is maximum frequency is 2 and there are 90, 92 and 96 so that means, multimodal case, 3 mode so; but, as the data set is very small, so whatever maybe there fine. So, this is what is mode.

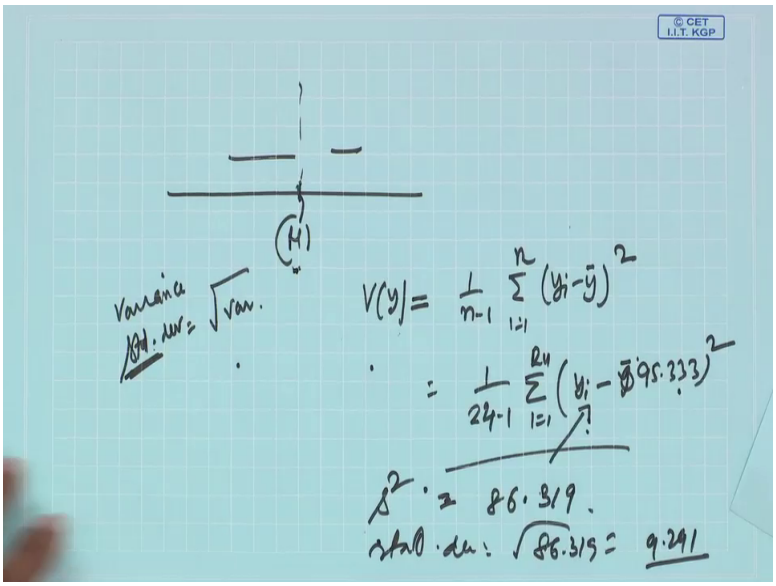
(Refer Slide Time: 19:40)



So, that means essentially what you now know? You know the given a data set how to compute mean, given a data set how to find out median, given data set how to find out quartile, given data set how to find out mode, all those things you know.

Now, quartile means you are dividing the data into 4 parts and as a result first quartile, second quartile, third quartile and fourth quartile; Q 1, Q 2 and Q 3. So, this Q 1 minus Q 3 is known as IQR – Inter Quartile Range, is Q 3 minus Q 1 this is known as inter quartile range. So, given this data as we know that the quartile values are Q 3 is 101.50 and Q 1 is 90, so that means, inter quartile range is 11.50. I hope that given data you will be able to compute all those things.

(Refer Slide Time: 20:49)



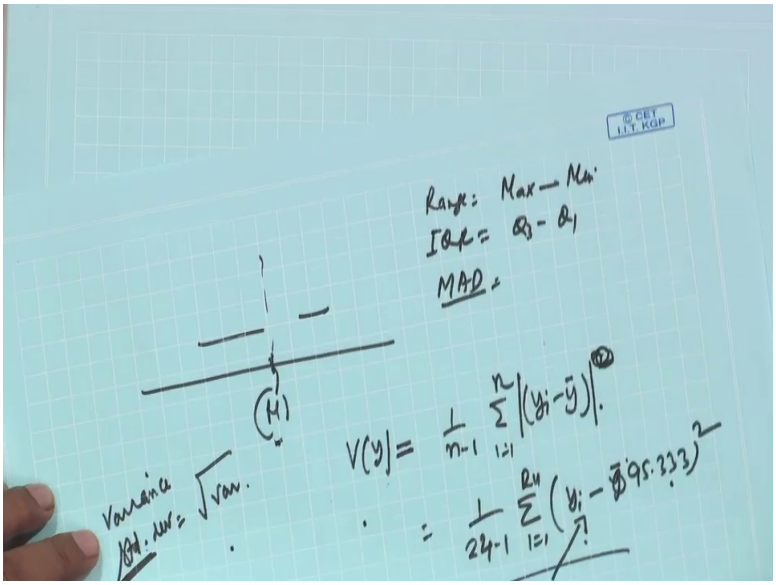
So, now, we will discuss about dispersion. As I said a dispersion means, now, this is my data. Suppose, you consider any central tendency for example, let the mean value is here mu. Now, we want to know how the data spread left to right with reference to this. So, this is known as dispersion. Now, it can be mu, can be any other target value also.

The measure of dispersion is variance and whose and standard deviation; standard deviation is actually the measure of dispersion and which is basically standard deviation is basically a square root of variance. So, how to compute the variance of y from the sample, it will be 1 by n minus 1, sum total y i minus y bar square, i equal to 1 to n. In our example n is 24. So, 24 minus 1, i equal to 1 to 24 and y i values are known and y bar value we computed, what is the

y bar value, y bar value I'm not computed, some value will be there y bar. So, y bar value I think 95.333.

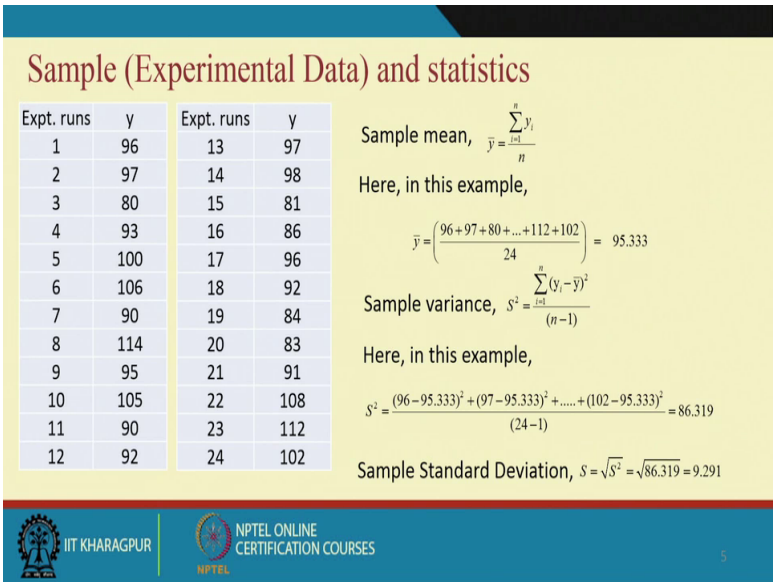
So, this square, you put all the 24 observed values here and subtract it from the mean and what you get that will be your standard that will be your variance. We are saying that V of y, which we basically defined in terms of s square, sample variance s square and if you use this formula you will get it 86.319. Then standard deviation is equal to square root of 86.319 which is 9.291.

(Refer Slide Time: 22:46)



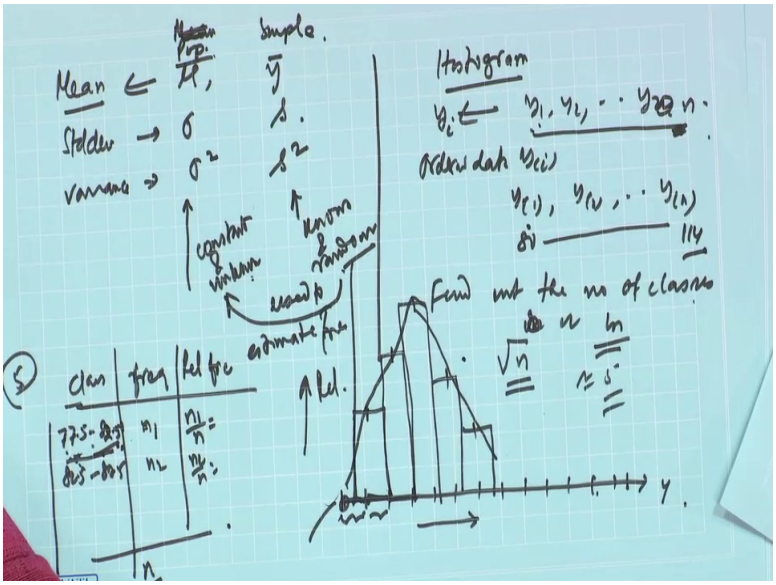
So, apart from variance and standard deviation the other measure of dispersion is range, that is, max value minus min value. IQR - inter quartile range, that is, Q 3 minus Q 1. There is mean absolute deviation; MAD – Mean Absolute Deviation, from the deviation from mean sheet of square, you just find out mean absolute deviation that may be mean of this one. That square will not be there, that absolute value, whatever value you got this is one also another measure of this thing.

(Refer Slide Time: 23:22)



So, in DAE design analysis experiment, we will be primarily concerned with mean with respect to population it is mu, with respect to sample it is y bar.

(Refer Slide Time: 23:37)



So, mean population and sample. We will be interested to know the standard deviation. So, with respect to population we say it is sigma and here we say it is s. We will be interested to know variance with respect to population is sigma square and with respect to sample it is s square. So, please keep in mind these are population parameters. Population parameters are

constant and unknown and these are the sample statistics. Sample statistics are known and random.

The sample statistics are used to estimate the population parameters. As population parameters are not known, so we require an estimation of that; where population parameter parameters are known I do not think the statistics is useful there. Now, I will tell you some graphical plots; first one is histogram. You see this data set and this is the histogram it is it is basically drawn in many time.

Now, what is the procedure of drawing a histogram? First one is, you have data; suppose data y, here it is 20, y 12 to y 22 to y 24, actually, I will write y n. So, what you do, you arrange them ascending order then. You find out ordered data. So, instead of, if I write y i is the data, ordered data we can write y within bracket i. So, that mean it will be y bracket 1, y 2 and like y n. Remember, these data and these data are not same. These are y 1 is not y within bracket 1, this may be some value, here it is 80 and it is 104 for the order case, but for the un order case it is different. It may be something else y n value, whatever order the data.

Then, find out the number of class or number of classes or means; mean what you want actually, when you develop histogram, you have, it is your y axis, instead of taking considering one every data point one after another you basically you know the min and max; you want to make it some equal intervals, large number of intervals like this. And, you want to find out, suppose, this is a value you choose a lowest value and higher value and this is known as first class.

So, this is my class 1 or mean 1, this has mean 2, mean 3 like this and every class has the lower boundary and upper boundary and then what you required to do, you select class; for example, in this case we may select that 77.5 to 80 maybe or 82.5, that may 72.5 to 82.5; that means, difference is 5. So, this what is the number of what is the frequency how many data points falls under this let it be n 1, then 82.5 to 87.5, let it be n 2 like this.

So, first what you do, we find out the class. So, you use root n, if n is 24 it will be around 5. It is just a trial and it is just a, what I can say guess it is a starting point. You may not be happy with a 55, you may go for some other values also, but whatever may be the thing, you have class, you find out all the class and their lower and upper boundary and then in each class what is the frequency. So, this side will be frequency; what do you do basically, here suppose this is my first class, you plot the frequency and draw a bar like this, for the second class

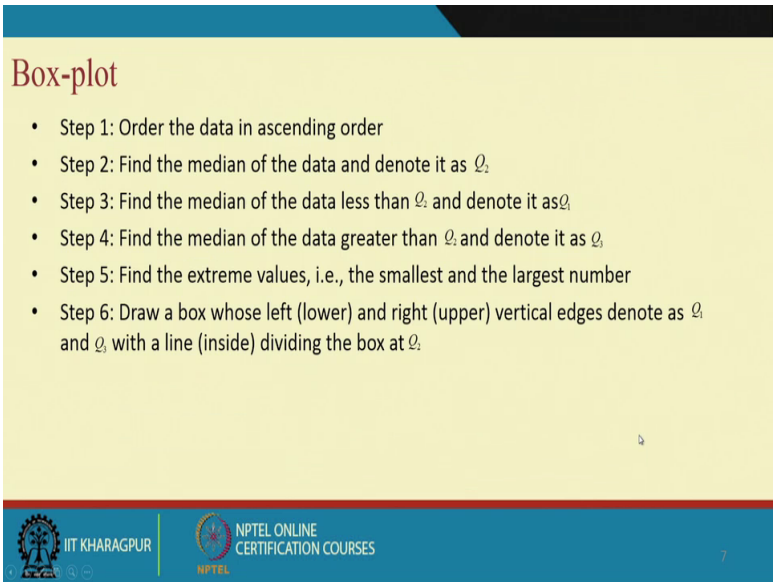
suppose this is the frequency draw like this, third one maybe like this, fourth one may be like this, fifth one may be like this.

So, that means, x axis the y value, y axis is the frequency. Sometimes what happened, you will use relative frequency; means, suppose you have total n number of observations, so, n 1 by n is the proportion, n 2 by n is the proportion, this side also relative frequency also you can use. So, if you use relative frequency that is also histogram.

Many a times what you do, we basically join the middle point of the bar and find out a graph. This is known as frequency polygon and if you smoothen this frequency polygon you will get some idea about the distribution of the y variable. So, with reference to this data this is what is the histogram and it is developed using many tab. You see that the few things that this is the middle value 80, so, this is 77.5, 82.5. So, like this 1, 2, 3, 4, 5, 6, 7, 8 means or 8 class and there are class interval, but all are equal interval. And all the middle points this point and these point if you join you get poly frequency polygon, but here happen that this frequency polygon is more than further and you got a curve like this, which it is resembling that the data may be normally distributed.

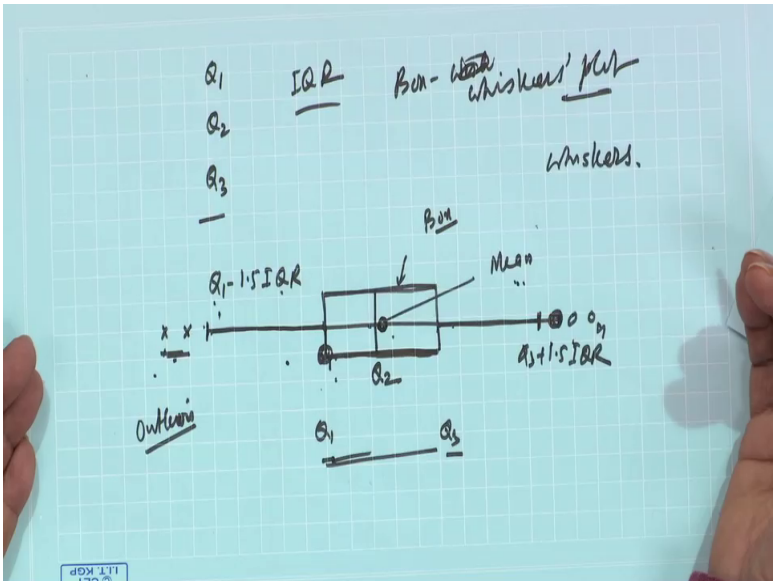
Here, you see that the mean type also given like this mean, standard deviation, variance skewness, kurtosis, so, these are other measures of this; where, plot skewness means how's, what is the is it symmetric or it is tail to left or tail to right, that is basically skewness. Kurtosis means basically it talks about the frequentness of the value; means, is the mode value is very, what is the mode? Whether the number of observations at the mode level is very high compared to the other levels or it is basically it is a smoothened one, not that much of high frequentness. So, those things are coming through kurtosis. But, here another important plot is shown here, you see that down this is known as box plot. So, I want to show you how this box plot is developed because it is very important plot for all of us.

(Refer Slide Time: 30:58)



Now, in order to draw box plot what you require to do the steps are given here. But, you required to find out all the quartiles that is Q 1, Q 2, Q 3. I already told you how to draw Q 1, Q 2, Q 3.

(Refer Slide Time: 31:11)



So, that mean this is my data axis y. So, I know that this is my Q 1 and this one maybe your Q 3. Suppose, this is Q 1 and this one is Q 3. So, what do you do, you develop a box here Q 1 and Q 3 with certain width. So, this is known as box. Then you also plot the median value

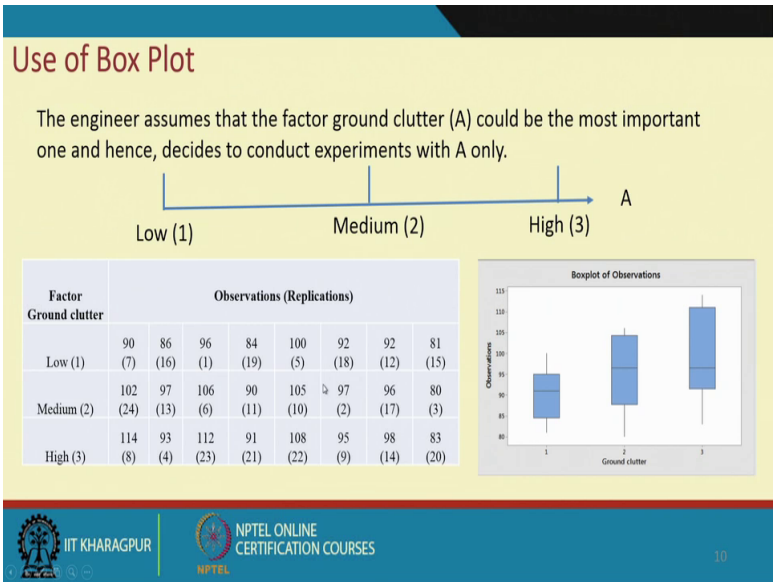
there, maybe median is somewhere here. Let the median is here, this is Q 2. You may plot mean value also just like a small circle, you can plot mean also; but, usually you do not put mean here. We put all Q 1, Q 2, Q 3, then you find out what is the value Q 1 minus 1.5 times IQR.

IQR already you have seen. IQR is the difference between Q 3 and Q 1. When this is first quartile, minus 1.5 IQR, similarly, this one this may be that Q 3 plus 1.5 IQR. So, my box or in the box plot these are not only box, there is some another line the spread line, left to Q 1 and right to Q 3 these are known as whiskers. That is why this plot is also known as box and whiskers plot. So, here some of the steps are given. So, you see that we said that first find Q 2, then find Q 2 left to Q 2 data set, that is, Q 1 and find Q 3 then find the extreme values. It may so happen that there will be some values here, some values here also.

So, these are known as extreme values; means the values y values which are beyond Q 1 minus 1.5 IQR or here below this value or above Q 3 plus 1.5 IQR, these are extreme values, these are known as outliers. So, usually we do not consider, this does not belong to the data. The advantage of having box plot is, that it not only tells you that where the median lies, it also give you that where the most of the data fall on which range because this IQR this is the range where most of the data falls. And it also give you that what is the spread this side and this is in the variability beyond that also. So, as mean can be proved, mean can be placed here also. If we place mean this will give you the location, mean, median, quartile. This will give you the dispersion, that is, the spread of the data set. From one plot you are getting central tendency and the spread or the dispersion in the data.

So, it is a very important plot and in fact, when you conduct experiment all the time what we will do after getting the experimental data, we will go for box plot to see that whether the factors level is really affecting the mean of the response variable or not so that the one diagram I will show you now.

(Refer Slide Time: 35:14)



You see, that last lecture with one factor complete randomized design, we found that the ground clutter has 3 levels and in each level there are 8 number of replications or data points; now and what we have done here, we have developed 3 box plots one for y, that is the response the detection level when ground clutter is at low level, one for y when ground clutter at medium level and another for high when ground clutter at high level. So, you see what happened here if I develop box plot for first one, second one and third one with reference to the ground clutter, low, medium and high and this is the box plot. What can you infer from this box plot?

You can simply infer that the median value here and the other 2 they are different. Here, median is lower than these 2, for this low and this value is around just above 90. Here, it is above 95. If I consider mean will be somewhere here, mean will be somewhere here, mean will be somewhere here; so, that means, if you change the clutter ground if you when you use the radar scope when the that environment condition in terms of ground clutter it is low, you will detect this quickly, because the signal detection level is lower. If ground clutter is high your mean detection time will be more, this is one.

A second interesting phenomenon is there; the variability which is represented by the box, length of the box. See, when ground clutter is at low level this length is, this is a small r compared to medium ground clutter level, compared to high ground clutter level. So, that mean if the ground clutter environment condition, the ground clutter level is high; it means

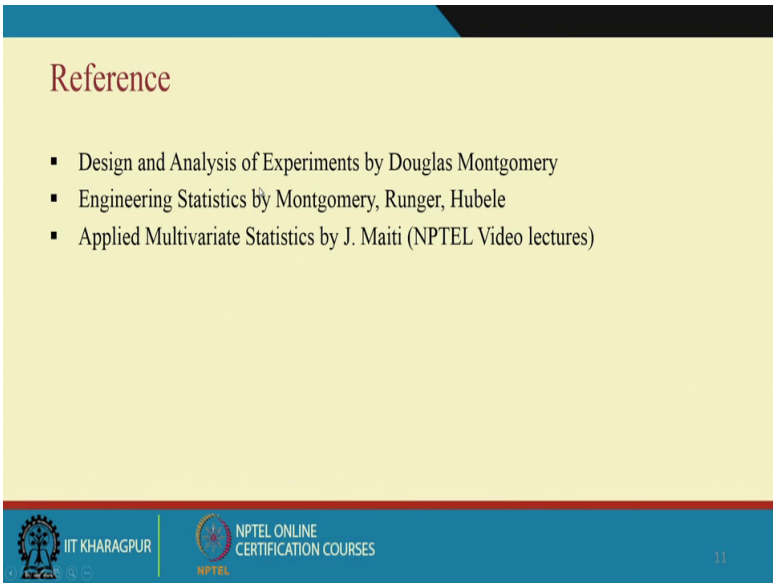
that not only the mean response time is high, in addition the variability is also high; that means, suppose you are using the radar scope when you detect the target if I go and use the radar scope I will take may be time much different than for both of us when the ground clutter level is at low level that is the one level.

So, from this plot we can say that there is difference, but whether this difference is significant or not statically significant or not this will be I can say tested by some techniques. We will discuss analysis of variance, that time I will show you how this will be tested, whether the mean response for the 3 ground clutter level different or not. If it is different which of the pairs are different, all those things will be tested.

So, with this I conclude that when you have data you must know how to represent the data in simple tools; like box plot, like histogram, there are pie chart, bar chart other things are also there, Pareto chart and also you must know that what is the tendency of the data; in terms of central tendency, that is, mean, median, mode, quartile all those things. Also you must know what the spread of the data is.

Depending on the sample size, these tendencies or spread can be thought of representing the population central tendency and population spread. So, whether they are really representing population level information or not that will be done through different kind of testing later on.

(Refer Slide Time: 39:29)



So, for the time being, I again tell that all those, whatever things I have given to you these are taken from different available books and some available videos, like here I have used Montgomery – Design and Analysis of Experiment, Engineering Statistics by Montgomery, Runger, Hubele and Applied Multivariate Statistics video lecture of mine.

Thank you very much, thanks.