Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

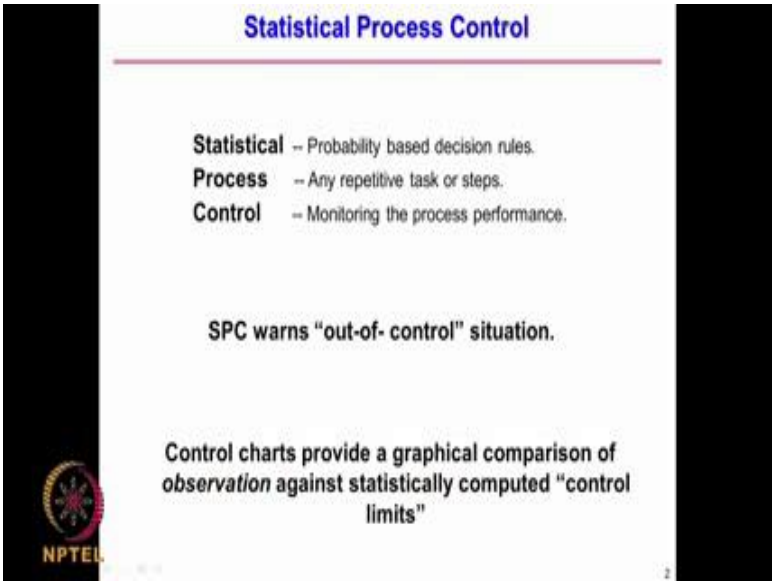
Lecture - 40
Control Charts

Welcome to the course on Biostatistics and Design of Experiments. Now, we are going to talk about something called Control Charts. These Control Charts are very useful, especially in manufacturing, especially if you are running bio process, when you are manufacturing bio materials continuously, day in and day out, you want to know whether you are able to achieve everyday, products within that specification, is, there is a slow drift from within limits, going out of limits and when should I get worried?

So, Control Charts in very, very important in industrial scenario, be it bio process, be it chemical process, be it manufacturing of goods, manufacturing of bio materials, even manufacturing of bolts and nuts and so on, actually. So, it tells you, it warns you apriori, whether you are going to go out of control; because, ultimately, I want to make every day, the same product with the same specifications; I do not want to go out of my specification, because, that is not good, number 1. And, number 2, customer, if he gets that sort of products which are outside the specifications, he will consider it as a defective product. So obviously, he is going to return it back to you. Which means, you may have to rework it; that means, it is cost or if you just dump it, that is a waste. So, you are going to lose lot of money.

So, Control Charts have become very important in the past 20 years and Japanese, who always believed in the concept of quality, believed that whatever products that are leaving their company should be within that control limits, within the specification limits, so that the customer does not get worried, customer gets satisfied, customer does not return products which are not under specific, in specification, which means, you are losing money, you are losing your profits. So, this statistical process control, this is a probability based decision rules, you are looking at the process; that means, any repetitive task.

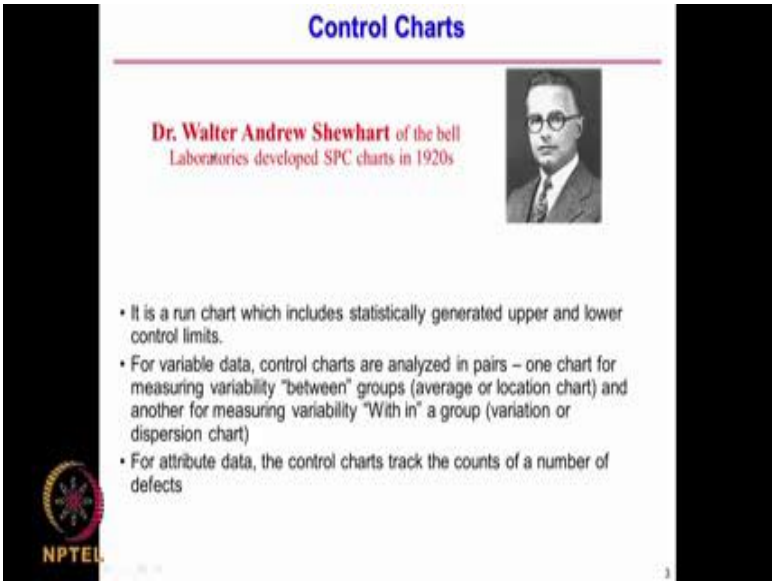
(Refer Slide Time: 02:24)



You are manufacturing a metabolite day in and day out, you are manufacturing antibiotic everyday for years and years, you are manufacturing a diaphragm valve everyday, you are manufacturing thousands of bolts everyday for years together, that is a repetitive task. So, you want to manufacture within the limits, of your control limits, specification limits. Do you understand? So, if I am saying that my diaphragm valves I manufacture will last for 400,000 cycles, that means, it should last for 400,000 cycles. Control, which is monitoring the process performance, that means, you want to monitor and keep it, see whether it is going to get bad.

So, the statistical process control or s p c, it warns out of control situation. So, if you are going to go out of control, it tells you, hey better watch out. So, it gives you a graphical comparison of observation against statistical computed control limits. So, over a period of time, it can tell you how the average size of the screws change over a period of time, within + or -1 sigma, or within + or - 2 sigma, or within + or - 3 sigma; is it slowly going up, up and up, or is there a randomly its varying; or sometimes, it goes out of 1 sigma, 2 sigma, 3 sigma or all the time it is within that regions. So, it tells you a graphical picture and it is very useful in the manufacturing scenario.

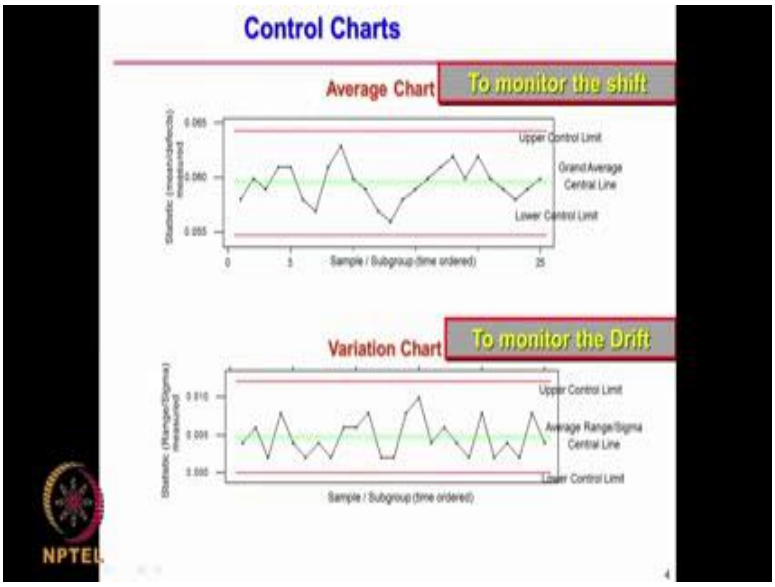
(Refer Slide Time: 04:08)



Control Charts. Walter Andrew Shewhart of the Bell Laboratories originally developed these charts in the 1920s. It is also called run charts. It gives you a upper and lower control limits; like I said, we can have + or - 1 sigma, then + or - 2 sigma, + or - 3 sigma, and everyday you plot the average flexural strength of a material. So, you keep on plotting it, and see whether it is all the time lying within that plus or minus 1 sigma, or does it go sometimes 3 sigma, or does it go beyond that, and so on. For variable data, control charts are analyzed in pairs. So, one chart for measuring the variability between groups, and another, for measuring variability within a group; that is variation or dispersion chart. So, variability between is average; other one, you call it dispersion. That is, variable data control charts. So, you take 2 data points; so, you take the average, so everyday's average you plot, everyday, day to day, tomorrow, day after, and so on, and the same time, you have another graph which gives you the difference, or the range, or the variability and keep on plotting that. So today, suppose, I got a bolt of a 10 and 10.5 diameter. So, the average will be 10.25; the variability will be 0.5. So, I will have 2 graphs. If tomorrow I am getting a 10 and 10.3, average will be a 10.15, and the variability will be 0.3; like that, I keep plotting. For attribute data, attribute data is yes no that sort of thing, control charts tracks the number of defects. So, if I take 10,000 bolts everyday, and see how many defects are there, today I got 20, tomorrow I got 19, day after tomorrow I got 18. So, I plot a graph; x axis will be time, y axis will be number of

defects; attributes; how many defects. So, these are various types of control charts which we can plot, ok.

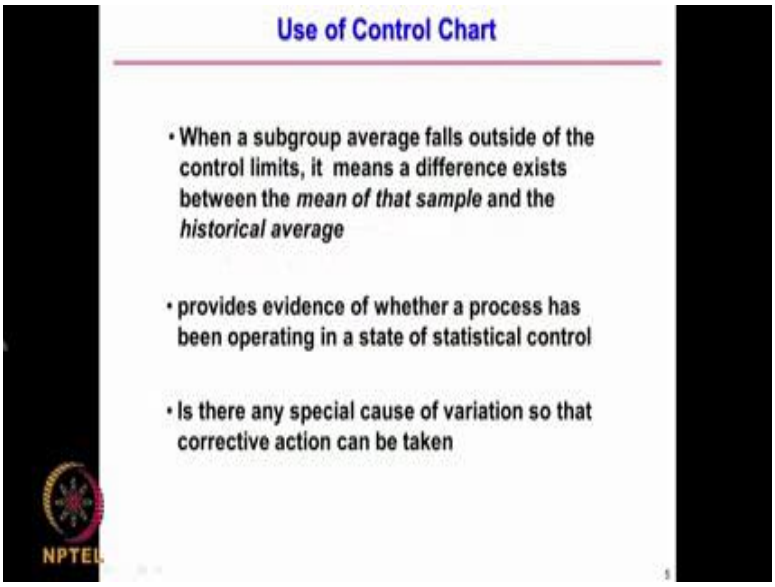
(Refer Slide Time: 06:13)



So, this a typical control chart. So, I am measuring some data here; this is an average chart, right. So, as a function of time here; this could be today, this could be tomorrow, the day after tomorrow, this, or it could be today, this is 12 hours from now, this 24 hours from now, then 36 hours, 48 hours; so, the x axis will be like that. This gives you an average of your, either the diameter of a part, or it could be flexural strength, or product percentage, or it could be biomass produced; so, as a function of time, or antibiotic concentration, or titer value and this is the lower control limit, this could be upper control limit. You can say, I will have this l c l, lower control limit, upper control limit as 3, + or - 3 sigma. So, as you keep monitoring, over a period of time, as you keep doing that, you can see some, what is happening to the results. Am I having some big variations, am I having small variations, is this variation only random or is there a pattern that is happening; that sort of understanding I can do by looking at it. So, this is the average. We can look at average. We can also, this could be the variation. So, like I said, if the 2 bolts, I take one is 10 and another is 10.1, the variation could be 0.1. In the next sample, one could be 10 and another could be 10.3, variation could be 0.3. So, like that, I could plot and I could look at how they change as the function of time. So, we can look at the

drift, how the data is drifting, is it slowly going down, or it is slowly going up, or is it random. So, we can see so many things by looking at it pictorially over a period of time. Like I said, it could be everyday, it could be every 12 hours, it could be every week, it could be every 6 hours, depending upon how I am doing my sample collections, and how important is the x axis.

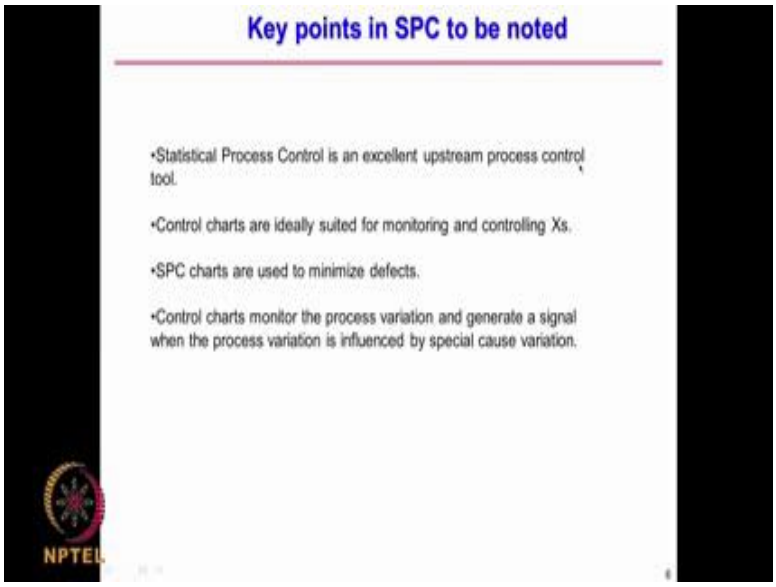
(Refer Slide Time: 08:23)



So, when a subgroup average falls outside of the control limits. So, if there is data point coming here, then I should get worried, actually. It means, a difference exists between the mean of the sample and the historical average. It gives you an evidence of whether a process has been operating in a state of statistical control. So, if it is all lying within this nicely, obviously, we can say it is within statistical control. This is very, very important, as I said, in many, many situations, where there is a manufacturing happening, chemical manufacturing, bio chemical manufacturing, antibiotic manufacturing, your bio material manufacturing. So, these types of control charts are extremely important. Is there any special cause of variation, so that corrective action can be taken? And suppose, data is going like this, and suddenly, the data all shoots up and it is going all like this. So, obviously, something has happened. So, immediately, I can check my plant; I can check my manufacturing and find out what is the reason for that; may be my raw material has changed; raw material quality has changed; may be some valves have start, well, some

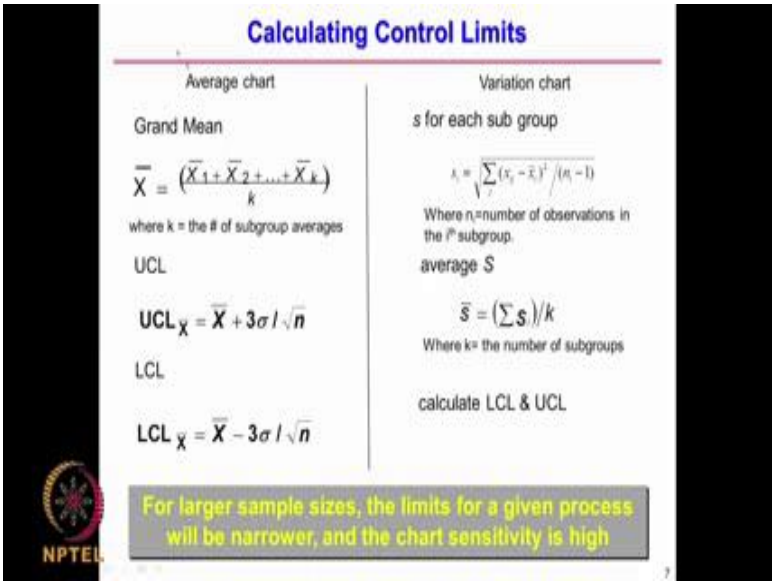
measuring instruments have stopped performing properly; its malfunctioning of instruments and so on, actually.

(Refer Slide Time: 09:41)



So, it is an excellent upstream process control tool. Control charts are ideally suited for monitoring and controlling x's. These are all x's. They are used to minimize defects, monitor the process variation, and generate a signal when the process variation is influenced by special cause. So, like I said, suddenly, the result, the results are changed dramatically and it goes to a new set of values, obviously, there is some special cause variation. It is not a random cause variation. I introduced the term this special and random cause long time back.

(Refer Slide Time: 10:10)



So, there are many types of charts; we look at it, average chart. So, I take 10 samples; I take an average and then, I plot the average against UCL and LCL; UCL is upper control limit, LCL is lower control limit; sigma /square root of n, right. You remember this, standard error. Or, it could be a variation chart. Variation chart is nothing, but you take

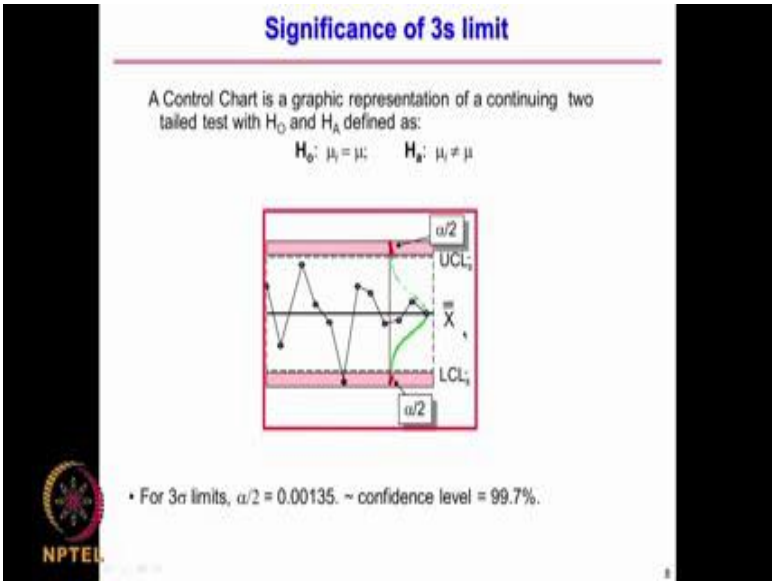
$$s_i = \sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 / (n_i - 1)}$$

, n is the number of observations, and then, we could take a

$$\bar{s} = (\sum s_i)/k$$

, where k is number of the subgroups. Then, we can calculate (Refer Time: 10:57). So, you can have average chart, we can have variation chart, ok.

(Refer Slide Time: 11:00)



So, it is a graphical representation. So, it is like, H_o: $\mu_i = \mu$,

H_a: $\mu_i \neq \mu$

. So, for 3σ , what do you do? 3σ , you take

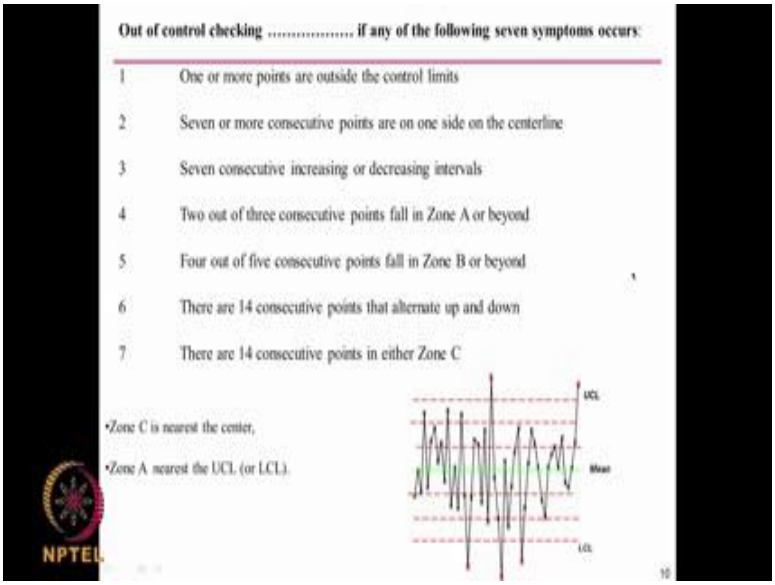
For 3σ limits, $\alpha/2 = 0.00135$

that is, you get 99.7 %, right. So, we plot a 3σ , or we can have 2 sigma, or we can have 1 sigma in the upper and lower bounds, actually.

(Refer Slide Time: 11:31)

So, this is the lower control limit, that could be - 3σ; the upper control limit could be + 3σ. Then, you will also have upper specification limit, lower specification limit. What is this control limits? Control limits is what you try to control in your manufacturing facility. Specification limit is what you tell your customer. He may tell the diameter of a bolt is a 12 mm + or - 0.1 mm. So, this specification limits, - 0.1 mm, + 0.1 mm; that is the specification. But, when you are manufacturing inside, you may have a better control; you may have control which could be plus or minus 0.05 mm. So, your lower control limit will be minus 0.05; upper control limit could be plus 0.05. You see, UCL and LCL is always inside the USL and LSL. This is what you tell your customer, and this is what you control in your manufacturing facility. Specification control limits are for averages, specifications of individual values. So, if a customer, if you are selling bolts to the customer, he picks up 1 bolt and measures its diameter, it should fall within the USL and LSL; whereas, you use this for controlling the averages.

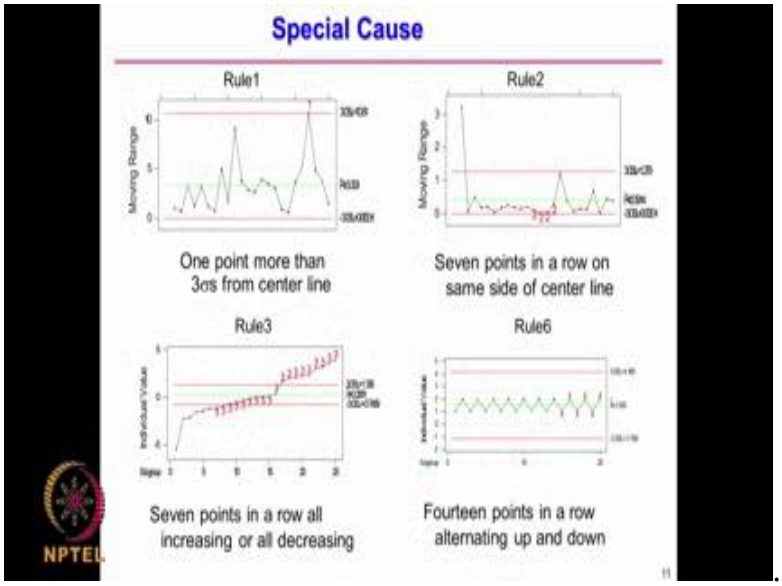
(Refer Slide Time: 13:06)



Now, you are having a control chart, you are looking at the control chart and then, when do you say if the process is going well, it is within control? Or, when do you say the process is not going well, it is going out of control? There are some symptoms. Like, there are some symptoms when you get some fever, right; your body temperature goes up, or you have a dry cough. So, you look at those symptoms and then, you say, probably you have some sort of a viral fever going on. So, those are the symptoms. So, you look at the data points and from those data points, you say, yes, my process is going to go out of control. Before it goes out of control, you try to tell that, the process is going to go out of control. So, one or more points are outside the control limits, of course. Suppose, you have points like this, then obviously, it is gone, no hope. 7 or more consecutive points are on one side of the center line. This is your center line. If 7 points are on one side, know, like this. 7 consecutive increasing or decreasing intervals; up, down, up, down, up, down, up, down. 2 or 3 consecutive points fall in zone a; zone a is this. Suppose, you have this as 3, + or - 3σ, you can have this as +or - 2 sigma, this as + or - 1 sigma. So, this is called zone a, 3 to 2 is zone a, 2 to 1 is b, 1 to 0, that is mean, as c. So, 2 or 3 consecutive points fall in zone a, 4 or 5 consecutive point fall in zone b, 3, 4 consecutive points alternate up and… There are 14 consecutive points that alternate up and down. There are 14 consecutive points in zone c, understand.

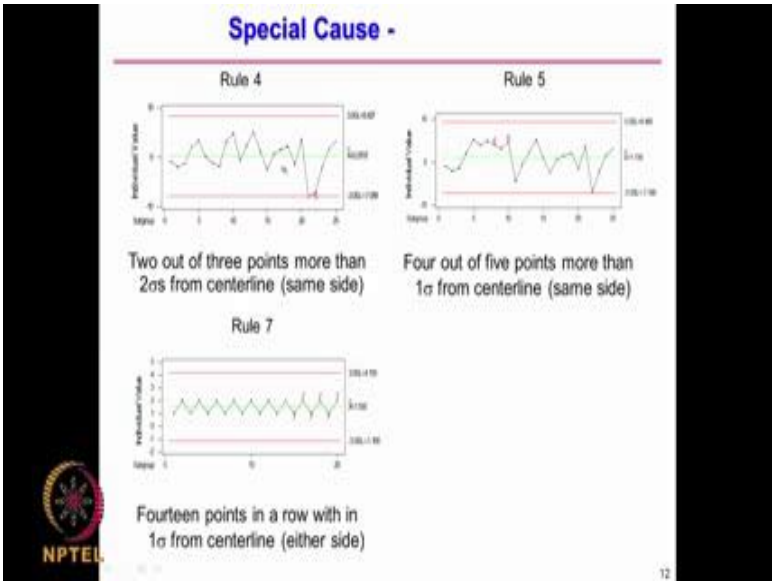
So, obviously, what it means is, if there are too many consecutive points in one zone, that means, that data is not moving random. A random cause will always have data points going up and down. So, only if there is an assignable cause there will be a pattern. So, these are the patterns. You may ask, why 7, why not 6, why not 8, yes. But, these are just taken it as the rule of thumb. So, you do not have to question that. So, 2 or 3 consecutive points in zone a, 2, 4 or 5 consecutive points in zone b, 14 consecutive points in zone c, or 14 consecutive points that go up and down, up and down, up and down, like that. So, zone c is nearest; zone a is nearest the UCL or the LCL.

(Refer Slide Time: 16:07)



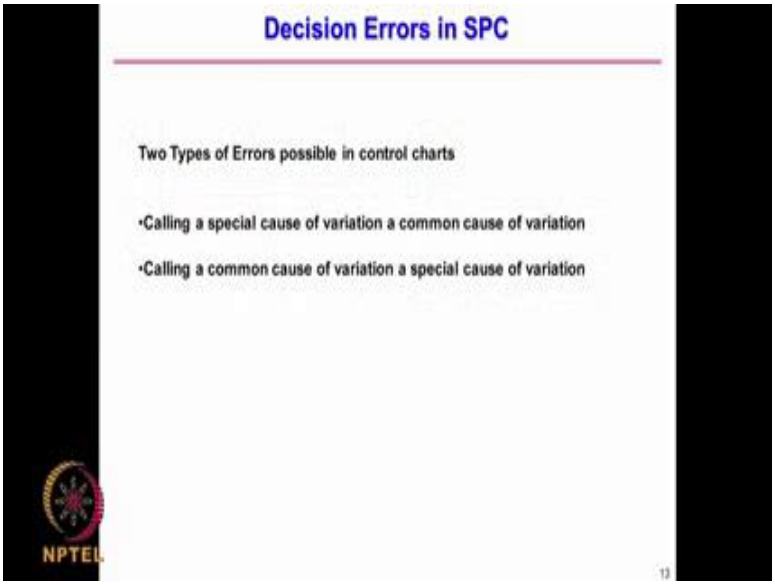
Let us look at some situations. So, obviously, 1 point more than 3 sigma, obviously this is a special cause variation. 7 points in row on the same side of the center line, so, this is the rule 2; obviously, there could be some problem here. 7 points in a row, all increasing, or 7 points in a row, all decreasing, like I mentioned here, increasing or decreasing. Obviously, there could be some special cause. 14 points, it is going like this, ting tong like this, you know, up and down, up and down, up and down, with respect to the mean. So obviously, there is a problem. You do not have to question, why 14, why not 16; yes, it could be, but as I said, it is just taken based on experience.

(Refer Slide Time: 16:57)



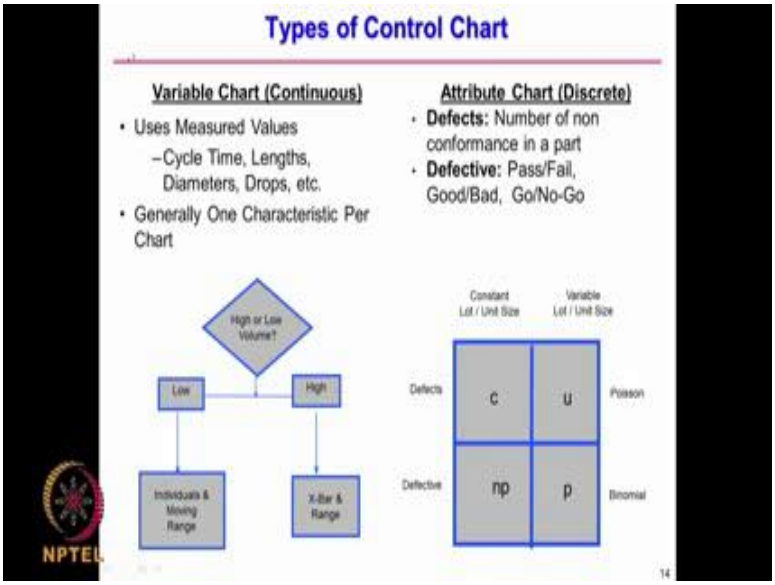
2 out of 3 points more than 2 sigma, 4 out 5 points more than 1 sigma from the center line, 14 points in a row within 1 sigma from the center line. So, all these mean, if you have such a situation, obviously, there is something going, happening with your system. There is a special cause variation. You better check out your process, so that, you do not have the entire process exploding in front of you. So, this is a very important point you need to keep in mind.

(Refer Slide Time: 17:37)



So, what are the different types of charts, you know. Of course, you remember the α and β error, 2 types of errors possible in control charts, calling a special cause variation a common cause that is random; calling a common cause that is a random variation as a special cause variation. So, both are possible, actually.

(Refer Slide Time: 17:53)



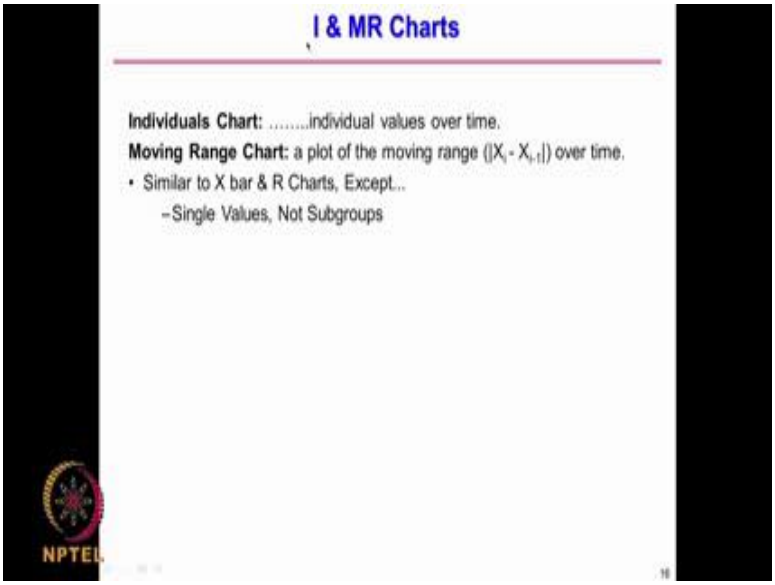
So, you have a variable chart, you have the attribute chart. Variable chart is continuous, measured cycle time, lengths, diameter, particle, or droplets. It is a one characteristic per chart. So, we can have low, individual access, moving average, range x bar chart, that means, average, or you can give range chart. What is range chart? So, you collect 2, 3 samples, and you look at the lowest and the highest, subtract, that is the range part. x bar is the average of the sample set. Moving range, that means, every time, you calculate the range based on the new set of data we have collected; individual values are also there. Now, you go to attribute chart, there is something called defect chart; number of non-conformance in a part, that is called defects chart. Defective, that is pass or fail, good or bad, go no go sort of situation. So, we can have different types of charts.

There are something called C chart, U chart, P chart, NP chart. So, if you have a constant lot or unit size, that means every time I take 3 samples and do calculations, how many C, how many defects are there, then you use a C chart. But if you have a variable lot, and then calculate the defect, then you have U chart. Again, you have constant set of sample taken, and find out the defective, that is called the NP chart. If you have variable set, and then look at the defective, that is called the P chart. So, here you use a binominal, use a Poisson here, you use a defects, defective. So, these are some of the charts that are also possible for you to work on.

(Refer Slide Time: 19:33)

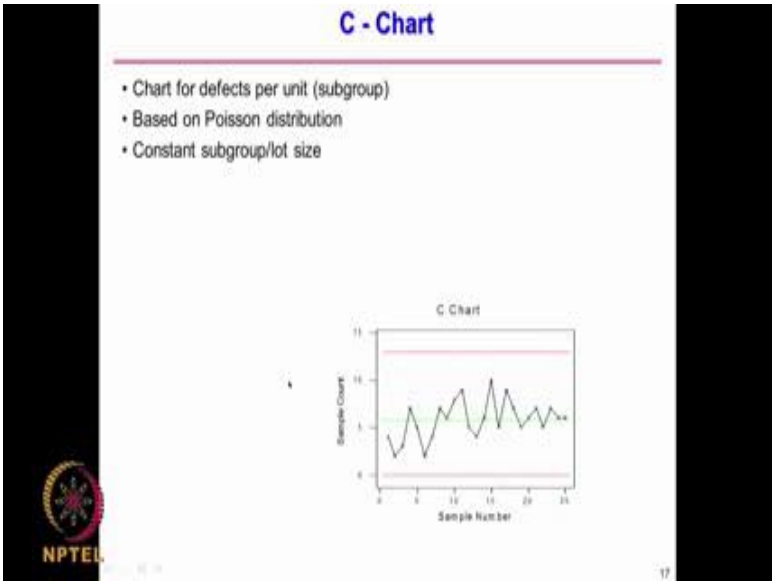
So, X bar chart and R chart. X bar is the average; R chart is the range. So, as the name implies, it is very straight forward. So, you have the X bar and this is the range. So, you plot them with +or - 3σ; points move up and down around the center line and stay inside the control predictable process; only this is called the random variable; does not indicate a best process, but still, it is a stable process.

(Refer Slide Time: 20:01)

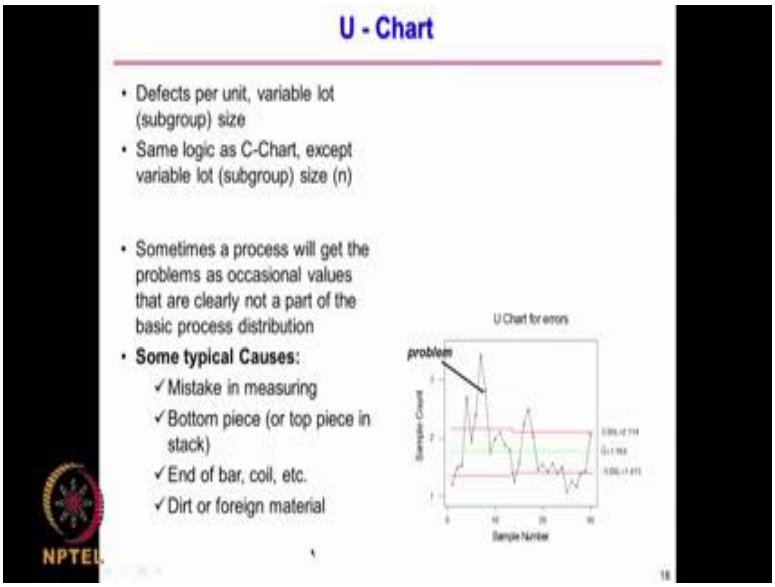


Then, you have the individual chart, that means, you plot the individual values over time. Suppose, I am measuring the pH as function of time, I put the pH as a function of time and do that. Moving range chart, so, what you do? We use x i -, x i - 1 bar over time. So, when I take another 2 samples, the x i will change, x i - 1 bar will change. So, that will have another range. So, that is called a moving range chart. This is similarly to X bar and R charts, ok, except that, it is single value, not subgroups. It is called the I and MR; I is the individual, MR is the moving range.

(Refer Slide Time: 20:46)



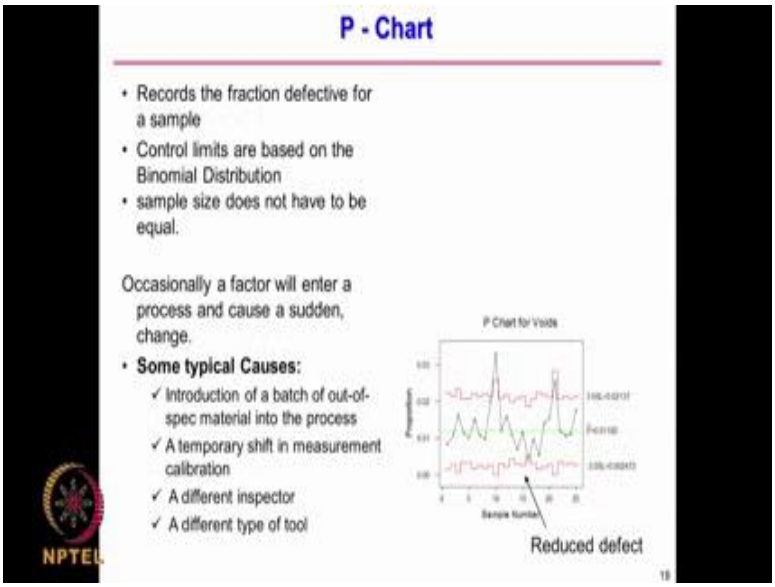
Now, we have the C chart. This is charts for defect per unit; like I said, the C chart here, if you look at this figure, a constant lot size. I take, say 10 samples and then, I calculate the defect, chart defects per unit; this is how it goes. So, every set of samples I take, constant subgroup lot size and then, calculate how many defects per unit; that is called the C chart.



Then, you have the U chart. When you have the variable lot, that is called the use. This is

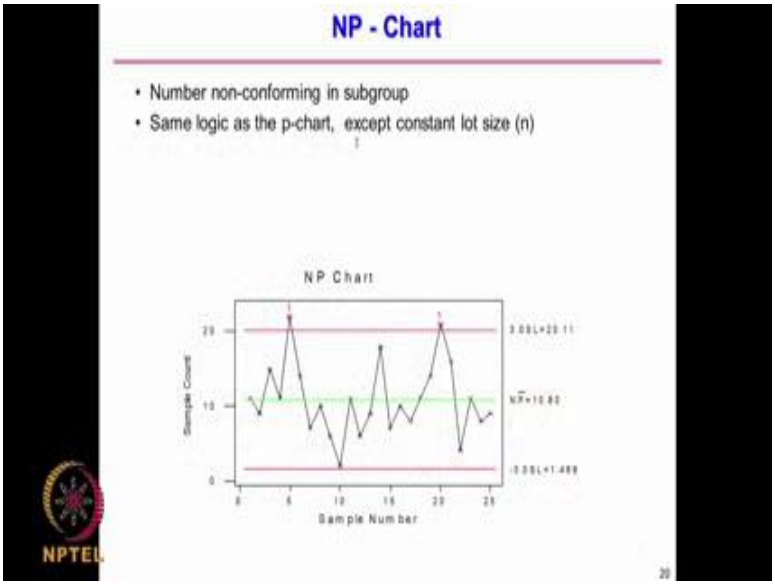
same as C chart, except this size, lot size keeps changing. So, sometimes, the process may go like this, you know, suddenly the variable, sometimes the process will get the problem, as occasional values that are clearly not a part of the basic process. So, mistake in measuring, or bottom piece, or top piece in stack, end of bar; so many different reasons; one point may be going up. So, the U chart is almost like C chart, only thing is, you do not, the number of samples you take in each lot may vary; whereas, in C chart, every time you take the (Refer Time: 21:53) same set of samples in a subgroup.

(Refer Slide Time: 22:03)



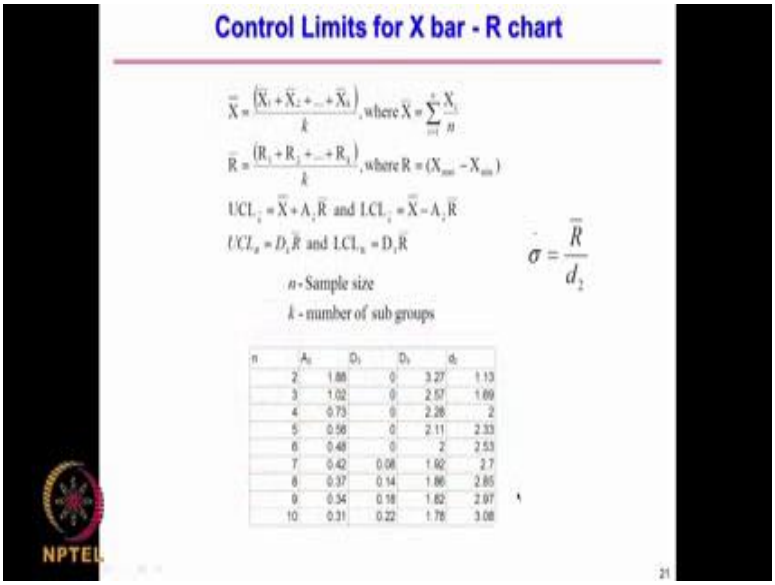
Then, you have the P chart, measures fraction defective, whereas, the other one is, C chart is actual number of defects; whereas, P chart is fraction defective, it is a proportion. Control limits are based on the binomial distribution; sample size does not have to be equal, here in the P chart also, because you are calculating fraction defective. Occasionally, some data may go up and down; here also, you can plot the 3σ, but this keeps, will keep varying.

(Refer Slide Time: 22:40)



Then, you have the NP Chart. NP Chart is nothing, but if you remember, NP Chart is constant lot size; you are looking at the defective here; number non-conforming in subgroup; same as the P chart, except here, lot size is constant. Fraction, NP Chart. So, when you do NP Chart, probability into n, so obviously, it becomes number of defective items. Do you understand? So, P chart is the fraction, whereas, when you multiply the n number of samples in the subgroup, then, it becomes a number, and that gives you sample count which are defective. So, you are plotting this, and you are seeing how many defectives over a period of time. So, this is going too much, or it is within this range, that is what you are studying in the NP Chart.

(Refer Slide Time: 23:48)



So, you have so many different types of charts which is talked about, and there are different types of formulae to calculate, and there are many soft wares which can do that also. But, this particular set of statistical process control, giving a pictorial representation of what is happening to the process, is extremely useful, because, before a real problem hits you, you can, from the data, identify whether a problem is looming behind the scene, whether you need to take any corrective action. So, the upper control limit and the lower control limit are the two extremes within which you would like to have your process. Not only that, but if there is some sort of a pattern that is happening, then, you need to better watch out; that means, if you have too many points near about to sigma, or if you have too many points above 1 sigma, you better watch out; or, corresponding - 2 and - 1 sigma. And also, if you have 14 points consecutively going up and down, if you are having 7 points consecutively going up, or 7 points consecutively going down, then obviously, you need to be very careful, and you need to assume that, there is some problem that is looming around.

So, there are many soft wares as I said, which can plot these various types of charts. As I said, the average chart, the range chart, then, you have the defective chart, defects chart with varying lot size, constant lot size, and so on. And, as I said, this is very useful in a manufacturing scenario like bio process, or chemical process, or if you are

manufacturing bio materials, bolts, nuts, anything, where you are trying to produce the same item, with the set of, with the range of product quality, that is the specification limits for the customers, and control limits for the manufacturing, you would like to always keep this range, keep the product qualities within this range.

So, that completes the course on biostatistics and design of experiments. We have been looking at lot of things in the past 40 lectures. Statistics was always a boring subject, but once the manufacturing industries, especially the chemical and the engineering manufacturing industries, car manufacturing industries came into being around 2nd World War, product quality became very very important and customer satisfaction became very important. So, one needs to look at the consistency of the product. So, statistics became very important in the manufacturing industries. Then, Japanese came into the world scenario, who introduced the concept of quality. So, products that were coming out from Japanese industries were very particular about highest quality. So, they adopted lot of these statistical techniques, design of experiment techniques, the analysis techniques, ANOVA, t test, and so on, for comparing different processes, for mentioning whether a product quality is within limits, outside limits. So, the Japanese made the statistics and statistical analysis very very important. Then, other companies started following. American companies realized that, Japanese were overtaking them, because Japanese were very particular about quality. So, the American companies like General Electric, and many other companies also adapted the concept of quality and statistical analysis.

Biology is a subject which also handles lot of data, lot of variations; so, biostatistics became very important, because, many of these statistical principles could be adopted into biostatistics for comparing different processes, bio processes and performance. Then, once the clinical trials came into existence over the past 25-30 years, the modern clinical trial approach started coming in. Then, statistical tools became very important for comparing results from different locations, different size, different types of drugs, and so on. And, without following these statistical analytical procedures and principles, one cannot conclude whether the drug is superior than the older drug, or whether the drug is as good as the older drug. So, we looked at lot of statistical techniques, tests, different types of tests we looked at. We also looked at different types of distributions, log normal distribution, beta distribution, Poisson, binomial and so on. All these distributions have

some bearing in the area of Biology, that is why this course is called Biostatistics. And, we did quite lot of problems over the period of 40 lectures, and as you can see, we had to use quite a lot of these tables, tables for z, tables for t, tables for f, tables for chi square, tables for non-parametric like rank test, sign rank test, and so on, actually. Some of these could be done using excel; some of these could be done using freely available software.

And of course, if you buy a commercial software, obviously, many of them could be done. Whereas, I did not show you any examples using commercial software, because one may have, there are so many hundreds of commercial softwares, each one having their own advantages and disadvantages, and I do not want to dwell on them. We tried to solve this problem using fundamental approach, so that, you will understand what is the underlying mathematics behind it; what is underlining philosophy behind it. So, even if you use a software on a later date, and you find that you get lot of results, at least you will be able to understand what these results meant to be. So, I hope you enjoyed this course.

I hope you benefited from the course, and every after end of every week, that is after end of every 5 videos, I also tried to have a small quiz with problems solving, so that, that will brush up the theory which I taught in the previous 5 lectures. And of course, some of these calculations could be quite intensive, but of course, if you use excel, it is not very difficult; even using a calculator, it is not very difficult. But then, it gives you a fundamental understanding of the whole concept of Biostatistics. Then, we also looked at design of experiments, large number of designs; the full factorial design, fractional factorials, 2nd order designs.

So, all these designs are very useful for you, to sort of plan how to go about varying different parameters, so that, you have minimum number of experiments, but at the same time, get maximum information out of the entire process. So, the idea is, minimum experiments, maximum knowledge out of it, rather than, maximum experiment, not so much, enough knowledge about it. And, each of these designs were well planned out. They are very symmetric, as I explained in many situations, they are extremely balanced; that is very, very important. In statistics, bias is one which you need to avoid, because that will mask the results, as well as your experimental findings, ok.

So, I hope you enjoyed these 40 lectures, and I hope you benefited from this. And, you will be making use of it quite a lot, especially, this is very useful for biologists, bio process engineers, clinicians, clinical trial scientists. And, some of these concepts, of course, can be used by other engineers and technologists as well. Good luck.

Thank you very much for your time.