

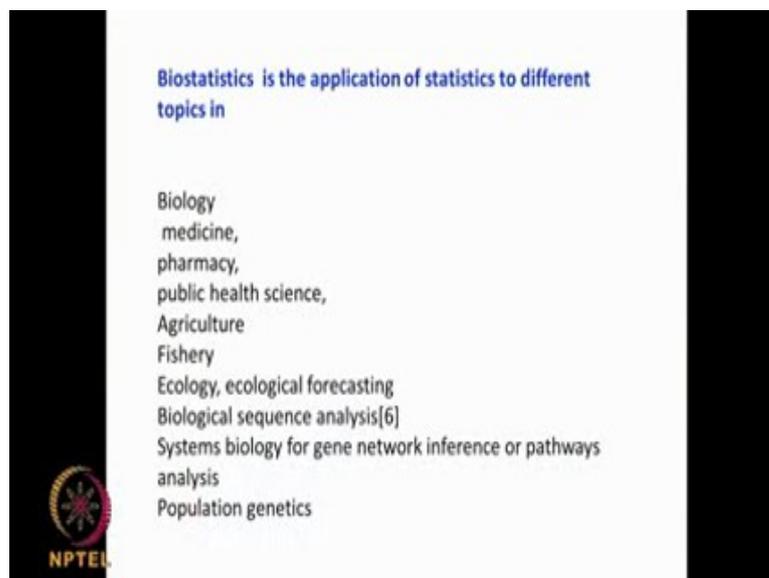
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 01
Introduction

Hello everyone, I am going to talk about Biostatistics and Design of Experiments for 20 hours, spread over 40 lectures, and spread over 8 weeks.

Biostatistics and design of experiments are interrelated with one another. So, when you carry out some experiments, you are going to analyze the data, compare the data with literature, look at the significance of the data, and tell whether the data is very important or not. So that way both are highly interrelated and that is what this course is all about actually. It is going to be a 20-hour lecture with quizzes in between, almost after every 5 lectures you will have a quiz, and then we will also have a final exam.

(Refer Slide Time: 00:59)



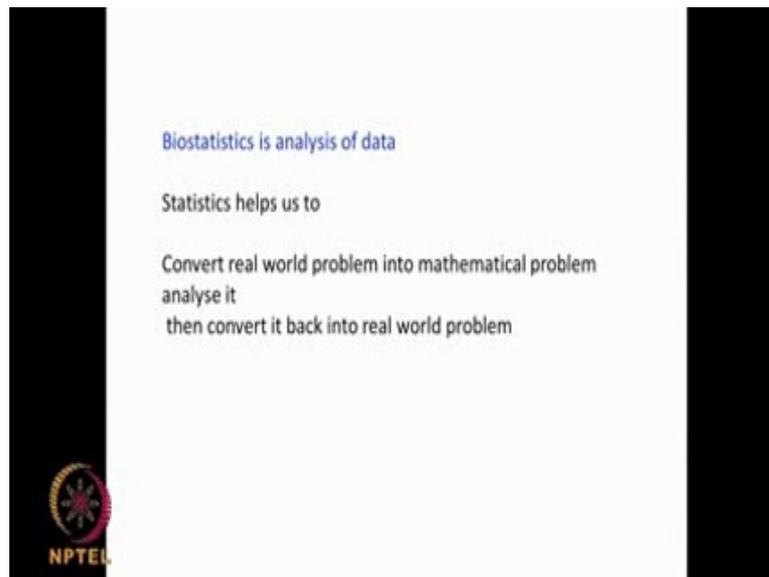
So, biostatistics is application of statistics in several topics. It relates topics in biology, medicine because now-a-days in medicine, clinical trials there is lot of data analysis, data collection, so you need to have it. In pharmacy when you are looking at new drugs, when you are comparing drugs, in public health science you are looking at disease, disease progression. Agriculture, when you are coming up with new strains, new plants, you are comparing old plants and old yields with new plants and new yields. Fishery, when you are doing a lot of

changes to a population of fish and you are trying to look at their growth pattern as well as the yield. Ecology, ecological forecasting; again changes in climates, climatic conditions, temperature, temperature of the sea and so on actually. And then sequence analysis, we were also going to look at some problems there. Systems biology, gene networking, pathway analysis, because in biology and bio informatics you are dealing with large amount of data, the sequence of genes - you are comparing two genes sequences and trying to say whether there is any statistical significant difference between them. Then, based on that you are going to analyze these pathways, and then compare the performance or the mechanism of action. Population genetics, you are looking at different types of population, how the diseases are prevalent in certain populations, then what are the habits of certain populations, what are the changes that happen to the population.

So, in all these areas where you are dealing with large amount of data, and you are making comparison between one data set and another data set, maybe one data set could be very old and the other one could be new. When you are comparing two drugs, a drug, a very old drug and a new drug which you are trying to introduce. You are comparing ecological systems - one part of their globe with another part of the globe. So, wherever there is lot of data, wherever you want a make useful scientific conclusions then the biostatistics come into play.

Statistics is very very important in all fields of engineering, and science, and in biology it is much more relevant, because you deal with large amount of data. So, the techniques are taken from the general statistics, but applied to biological system. So, one could use these techniques again vice versa into other engineering and science related systems.

(Refer Slide Time: 03:37)



So, basically it is analysis of data as I said. So, it can convert real world problem into mathematical problem, then analyze the results, and then again convert it back into real world problem.

For example, I want to say drug A is better than drug B. So, we convert that into your statistical problem, we do some analysis using statistical techniques or tools, and then again we end up saying – yes, drug A is better than drug B; or no, drug A is not as good as drug B; or drug A is as good as drug B; or drug A is less than drug B and so on. So, that is the real world problem. So, we are comparing two types of populations, and then trying to make a statistical analysis, and then say – yes, this population is different from that population or there is no statistical difference.

I am comparing two classes - classrooms of students - and then I would say their heights are statistically different; or no, their heights are not statistically different. Or I may be comparing a certain chemical in the neuronal cells of ancient population and same chemical in an African or a Chinese population, and then, yes, I would say there could be a statistical significance difference or no there is no statistically significant difference.

So, especially when you are handling lot of data, and you want to make a very reasonable conclusion, statistics come into play, and that is where the biostatistics is all about. So, basically it deals quite a lot with real world problem. Whatever examples I am going to talk

about is very very real, and they are very relevant in the area of biology, and finally, we need to give the answers again back into the real world problem.

(Refer Slide Time: 05:28)



So, it involves analysis of data from experiments. So, you are doing lot of experiments, you may be doing clinical trials with drugs towards inflammation, maybe doing clinical trials with drugs towards treatment of some cancer or you may be doing a fermentation where you are changing ph or changing temperature and looking at yield of some metabolites. So, you are doing lot of experiments. So, analyzing that data, statistics come in to picture.

Interpretation of the data, you have so much data, I need to interpret. I need to say whether pH has an effect on the yield or I would say ph has a direct relationship, pH has an inverse correlation, temperature has no relationship. So, interpreting the data and drawing conclusion from the data; again statistics come into play.

Distribution of data, can I group them into different categories? High active, low active, medium active, no active. So, I can look at performance of drugs and I can put them into various categories. I can look at some bacterial systems, large number of bacterial systems, and I can look at the metabolite profiles, and then, I can say high yielding, low yielding, and medium yielding.

Sample size, I want to decide how much of sample I need to test to have a meaningful or a useful conclusion. Is it enough if I take 10 samples? Suppose I am performing a clinical trial

on a drug, is it enough if I test on 10 patients? Or should I test on 100 patients? Should I test on 1000 patients? If I go up and up, of course, it is more accurate, but then the amount of the work that needs to be done, amount of resources that is needed is going to be very high. So, ideally you would like to keep the sample size less, but at the same time get lot of useful information. So, this is a very important contribution of statistics.

Significance is the results are significant or the results are not significant. I see a change when I give a drug to a patient. Is it much more than a placebo effect? That is what it is significance.

Data reduction, I am having lot of data, I want to reduce the data. Because handling lot of data requires very high end computing. Can I reduce the data so that they are manageable? Can I group the data into categories, so that then it becomes much more easy for me to analyze the data. So data reduction, there are many techniques available for data reduction. Things like principal component analysis, partial least square method. So, all these are data reduction terms.

Regression analysis, I want to draw a relationship between temperature and yield; as the temperature increases, yield also increases. So, is there a linear relation? Is there a non-linear relation? So, I can use statistics to develop a relationship between temperature and yield.

Comparison of performance of drugs. As I said, you know, I am testing drugs in the clinical trials. I am testing drug a new drug which I have discovered and I am comparing with the existing drug. I may compare with placebo; I may compare it with some other standard method. So, comparing different drugs, placebos, standards, already existing in the market and so on. Very, very useful; biostatistics is very, very useful there. Comparison of two crops, based on their yield. I am testing a fertilizer and currently some other fertilizer is there, so I want to say my fertilizer is better. So obviously, I am going to compare two fertilizers. I am coming up with a new bio fertilizer, so I will compare it with the existing chemical fertilizer. I am coming up with a new genetically modified plant; I want to compare the yield of this plant with the existing plant; so, again there are comparison.

Rainfall data, I am looking at the rainfall in say Chennai, in this particular 2015, and I am comparing it with 2014 data, 13, data, 12 data and so on for the past 50 to 100 years, and tell the rainfall in average is higher, rainfall average is lower or there is no difference in the rainfall. I am comparing say (Refer Time: 09:59) the rainfall of Chennai with other coastal

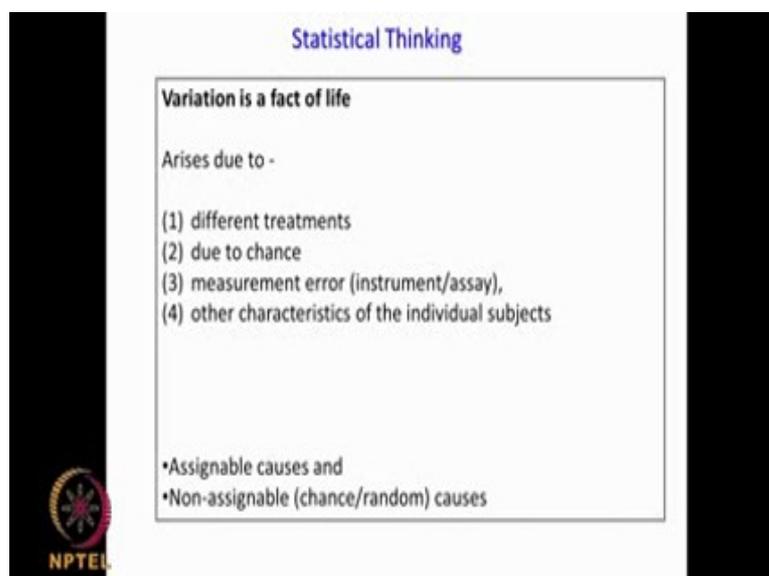
regions like Bombay or Visakhapatnam and Calcutta and so on. Then, I may make a comparison on whether the rainfall in Chennai is lower than rest of these places.

Accidents between cities, as you know road accidents are very common and quite prevalent in India. So, I am comparing road accidents in various metros like Chennai or Bangalore or Mumbai, Delhi, Calcutta, and say, yes, this particular city has much higher road accident per month when compared to the other city. So, can I find out what are the good practices or best practices they follow? So, there I need to use statistical techniques. You have something called Poisson Distribution which can help you to find out and compare these type of accidents.

Performance of class, you have students in your class, 100 students, and they have done some exam, and you know the average marks, and the spread or the standard deviation, you want to know whether these students performance is as good as the students from the previous batch or the batch before or you want to compare it with rest of the university standard and so on. Again, we can use this type of statistical tools test, there are something called T-test, different types of T-test are there - one sample T-test, two T- test and so on actually.

Gene sequences, I am comparing a sequence of a particular gene from a mammal with the human and trying to see what are the differences, so again statistics is very useful. So, as you can see, a large number of real world problems use statistical tools to do these types of analysis and that is what we are going to learn in the next 40 lectures.

(Refer Slide Time: 11:58)



Statistical Thinking

Variation is a fact of life

Arises due to -

- (1) different treatments
- (2) due to chance
- (3) measurement error (instrument/assay),
- (4) other characteristics of the individual subjects

• Assignable causes and
• Non-assignable (chance/random) causes

 NPTEL

As you know variation is part of life. You cannot have a situation where there is no variation. If, I have to reach my office every day at 10'O clock, I will reach some days at 10 'O clock, some days at 10:01, some days at 10:02, some days at 9:59, some days at 9:58. So, these variations are going to be there, it is part of life. Because, it depends on how the bus or car comes, what is the traffic situation, and was there any rains or snow during my travel and so on. Small, small, variations are going to there. These are called Statistical Variations, these are called small errors, these are called Noise Variables. In statistics these are called Noise Variables.

But then one day I came at 10:15, which is not really a small statistical variation, it is a large variation. Then we can call it an assignable reason. That day I came late because I missed the usual bus or that day I came late because there was a heavy traffic jam because of VIP movement. So, that is called Assignable Cause. But that small variations instead of coming at 10:00 if I come at 10:01, 10:02, 9:58, 9:59 and so on, that is a random, that is a chance and that is called Non-assignable. And as I said, the random variations are always going to be there in any situation. Whereas, the assignable causes which are very large changes, like me coming one day at 10:15, or me coming at one day 9:45, you can always assign it for something else. I am coming, one day I came at 9:45 probably because I caught the previous bus, I got up early or there was absolutely no traffic or it was a holiday for schools and government offices, so there was no traffic. So, you can assign a reason for that. So, that is called an Assignable Cause.

Any time you can assign a reason that is called an Assignable Cause, whereas small variations are always going to be there; you cannot prevent that; that is a random or a chance. I take a pH meter and measure the pH of a solution, I dip it inside, I get a pH of 3.00. Then I take it out, wash it again, dip it inside, I may get a pH of 3.05. I take it out, wash it, and put it, it could be 2.98. So, small variations are happening, that is a random variation, that is always going to there.

So, these variations arise because of different treatments; now I use a different treatment strategy. Due to chance, this is called the Random Measurement error, I may be having an instrument or an assay method which may always lead to error. I have to look at the colour change, sometimes I might not be able to exactly judge the colour changing from pink to white. Colorless, you must have all done experiments, acid base titration and we will see changes like colorless to light pink or slight pink to colorless. So, there could be always be

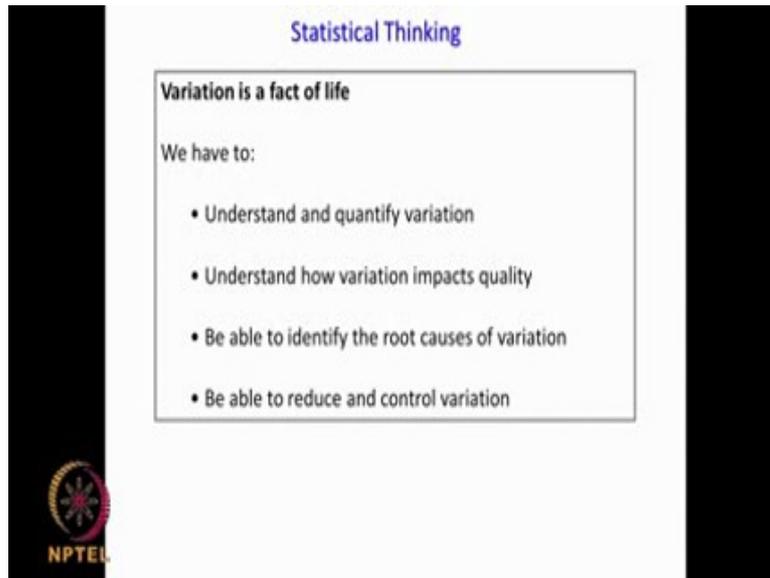
variation the way you see it. So, there will always be one drop of the acid going plus or minus that is the assay. Sometimes the instruments will always give small changes because of the environment, because of the voltage fluctuations, and so on.

And then, there could be variations arising because of the individual; when I do the experiment and when you do the experiment there is always going to be there -differences. That is called the Individual Errors. The way we see at the instruments, the way we take samples or the way we weigh it, and so on, those variations. So, variations are always going to be there, and many of them are because of chance, and if I want to say a drug A is better than the drug B, that differences I see between the performance of drug A and B should be much larger than this statistical random variation. So, if the drug A reduces the tumor size by 10 % and drug B reduces the tumor size by 9 %, so the 10 and 9, that 1 % difference should be much, much, much larger than the statistical random and white noise type of variation. So, if the random variations with the instrumental techniques is going to be 1 % or 2 %, then we will not be able to say whether 10% because of drug A and 9 % because of drug B is really a remarkable performance of drug A as against B or it is because of chance. So, most of the statistical tests are based on that.

So, there (Refer Time: 17:01) changes or the differences you observe between a performance of one group or one sample as against another group or another sample should be much, much larger than this type of random variation or white noise. That is very very important. That means, you should have good, accurate ways of measuring your results. You should have good well trained operators. You should have good instrumental methods, techniques, instruments which give very low error, very low error bars, which gives very low white noise. Then only I will be able to see real differences between drug A versus B, method A versus method B or student A versus student B.

So, we need to always consider variation is part of life, we cannot prevent a variation, and you need to understand what is variation because of assignable causes, what is because of non-assignable causes. Assignable causes, I should be able to address and prevent it. If I miss the bus and come to the office at 10:15, then you can easily prevent that. In future, you will always make sure that you will not miss the bus. So, that is an assignable. But small variations are always going to be part of life, and whatever studies you do, whatever comparisons you do, the differences between situation A versus situation B must be much larger than this statistical small variations or error.

(Refer Slide Time: 18:37)



Statistical Thinking

Variation is a fact of life

We have to:

- Understand and quantify variation
- Understand how variation impacts quality
- Be able to identify the root causes of variation
- Be able to reduce and control variation

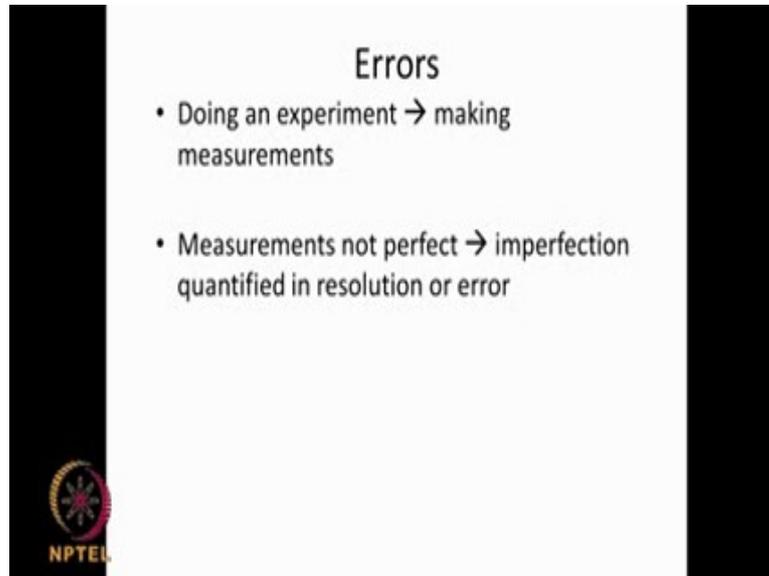


So, we need to understand that variation is fact of life. We have to understand and quantify these variation, and we have to understand how these variations impact quality. For example, I have a fermenter. I am doing reaction at 30° and the temperature will fluctuate between 29 and 31, because I have a very poor heating system. Now how much these changes 29 to 31 is really affecting the yield of my product, I need to know. Is it effecting by 10 %? Or is it effecting in a small way? So, you will know the extent of the effect of these variations on my final product quality.

Be able to identify the root cause of variation. What causes this variation? I come sometimes very late, sometimes at very early. What is the reason for me coming very late or coming very early? Is it the bus, is it the route I take which has lot of traffic, and so on actually. We should be able to reduce and control the variation. So, we should always try to have a better instrument which will give us less variation, our assay method should be very robust so that it does not give variation. The operator should be very well trained so they do not make mistakes quite often. Their hands are so set, they are very good that they do not make any mistakes. These are important points one need to plan, if one is designing experiments so that they have minimum variation. You should have well-trained operators, you should have good instrumental tools which give you less variation, you should identify what are the places where you could have errors or variation, address them, how much variation will impact the quality of my final product, and so on.

So, that is part of the design of experiments that is we need to keep all these points in mind. So, what are the variables that lead to variation, and how do I address these, and which ones cause big problems and which one cause small problems.

(Refer Slide Time: 20:47)



So, errors can come because of so many reasons. When we are doing errors, when you are doing measurements, we are measuring temperature, we are measuring pH we are taking out samples, we are measuring heat produced, we are measuring what is the amount of liquid flowing in, what is the amount of a liquid coming out, so all these are measurements, and there is always going to be some errors coming into it. Instruments will have errors, people will have errors. So, we need to know, and measurements are not very perfect. As you know instruments are not very perfect, human is not perfect, assays are not perfect, human training is not complete. So, this leads to error. So, these are the areas where errors are prone - the experimental techniques, the human operators, the instruments which we use. So, we need to focus on them to reduce errors. If you can reduce error then your statistical analysis becomes much more robust actually.

The second part of my talk I said, is called Design of Experiments.

(Refer Slide Time: 21:54)



Design of experiments is planning experimental strategy,

- screening a large number of parameters and selecting the important ones,
- determining the minimum number of experiments and deciding on the mode and manner in which experiment have to be conducted.
- screening design and second order designs
- Modelling/regression analysis
- Data reduction



This is planning experimental strategy. One strategy I told you before all these time, instruments which will have less error, assay methods which will have less error, operators who are trained very well. In addition, you need to consider other factors also and that is what we are going to talk in design of experiments.

We are screening a large number of parameters. For example, if you take fermentation, aerobic fermentation pH may affect yield, temperature may affect yield, inoculum size may affect yield, the volume, aeration rate, agitator, rpm, amount of the carbon you are adding, amount of nitrogen you are adding, amount of micro nutrients you are adding; so, all these have an effect on your final product actually. So, am I going to look at all these parameters? That may take a long time. Or am I going to look at very important parameters only, not all the parameters. So how do I decide on which parameters are important, which parameters are not important.

Selecting the important ones that becomes are very, very important. How do I do minimum number of experiments but get maximum information? Everybody would like to do minimum experiments because of the resource involved, time, money, waste you are generating, man power requirements. So, you would like to do minimum, but then you would like to get as much information as possible. So, minimum experiments maximum information. Design of Experiments tells you. We are going to look at screening designs, fractional factorial designs, and so on. How to reduce the number of experiments? How to reduce the number of variable?

So, finally, you go into the best set of variables and then you do a detailed experiment. So, Design of Experiments will tell you.

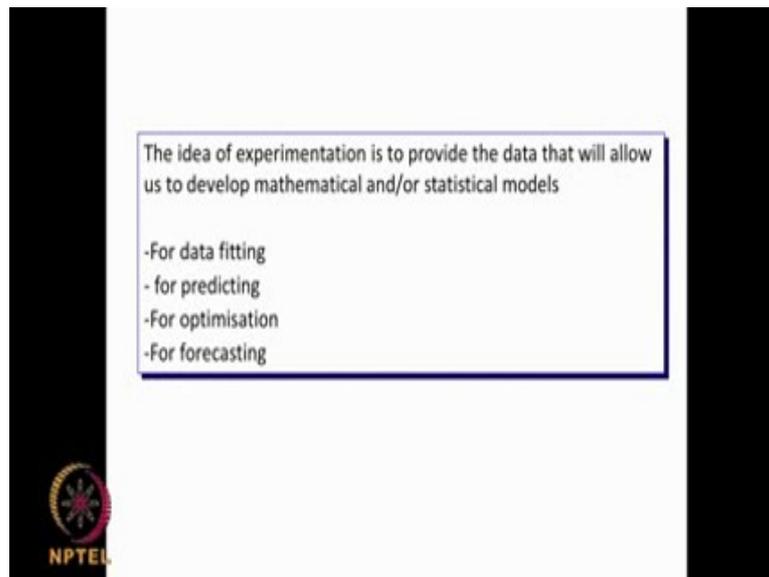
So, we have something called Screening Designs. You may take large number of parameters, we may take even 25, 30 parameters. Like I said temperature, pH, inoculum size, carbon amount, nitrogen amount, agitator, rpm, amount of oxygen that is bubbled in to the vessel, the time, amount of micro nutrients you are adding, type of micro nutrients you are adding, and so on. You can have 20 variables, but then you cannot do detailed experiments, but you do minimum experiments that is called Screening Design. And from this you eliminate some of those variables which you think is of no use. Then take may be 3 or 4 which are very important, and then you do a detailed design, where you do second order designs. That is very useful. You take large number of parameters, do some experiments, screening, eliminate variables which you think are not useful. Take a (Refer Time: 24:44) small number of parameters, and then do the experiment, and then do as many experiments; get a complete understanding; that is called Second Order Design.

Once you do the experiment, you can do a mathematical modeling, you can do a regression analysis, you can do different types of curve fitting, surface fitting. So, you can get a mathematical relation between temperature and yield, or pH and yield, the amount of carbon you are adding and yield, amount of oxygen that is bubbled in and yield, and so on actually.

Or you can do a something called Data Reduction. You will get, generate so much data that you would like to reduce the data which is manageable. Because sometimes if you look at some, you are doing genome analysis, you will get thousands of data. When you are doing high throughput screening, for example you are testing drug against a number of genes you will get 10,000 to 100,000 set of data. But then it becomes very difficult to analyze the data and there is something called Data Reduction. There is some technique called Principal Component Analysis. You can have methods like grouping the data into different types of characteristic groups and so on.

And then you manage the smaller number of data and do different types of analysis, that is called Data Reduction. So, all these I will be covering in the Design of Experiments and that is what we will do in the second part of our course.

(Refer Slide Time: 26:10)



The idea of experimentation is to provide the data that will allow us to develop mathematical and/or statistical models

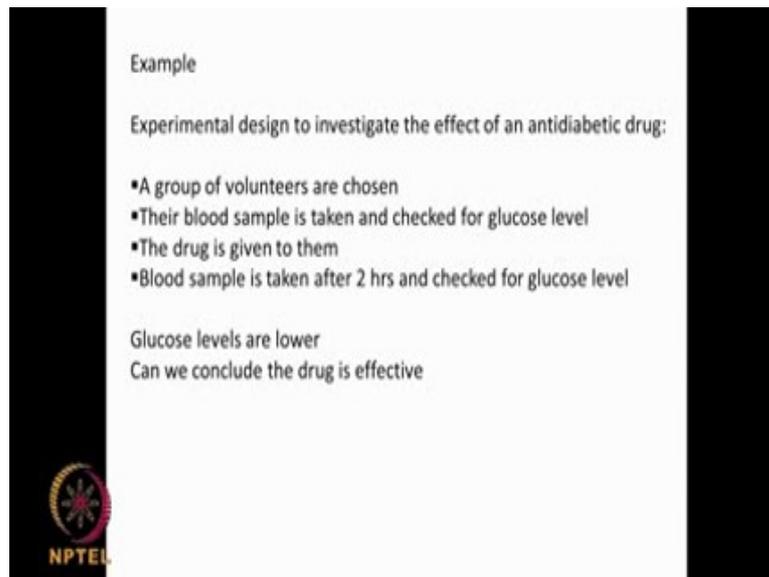
- For data fitting
- for predicting
- For optimisation
- For forecasting

NPTEL

So the idea of experiment, why do you need to do experiment? As I said, you want to develop mathematical model or statistical models it is useful for data fitting. We can use it for predicting. Suppose, I do experiments with different temperatures and pH, I can try to predict what happens if I am doing experiment at a higher temperature? What happens if I am using more carbon amount? So, I can use it for predicting because now-a-days prediction is very, very important; you cannot do all the experiments that are possible in life, but later on you may like to predict. If I change the raw materials what will happen. If I change some new conditions what will happen.

Optimization, so I want to modify the parameters so that I get maximum yield. I want to modify the parameters where I get very low impurity profile. I want to modify the parameters so that the growth rate is higher. I want to modify the parameter so that the colony that are formed is higher. I can use it for forecasting. What will happen? Especially in share market, mathematical modeling, statistical modeling is used in large number for forecasting. How will the share price of certain company change over the period of next one year, two years? So, this is very, very useful for forecasting also.

(Refer Slide Time: 27:35)

The slide contains the following text: Example
Experimental design to investigate the effect of an antidiabetic drug:
▪A group of volunteers are chosen
▪Their blood sample is taken and checked for glucose level
▪The drug is given to them
▪Blood sample is taken after 2 hrs and checked for glucose level
Glucose levels are lower
Can we conclude the drug is effective
NPTEL logo is visible in the bottom left corner of the slide.

Example

Experimental design to investigate the effect of an antidiabetic drug:

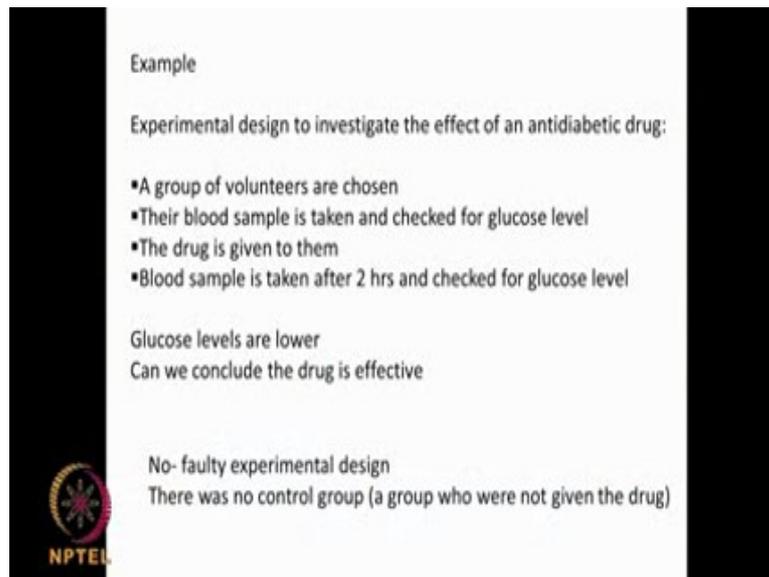
- A group of volunteers are chosen
- Their blood sample is taken and checked for glucose level
- The drug is given to them
- Blood sample is taken after 2 hrs and checked for glucose level

Glucose levels are lower
Can we conclude the drug is effective



For example, I want to look at design, investigate the effect of an anti-diabetic drug. Now we take a group of volunteers. Then, the blood samples are taken, and then you check their glucose level. After sometime, one hour, you give the drug to them; again blood samples are taken, after two hours, and then check their glucose level. Now the glucose level is lower, can we conclude the drug is very effective? So what you do is, you take some volunteers, you take that blood, check their glucose level, and then you give the drug to them. After two hours, you again check there glucose level. Now their glucose level in the blood has gone down. Can we conclude that the drug is very effective? No, the experimental design is at mistake.

(Refer Slide Time: 28:35)



Example

Experimental design to investigate the effect of an antidiabetic drug:

- A group of volunteers are chosen
- Their blood sample is taken and checked for glucose level
- The drug is given to them
- Blood sample is taken after 2 hrs and checked for glucose level

Glucose levels are lower
Can we conclude the drug is effective

No- faulty experimental design
There was no control group (a group who were not given the drug)



Because there was no control group. You never had a group where no drug was given to the patients, and simultaneously you check their glucose level, because that is very, very important. Should always have a control group when you are performing experiments with the new type of drug. So, you need a control group where you do not give any drug, but you check their glucose level at the same time simultaneously when you are checking the glucose level of patients to whom drugs are given, and then, simultaneously you see whether their blood level glucose is going down or not. Because glucose levels may go up and down depending upon the time, depending upon the food they have taken, depending upon the environment, depending upon the exercise pattern, depending upon their obesity level, depending upon the age, gender, and so on.

So, you have to be very careful. You should always have a control group, which exactly replicates that group which is called test, but you don't give the drug to the control group. That is very, very important when you plan an experiment. So, we will continue more of it as we go along in the forthcoming classes.

Thank you very much.