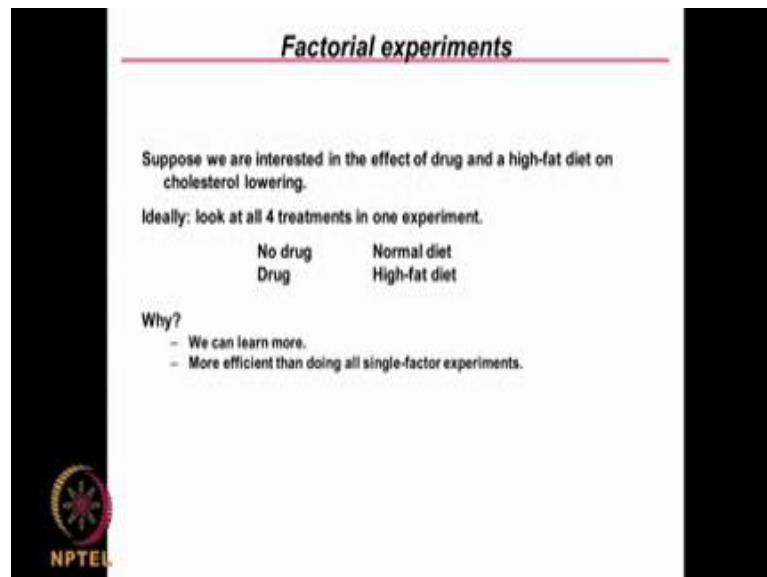


Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 33
Design of Experiments (DOE) - Factorial design

Welcome to the course on Biostatistics and Design of Experiments. We will continue on this topic of design of experiments, we are going to talk about factorial designs today.

(Refer Slide Time: 00:23)



Factorial experiments

Suppose we are interested in the effect of drug and a high-fat diet on cholesterol lowering.

Ideally, look at all 4 treatments in one experiment.

No drug	Normal diet
Drug	High-fat diet

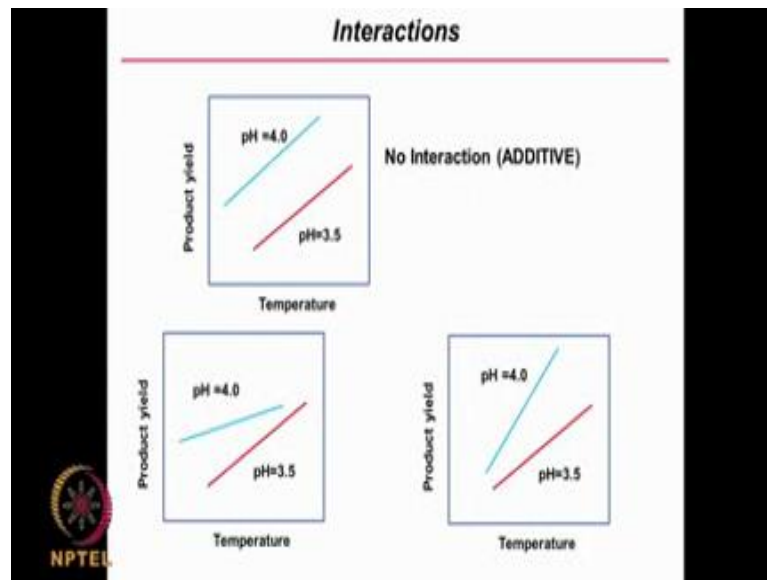
Why?

- We can learn more.
- More efficient than doing all single-factor experiments.

NPTEL

I mentioned about 2 parameters one is the drug, the other one is the diet each at 2 levels drug has a level of no drug, drug and same thing with diet also has normal diet, high fat diet. We could conduct 4 experiments, the first experiment could be with no drug and normal diet then second experiment could be no drug and high fat diet, third experiment could be drug and normal diet, the fourth experiment could be the drug and high fat diet. We have 2 parameters 2 levels **2 into 2**, 4 experiments. Do you understand? By doing this we are achieving many things because we will be able to even see things like interactions. Is there any interaction between drug and diet, if the person who is taking drug and high fat diet behaves very differently when compared to the normal diet or no drug and so on actually? We also talked about what is interaction?

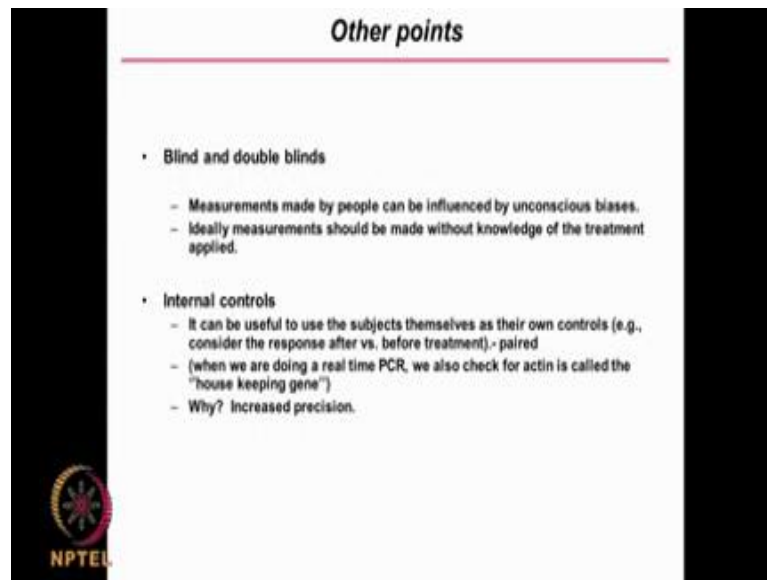
(Refer Slide Time: 01:27)



Before let me recollect, interaction is, if 2 variables they behave in an additive fashion for example if I am looking at the product yield as a function of temperature in pH at a particular value of pH 3.5 as I change temperature or increase temperature product may go up so at another pH, again the product yield may go up. These 2 lines look almost parallel, so the effect of temperature and pH on the yield is additive.

Whereas you can have situations like this you know, so at 3.5 pH product yield is going up but at pH is equal to 4 it is going up with the less slope like this or at **pH = 4** it may be going up with higher slope unlike a parallel situation, then we can say the pH and temperature are interacting with the each other in arriving at the product yield. The effect of temperature is not same at all pH values but it depends on the pH values, for example some drugs perform very differently with Europeans as against Africans or Asians then you can say the drug and the race are interacting. If you want to study interactions we need to definitely perform factorial type of experiments that is very, very important.

(Refer Slide Time: 03:04)



The slide is titled "Other points" and is framed by a black border. It contains two main bullet points. The first is "Blind and double blinds" with two sub-points: "Measurements made by people can be influenced by unconscious biases." and "Ideally measurements should be made without knowledge of the treatment applied." The second is "Internal controls" with three sub-points: "It can be useful to use the subjects themselves as their own controls (e.g., consider the response after vs. before treatment) - paired", "(when we are doing a real time PCR, we also check for actin is called the 'house keeping gene')", and "Why? Increased precision." In the bottom left corner, there is a circular logo with a colorful wheel and the text "NPTEL" below it.

Other points

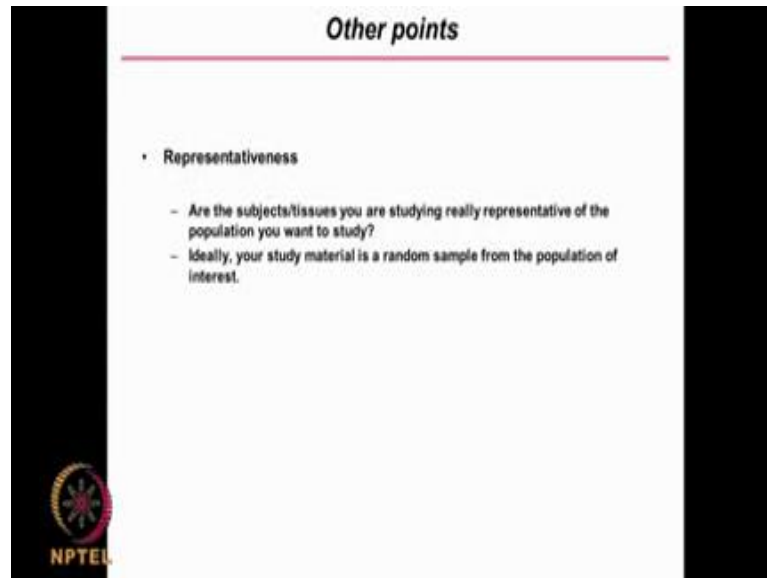
- Blind and double blinds
 - Measurements made by people can be influenced by unconscious biases.
 - Ideally measurements should be made without knowledge of the treatment applied.
- Internal controls
 - It can be useful to use the subjects themselves as their own controls (e.g., consider the response after vs. before treatment) - paired
 - (when we are doing a real time PCR, we also check for actin is called the "house keeping gene")
 - Why? Increased precision.

Then another important point which I had mentioned before is performing the blind and double blind test because generally people will get influenced in a very unconscious manner they will get biased, the medical practitioner that is physician as well as the volunteers or patients. If a patient is told he or she is being given a drug for treating their disease or there could be some biases in the way they react, so generally they are not told whether they are being given a placebo or a drug that is a blind but then even the medical practitioner or the clinician who is giving the drug may be biased when he or she is giving it to the patient. The doctor is also not told whether he is giving a drug or a placebo to the patient who also does not know whether he or she is receiving a drug or a placebo that is called a double blind that is very, very important to do actually.

Internal controls. This is very important to have, sometimes we use the subject themselves as their own controls, right, **pair t test** for example, subject themselves are used as a control. Sometimes, especially when you are doing real time PCR you all know in molecular biology we do quite a lot, we look at the house keeping gene like actin and then compare the performance, the up regulation and down regulation of rest of the gene with respect to the house keeping gene that is an internal control. It is very, very important that you all have internal control in whatever you do because that sort of normalizes the variations we see. Suppose I am doing analysis with chromatography, I am looking at a particular peak it may be changing either it is going up or it is going down but then if I have another peak which does not change or which also changes then I

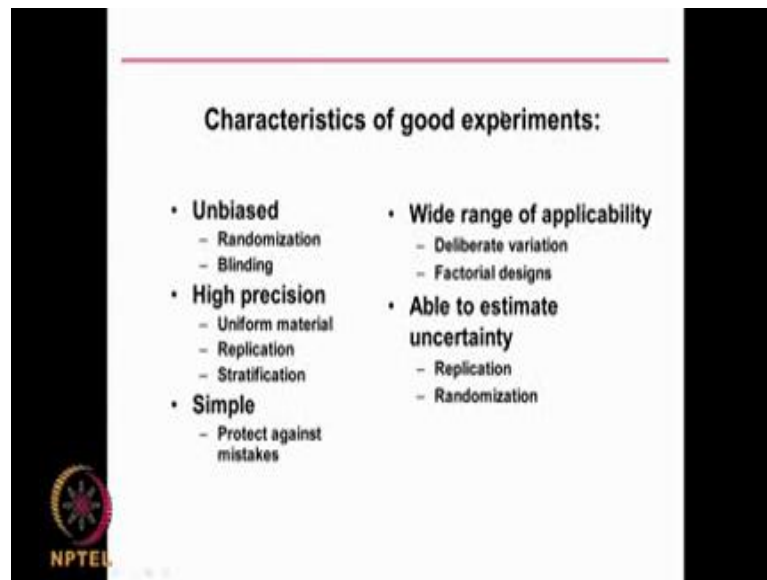
can take the measurements of the unknown with respect to the control that sort of ratio is always better than dealing with absolute values, that way internal controls are very, very important. In many situations, you always do that especially biologist know very well where we use internal controls molecular biology some times in biochemistry, product purification, chromatography and so on actually.

(Refer Slide Time: 05:43)



Representativeness is very important especially when you are doing the clinical trial. The subjects you are taking as volunteers, they should be representing the entire population rather than one strata of society. The tissue samples which you are taking that should be representative of the population. You should generally study the material in a random population of interest rather than being only specific to one group of people that is very, very dangerous.

(Refer Slide Time: 06:15)



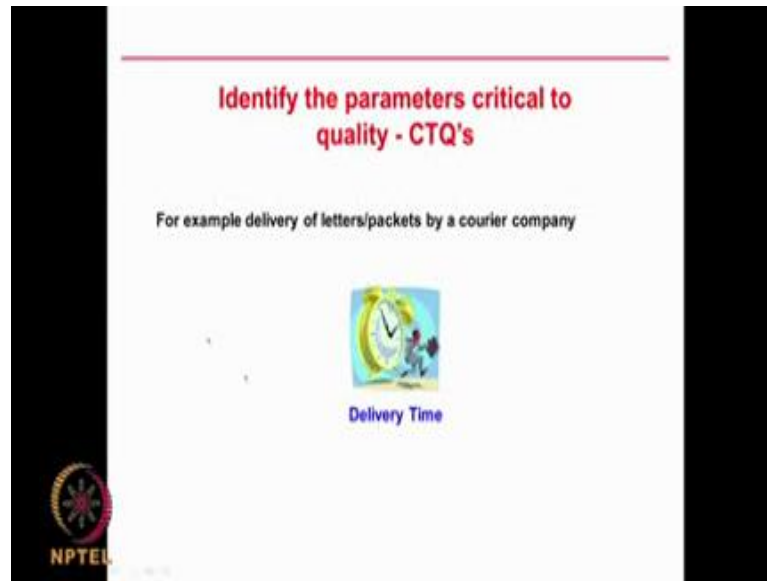
Characteristics of good experiments:

- **Unbiased**
 - Randomization
 - Blinding
- **High precision**
 - Uniform material
 - Replication
 - Stratification
- **Simple**
 - Protect against mistakes
- **Wide range of applicability**
 - Deliberate variation
 - Factorial designs
- **Able to estimate uncertainty**
 - Replication
 - Randomization

NPTEL

A characteristics of good experiments should be unbiased that is the randomized, blind, high precision, uniform materials, same raw material, replication, blocking, simple, so that it will protect against mistakes, wide range of applicability, so you change factorial designs we are going to spend lot of time on this. We should be able to estimate uncertainty that is, error we should be able to get the confidence limits. How do you do that? We replicate, we calculate standard, error we have been spending lot of time on all these actually, that is a measure of uncertainty that is very, very important. Randomization is also very important if you want to bring in uncertainty rather than being very specific.

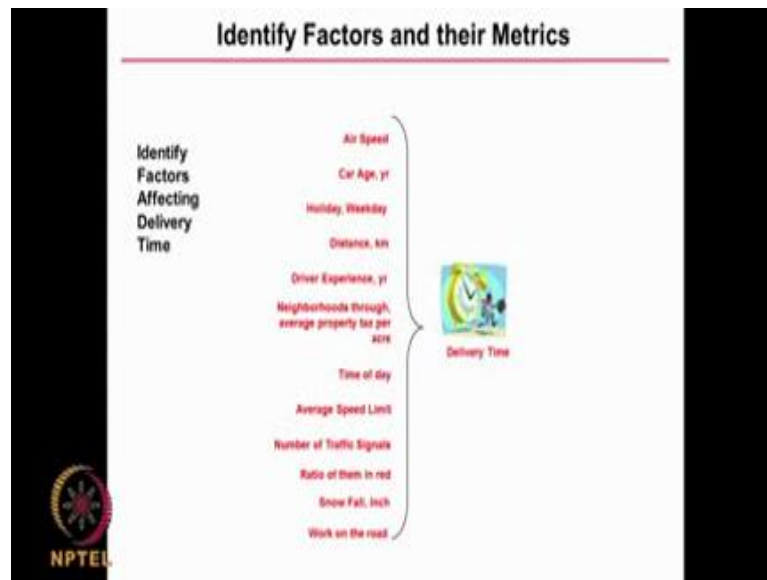
(Refer Slide Time: 07:05)



Let us look at some new terms this is called the critical to quality. We need to identify which parameters are very important which are not, for example in your product, in a bio process product development you can think of 100 of parameters carbon concentration, nitrogen concentration, various micro nutrition concentration, the pH temperature, the agitator r p m, the amount of oxygen you are bubbling, so many parameters but only few of them will be very important that is called the critical to quality. The whole goal is to identify only those few of them and focus and try to control them that is very important that is called critical to quality. So, you need to find what are the parameters which are critical to quality?

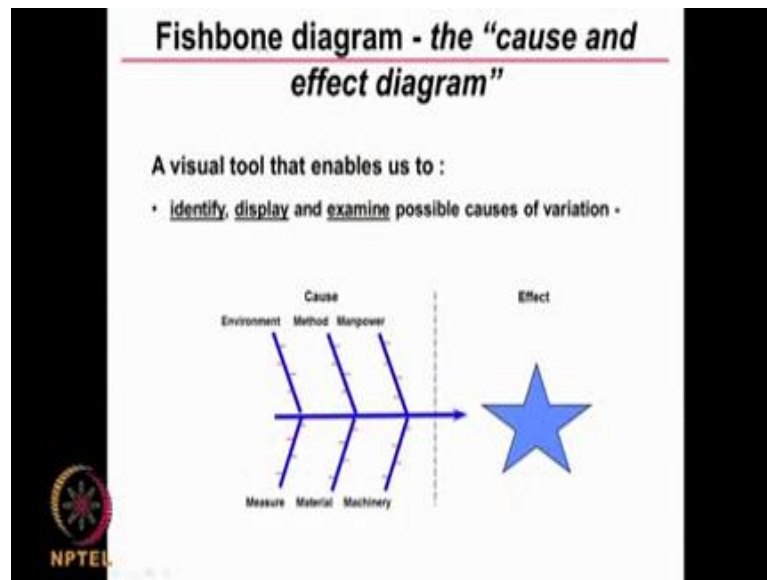
When you do the design of experiments initially you may take many variables and then you identify only few variables that are critical to quality, that type of experiments is called screening design. When you do a screening design you take in many parameters and then you come down to only few 3, 4 which are critical to quality and then you spend more time doing a very detailed design with that. Generally, we will do a screening design followed by a detailed design. For example, let us take the letter packets delivered by a couriers company and so here delivery time is the most important parameter which one looks at actually because the courier even if you think about the pizza delivery, time taken to deliver is the most important that is what everybody looks at, every customer looks at. So, what are the parameters that affect?

(Refer Slide Time: 08:51)



What are the factors that effect may be if he is driving a two wheeler air speed, if you are driving a car the age of the car, how well it is maintained? Is it a working day or holiday? How far you have to deliver? What is the driver's experience? How good the neighborhood is? Is it well laid out roads? What time morning or afternoon or evening or night average speed limit allowed? Some roads allow **70 kilometers / hour**, some roads allow only 25, 30 number of traffic signals and then how many of them red at that point? Here we talk about snow or rain or any other factor. Is there any road work going on the road? All these are going to affect your delivery time, right. Many of them may be very important parameters some of them are not very important but you have large number of parameters which are to be at least initially considered if your going to do a proper study.

(Refer Slide Time: 09:56)



That is where we have something called a **Fishbone diagram**, this is called a cause and effect diagram. It is visual tool that enables us to identify, display, examine possible causes of variation. So effect, this could be your product yield, this could be the time taken to deliver a pizza, this could be amount of bio mass produced but what are the factors that effect.

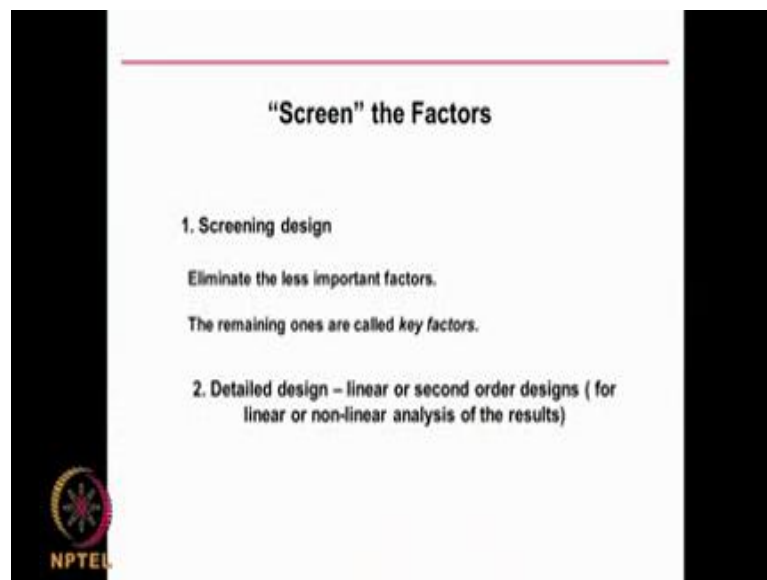
Here many things you know the type of raw materials you use, type of machinery that means fermenter or bio reactor you use and then type of people who are working on how skilled or unskilled, what are the various techniques the fermentation techniques you are using? What are the environment conditions? What are the measurement tools we use? So on lot of parameters will come into and all of them will have some say in it some of them will have more say some of them will have less say in the product yield this is called the Fishbone diagram. And it is very useful because initially when your planning an experiments you can put down all the parameters that is called the brain storming you put down as many parameters complete whatever you think and then we will be able to eliminate some of them thinking that either they do not contribute much or their contribution is minimal thereby we can just focus on only a few of them.

This Fishbone is a good idea to prepare especially when you are planning a DOE, so before we do a DOE you can use this. But of course, in a bio process fermentation we all know what are the parameters that is going to affect but if you take a much more

complex process like manufacture of a biosensor there could be raw material related, there could be manufacturing related, issues there could be type of instrument used and so on or if you take a clinical trials study and you want to look at some effect that is very complicated. If you are talking about clinical trials you have our doctors, medical practitioner is coming into picture, you have nurses coming into picture, you have the volunteers coming into picture, the type of drug, the type of acids, the different type of analytical tools used all these come into picture. It is extremely complex we can put down as a pictorial form, a cause and effect diagram is called a Fishbone diagram because this looks like a fish bone and some of the parameters may not be in your control at all some of them may be in your control.

For example, if you take your pizza delivery the road work you have no control, if there is a heavy rain or heavy snow you have no control, the number of red lights coming on the way you have no control but you have control like the driver's experience, the cars, quality all these you have control, right. There are some factors which you do not have control and some you have control.

(Refer Slide Time: 13:06)

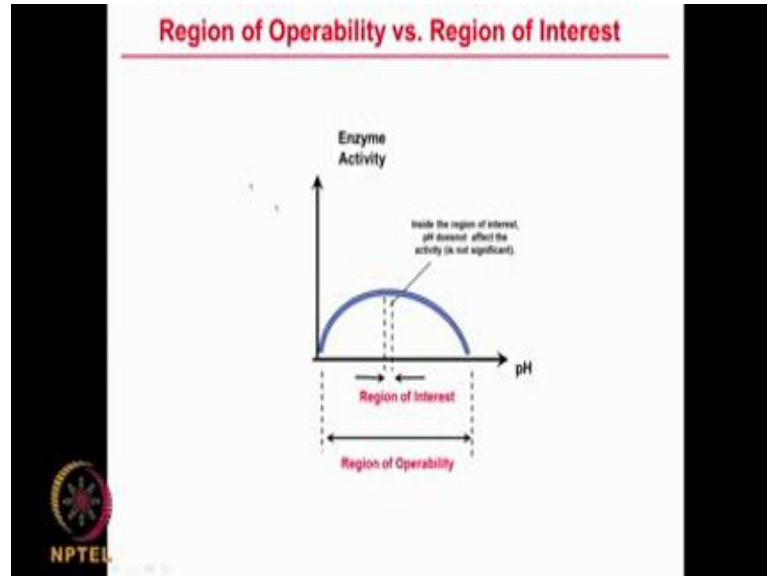


The slide is titled "Screen" the Factors. It is divided into two main sections. The first section is "1. Screening design" which includes the instruction "Eliminate the less important factors." and a note that "The remaining ones are called key factors." The second section is "2. Detailed design – linear or second order designs (for linear or non-linear analysis of the results)". The slide features a red horizontal line above the title and an NPTEL logo in the bottom left corner.

Initially you do something called as **Screening design**, you eliminate less important factors and the remaining factors are called key factors. Then you do a detailed design you have linear design, second order designs so that you can fit linear or non-linear type

of a regression mathematical equation. So you have factors which are key factors which I shortlisted from the entire set of factors using a screening design.

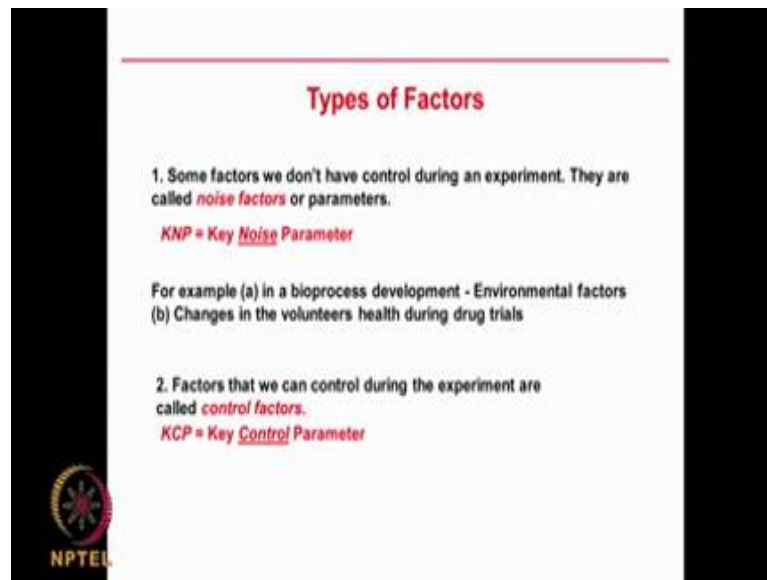
(Refer Slide Time: 13:43)



Another important parameter you need to consider is Operability versus Region of interest. For example, if you take pH versus an enzyme activity, generally you all know it is a bell shaped curve or it is a curve going up and coming down and so on but at different pH's the activity may vary. But if you are operating at only a very short range of pH like 3.5 to 4 this looks almost linear and we can say the enzyme activity does not change with pH but in reality it changes which is the Region of Operability. You can operate at any time but if you are interested only in a very narrow range then obviously the enzyme activity does not change in that particular narrow range.

It depends on which regions you want to study, so temperature will affect your reaction rate, conversion process but then if you think I can work only between only 35 and 40 ° then that range of temperature will not have any effect on your activity. But in reality if you work at a very large Region of Operability for temperature then definitely the activity will vary with the temperature, so it depends on the Region of Operability and the region of interest. In the region of interest in which you are studying that particular parameter will not have any effect on the activity we need to keep that in mind.

(Refer Slide Time: 15:16)




Types of Factors

1. Some factors we don't have control during an experiment. They are called **noise factors** or parameters.
KNP = Key Noise Parameter

For example (a) in a bioprocess development - Environmental factors
(b) Changes in the volunteers health during drug trials

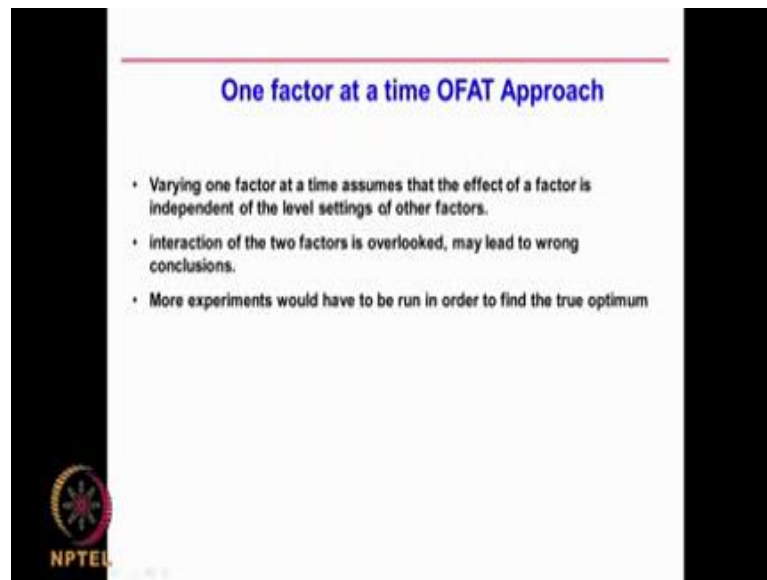
2. Factors that we can control during the experiment are called **control factors**.
KCP = Key Control Parameter



As I said there are some parameters which you do not have any control, there are some parameters which you have control, the parameters which you do not have control like the traffic lights or the road over that is going on or heavy rains so they are called the key noise parameters. The parameters which you have control or the key control parameters that means temperature I can change temperature in a fermenter I can change. For example, environmental factors the humidity conditions, the temperature, the ambient temperature in a bio process you have no control, the volunteers' health change during drug trails.

You have started a drug trail and sometimes the drug trail will last for 2 months, 3 months and the suddenly the volunteer may get a fever, you have no control, right, that is called the key noise parameters. Key control parameters you can exactly control in a fermenter temperature pH I can exact amount of nutrient added, amount micro nutrient added, you have exact control on that actually that is called the key control parameters and that is what is very important noise parameters we have no control on it actually. So, originally I said we can change one factor at a time and do a study that is called the one factor at a time.

(Refer Slide Time: 16:43)



The slide is titled "One factor at a time OFAT Approach" in blue text. Below the title, there are three bullet points: "Varying one factor at a time assumes that the effect of a factor is independent of the level settings of other factors.", "Interaction of the two factors is overlooked, may lead to wrong conclusions.", and "More experiments would have to be run in order to find the true optimum". In the bottom left corner, there is a circular logo with a gear-like design and the text "NPTEL" below it.

For example, I am looking at temperature I vary temperature from 30 to 40 ° in steps of 2. Later on I take pH I vary pH from say pH 4 to pH 7 in steps of point 2, that is called one factor at a time design, that is not a factorial design. When you do that you will not be able to see interaction so if I have the pH and temperature and I first study only pH effect, later on I will study only temperature effect then that is called one factor at the time design I will not be able to see interaction between temperature and pH, whereas if I change both temperature and pH in a factorial way then I will be able to see the interactions, another thing is when you are doing one factor at a time design it may be more difficult to arrive at the optimum condition as against changing many factors at this time simultaneously. So one factor at a time design is extremely not correct adapt, please remember next time you are running fermenter and you want to change temperature pH it is very bad to just change temperature at 5 different values look at the results then change pH at 5 different values and look at the results that is very very very bad and very inefficient. You will not be able to see the interaction effect so the best thing is to go for some sort of a factorial design where you change temperature and pH very simultaneously, very well planned out design strategy there you can see the effect of both temperature and pH individually as well as when they are interacting. In future always use some sort of a design methodology if you want a change more than one factor. So the one factor at a time design is completely wrong. Let us look at a factorial design experiment.

(Refer Slide Time: 18:56)

A factorial design Experiment

- e.g. 2 factors A & B at 2 levels: $2^2=4$ runs
- It is the most efficient experimental design for two-level factors
 - It will give averaged effects without the need for replication.
 - It will account for interaction effects.
 - It facilitates a more rational optimization scheme - steepest ascent.

Table

RUNS	FACTORS	
	A	B
1	Low	Low
2	High	Low
3	Low	High
4	High	High

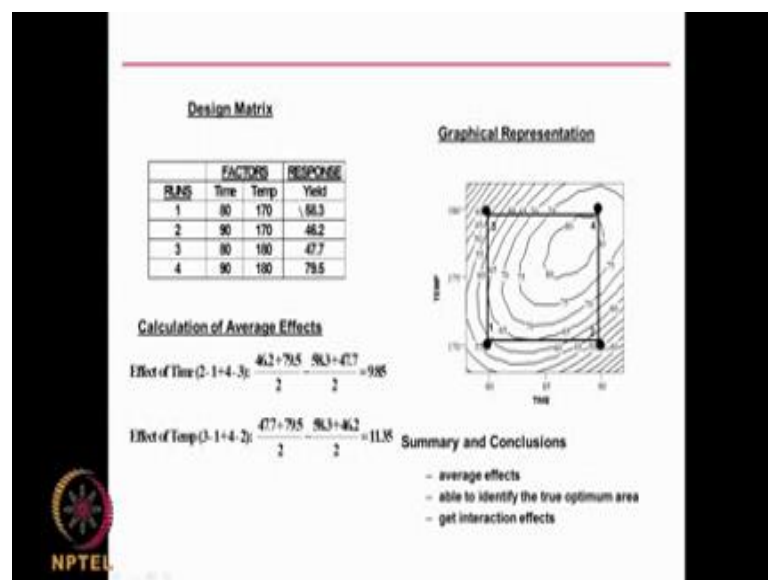
NPTEL

Imagine I want to change like in the previous example I said drug, no drug that is called 2 levels then diet, normal diet and high fat diet that is based on diet 2 levels, so I have 2 parameters 1 is drug, 1 is diet and each of them had 2 levels that is called $2 * 2, 2^2$ design. If I have 2 factors A and B at 2 levels then I will call it 2^2 design, so this 2, the bottom 2 is an indication of the number of levels that means diet and a high fat diet that is the level, you know and these 2 tells you how many factors or parameters I am running, if I have 3 factors A B C then I will call it 2^3 design, understand that is $2 * 2 * 2, 8$ experiments. If I have 4 factors A B C D that means, temperature pH carbon nitrogen 4 factors each one I am studying at 2 levels then it will be 2^4 design, 2 levels, people does studied at 3 levels also but let us look at 2 level design as of now, so temperature, pH, carbon, nitrogen these 2 at different concentrations each one at 2 levels, so the temperature could be at 30 and 40, pH could be at say 4 pH and 5 pH, carbon concentration could be 1 % and 2 %, nitrogen could be 0.5 and 1 %, so 2 levels that is 2^4 design that means $2 * 2 * 2 * 2, 16$ experiments that is called a factorial design. Factorial designs are very efficient it will also give averaged effects without the need for replication. It can also look at interactions because as I said we are going to change many variables at the same time, it gives you nice idea later on to optimize. Ultimately I want to optimize my process.

Now let us look at this 2 factors A B at 2 levels, 4 experiments. How do I do the 4 experiments, I will keep A at low, B at low then I will change A at high, B at low then I

will change A at low, B at high then I will change A at high, B at high, so I have done 4 experiments and both the factors have been looked at low level and high level and as you can see in some cases I am changing both the factors, so that I can even look at interactions here that is the beauty of this type of factorial design. It is extremely powerful, so 2^2 , 4 runs, I need to find the first experiment, suppose this is temperature and pH and temperature is 30° and 40° levels I want to study, pH I want to study at 3 and 4 pH, what do I do the first experiment will be 30° temperature and pH 3, second experiment will be 40° temperature and pH 3, the third experiment will be 30° temperature and 4 pH, the fourth experiment will be 40° temperature and 4 pH, I have achieved all the 4 in different combinations. Do you understand? The low means the lower level for A we have chosen this case 30° temperature, high means higher level for A we have chosen, 40° temperature, low for pH means say pH 3 high for pH means pH 4, that is how you do these experiments, understand. You have 2 level, 2 factors or 2 parameters that gives you 4 runs so you can have 2^3 design, 2^4 design, 2^5 and as you keep on increasing the parameters or factors actually.

(Refer Slide Time: 23:09)



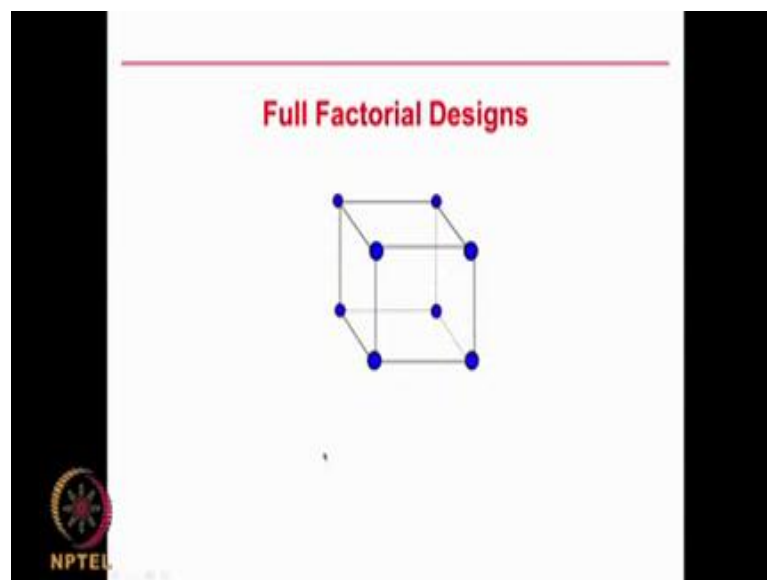
If I know I am looking at, say time and temperature and I am getting the yield here, what do I do? What will be the effect of time? For example, I would done say experiments one and 3 makes use of time at 80 and experiments 2 on 4 makes use of time at 90, what I do I add these 2 results $\div 2$ then I add these 2 results $\div 2$, subtract 1 fr from another that will give you the effect of time. Do you understand? Because there is 2 experiment where we

have done at 80, there are 2 experiments were I have done at 90, so what you do the results of you take

$$\frac{46.2+79.5}{2} - \frac{58.3+47.7}{2} = 9.85$$

, this is the effect of time. The yield has changed increased by 9.85 because of change in time from level 1, 2 that is 80 to 90. Now look at temperature these 2 experiments are done at lower temperature 170, these 2 are done at higher temperature 180 you add up these 2 results $\div 2$, add up these 2 results $\div 2$ and then subtract this from that so $47.7 + 79.582 - 58.3 + 46.2 / 2$ is your 11.35 that means increasing temperature by 10° is increasing my yield by 11.35, increasing the time from 80 to 90 increases my yield by 9.85, so it is extremely powerful we can analyze the results and we can get the effect of mean parameters, we can even get effect of interactions also. I will talk about it later, we will look at that also in due course but this is a very good experiment it is a 2^2 design in the bottom 2 indicates the levels, this 2 the top 2 indicates the number of factors or parameters we have.

(Refer Slide Time: 25:27)



We can have different types of designs and we will spend more time as we go along and next will be the full factorial designs.

Thank you very much for your time.

Key words, - Design of Experiments (DOE) - Factorial design, interaction, variables,
factorial type of experiments, blind and double blind test, critical to quality, Operability
verses Region of interest, Operability verses Region of interest, screening design