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

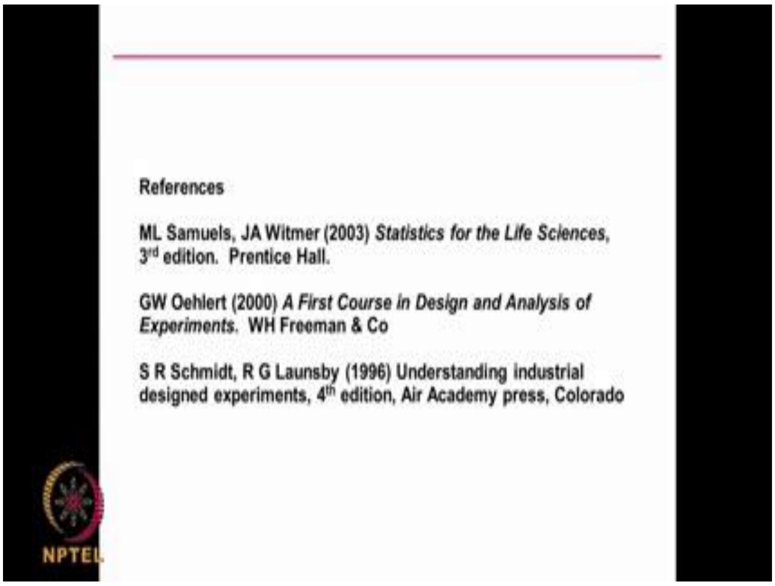
Lecture - 32
Design of Experiments (DOE) - Introduction

Hello everyone, welcome to the course on Biostatistics and Design of Experiments. Today, we are going to talk about design of experiments, it is also called DOE.

Design of experiments are extremely important if you want to do a well-planned out study of a very complicated system. If you do not plan your study properly, then whatever data you collect, will be completely wrong. You will not have a statistical basis for analysis and statistical basis for coming to a conclusion. So, design of experiments is very important and it is not taught in many courses.

Many of the software, but have facility to give out designs of various types and most of you might not be aware how each software spews out these different types of designs. So, we are going to talk about in the next few classes, how one goes about designing or planning the experiments and how do you vary the variables and so on, actually.

(Refer Slide Time: 01:17)



References

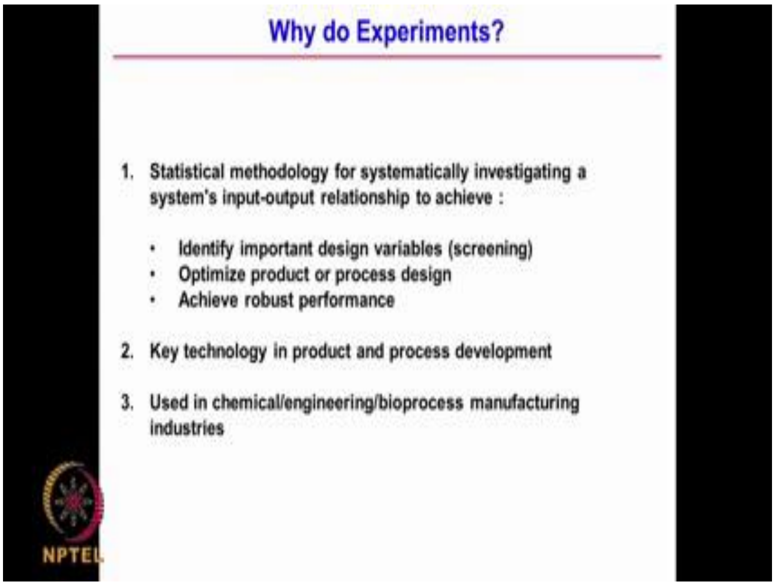
ML Samuels, JA Witmer (2003) *Statistics for the Life Sciences*, 3rd edition. Prentice Hall.

GW Oehlert (2000) *A First Course in Design and Analysis of Experiments*. WH Freeman & Co

S R Schmidt, R G Launsby (1996) Understanding industrial designed experiments, 4th edition, Air Academy press, Colorado

So, some of these references, and I have listed out here. So, if you have access to these references that will be very useful for you. There is a reference relates to life sciences and then, there is one on design and analysis; there is also understanding industrial designed experiments. I do make use of this book also, this is quite simple and very practical and so, I think you should have a book for yourself so that you do not just rely on a software all the time. You should get the philosophy of how the designs are done and if you understand it, it is extremely interesting and very fascinating actually.

(Refer Slide Time: 02:09)



Why do Experiments?

1. Statistical methodology for systematically investigating a system's input-output relationship to achieve :
   - Identify important design variables (screening)
   - Optimize product or process design
   - Achieve robust performance
2. Key technology in product and process development
3. Used in chemical/engineering/bioprocess manufacturing industries

So, why do we need to do experiments? Actually, this is a fundamental question, why should I do experiments? Why should I have a design of experiments?

So, design of experiments is a statistical methodology for systematically investigating input-output. So, you may have several inputs and you may have several outputs also. Like, for example, my carbon concentration, nitrogen concentration, the pH, the temperature, the agitator, rpm, the amount of oxygen bubbled, these could be input. My output could be, amount of, say, biopolymer produced, amount of biomass produced, amount of secondary metabolites produced, so lot of outputs. So, you could have several inputs, several outputs and each of them may behave differently for different inputs. These inputs are called x's, independent variables, parameters and so on. The output is called generally the dependent variable, the y.

So, we do these experiments to identify important design variables. You may have hundreds of variables, but only few of them may be important. So, if you are running a plant, you are interested to know, which ones I should focus on. Which x’s should I think about having a good control on? So, I do not have to spend money on looking at other x’s, so I focus only on the important x’s.

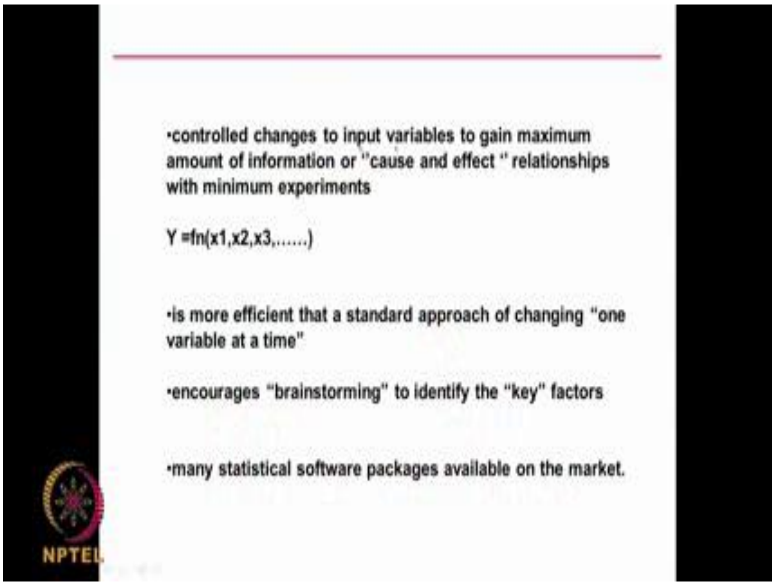
Optimize my product and process design, this is very important. Ultimately, you want to get the best out of your plant, you want to minimize the energy usage, raw materials usage and get maximum amount of your desired product. Whether it is a biopolymer or whether it is a secondary metabolite or whether it is an antibiotic, I want to maximize its production and minimize my raw material usage, that is obvious, right. That is called optimization. And similarly, if I am doing a product design, I want to improve the quality of the product. Product which will have the best, say, tensile strength or compressive strength or flexural strength or maximum reliability and so on. So, that is called the optimizing the product design.

Achieve robust performance, ultimately we want the, say, bioreactor to be robust. It should not go out of control for small changes in your x’s. You know, the temperature changes by one degree, we do not want a very large change in my product amount and quality. So, that is called a robust design. How the process is able to absorb small, small changes in your inputs. For example, raw materials can have different amounts of impurities, will that affect too much on my product concentration, product purity? If it affects too much, then I need to have a very pure raw material. So, even for small variations in the raw material concentration, if my product concentration or yield changes a lot, then it is not very robust. But whereas, if it can absorb the concentrations of the impurity present in the raw materials and still give me the desired amount of product, desired quantity and concentration, then that is called a robust design.

This is, design of experiments is very, very important in product process development. So, if you are moving from a small scale, that is, lab scale going right up to a manufacturing scale without performing a design of experiments, you cannot just jump and start making in a large scale. This is very commonly used by chemical engineers, by bioprocess engineers in any manufacturing. Whether you are manufacturing a chemical, whether you are manufacturing antibiotics, whether you are manufacturing metabolites, secondary metabolites, whatever be it, unless you do a proper design of experiments, you cannot move from small scale to large scale. You cannot expect to have an optimum process with the minimum raw material and energy usage and maximum product yield and desired product concentration.

So, that is what we are going to talk and I will be talking about how one varies the various x’s or various input parameters to achieve the maximum information as well as maximum output, desired output.

(Refer Slide Time: 06:22)



- controlled changes to input variables to gain maximum amount of information or "cause and effect " relationships with minimum experiments

Y =fn(x1,x2,x3,......)

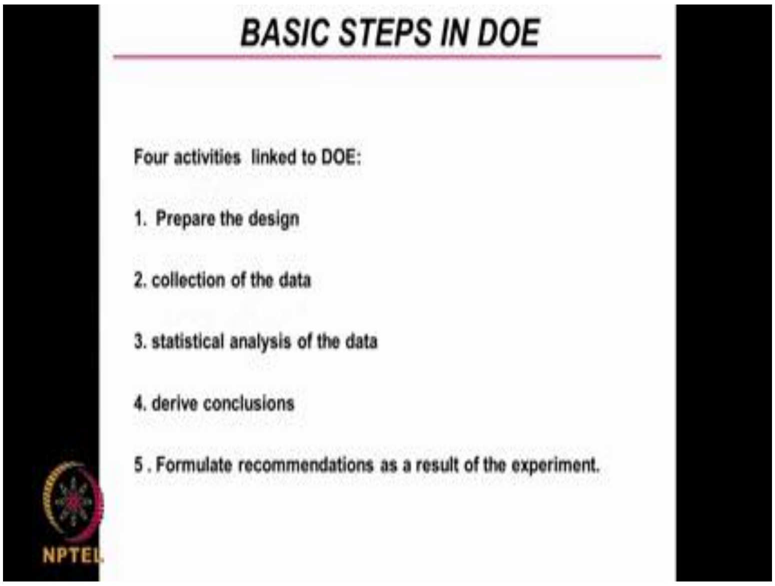
- is more efficient that a standard approach of changing "one variable at a time"
- encourages "brainstorming" to identify the "key" factors
- many statistical software packages available on the market.

So, we are going to controlled changes to input variables to gain maximum amount of information, this is called a cause-effect relationship. We need to have design of experiments performed, so that we can develop regression relationship. We will talk about regression also later. So, we want to develop equations like, yield of my desired product is equal to function of various input parameters, right. So, in order to derive such an equation, I need to perform experiments so that gives you a cause and effect. You know, I may develop equations like this, right, the yield is equal to function of temperature, pressure, dissolved oxygen and so on. It may be a linear relation, non-linear relation, it could be anything actually.

Now this is more efficient, design of experiment is more efficient then changing one variable at a time. Imagine I want to look at temperature, pH and rpm, that is, agitator rpm. It is not very intelligent just to do experiments by changing temperature alone, few experiment changing temperature alone, then keep everything constant, then now keep temperature also constant, change pH alone, different values of pH, then keep all of them constant, then change rpm alone, different values of rpm. That is called one variable at a time or one factor at a time and that is not very, very efficient because it will not be able to identify interactions. You know what is interactions? I talked about interactions many times in ANOVA, two way ANOVA, three way ANOVA.

So, when you change only one factor, you will not be able to identify whether there is an interaction between two factors like temperature and pH maybe having interaction. Unless you simultaneously change this, you will not able to study those effects, ok. Also, statistical software will also have in the market these design of experiments. I, like I said, you know, it can spin out different types of designs, these packages can do that actually. So, it does not require much intelligence at all.

(Refer Slide Time: 08:32)



BASIC STEPS IN DOE

Four activities linked to DOE:

1. Prepare the design

2. collection of the data

3. statistical analysis of the data

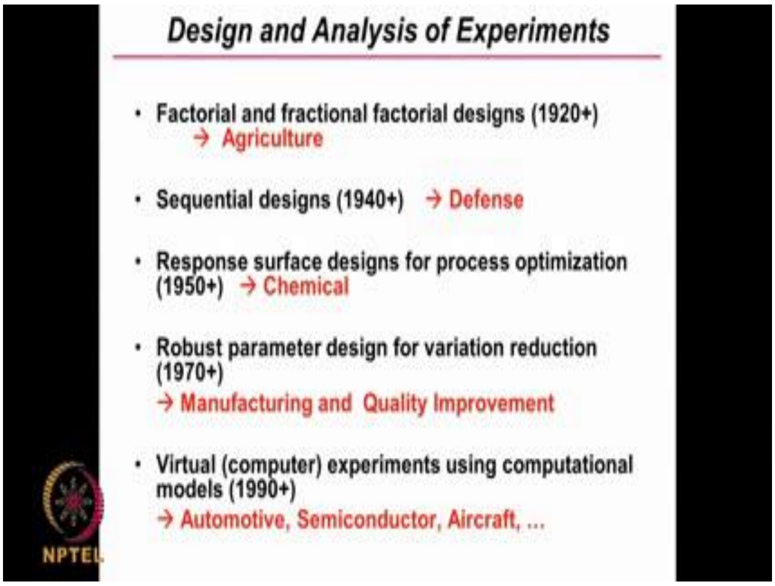
4. derive conclusions

5 . Formulate recommendations as a result of the experiment.

So, what are the activities involved in DOE? First, you need to prepare the design. We will talk about it in the next few classes, how do you prepare. Once we have the design, which gives you the different levels of the input parameters, then you go to the lab or plant and collect the data. If your output or desired dependent variable is biomass, so you measure biomass at different input values or input variables, then you statistically do the analysis of the data. You may use T test, F test, we looked at so many tests in the past, say about 30 classes and then you derive conclusions.

Based on that we will say, we will accept null hypothesis or we agree to reject null hypothesis then. So, we agree on alternate hypothesis. Then, we develop mathematical relation between various input parameters with the output parameter and then we formulate recommendation because of all these actually. So, we decide, that temperature should be only between 35 and 37, pH should be always (Refer Time: 09:51). So, these types of recommendations we make based on our design study actually. These are the basic steps in design of experiment.

(Refer Slide Time: 10:04)



Design and Analysis of Experiments

- Factorial and fractional factorial designs (1920+)
  → Agriculture
- Sequential designs (1940+) → Defense
- Response surface designs for process optimization (1950+) → Chemical
- Robust parameter design for variation reduction (1970+)
  → Manufacturing and Quality Improvement
- Virtual (computer) experiments using computational models (1990+)
  → Automotive, Semiconductor, Aircraft, ...

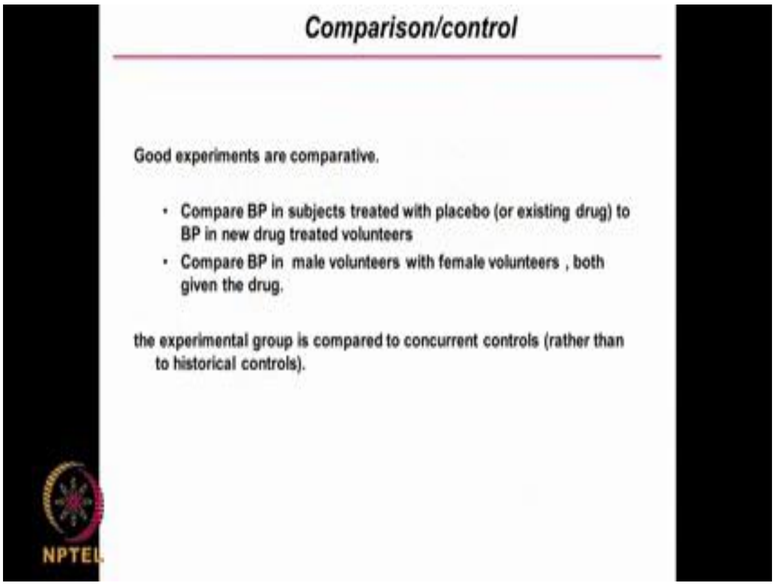
If we look at design of experiments historically, it has been there from 1920s, early 1920s. So, it was used in agricultural and factorial designs were developed during agricultural studies. For example, studies were carried out to see whether this particular fertilizer is better than that or this treatment of pesticides was better than that and how they performed on different types of land areas and how they performed with different plants. So, we had many parameters and you cannot do too many experiments, so design of experiments was thought of at that point of time, that is, 20s.

Then, came sequential designs in the area of defense and of course, by around 50s chemical industries started using these different types of designs. This is called response surface designs, which was used for process optimization because ultimately in chemical industries, they want to maximize the production of the desired product, minimize the usage of chemicals. So, the design is called response surface designs were incorporated in the early 50s.

Then, came the robust parameter design. As I said, I do want my product quality or product performance to change too much with respect to my input values. So, it should be able to absorb these variations and that is called the robust design that came into manufacturing and quality control, ok. Even if, for example, the quality of my fuel varies in a range, the performance of the car should be so robust enough to give you the same mileage per liter of the fuel. That is called a robust design.

Then, came virtual experiments using computational models design of experiment were also used in computer simulation, especially for simulating semiconductor performance, aircraft performance, automotive performance. So, design of experiments was also started being used in mathematical modeling and simulation also. So, it has been there, it is being used in almost many fields of science and engineering. And biological engineering also has taken it and they have started using the various design of experiments tools in the biological research.

(Refer Slide Time: 12:39)



Comparison/control

Good experiments are comparative.

- Compare BP in subjects treated with placebo (or existing drug) to BP in new drug treated volunteers
- Compare BP in male volunteers with female volunteers , both given the drug.

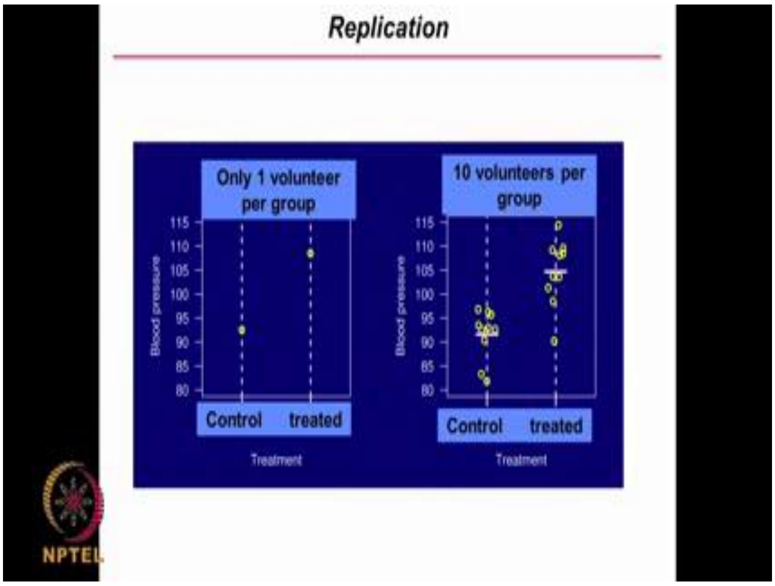
the experimental group is compared to concurrent controls (rather than to historical controls).

Let us go forward. So, good experiments are always comparative, you know. If you are, say, comparing BP in subjects treated with placebo to BP in new drug. So, if we are looking at a drug, I will always compare it with the placebo. We talked about it in many times in the course of these weeks, so either placebo or existing drug. So, if I want to say, this new drug is better or as good with respect to placebo or existing drug, so we need to do that. So, you may compare say male volunteers with female volunteers on the performance of a drug. So, always good experiments are comparative.

We never take historical controls and then compare it that is very, very rare. So, if I want to introduce a new drug into the market, I will always carry out clinical trials with the old drugs, with the set of volunteers and new drug with set of volunteers and make a comparison, ok. That is always done. I will never take historical data. The data performance of the old drug is given in the literature, so I will take that and do it; that is not a good idea at all. So, it is always good to have set of volunteers for control or for old drugs if you want to introduce a new drug into the market. So, comparison and control are very, very essential.

We have being looking at many problems in this idea. Never, never compare with the historical controls. That is not a very good idea unless you do not have a control. For example, you can say, the life span of people have increased from, say, 40 years in the 19th century to almost 70 years. So, if I want to do that sort of study, I may get volunteers in the current age, but I will be not able to get volunteers from the 90s, 90s, 19th, right, so that is a problem. So, in such situations, of course, we cannot have a comparison. The current, concurrent controls, we have to make use of the historical controls only in such situations, but otherwise it is always good idea to have concurrent control, be it placebo, be it old drug, old assay, old volunteers and so on, actually.

(Refer Slide Time: 15:03)



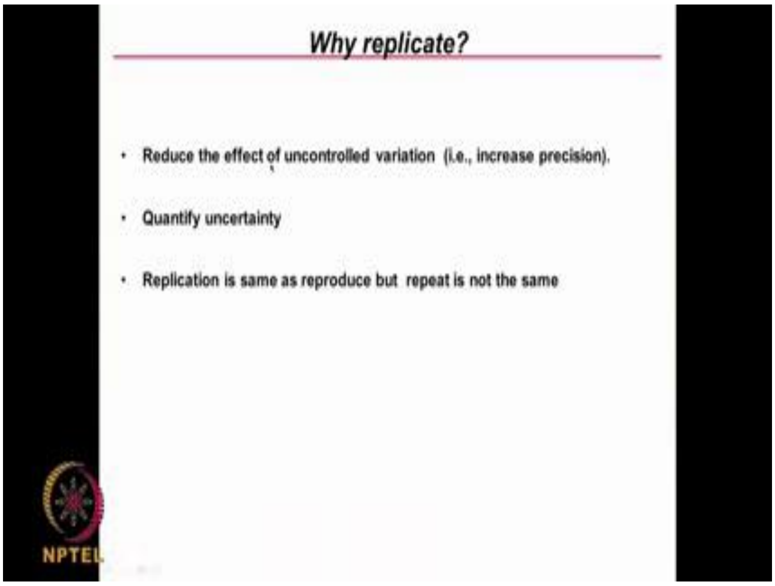
Replication

So, then next comes replication. We talked about replication or reproduction that is very, very important. That means, you carry out the entire experiment not just once, may be twice, thrice, four because that gives you an idea about error and if you want to get error sum of squares without replication, it is very, very difficult.

So, suppose I am looking at blood pressure on control group and those we treated, it is very bad idea to just do experiment with only one volunteer, one of each, that is very bad because we have no idea about the error involved. But it is always a good idea, say you take 10 volunteers per group, so the blood pressure may vary of the control from, say, 85 to 97 and the treated could vary between 90 to 115. So, we have a range of a values. So, we can calculate variances for the control, we can calculate variances for the treated, we can perform F test and so many things we can do. But with this we cannot do anything. Actually, it is just a single point control.

So, replication of experiments is extremely crucial. And I also showed you before, that when you do not have replication, it becomes very, very difficult to understand error sum of squares or even sometimes it is very difficult to understand confounding or interactions.

(Refer Slide Time: 16:37)

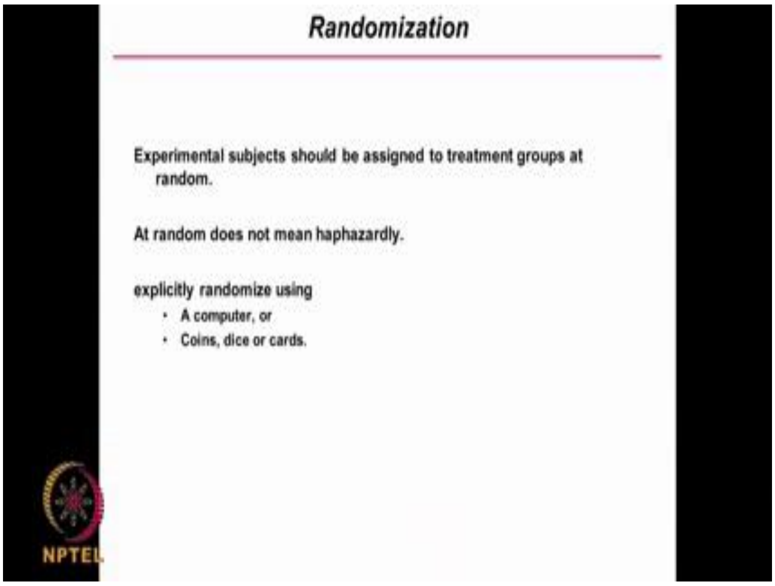


Why replicate?

- Reduce the effect of uncontrolled variation (i.e., increase precision).
- Quantify uncertainty
- Replication is same as reproduce but repeat is not the same

Why replicate? Reduce the effect of uncontrolled variation. So, we increase a precision, quantify uncertainties because say, any assay, any methodology will always have an error. So, replication helps you to find out what is the error margin. So, replication is same as reproduce like I said, but it is not same as repeat. Repeat is just taking a sample and repeating the measurement in the instrument three times, but replication is by performing the entire experiment with the x’s; that is replication.

(Refer Slide Time: 17:16)



Randomization

Experimental subjects should be assigned to treatment groups at random.

At random does not mean haphazardly.

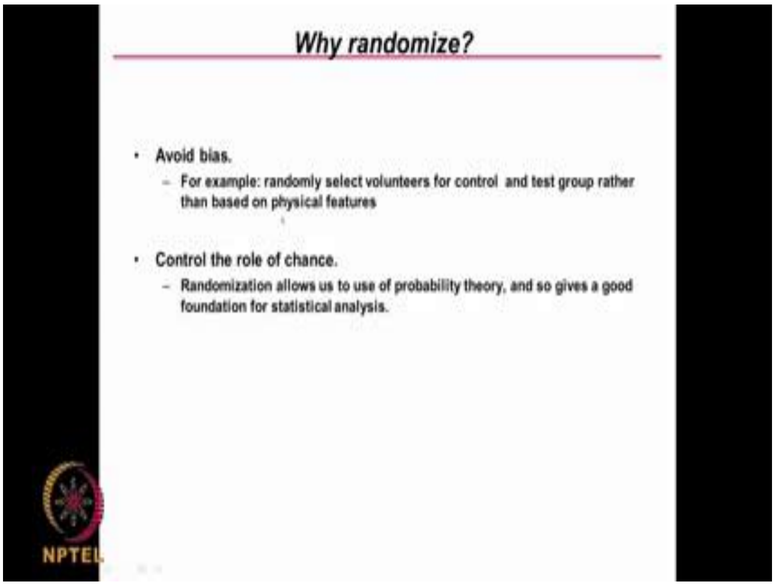
explicitly randomize using

- A computer, or
- Coins, dice or cards.

Randomization, this is also very important. We have to randomize otherwise we will always have a bias. If I am going to take, say, 20 volunteers, I will put some of them into placebo and some of them in the drug. I will randomly pick volunteers and put into these two groups. I will not go with certain bias, I will not take people who look healthy and put them into placebo or vice versa, that is not correct, that is called biasing.

So, we can randomize using a, there is a random number generator software was there, table was there. So, if there are 20 volunteers, you can make them, ask them to stand in a queue and then, use a random number generator or even toss a coin and pick them randomly and put them assigned them into these two groups. That is the correct way of doing it rather than bringing in a bias, otherwise that is very, very dangerous. So, randomization is very important when we perform experiments.

(Refer Slide Time: 18:20)



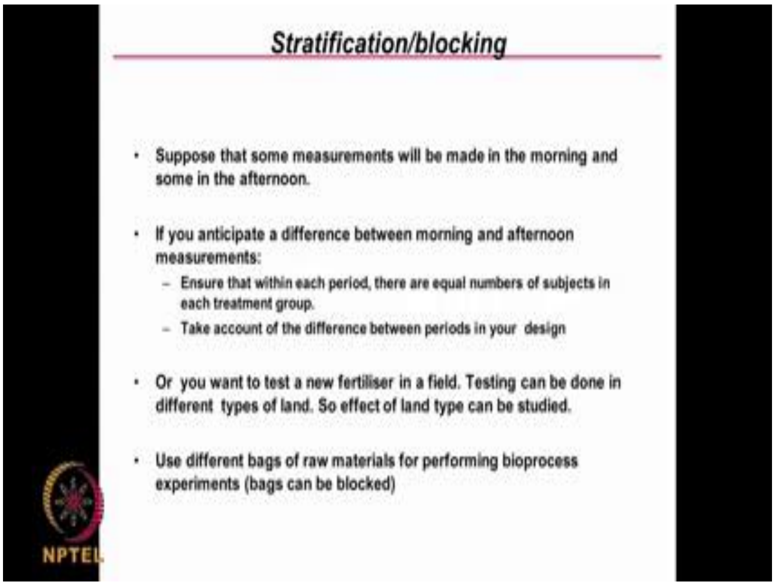
Why randomize?

- Avoid bias.
  - For example: randomly select volunteers for control and test group rather than based on physical features
- Control the role of chance.
  - Randomization allows us to use of probability theory, and so gives a good foundation for statistical analysis.

Why randomize? It avoids bias. So, randomly selected volunteers for control and test group rather than based on physical features, like as I said, you know, we look at people who look healthy and put them in control. That is not correct; that is bias and if you look at healthy volunteers or unhealthy volunteers and put them into test where we are going to give the drug again, that is not correct actually. That way we have the chance.

Randomization allows you to use the probability theory because the entire probability theory is based on random tossing of coins, tossing of dice and so on, actually. So, entire statistical analysis techniques can be applied if we use a random method rather than a biased method.

(Refer Slide Time: 19:07)



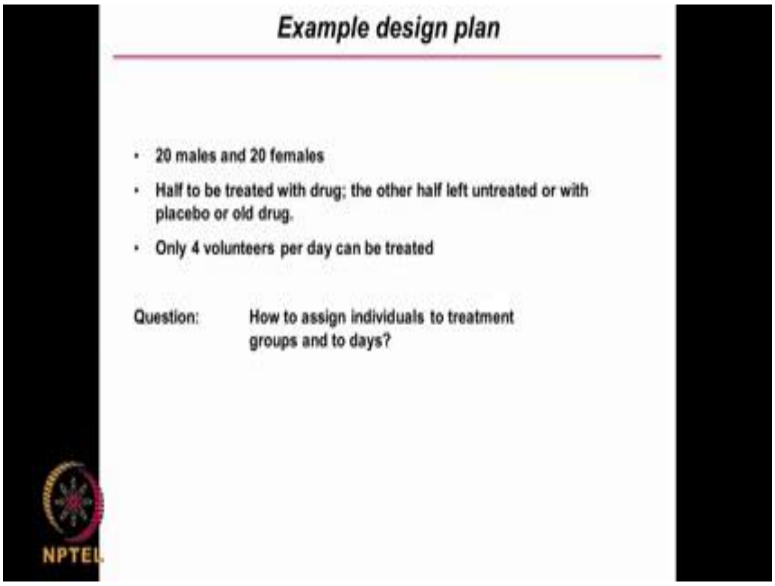
Stratification/blocking

- Suppose that some measurements will be made in the morning and some in the afternoon.
- If you anticipate a difference between morning and afternoon measurements:
  - Ensure that within each period, there are equal numbers of subjects in each treatment group.
  - Take account of the difference between periods in your design
- Or you want to test a new fertiliser in a field. Testing can be done in different types of land. So effect of land type can be studied.
- Use different bags of raw materials for performing bioprocess experiments (bags can be blocked)

Next comes blocking or stratification. So, for example, I am taking some, say, blood glucose measurement or blood pressure measurement of volunteers with test group and control group. These may data will be made in the say, morning or afternoon. So, if you think there is going to be some differences when I take data in the morning or in the afternoon that is true with blood pressure or even with glucose. For example, blood pressure may be low in the mornings, whereas it could be high in the afternoon. So, in such a situation we can have equal number of subjects in each group, you know, that is called blocking. That way we can take account of the differences between periods in your design. So, you do not have to worry their morning data collected and afternoon data collected is going to give you problems.

For example, you are testing a fertilizer in a field, there are different types of field. So, you do not, you are not very sure, that whether that is going to affect your, the performance of a fertilizer, then we can sort of, different types of lands could be blocked. Similarly, if you have different bags of raw materials for performing bioprocess experiments, suppose I take samples from one bag and do some experiments and take samples from another bag and do experiments. If I am worried, that each bag may have some variations, which may affect your results, then I can use bag as block. So, I will control, I mean, sorry I will perform a, measurement, calculations only in each individuals block and we can also later on do between block analysis to see whether block has a effect, that is called blocking.

(Refer Slide Time: 21:10)



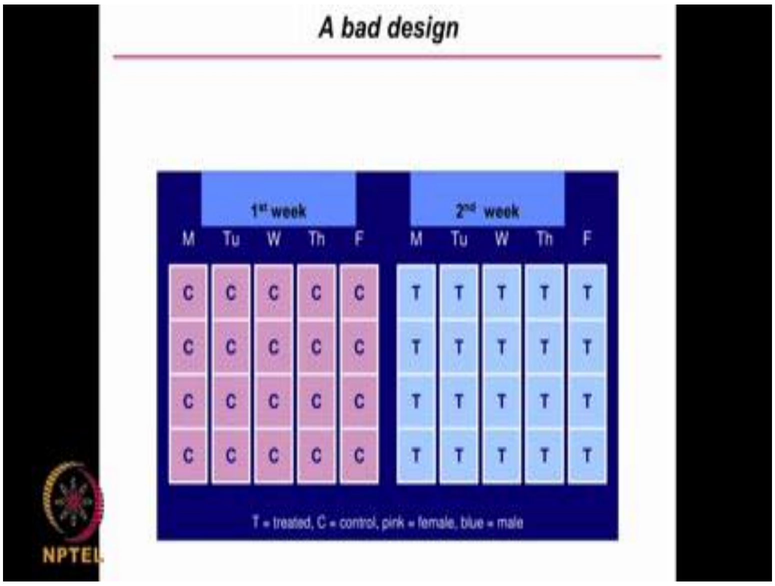
Example design plan

- 20 males and 20 females
- Half to be treated with drug; the other half left untreated or with placebo or old drug.
- Only 4 volunteers per day can be treated

Question: How to assign individuals to treatment groups and to days?

So, look at this, 20 males and 20 females. I have, half of them are going to be treated with drug, other half left untreated or with placebo or old drug. I can do the treatment only for 4 volunteers per day. So, Monday to Friday only I am going to do the work. So, how will you assign individuals to the treatment groups in two days? So, I have 20 males, 20 females and half of them in each group will be controlled, half of them in each group will be the test. So, how am I going to perform this design plan?

(Refer Slide Time: 21:48)



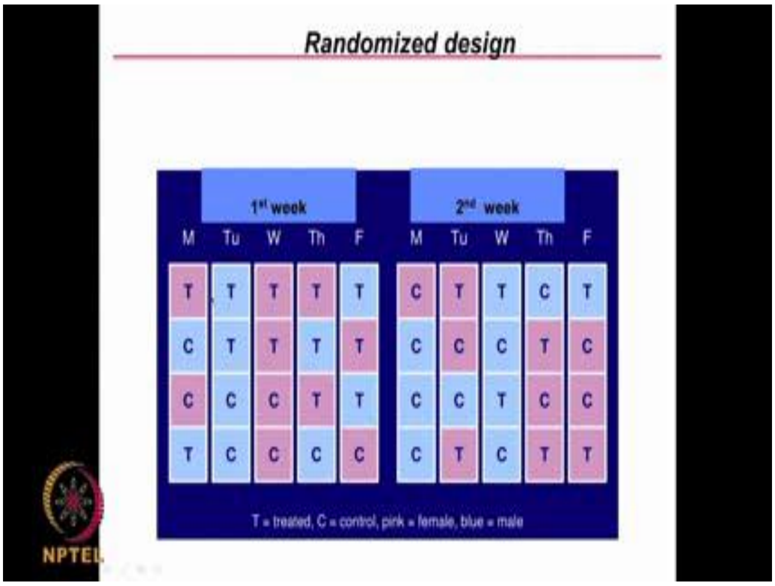
A bad design

1st week					2nd week				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T
C	C	C	C	C	T	T	T	T	T

T = treated, C = control, pink = female, blue = male

One design plan, Monday I will have a control, control, control female and then control, control, control, again female on Tuesday, like that. And then, later on, in the next week I may have the treated, treated, treated male. This is a very bad design; this is extremely bad because you are completing all of one set and then all of second set. There is no randomization; there could be bias coming into the picture. So, that is a very bad design. So, another alternate will be randomize design.

(Refer Slide Time: 22:28)



Randomized design

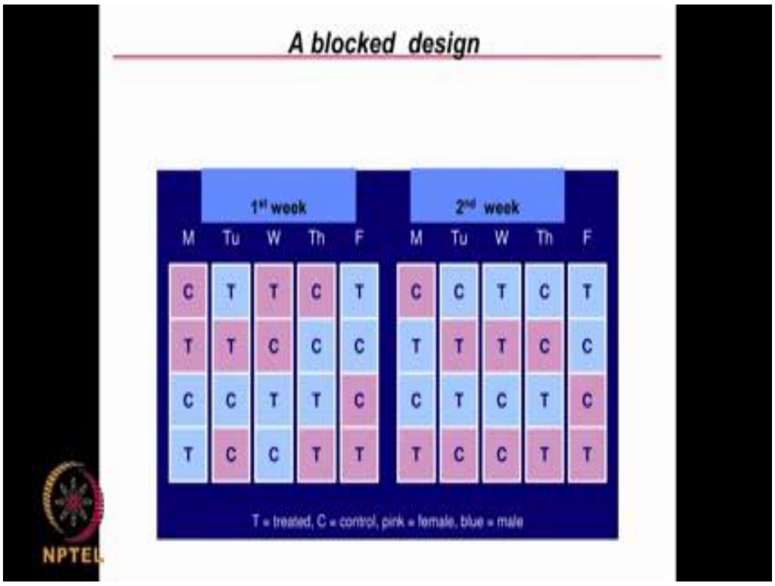
1st week					2nd week				
M	Tu	W	Th	F	M	Tu	W	Th	F
T	T	T	T	T	C	T	T	C	T
C	T	T	T	T	C	C	C	T	C
C	C	C	T	T	C	C	T	C	C
T	C	C	C	C	C	T	C	T	T

T = treated, C = control, pink = female, blue = male

So, what we do is, we may take a treated person, a drug female and then we could take a control male, then we could take a control female and then we could take a treated male, drug treated male. So, we have different types. We have a female and male taken here because you have pink and pink and blue and blue, but you also have treated control, control treated, that is, on Monday. It is quite random.

Next day, we may take two treated male and two treated control male. Next day, we may have two treated female, two control female, like that. Now, this is quite random. As you can see, it is randomly done. There is no pattern at all coming into the picture. This is called a randomized design.

(Refer Slide Time: 23:35)



A blocked design

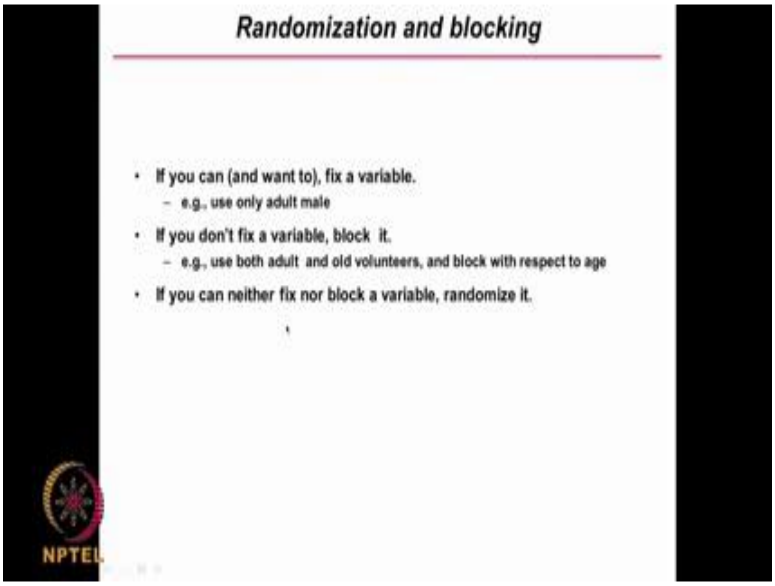
1st week					2nd week				
M	Tu	W	Th	F	M	Tu	W	Th	F
C	T	T	C	T	C	C	T	C	T
T	T	C	C	C	T	T	T	C	C
C	C	T	T	C	C	T	C	T	C
T	C	C	T	T	T	C	C	T	T

T = treated, C = control, pink = female, blue = male

If you want to block it also, then we can do it like this. So, we will have the female control and test together, then we have a male control and test together, like that, you know, we have some blocking. So, this is a block design, like that we can do.

So, as you can see, never, never have a design like this where the complete one set of all the female control, then you go into treated and so on. This is a very bad approach to do, whereas this a much better randomization and this is blocking of the data of male and female together.

(Refer Slide Time: 24:18)



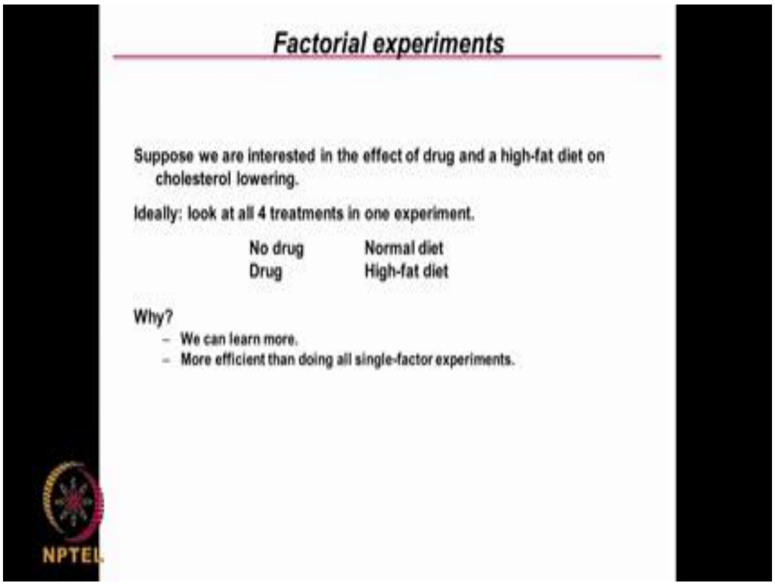
Randomization and blocking

- If you can (and want to), fix a variable.
  - e.g., use only adult male
- If you don't fix a variable, block it.
  - e.g., use both adult and old volunteers, and block with respect to age
- If you can neither fix nor block a variable, randomize it.

So, if you can fix a variable, like if you want to do only adult male, then it is ok, but if you do not fix a variable, then block it, that is, if you are going to take both adult and old volunteers, then we can block with respect to age. So, we, and have some group of volunteers adult, some group of volunteers who are old and then you perform the experiments and then, later on, you can also look at effect of age also. That is a good (Refer Time: 24:56), but if you can get only a adult male between the age of 30 to 45, then no problem, age will not come into the picture.

If you can neither fix nor block a variable, then better to randomize it, because there could be situation where you might not be able to get all adult and old people. Suppose, if you are testing some drugs for sudden treatment, most, some disease may happen only in certain type of population and so on. Then, say, you just randomize it. So, this is how we do plan the experiments.

(Refer Slide Time: 25:34)



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.

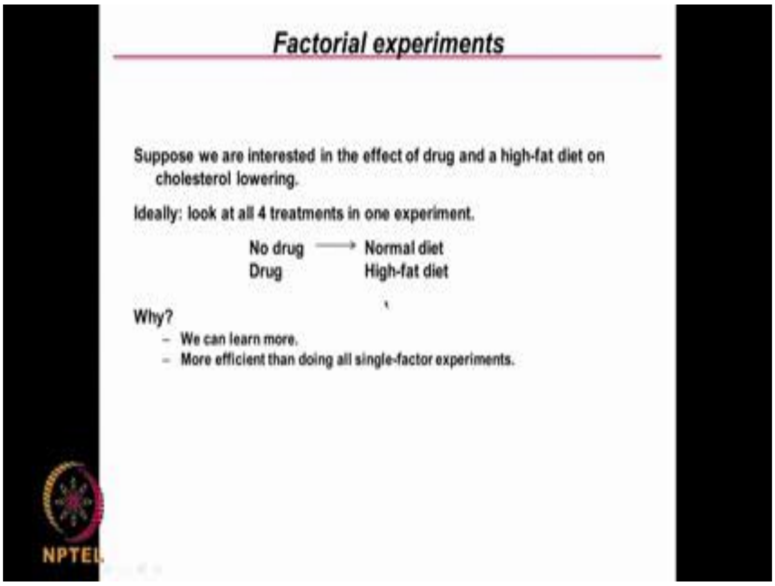
No drug	Normal diet
Drug	High-fat diet

Why?

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

Now, there is something called factorial experiments. We will look at these factorial, you are going to come across this word factorial quite often. So, imagine, I am looking at a drug and diet for cholesterol lowering, so you could have no drug, drug and then normal diet, high fat diet.

(Refer Slide Time: 26:00)



Factorial experiments

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

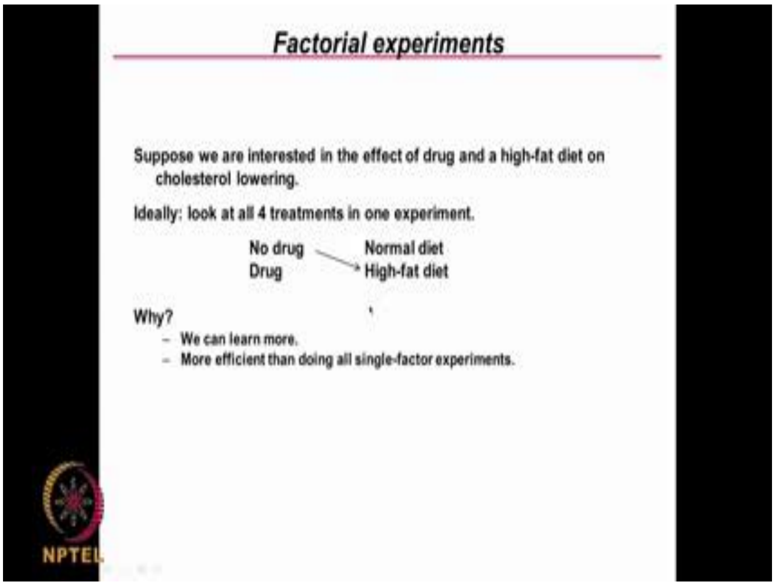
Ideally: look at all 4 treatments in one experiment.

No drug → Normal diet
Drug High-fat diet

Why?

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

(Refer Slide Time: 26:09)



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.

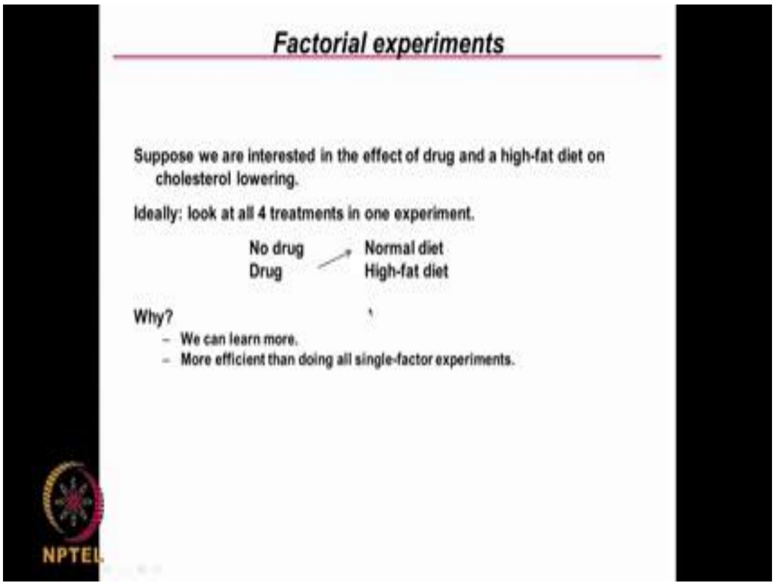
No drug → Normal diet
Drug → High-fat diet

Why?

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

So, you can have four different treatment strategies, right. No drug, normal diet; no drug, high-fat diet.

(Refer Slide Time: 26:12)



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.

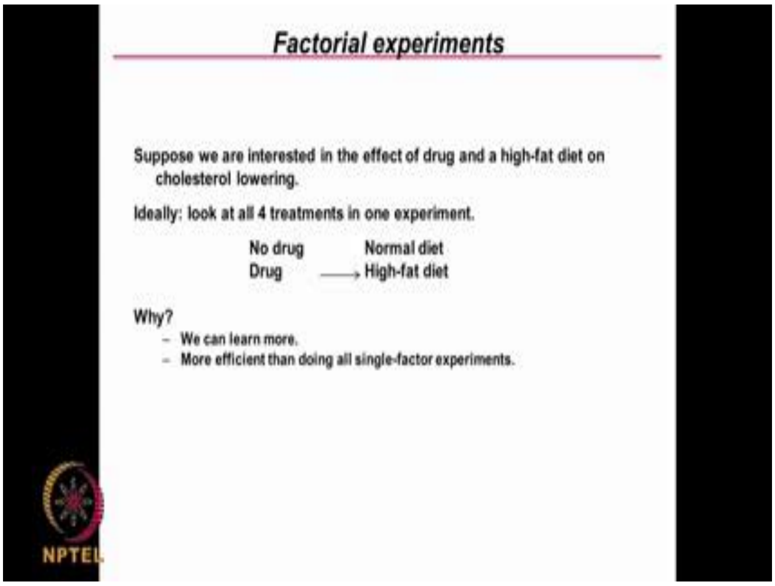
No drug → Normal diet
Drug → High-fat diet

Why?

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

.

(Refer Slide Time: 26:14)



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.

So, we can have a drug, normal diet. Then, finally, drug, high fat diet.

So, we have four different situations because we have two factors: no drug, no drug, normal diet, high-fat diet. So, 2 into 2, 4. So, by doing this we can learn more, we can look at effect of the diet, we can look at effect of drug, we can even look at effect of, that is, each one is a single factor and then we can even look at effect of drug and diet combined together also. So, that is an advantage. So, this is called the factorial experiment.

We have two factors, that is, drug is one factor, diet is another factor and each at two levels, that is, no drug, drug; other one is normal diet, high-fat diet. So, 2 into 2, 4. So, we will be doing 4 experiments. So, it is always better to look at four different types of experiments. How do you do?

We will take, first experiment will be no drug, normal diet, that will the first experiment and see the performance; no drug, normal diet. Next experiment could be: no drug, high-fat diet. Third experiment could be drug and normal diet. Fourth experiment could be with drug and high-fat diet. So, we are combining both these factors and getting four experiments. So, it is much better than doing single factor experiment.

For example, single factor experiment could be, one experiment be no drug, next experiment could be drug, next, third experiment could be only with normal diet, fourth experiment could be high-fat diet, no change in the drug pattern. Whereas, the factorial experiment, we are changing both simultaneously in some situations, that way we will be able to look at even interactions very efficiently.

So, many design of experiments makes use of factorial experiments or factorial designs, so we are going to look at factorial designs. So, this is called a two-level factorial design because we have two, two levels: no drug, drug or normal diet, high-fat. And we have two variables here or two parameters here: one is called the drug parameter, other one is called the normal diet, high-fat diet that is another parameter, that is, diet as another parameter. So, we will talk about this factorial experiment in the subsequent classes.

Thank you very much.

Key words,

Design of experiments, variable,factor, ANOVA, Interaction, experiments, Excel, replication of experiments, Randomization, blocking or stratification, One design plan, factorial experiments