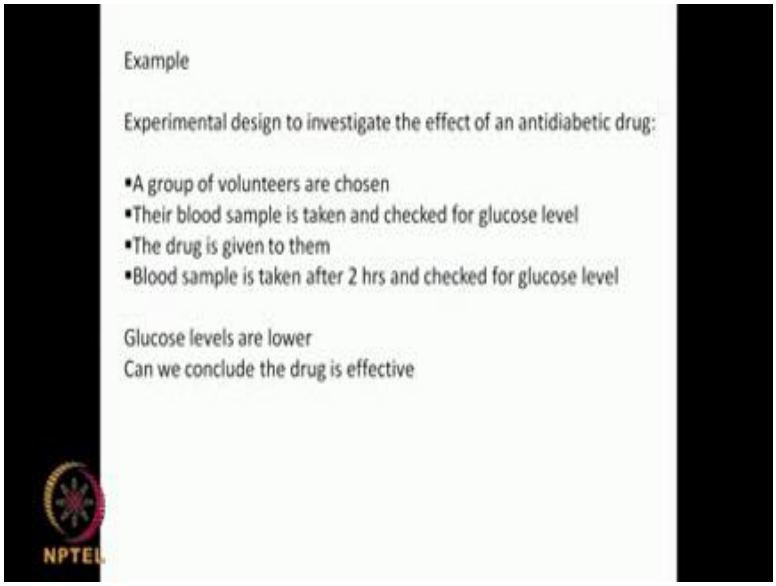
Biostatistics and Design of Experiments
Prof. Mukesh Doble
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 02
Experimental Design Strategy

Welcome back to the course on Biostatistics and Design of Experiments. As I said biostatistics deals with the analysis of data and trying to reach at some statistical conclusion based on the data. The second part deals with the design of experiments, how do you go about designing experiments in a correct way, how to reduce the number of experiments I would like to do and so on. So, there are many experimental design, philosophy or strategies one needs to consider, and sometimes we make mistakes if we do not plan a well thought out design strategy, and at the end of the experimental work we may realize that we have made a big mistake, and we have lost time as well as resources. So, you need to keep many points in mind before you start your experimental, real experimental, work. Let us look at some examples.

(Refer Slide Time: 01:04)



For example, I want to investigate effect of an antidiabetic drug. What I do is, I get a group of volunteers, then their blood samples are taken, and you check their glucose level. Then you give the drug to them and again after 2 hours you check their glucose levels. So, now the glucose levels are lower; the glucose levels in the blood are lower.

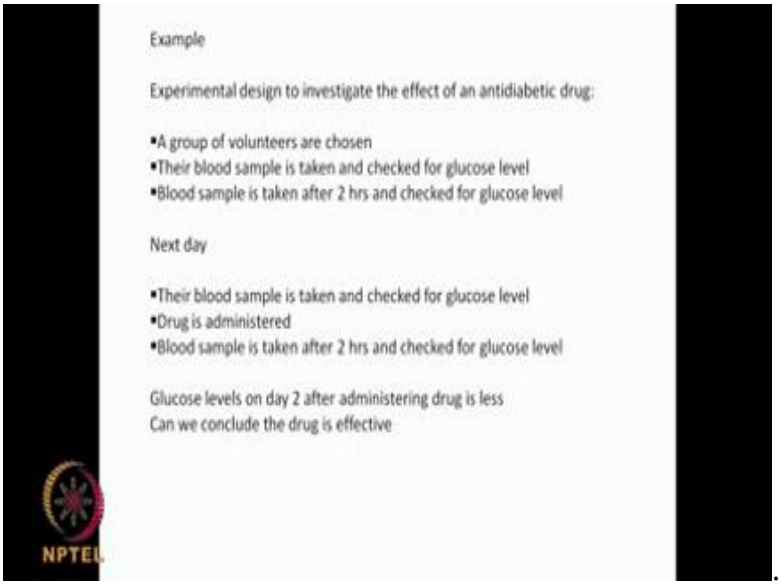
So, can we conclude the drug is effective? Definitely not, because there is a very big mistake which we have made. What is the mistake? We have not considered a control group. So what is the control group? We need to have a group of people, who follow the same pattern of activities just like the other group that is called the test group, where you are testing the drug but the control group they will not be given the drug. So they will not be given the drug but they will undergo the same process, may be the food eating pattern, the staying pattern, the other activities - they are called the control. So at the end of the time, you need to collect blood samples from the control group - where they have not been given the drug, as well as on the test group for whom the drug has been given, and then you have to make a conclusion. Then only the experiments are correct; otherwise, they are completely wrong. So, you always should have a control, that means a group where everything is same but they do not get the drug and the test group where you give the drug.

So, if you are testing an anti-bacterial drug or a diabetic drug or anything you need to always have a control group. Even if I am testing some anti-bacterial compound on some bacteria, I need to have a control set of plates where I will grow the bacteria but I will not give the treatment; and I will also have a test where I grow the same bacteria, at the same conditions but I will give the anti-bacterial drug. So, that way you can make a comparison between these two.

Let us look at another experiment. You are again looking at an antidiabetic drug. What you have done is, you have a taken a group of volunteers, their blood sample is taken and checked for glucose level; then, again you give the drug and then after another 2 hours you do the same test. For example, in this particular case glucose levels are the same. Can I say the drug is not effective because the glucose levels are the same, right? No, again you have a fault. So you need to have a control group, and then you need to check their glucose levels as well. You check their glucose levels, and see whether there is change in the glucose level, and you also check the glucose levels on the people on whom you have given the drug, and if the glucose levels have changed on the control group whereas, that means the glucose levels might have gone up in the control group whereas in the test group it has remained same, then you can say the drug is effective. Whereas if that glucose levels are the same in the control group as well as in the test group, then we can say the drug is ineffective.

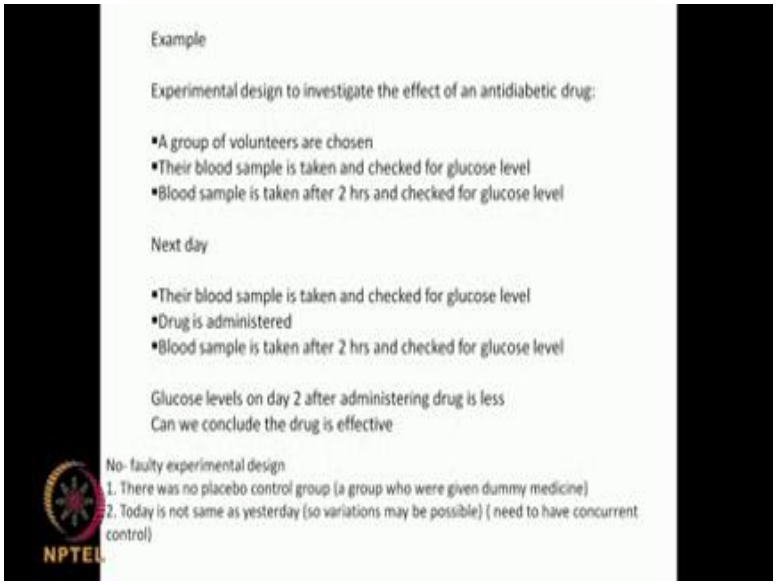
So, control group is very, very important. You looked at two examples, right? When you are having a glucose level reduced, still we cannot say the drug is effective unless you have a control group, and if the glucose level remains same, still you cannot say the drug is ineffective, because you still need a control group where the glucose levels might have shot up if the drug has not been given.

(Refer Slide Time: 05:01)



Now, let us look at another scenario. You are trying to look at again an antidiabetic drug. So, you have a group of volunteers; that blood sample is taken and checked for glucose levels; again, blood sample is taken after 2 hours and checked for glucose levels. Now, next day they come, then the blood sample is taken and checked for glucose level, drug is given now; then the blood sample is checked after 2 hours for glucose. So, the glucose level on day two has gone down. So, can we conclude the drug is ineffective or effective? You see these differences - on the first day, you are using the same volunteers as control; that means you are checking their glucose levels for a long time at 2 hours, and the next day you are checking their glucose levels, then you are giving the drug and again after 2 hours you are checking their glucose levels to see whether there is a change. So the glucose levels have gone down. Can we conclude the drug is effective? No.

(Refer Slide Time: 06:17)



Now, we introduce something more complicated that is called the placebo control. That means, you also should have a group of volunteers on whom you give a placebo or a dummy drug, so that they do not have the feeling that they are getting a glucose lowering drug, so automatically it may be acting on these people actually. Obviously, we should have a placebo control; that means, you should not tell the patients on whom you are giving the drug, whether you are giving the drug or the placebo. So, we have now 3 different groups - a group which will be control, that means on whom no drug is given and then there will be 2 groups. One group you may give the placebo and one group you may give the drug. But you do not tell on whom you are giving the placebo, you do not tell the group on whom you are giving the drug. So that way the patients or the volunteers will not know whether they are getting a placebo or a real anti-diabetic drug; that is called a blind.

So you need to always have a placebo, especially if you are testing drugs, because even the placebo can act on people in lowering their problem if it is diabetic or if it is infection, so we need to get rid of that placebo effect and this is called a blind test. That means, you are not revealing whether you are giving a placebo that means or a dummy, or you are giving the real drug. And also a variation from yesterday, today, tomorrow can be possible actually, because the way, the type of food we eat over the period of time, the glucose levels may be going up and down, so you need to keep that in mind. So we need to have a placebo group and that is what is called concurrent control.

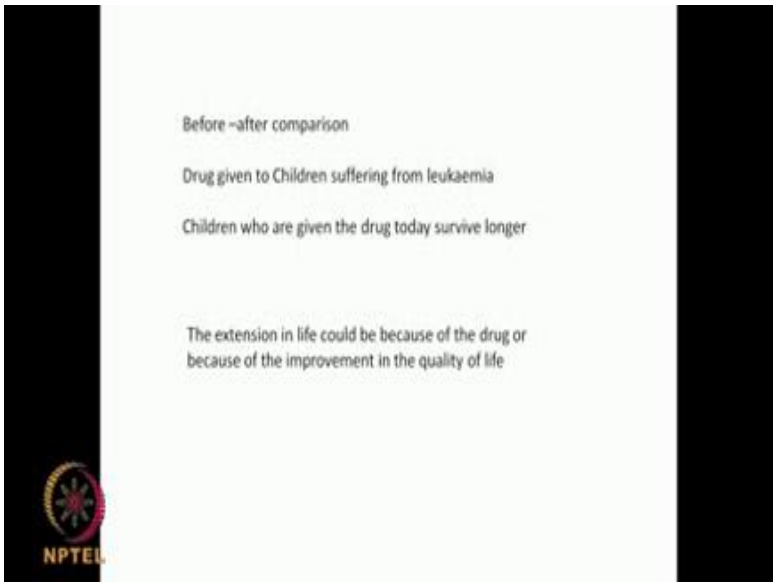
(Refer Slide Time: 08:22)

Now, as I said you need to keep the patients or volunteers blind. So, they will not know whether they are getting the drug or the placebo. Now, what is double blind? So, here the doctors who are administering the drug also will not know whether they are giving a placebo or the drug, because if the doctors know whether they are giving the drug, inadvertently or involuntarily they maybe passing on the information to the patients. So, the doctors also will not know whether they are administering a drug or a placebo, the patient or the volunteers also will not know whether they are getting a drug or a placebo. That is called a double blind test. If you want to do a good study of a drug we need to do a double blind type of test with proper control and where you could perform these operations in a concurrent matter. So, that means the patients will not know whether they are getting a drug or the placebo, doctors also will not know whether they are administering a drug or the placebo, so you get rid of all these variation and real change that happens is because of the drug. But still that is not enough as the examples I am going to tell you is much more complicated.

(Refer Slide Time: 09:51)

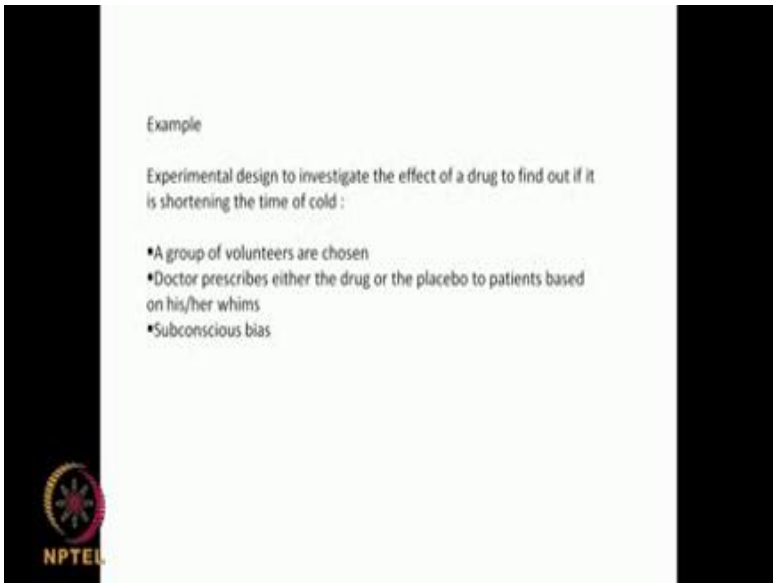
There is something called before-after comparison. For example, drugs are given for children who are suffering from Leukemia, so once upon a time may be 50, 60 years back there were no drugs for Leukemia, whereas now we have drugs for Leukemia. Now, the children are surviving longer with the drug today. So, can we say that the drug is the cause or reason why the children are surviving longer? No, because the life style has changed, the quality of life has changed, so the children have now more nutrition, the children of yesteryears - that means a couple of decades back or even 30 years back - vis-a-vis the children of today, the children have a better quality of food intake, better nutrition, they are able to have a better defense mechanism, so they are able to survive. So, that effect also comes into picture; it is not only the drug, but also the effect of defense mechanism of the children of today vis-a-vis the children of 3 decades or 5 decades back. So, that is called the before and after comparison. So, you need to be very careful that you cannot completely say the drug has prolonged the life of children suffering from Leukemia.

(Refer Slide Time: 11:17)



The extension in life could be because of the drug or because of the improvement in the quality of life. Now, the children get better quality of food and better defense mechanism and so on actually.

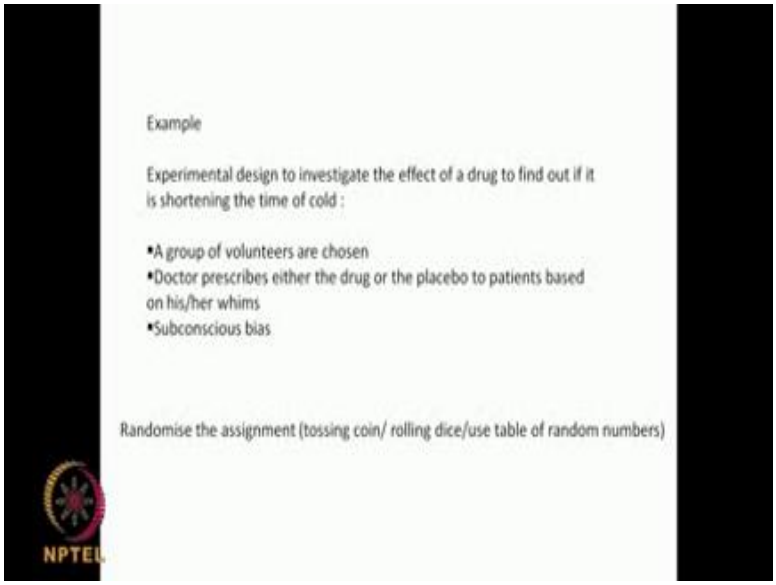
(Refer Slide Time: 11:27)



So another one. Now you are testing a drug which will shorten the time of cold; everybody gets cold. So, a drug is being tested, and the doctors are asked to give the drug to patients who have the cold, and not give the drug to people who do not have a cold or show a symptom of cold. So, a group of volunteers are chosen. Now, the doctor

prescribes either the drug or the placebo to patients, based on his conclusion or based on his observation. There could be some bias coming into the whole thing because the doctor could be biased looking at the symptoms of the people or he may think somebody has the cold whereas the patient might not be having a cold or he may think somebody is not having a cold whereas the person may be having a cold. When the doctor by mistake prescribes the placebo when a patient has a cold, while the doctor prescribes a cold medicine or the drug which is being tested when the person does not have, so there is a mistake. There is a subconscious bias which is entering into this experimental design strategy.

(Refer Slide Time: 12:49)

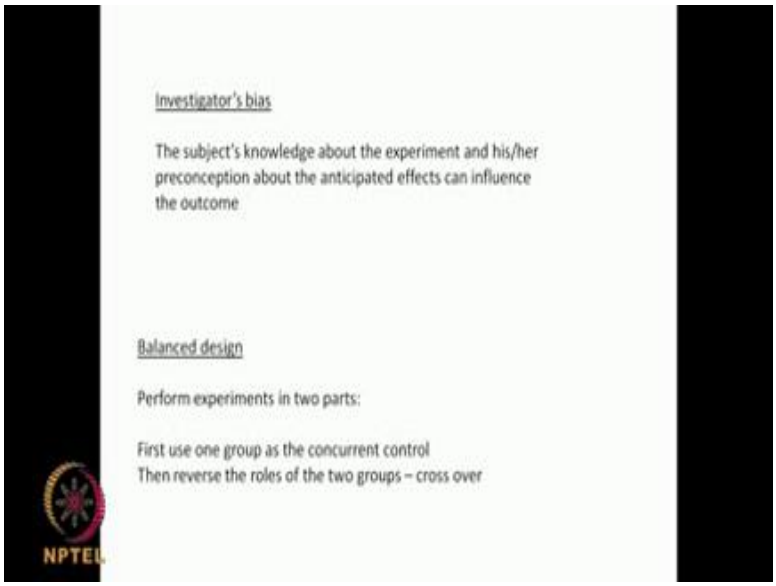


Let us look at another case. Now, you are looking at…So randomize the assignment. We need to be very careful whether the doctor is not making a mistake. Can we randomize the whole process rather than giving the entire decision for the doctor to make, so there could be lot of bias coming in to the picture? So the whole process of drug testing could be completely mistaken.

(Refer Slide Time: 13:17)

So we have something called Balanced design. That means, we need to perform the experiments in two parts. First time you can use, suppose you have a group of volunteers divide them into 2 groups. The first group may be the control, the second group could be the test; and then, you can reverse the roles, you can make the first group as test and second group as the control; this is called a Balanced design. So, that way you are very, very sure that if there is a change in the disease pattern, you can say it is because of the drug; you understand? You have the group A as a control, group B as a test, and then reverse the roles and make group B as the control and A as the test - that is called the cross over design. This is called a cross over design.

(Refer Slide Time: 14:11)



That way we will get rid of the investigator's bias. So, investigator who is performing some experiments always will be having. For example, if I am carrying out a bio transformation or a fermentation reaction, I feel that if I had more carbon the productivity goes up. So, that is a bias. That is the subject knowledge I have; so, I feel that the yield has to go. So, I also anticipate which may influence the outcome, even if I do not see any change in the yield, I will still feel that yield has to increase, because I know that when I increase carbon amount yield has to grow up. So that is called the investigator's bias.

That is another issue that can come into when you are doing a statistical study; actually, we all do that. We have some domain knowledge whether it is fermentation or drug testing, we have a chemistry knowledge or a bio chemistry knowledge or a fermentation knowledge. So, when we do the experiments, we always feel we expect the trend to be like that, if I increase temperature, reaction rate should increase; if I change pH the organism may not survive so I will get less colonies. So that is the bias I already have preconceived notion which I have, so that maybe affecting my entire experimental results also. So I need to be very, very careful about investigators bias.

So sometimes you need to do randomly experiment. So that sort of a bias does not come in or you bring in somebody else who do not have much of domain knowledge. So the bias of the investigator does not come into picture.

(Refer Slide Time: 16:00)

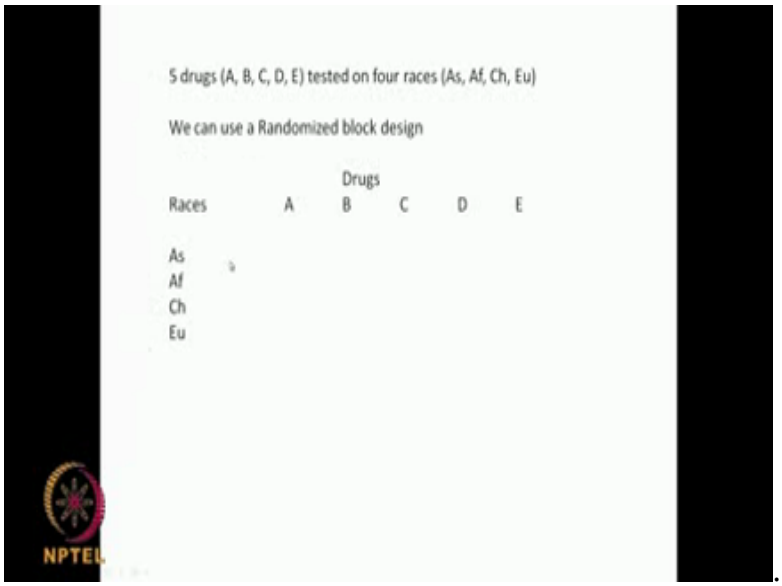
So cross over I talked about, testing the effect. For example, I am testing anti-diabetic drug. So I have 2 volunteers, A set and B set. The group A will receive placebo on day one and then drug on day two; whereas, group B will receive drug on day one and placebo on day two. So, the glucose levels are reducing when the drugs are given and remain the same when placebo was given in either groups, then we can be very sure that there is an affect because of the treatment. Of course, we can also have double blind; that means, here we do not tell the group A or B whether that they are getting placebo today or whether they are getting drug today and vice versa. And we can also have the doctors or physicians who are administering the drug in a blind state; that means we do not tell them whether they are administering placebo or a drug. So, we can ideally, in a drug testing scenario, you are going to have a cross over double blind type of design. So you need to keep that very, very carefully in mind.

(Refer Slide Time: 17:13)

If we have more than 2 groups, of course if you may be testing drug A, drug B, drug C, drug D, drugs which are existing already in market, drugs which you are introducing control. So all these are there, right. So we can call it treatment 1, 2, 3, 4 strategy. For example, I am testing different types of rice. I have 4 types of rice A, B, C, D and I may be having different treatments strategy. I may be fertilizing some with the higher nitrogen phosphorus ratio, I may be fertilizing some of them with less of potassium. So I may be changing the ratios of N, P, K So four treatment strategy, 4 types of rice, so how do I do? This is called a 4 x 4 Latin square. So, day one I have the rice A getting treatment strategy 1, and then day two I have rice B getting the treatment strategy 1, day three I have the rice C getting the treatment strategy 1, and day four I am having the rice variety D getting the treatment strategy 1 and so on actually.

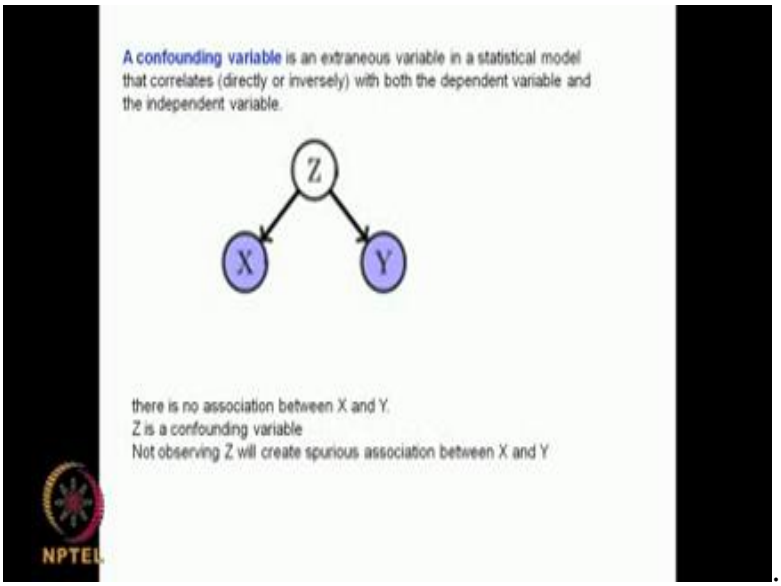
As you can see here, these are very symmetric, so on day one you will have rice A getting treatment strategy 1, then rice B getting treatment strategy 2, rice C getting treatment strategy 3, and rice 4 getting treatment strategy 4. So it is very nice, this is a very good cross over, this is called a Latin square design, this is very symmetric, you can design such Latin squares for even unsymmetric cases also. That means, if I have a 3 treatment strategies and 4 types of rice, we can design. We are very, very sure that the other effects which we talked about do not come into play and mess up our final conclusion as well as the result. So, this is called the Cross Over Symmetric Latin Square Design 4X4.

(Refer Slide Time: 19:23)



Now, I am testing 5 drugs on 4 races. That means, I maybe testing on Asians, Africans, Chinese, Eurasian or Europeans and I have 5 drugs. It need not be 5 drugs, A could be control, B could be placebo, C could be existing drug, D and E could be new chemical entities; I am testing that. I can have this type of strategy, I have A, B, C, D, E and I have Asian, African, Chinese and Europeans. I can give each one of these groups different types of drugs on each consecutive day and then I can see the affect of these drugs. So this is… we can have it in a very randomized fashion also so that you do not have a bias into entire thought process; it is called a randomized block design. So, block plays a very, very important role in drug testing, in agriculture sciences, in medicinal chemistry and so on actually. So whether it is symmetric system, whether it is unsymmetric set of data we can nicely design a block for the entire design strategy.

(Refer Slide Time: 20:47)



Now, I want to introduce another terminology which is called a Confounding Variable. Or this is a variable, which is extraneous or external, but it correlates with both X and Y. X could be independent variable, Y could be some other variable independent also, but actually there is no relationship between X and Y, but this confounding variable has a relationship between X and Y; so it appears as if there is a relation between X and Y. So you have 2 variables X and Y, they are not related, but there is another variable which is called a confounding variable Z, it is related to X and Y, so because of that you end up observing as if there is a relationship between X and Y. So it is a spurious relationship between X and Y but actually there is no relationship. So you need to know what the variables that is confounding; otherwise, you will end up creating new relationship between variables which do not have any relationship at all.

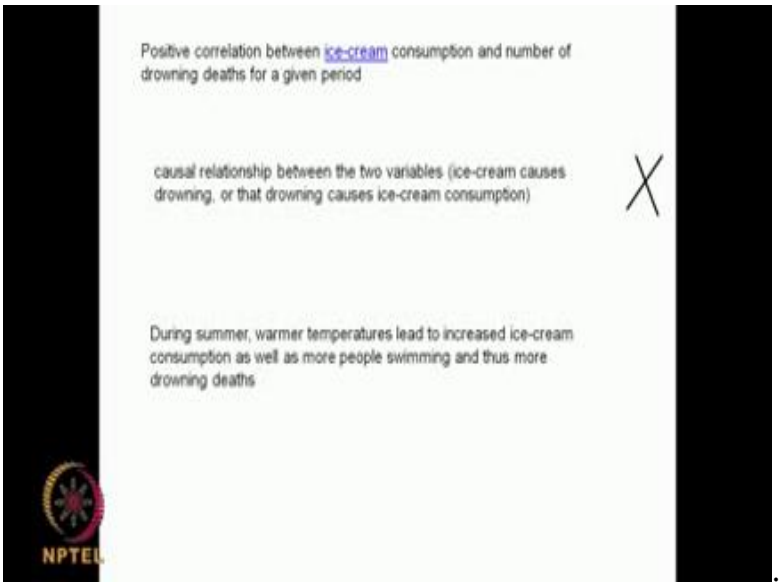
(Refer Slide Time: 22:02)

Let us look at some examples, very interesting examples, very, very interesting. There is a positive correlation between ice cream consumption and number of drowning deaths in a given period. It looks very funny right, how can ice cream consumption have a relationship between drowning deaths.

(Refer Slide Time: 22:22)

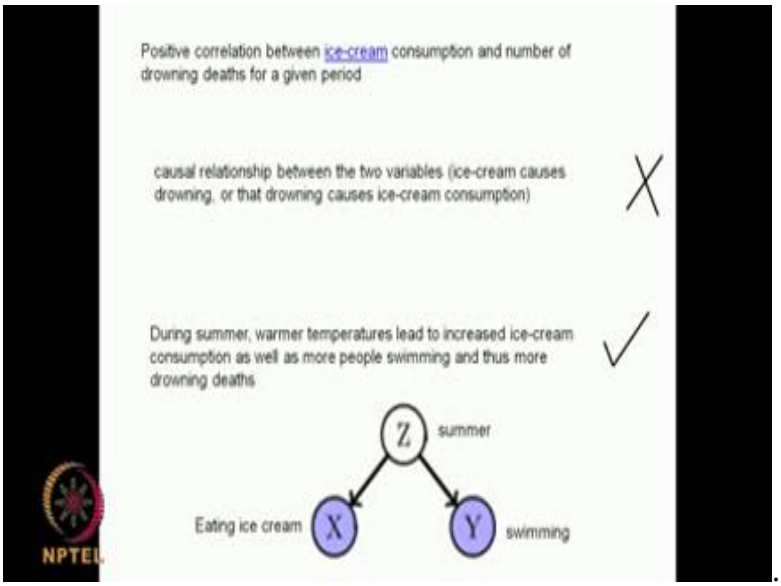
What does that mean? So causal relationship between the 2 variables. Does ice cream cause drowning? Or drowning causes ice cream consumption? This is very funny right, but then if you look at it very, very carefully, so obviously, this is wrong.

(Refer Slide Time: 22:36)



If you look at it very carefully summer, summer time is very warm. So, everybody eats ice cream. So consumption is high. Summer time is warm, so everybody goes for swim and hence there is more drowning deaths. So we see a relationship between ice cream consumption and drowning deaths, but actually the summer is the confounding variable which connects the ice cream consumption and the swimming, and hence, the drowning deaths; very interesting. So this is the correct thing.

(Refer Slide Time: 23:10)



So obviously, summer connects eating ice cream, summer connects swimming. We find

more people eating ice cream in summer; we find more people swimming in summer. So it appears, and hence drowning. It appears that there is a mathematical relationship or positive correlation between consumption of ice cream and drowning deaths.

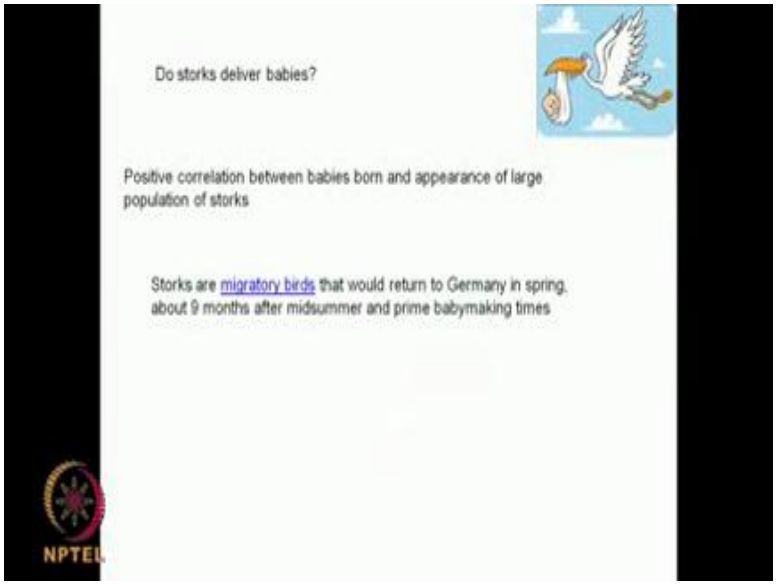
(Refer Slide Time: 23:38)

Now, let us look at another example. A study of smoking tobacco on human health. Obviously, smoking, drinking, alcohol diet, lifestyle activities - they are all related. If somebody looks at smoking and human health, but does not look at whether that person is consuming alcohol, whether the person has healthy diet or not, whether the person has good lifestyle or not. If the analysis completely neglects these, and looks at only smoking and human health, then it may appear as if smoking has a very, very serious effect on human health. But then one should look at other variables which are also going to have an effect on the human health like consuming alcohol, type of food the person eats, and also the lifestyle - sedentary or active activity. So, if we neglect all those, then you assume that smoking has a very, very serious effect on health.

(Refer Slide Time: 24:42)

Let us look at another example, very interesting. Do storks deliver babies? Because it has been found that there is a positive correlation between babies born and appearance of large population of storks. This goes back to several decades and there are many jokes that stork carries babies and bring them to the houses actually. So, obviously there is a confounding variable which if one neglects, the whole situation appears very funny.

(Refer Slide Time: 25:16)



Storks as you know are migratory birds, they would return to Germany in spring about 9 months after mid-summer, and most of the babies, prime baby making time is in the

summer. So, obviously they will be born 9 months after the summer, which is this spring time. Storks will be returning during that period in many European countries, and that is the time, that means during the times babies are born 9 months after the mid-summer time, so appearance of large population of storks matching with the birth of babies, both happen around the spring time.

(Refer Slide Time: 26:02)

So obviously, one may get confused about whether storks bring in, because the storks population in various European countries depends upon pollution, conservation efforts, habitat, size of the country, land area. All these have an effect on storks and they are migratory birds they go away during the autumn and winter. They come back again during the spring time; so you see lot of them coming back, that is one way.

In many countries when babies are born, in order to keep the house warm, people used to heat their houses - in the olden days they use to heat their houses with the wood; so obviously, the chimneys will be very hot and warm. So, obviously birds, migratory birds, also seek warm places, so they come and settle down on these chimneys which are emanating lot of warm gases. So, you will see that houses which have new babies born will be wanting to be warmer, so they use more of the heating charcoal or coal, so the chimneys will be hot and more migratory birds like storks settled down to keep themselves warm. So, there seems to be a relationship between babies and storks.

So the confounding variable is a very important concept which one needs to keep in

mind, otherwise we will try to find relationship between 2 independent variables which cannot have any relationship. We saw some interesting examples like ice cream consumption vis-a-vis the number of drowning deaths and so on.

So, when you are doing a statistical analysis, if we see that there is a relationship between 2 variables which cannot have some relationship, then you have to be very, very careful about the whole thinking. There could be a confounding variable or variables which may be imparting this type of relationship.

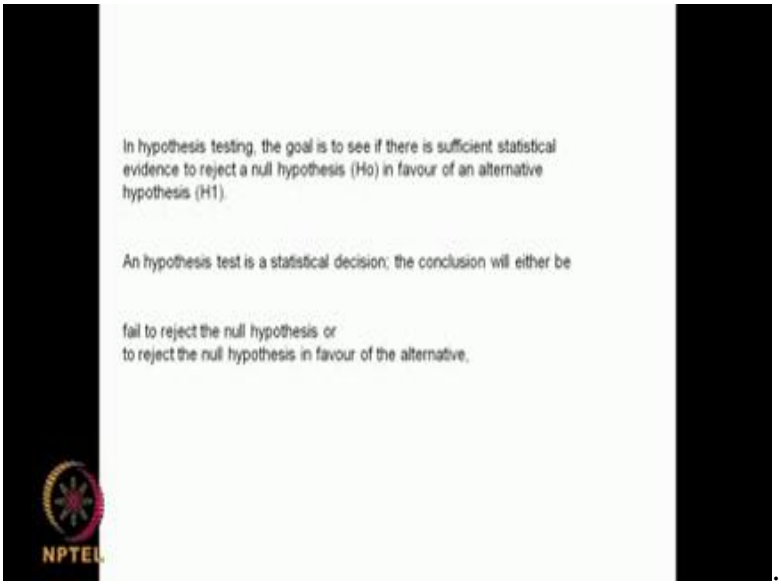
(Refer Slide Time: 28:04)

Now, let us go deeper and we will define a terminology called hypothesis. Hypothesis is an educated guess that attempts to explain a set of facts or phenomenon. It is used very commonly in fields of science, medicine, where we use a scientific approach for collecting data and testing the data. A statistical hypothesis is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables. So, we have set of data collected, we are observing the data, and then making a conclusion. So we originally have a hypothesis. When I am wanting to introduce a drug, I will say, this drug is going to be more effective than existing drugs, that is an hypothesis. Then, you test it out by collecting the data and then you prove or disprove the hypothesis.

So hypothesis is a very important concept in statistics. So you will see about lot of things hypothesis. One is called the null hypothesis, alternate hypothesis. Null hypothesis is status quo; that means, if you are going to introduce something, you say there is no effect

or no improvement of the existing drug; that is called null hypothesis. So, we will go deeper into that but hypothesis is a very important concept in statistics.

(Refer Slide Time: 29:33)



In hypothesis testing, the goal is to see if there is a sufficient statistical evidence to reject a null hypothesis. So if I am going to introduce a drug, the null hypothesis could be, there is no change in the existing drug as against the new drug. Alternate hypothesis could be - yes, the new drug has an affect. So, when you do a hypothesis test we make a decision or a conclusion, we say we fail to reject the null hypothesis; that means, the null hypothesis cannot be rejected or we reject the null hypothesis, so we accept the alternate hypothesis.

For example, let us go back to the drug. Now, I am going to introduce a drug B, there is already a drug A. So the null hypothesis could be drug B is as good as drug A. Then, I collect the data, and I analyze the data, I come to a conclusion. So I say I fail to reject the null hypothesis. That means, drug B is as good as drug A. So if I do certain statistical calculation, I prove that null hypothesis has to be rejected, I accept the alternate hypothesis, then I say drug B is better than drug A.

So, in any statistical analysis we have a hypothesis mainly the null hypothesis, that is status quo or no change, no difference, then we have the alternate hypothesis, we say there is a difference. Then we collect the data, we analyze the data, we have to reject the null hypothesis in order to accept the alternate hypothesis. If we do not reject the null

hypothesis, then we can say we fail to reject the null hypothesis, so obviously we cannot accept the alternate. If we fail to reject the null hypothesis, we cannot accept the alternate hypothesis. So this is very, very important.

(Refer Slide Time: 31:46)

For example, if I am testing a new drug, null hypothesis could be drug A is as good as placebo status quo, no difference, no effect. Alternate could be drug is active; drug is better in reducing say diabetic or bacterial infection and so on. I do collect the data now, I do some test, and I am not able to reject the null hypothesis, then I have to accept the null hypothesis which is drug A is as good as my placebo. I do some tests, some analysis, then I say I reject the null hypothesis, so I accept the alternate hypothesis, then I can say drug A is more active.

Now if I am comparing two drugs, say drug A and B, the null hypothesis could be drug A is as good as B, there is no difference. Now, alternate hypothesis could be drug A is different from drug B; that is drug A it could be more better or worse or drug A is better than drug B, drug A is less than or less active than drug B. So, in this case we can have 3 alternate hypotheses. Drug A is different from drug B; that means, we are not telling whether drug A is better than drug B or drug A is less than drug B, but we are just saying drug A is different from drug B.

For example, we can say students in this class are different from the students in another class. We can say the plants grown in this place are different from the plants that are

grown in Kodaikanal; that is what we are saying, whereas so we use different. Whereas if you want to say the drug which I have discovered recently is better, then obviously I have to say drug A is better than drug B. That is the alternate hypothesis that comes into picture. Or drug A could be worse than drug B. So we have in the alternate k three different situations here, but the null hypothesis remains same. There is no difference between drug A and drug B, that means status quo both remains same. So you have to prove that there is no difference; that means, the null hypothesis is valid. You have to reject the null hypothesis, then only you can accept the alternate hypothesis. That is how things work in statistics, and we use these null hypothesis and alternate hypothesis and test to differentiate between two different data sets and so on. We are going to use them in very large situations as well as the cases actually.

We will continue in the next class further on the hypothesis testing, and type of errors, and how does one go about analyzing the data and so on.

Thank you very much.