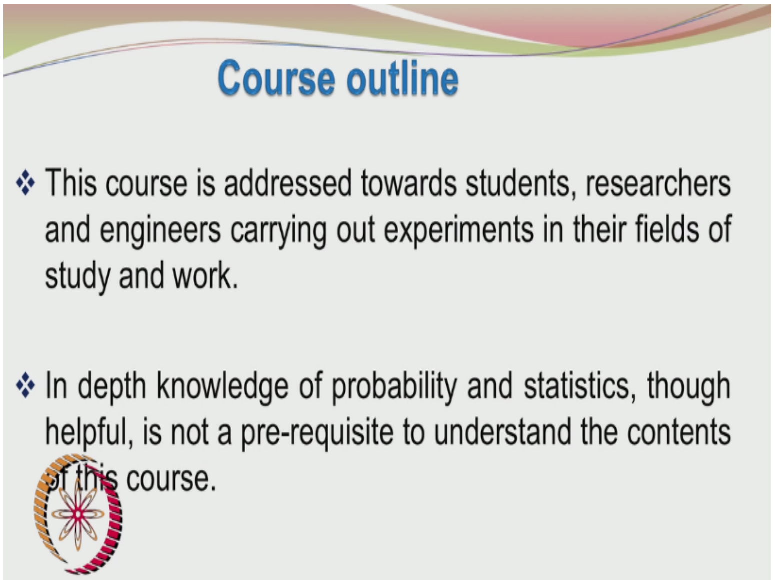
# Statistics for Experimentalists
**Prof. Kannan. A**
**Department of Chemical Engineering**
**Indian Institute of Technology – Madras**

## Lecture – 01
## Introduction

Welcome to the course on statistics for experimentalists. I am Dr. A. Kannan from the Department of Chemical Engineering, Indian Institute of Technology, Madras. This will be an introduction lecture. I will be giving a brief overview of the course.

**(Refer Slide Time: 00:43)**



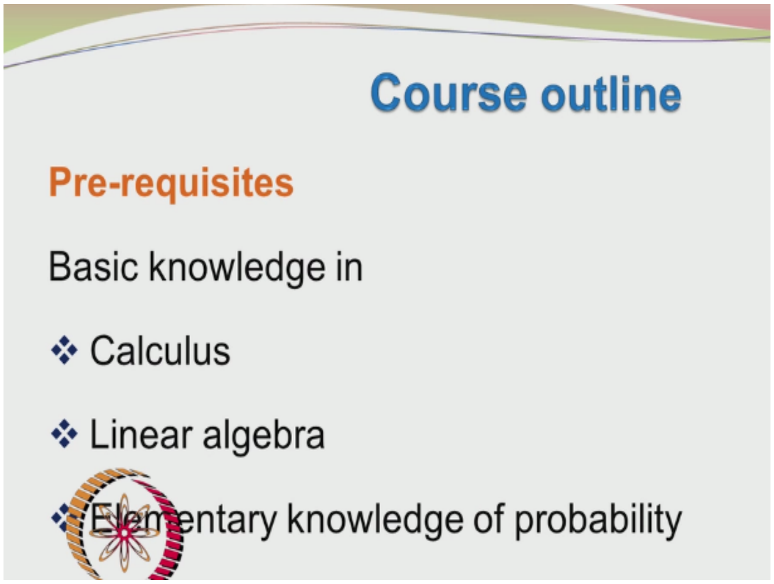
Course outline

- This course is addressed towards students, researchers and engineers carrying out experiments in their fields of study and work.
- In depth knowledge of probability and statistics, though helpful, is not a pre-requisite to understand the contents of this course.

This course is mainly addressed towards students, researchers, and engineers carrying out experiments in their fields of study and work. Many of you doing experimental work want to analyze the data in a scientific regress manner; however, you may be thinking that it requires a very strong foundation in statistics and probability. A lot of theoretical knowledge is required that may be your concern.

I would like to allay your concerns on that. All you require is an elementary knowledge of probability and statistics. For example, you should know that the probabilities can be only positive and the sum of the probabilities will be = 1. Even those people who have a very strong foundation in statistics and probability, will consider the course contents useful to them, because they can relate their knowledge to an application. In our case, the application is analysis of experimental data, design of experiments, fitting of empirical model to the data, and so on.

**(Refer Slide Time: 02:18)**



## Course outline

Pre-requisites

Basic knowledge in

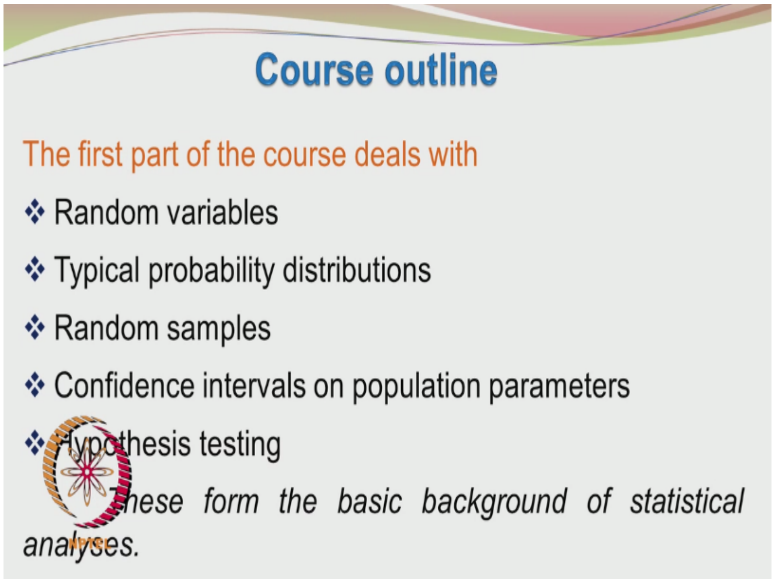
❖ Calculus

❖ Linear algebra

❖ Elementary knowledge of probability

What are pre-requisites for the course? You should have a basic knowledge in calculus. You should have an idea on differentiation, integration. You should also have knowledge in linear algebra, matrices, multiplication of matrices, inversion of matrices, and an elementary knowledge of probability. The course is divided into 2 parts. The first part is the fundamental part, where you would be exposed to the essential tools required for data analysis and design of experiments, so will be starting with an introduction to random variables.

**(Refer Slide Time: 03:11)**



## Course outline

The first part of the course deals with

❖ Random variables

❖ Typical probability distributions

❖ Random samples

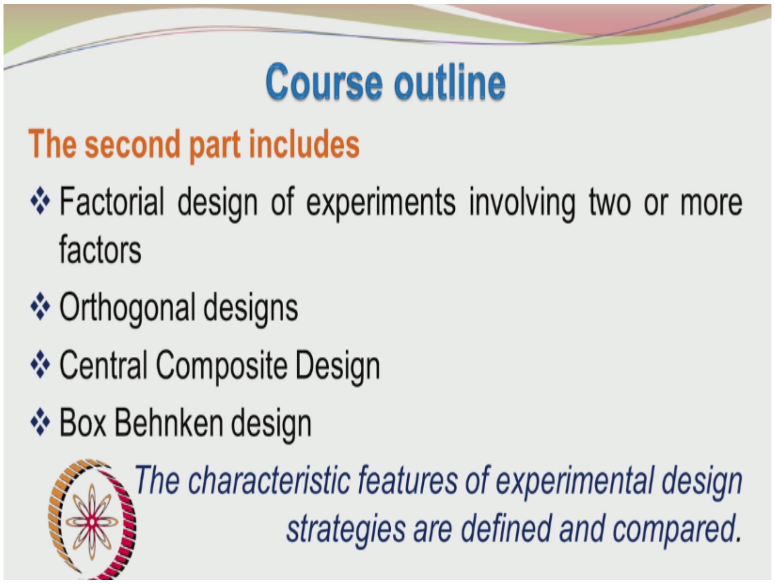
❖ Confidence intervals on population parameters

❖ Hypothesis testing

*These form the basic background of statistical analyses.*

Then we will be looking at typical probability distributions. Many of you may be knowing what is meant by a normal distribution or a Gaussian distribution. You might have heard about the T distribution, Chi square distribution, and so on, so will be looking at these popular distributions. They also have important applications and then we will be looking at random samples and followed by confidence intervals on population parameters. Then we will be looking at the hypothesis testing.

**(Refer Slide Time: 03:50)**



## Course outline

The second part includes

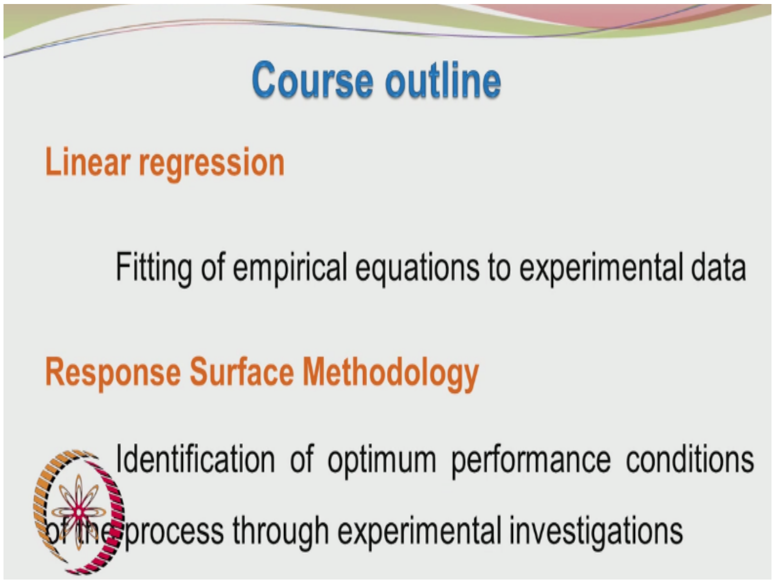
- ❖ Factorial design of experiments involving two or more factors
- ❖ Orthogonal designs
- ❖ Central Composite Design
- ❖ Box Behnken design

*The characteristic features of experimental design strategies are defined and compared.*

The second part of the course deals with these design of experiments. The most popular among them is the factorial design of experiments which may involve 2 or more factors. This would be followed by an introduction to orthogonal designs. Orthogonal designs are very convenient and they make our analysis very easy so I will be giving an introduction to the orthogonal designs.

Then we will be looking at some of the higher order designs like the central composite design, Box Behnken design. You might have seen some of these designs being used in several research papers. How to choose a particular experimental design? What are the important characteristics and features of different designs? Which is the most suitable design for you specific experimental work or the issues will be discussing here?

**(Refer Slide Time: 04:52)**



Course outline

Linear regression

Fitting of empirical equations to experimental data

Response Surface Methodology

Identification of optimum performance conditions of the process through experimental investigations

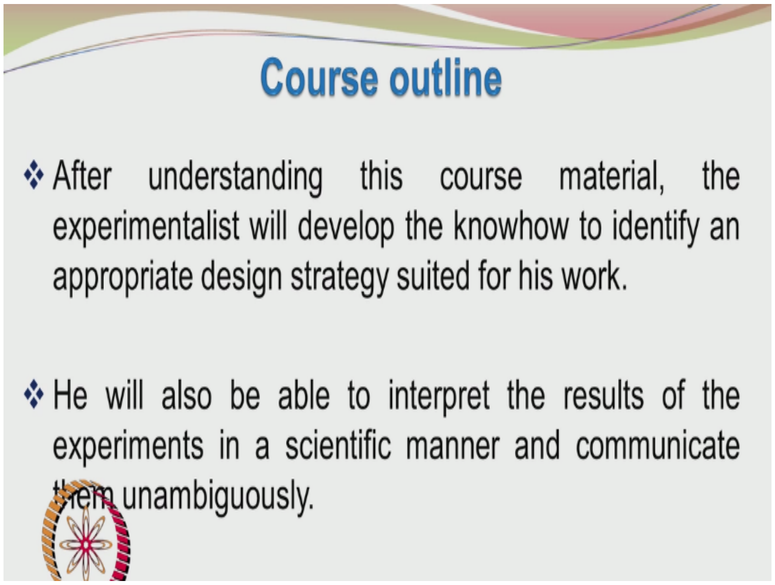
So continuing with our course outline, the next topic would be linear regression. Even without background in this course you might have fitted experimental data using spread sheet. You might have used the trend line in the spread sheet and fitted a linear model or a power law model and would have been very happy if you had got a R squared value of 0.99 or above. In this course, we will be looking at fitting of empirical linear equations to the experimental data.

We will also be looking not only at R squared, but also at other statistical parameters that are obtained from the data fitting exercise will be understanding their significance and will also be looking at which of those parameters are more useful than the remaining. The next topic would be response surface methodology. Suppose you have done some experiments okay involving several variables. You have got some results.

Now the management or the research supervisor tells you I want to increase the yield of the experiment. I want to improve the conversion or I want to get higher amount of products. So you want to know at what experimental conditions you should conduct the runs so that you get the desired output. Suppose you are having large number of variables and the range of these variables are also very high, then you are really at a loss as to identify the conditions where the experiments are to be performed.

Many of the real life experimental work may not have a regress mathematical model. You cannot then find the true set of optimum conditions unless you do the experiments. The response surface methodology is a very valuable tool to identify the optimum performance conditions of the process through experimental investigations.

**(Refer Slide Time: 07:30)**

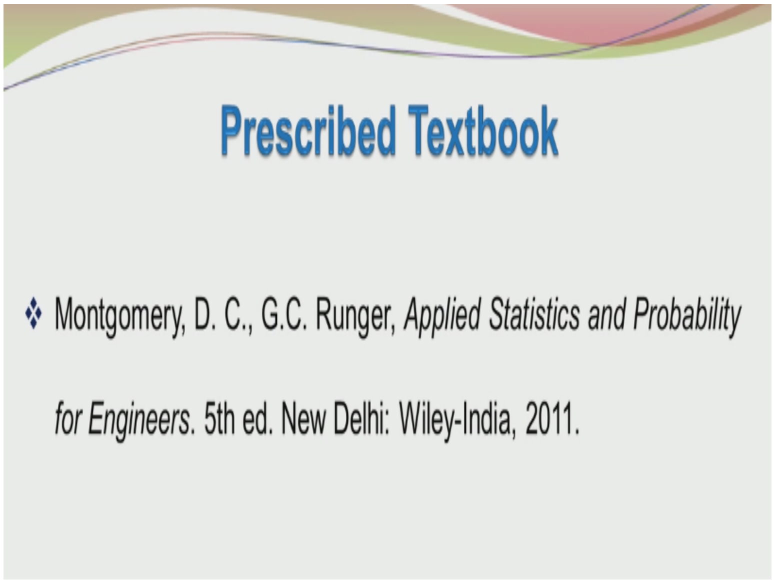


Course outline

- After understanding this course material, the experimentalist will develop the knowhow to identify an appropriate design strategy suited for his work.
- He will also be able to interpret the results of the experiments in a scientific manner and communicate them unambiguously.

So after you understand the course material, you will be able to first select an appropriate design strategy suited for your work and after using this design strategy you will get the results of your experiments. You will be able to analyze the experimental data in a scientific manner and communicate it to the outside world to research meetings, journal papers, Ph.D. thesis in an unambiguous manner.

**(Refer Slide Time: 08:22)**

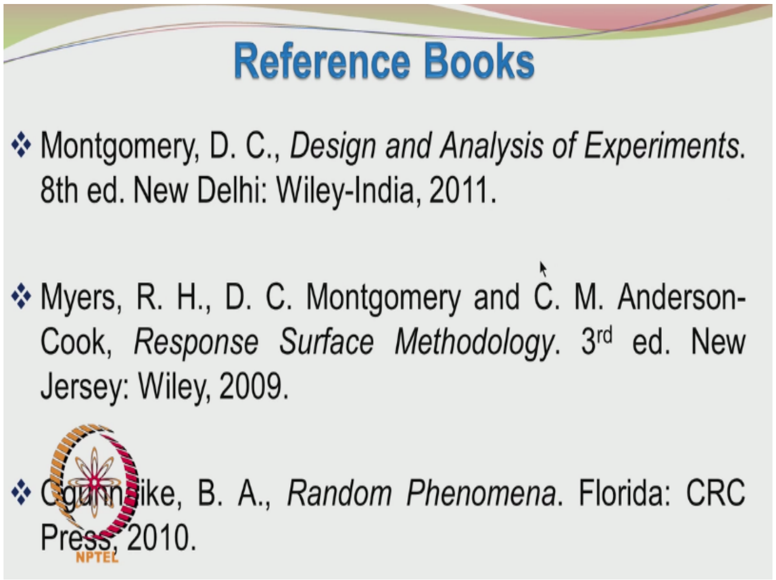


## Prescribed Textbook

- ❖ Montgomery, D. C., G.C. Runger, *Applied Statistics and Probability for Engineers*. 5th ed. New Delhi: Wiley-India, 2011.

There are several good text books available on these topics design of experiments, applied statistics, and probability. What we will be following is the book written by Montgomery and Runger. The title of the book is Applied Statistics and Probability for Engineers. The fifth edition is the latest one and the Indian edition is also available which was published in the year 2011.

**(Refer Slide Time: 09:04)**



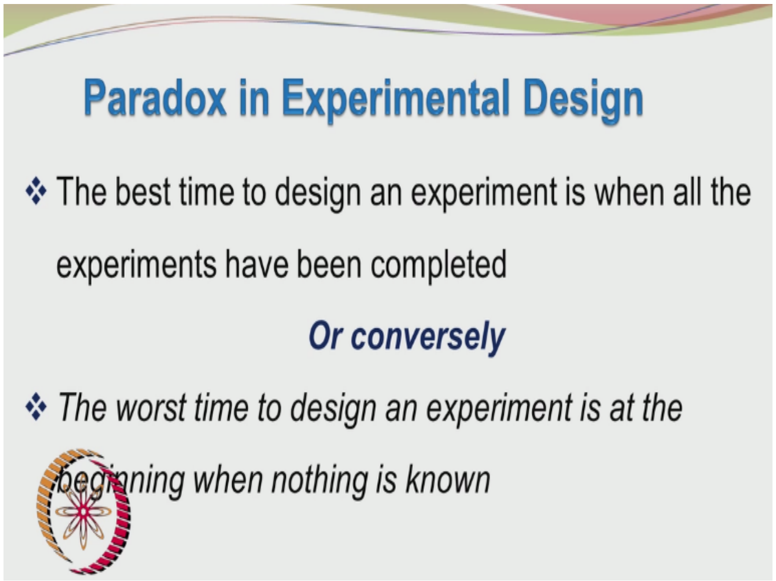
## Reference Books

- ❖ Montgomery, D. C., *Design and Analysis of Experiments*. 8th ed. New Delhi: Wiley-India, 2011.
- ❖ Myers, R. H., D. C. Montgomery and C. M. Anderson-Cook, *Response Surface Methodology*. 3rd ed. New Jersey: Wiley, 2009.
- ❖ Ogunnaike, B. A., *Random Phenomena*. Florida: CRC Press, 2010.

We will also be looking at other reference books, the book written by Montgomery on Design and Analysis of Experiments is a slightly more advanced book and it will be dealing specifically with the design and analysis of experiments, different designs, and how to analyze them. This is an advanced book on response surface methodology. Where it compares different designs, their pros and cons, which one to use so really very interesting book.

There is also a book written by Ogunnaike. The title of the book is Random Phenomena. This is also an excellent book where it gives in depth understanding of the various concepts you will be coming across in statistical analysis.

**(Refer Slide Time: 10:01)**



## Paradox in Experimental Design

- The best time to design an experiment is when all the experiments have been completed

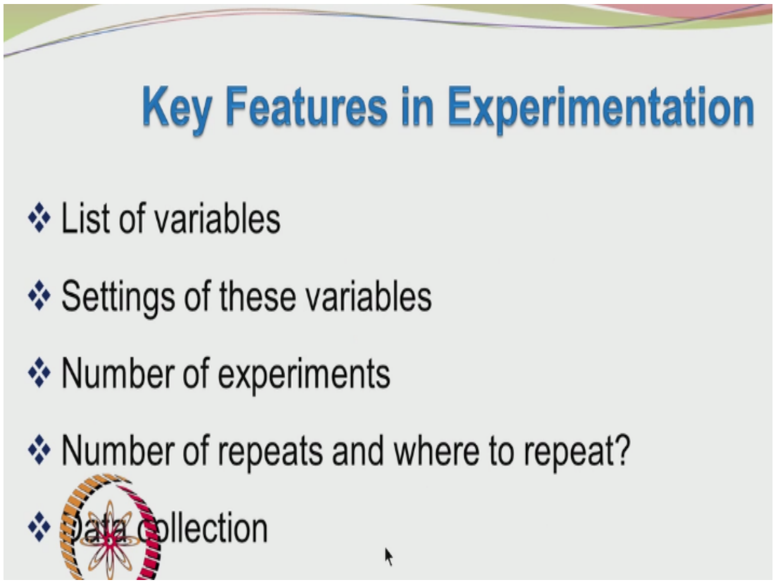
***Or conversely***

- *The worst time to design an experiment is at the beginning when nothing is known*

Now we will come to overview of what we are going to do. Well, when you are starting to do the experimental work, you are completely in the dark and that is the worst time to design the experiment, where nothing is known and after having completed all the experiments, you will be feeling oh if I had known this I would designed the experiment in a slightly different way. So the best time to do the experiment or design the experiment is, when all the experiments have been completed.

Let us look at some issues which we need to understand before we start designing the experiments and analyzing the experimental data.

**(Refer Slide Time: 10:52)**



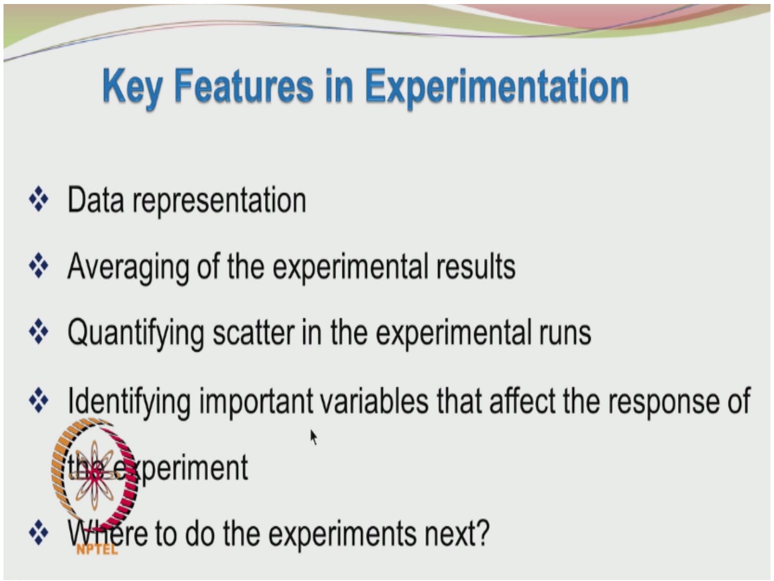
## Key Features in Experimentation

- List of variables
- Settings of these variables
- Number of experiments
- Number of repeats and where to repeat?
- Data collection

You have to make a lot of decisions when you embark on an experimental program. You have to first see what are the lists of variables and what are the values taken by these variables, how many experiments you should perform. It is not an endless or a limitless number of experiments you can perform. May be you can do a lot of experiments in the lab in a university.

But when you go to the industry, the manpower, the resources, time are all very valuable and you have to be very careful when using these resources. So the number of experiment is very important and whether your experiments are reproducible is another important question. You would out of curiosity repeat some of the experiments. You may want to even repeat all the experiments. So you have to decide on number repeats, and also where to repeat the experiments. Then comes the data collection, analysis, modeling interpretation, and so on.

**(Refer Slide Time: 12:13)**



## Key Features in Experimentation

- Data representation
- Averaging of the experimental results
- Quantifying scatter in the experimental runs
- Identifying important variables that affect the response of the experiment
- Where to do the experiments next?

So once you have got the experimental data, you have to represent the data in a concise manner. Many times when you repeat the experiments, you will be averaging the experimental results. Suppose you want to report the yield of an experiment or the percentage conversion you have done 5 repeats. You will average all these 5 values and then say by average yield is so much or my average percentage conversion is let us say 70%.

Very importantly the very fact that you have to average the response or the n result of you experiment implies that there was a scatter in your data. Otherwise, you would have reported the unique value you had obtained. You would have said my yield is 72.5%, but you have repeated the experiment 5 times and then you got some times 72, 74, 65, 67. So when you average those values, you report the mean value. So the scatter is responsible for you to do the averaging.

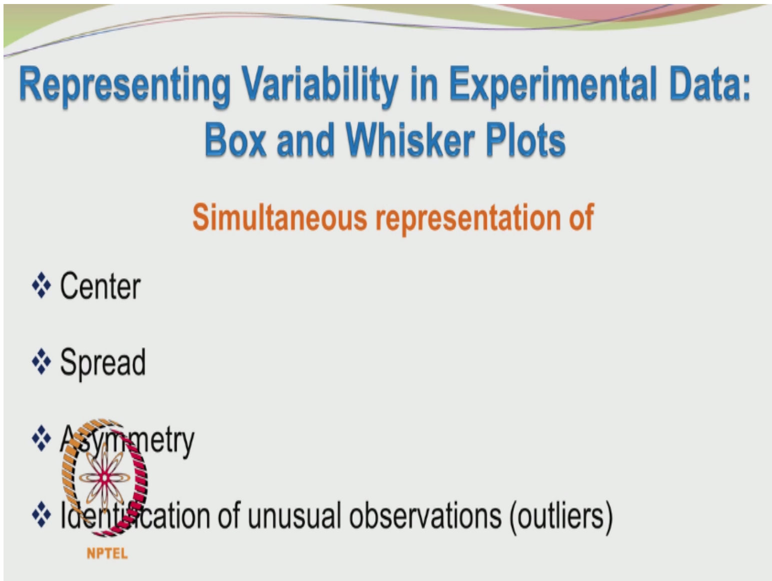
So it makes a lot of sense to quantify the scatter in the experimental runs. Experimental runs are always associated with scatter and it is very important for us to quantify the scatter and after having done that we will be able to identify which are the important variables that influence the response of the experiment. After having finished all these we can now investigate where to perform the next set of experiments.

**(Refer Slide Time: 14:18)**

DATA REPRESENTATION

So let us now look at data representation.

**(Refer Slide Time: 14:23)**



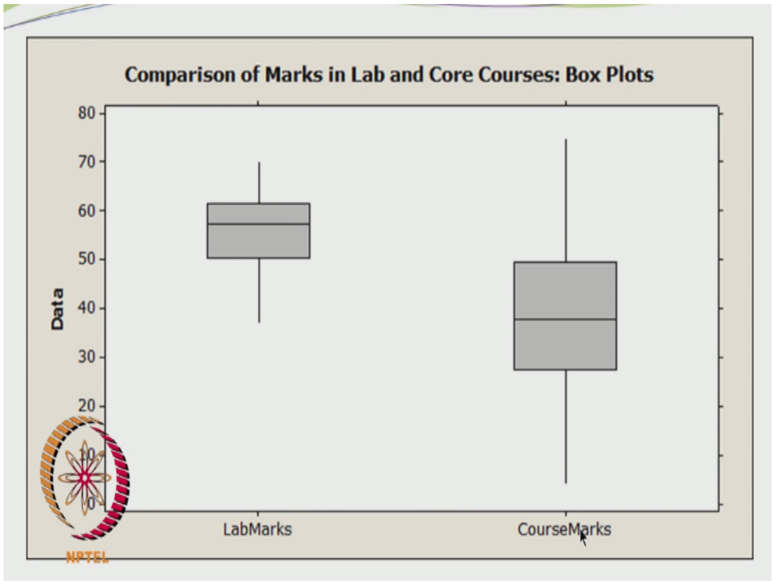
## Representing Variability in Experimental Data: Box and Whisker Plots

Simultaneous representation of

- ❖ Center
- ❖ Spread
- ❖ Asymmetry
- ❖ Identification of unusual observations (outliers)

One popular technique is the box and Whisker plot. In this you can simultaneously illustrate the center of the data, the spread of the data, the asymmetry in the data, and if there are any unusual observations which are termed as outliers.

**(Refer Slide Time: 14:52)**



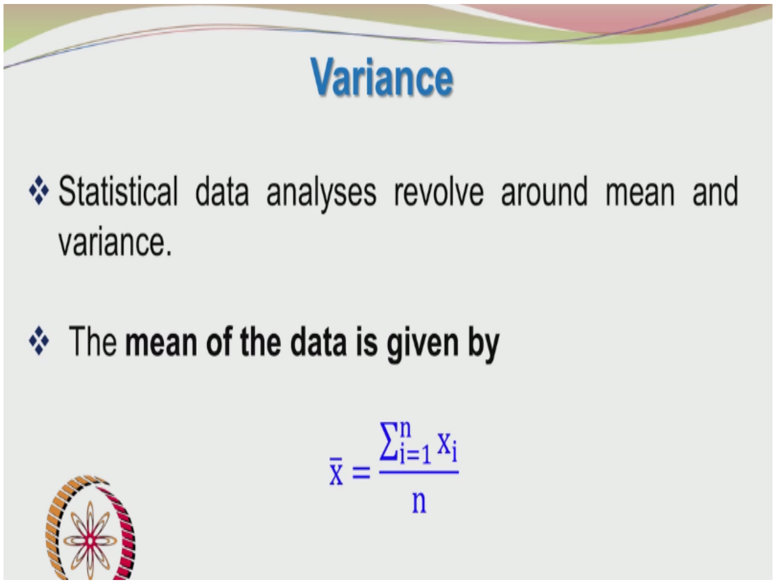
So what I have presented here is the representation of 2 subjects, one is the laboratory and another is the core course and I have given here the marks distribution in the laboratory exam and the marks distribution in the course. You can see that the laboratory marks are having a less spread. The core course subject marks are having a large spread. Lab work is usually a combined work at least for the reports part you do the experiments as a group of 2 or 3 students and that is also a part of the evaluation.

Since there is a combined activity involved, the marks are not as wide spread as the marks obtained in the core course and also the median mark is higher than the median mark for the core course and you can see that the maximum mark is only about 70 and whereas the maximum mark in the core courses about close to 80 and since it is a group activity, the minimum mark is only about 40 whereas for individual performance unfortunately for this particular exam, it has even gone to about 8 or so.

So in 1 graph we have been able to pack a lot of information. The first line in the box represents the first quartile. The second line in the box represents the median or the second quartile and the third line represents the third quartile. Each quartile represents the appropriate percentage number of students who have got marks <= the value. For example, the first quartile is about 52 is about 25% of the class have got marks <= 52, 50% of the students have got marks <= to about 60 and 75% of the students have got marks < about 62.

So you can see that there is not much variation between the second quartile and the third quartile in terms of the marks whereas between the second quartile and the third quartile there is a larger gap. So lot of inferences can be made from a single diagram. So you have to present large amounts of data in a concise or compact manner.

**(Refer Slide Time: 18:16)**



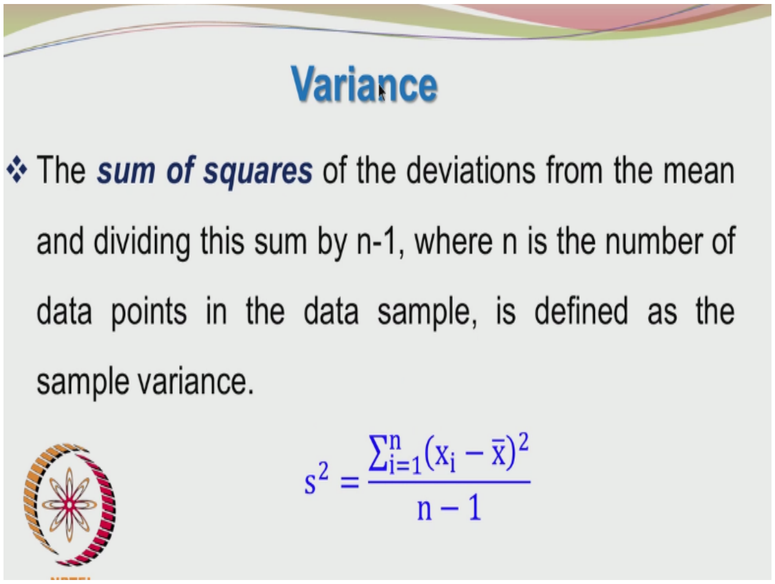
## Variance

- Statistical data analyses revolve around mean and variance.
- The **mean of the data is given by**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

So we will be now looking at the mean and variance of the data. The mean of the data is given by x bar = sigma I = 1 to xi/n. The sigma is the summation and we sum over all the values x1 + x2 + so on to xn and then divide by the total number of observations that will give us the mean. We use this mean to calculate the variance.

**(Refer Slide Time: 18:52)**



## Variance

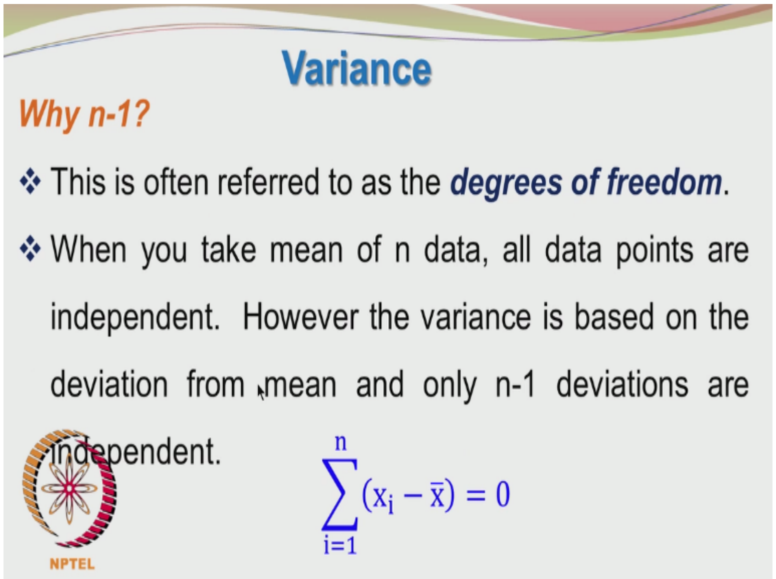
- The ***sum of squares*** of the deviations from the mean and dividing this sum by n-1, where n is the number of data points in the data sample, is defined as the sample variance.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

The variance is a very important quantity in statistical calculations. This is the heart of the experimental data and we use variance to analyze our data and make the appropriate conclusions. So let us now define variance formally. We have already seen how to calculate the sample mean x bar. Now we have to calculate the sample variance.

A sample is a collection of n numbers for example or n data points and so what we do is we take the deviation of the individual data point from the mean and then square it and then we add all such square of the deviations from the mean and then divide by n - 1, where n is the number of data points. This gives us the sample variance.

**(Refer Slide Time: 20:01)**



## Variance

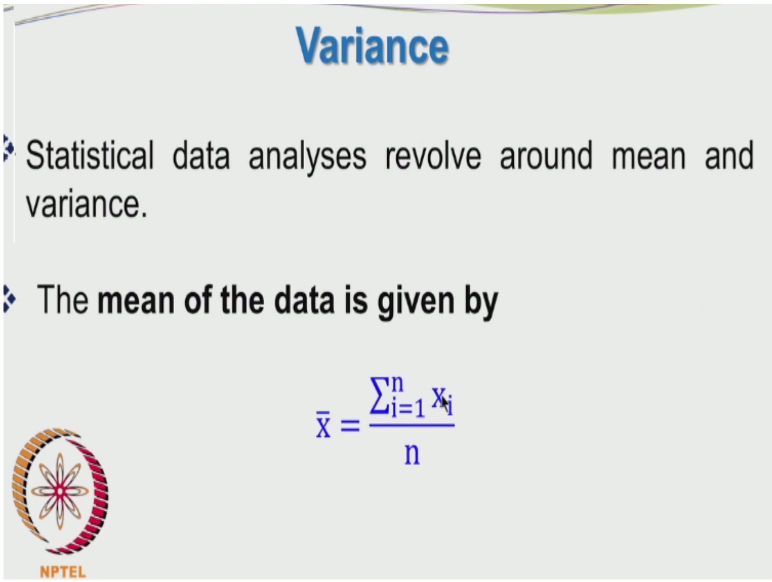
*Why n-1?*

- This is often referred to as the ***degrees of freedom***.
- When you take mean of n data, all data points are independent. However the variance is based on the deviation from mean and only n-1 deviations are independent.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

So in the case of mean, we have used n. In the case of variance, we are using n - 1. This leads us to another important concept which is called as the degrees of freedom very often will be encountering the degrees of freedom in the variance or sum of squares calculations. The degree of freedom refers to the number of independent entities.

**(Refer Slide Time: 20:42)**



## Variance

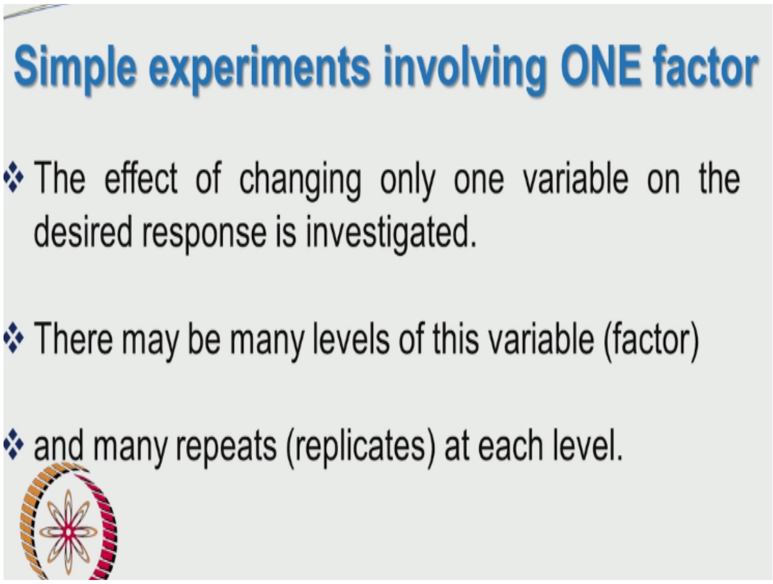
- Statistical data analyses revolve around mean and variance.
- The **mean of the data is given by**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

When you are calculating the mean, all the xi values were independent of each other. So you had n independent entities and so the n xi values for independent and so you put n in the denominator. Whereas in the case of variance not all the xi - x bar deviations are independent, they are actually dependent, because they are subject to the constraint sigma I equals 1 to n xi - x bar = 0.

This means that the sum of the deviations = 0. The positive contributions will cancel out with the negative contributions with respect to the mean. This is arising from the definition of the mean. The mean is somewhere at the center of the distribution and so some of the deviations about the mean will be = 0. So when this constraint is satisfied there will be only n - 1 independent deviation quantities. Hence we have n - 1 degree of freedom. So whenever we are calculating the sum of squares in our variance calculations we also look at the degrees of freedom.

**(Refer Slide Time: 22:21)**



## Simple experiments involving ONE factor

- The effect of changing only one variable on the desired response is investigated.
- There may be many levels of this variable (factor)
- and many repeats (replicates) at each level.

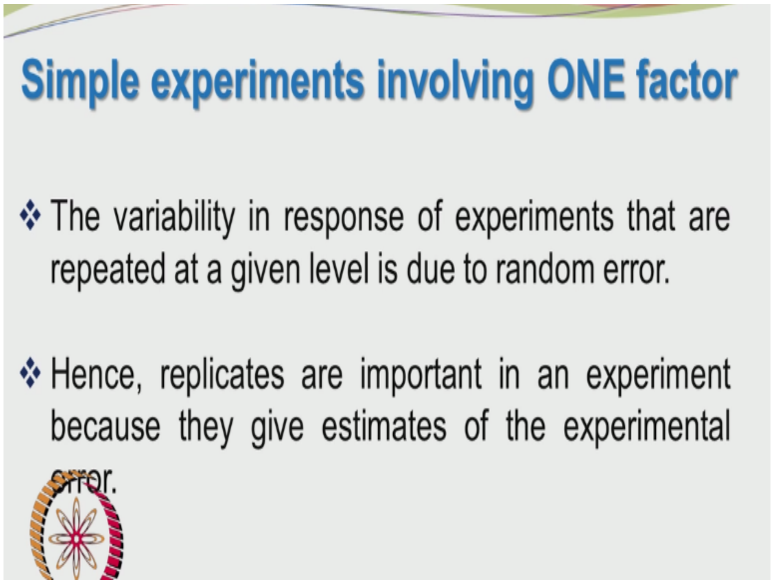
I will be now introducing briefly the concept of analysis of variance. We will take a simple case where our experiments are controlled or influenced by only 1 factor. This is a very simple case, but this gives us a lot of useful information which we will use in our further analysis. So we are going to change only one variable. Even though we have only 1 variable, we may have several levels of this variable. This variable is also termed as a factor.

So at each level or at each setting of this factor, you can conduct many repeats. The repeats are also termed as replicates. To summarize or repeat what I have said, we are now doing experiments which are influenced by only 1 variable, but that variable can take several values, they can take several settings or levels. So we can do those experiments, but we have investigated the experiment at only 1 value of the variable, at only 1 level.

What we should do is, carry out repeats at each level, at each setting, we may decide to do 2 repeats or 3 repeats so that we want to see whether the experiment is reproducible, whether we get more or less the same response, when we repeat at different settings. Suppose my experiment is involving the variation with temperature. So temperature is 1 variable and it can have different settings like 30 degrees, 50 degrees, 70 degrees, and 100 degrees.

So I am having 4 levels of temperature, 30, 50, 70, and 100 and at 30 degrees I may repeat the experiment 3 times. So the number of repeats at each level will be = 3. So just keep this in mind.

**(Refer Slide Time: 25:01)**



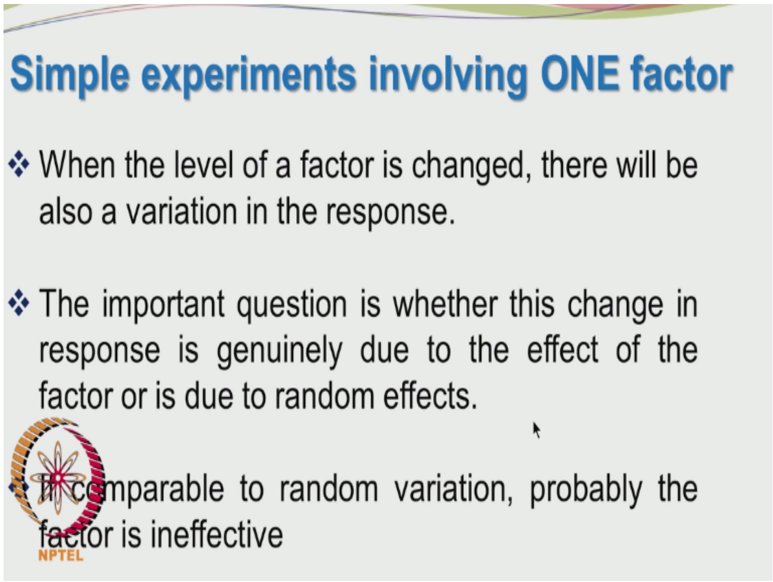
### Simple experiments involving ONE factor

- The variability in response of experiments that are repeated at a given level is due to random error.
- Hence, replicates are important in an experiment because they give estimates of the experimental error.

So whenever we repeat the experiment, we unfortunately do not get the same response. This is a cause for worry, annoyance, and when you do not get the same response you may you want to again repeat the experiment and then it becomes a never ending task, because not only you have to repeat the experiment at the particular setting, but then you have to go the next setting or level and again do repeats there and we do not know what is going to happen there.

So it is very important for us to analyze the variability in the experimental process. The reason why our experiments do not give identical values even though you have taken care to keep the level or setting at the same value is because of random effects. There may be uncontrollable factors in the environment which may be randomly influencing the outcome of your experiment. So we will do repeat to get an idea how much is the random variability in our experimental response.

**(Refer Slide Time: 26:27)**



## Simple experiments involving ONE factor

- When the level of a factor is changed, there will be also a variation in the response.
- The important question is whether this change in response is genuinely due to the effect of the factor or is due to random effects.
- If comparable to random variation, probably the factor is ineffective

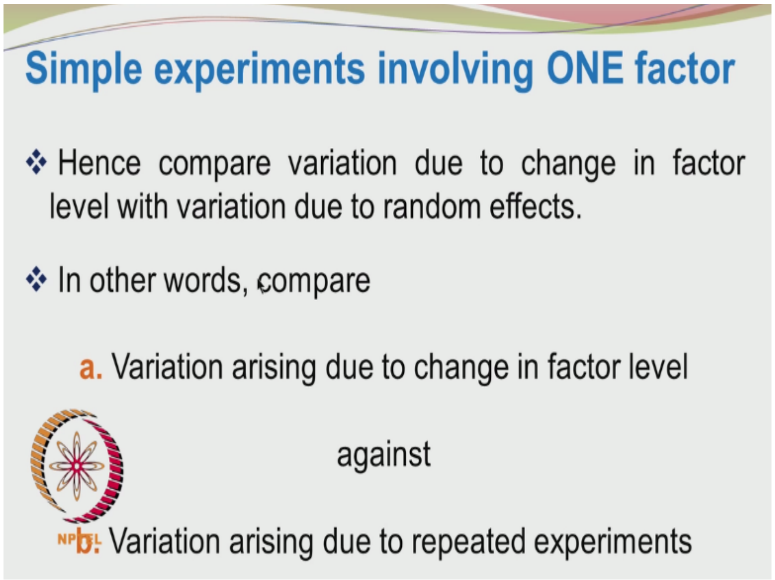
We are going to change the level of a particular variable or a factor and when we do that we notice a variation in the response. If you recollect the information presented in the previous slide, I told that even at a given setting or level the experimental response may be different and this was because of random error or random fluctuations. You have noted those fluctuations and you have seen how much is the difference in your experimental values?

Now you are going to the next setting and you perform the experiment. Again you are going to get some response and when you do the repeats you are also going to get different responses. Now the important question we have to face is whether the change in the response is because of changing the level of the factor from 1 value to another value or it was because of the random fluctuations.

We have to compare the change in response when the variable level was changed with the variation due to experimental error or variation due to random effects. If the variation when changing the level of the factor is comparable to the variation due to random error, then changing the level of the variable actually did not cause a significant change in the response. So you have to compare the change in your response with the change in the variable setting and also with the error effect or the random error effect.

So if the response change is comparable to the random variation, then the factor is ineffective.

**(Refer Slide Time: 29:06)**



## Simple experiments involving ONE factor

- Hence compare variation due to change in factor level with variation due to random effects.
- In other words, compare

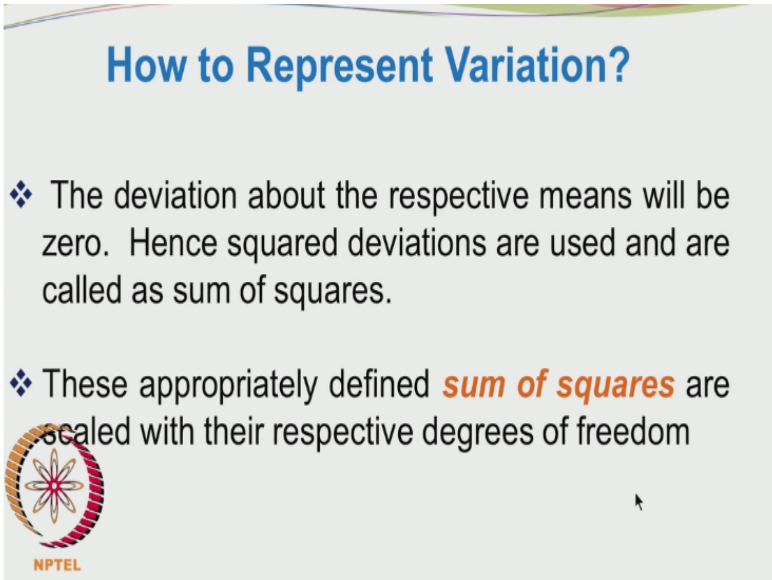
a. Variation arising due to change in factor level

against

b. Variation arising due to repeated experiments

So what we have to do is, compare the variation due to change in factor level with variation due to random effects. Compare variation arising due to change in factor level against variation arising due to repeated experiments.

**(Refer Slide Time: 29:30)**

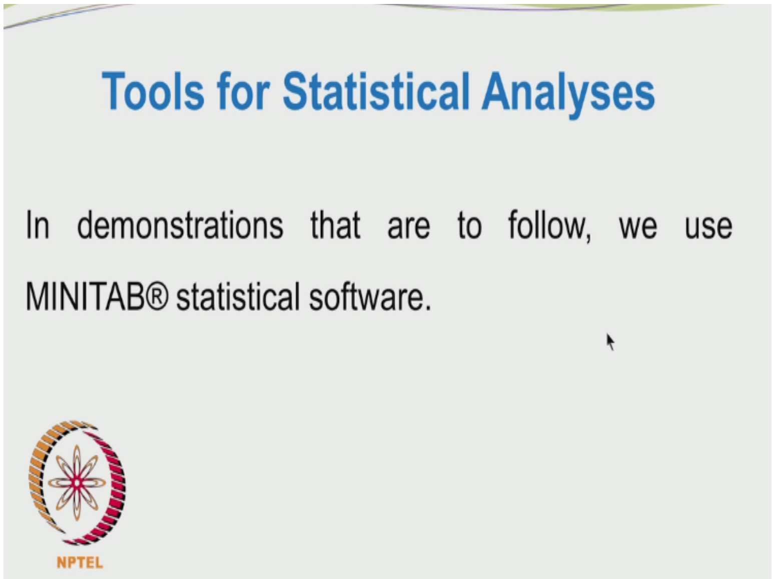


## How to Represent Variation?

- The deviation about the respective means will be zero. Hence squared deviations are used and are called as sum of squares.
- These appropriately defined ***sum of squares*** are scaled with their respective degrees of freedom

So as I said earlier we will be looking at the square of the deviations from the means and the square of the deviations from the means are summed and we call it as the sum of squares.

**(Refer Slide Time: 29:51)**

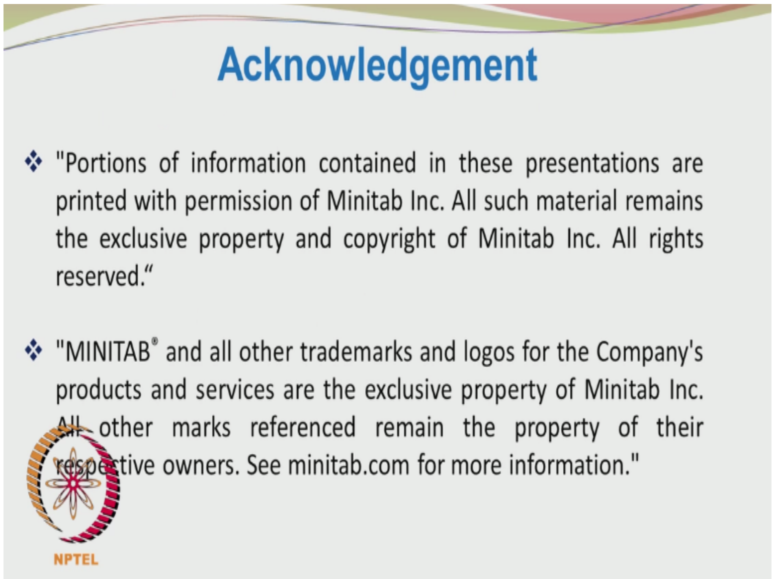


Tools for Statistical Analyses

In demonstrations that are to follow, we use MINITAB® statistical software.

So I will be now demonstrating a simple example to illustrate these points and for that purpose I have used the MINITAB statistical software.

**(Refer Slide Time: 30:07)**



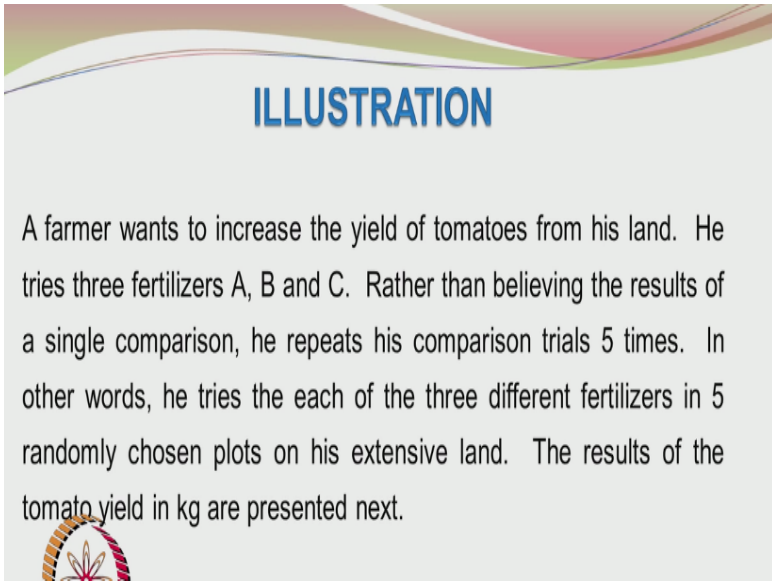
Acknowledgement

- "Portions of information contained in these presentations are printed with permission of Minitab Inc. All such material remains the exclusive property and copyright of Minitab Inc. All rights reserved.“
- "MINITAB® and all other trademarks and logos for the Company's products and services are the exclusive property of Minitab Inc. All other marks referenced remain the property of their respective owners. See minitab.com for more information."

Here I would like to acknowledge the use of Minitab software. The portions of information contained in these presentations are printed with the permission of Minitab incorporated. All such material remains exclusive property and copyright of Minitab incorporated, all rights are reserved and Minitab and all other trademarks and logos for the companies, products, and services of the exclusive property of Minitab incorporated.

All other marks reference remained the property of the respect to owners. If you want further details, please see the Minitab.com website.

**(Refer Slide Time: 30:49)**

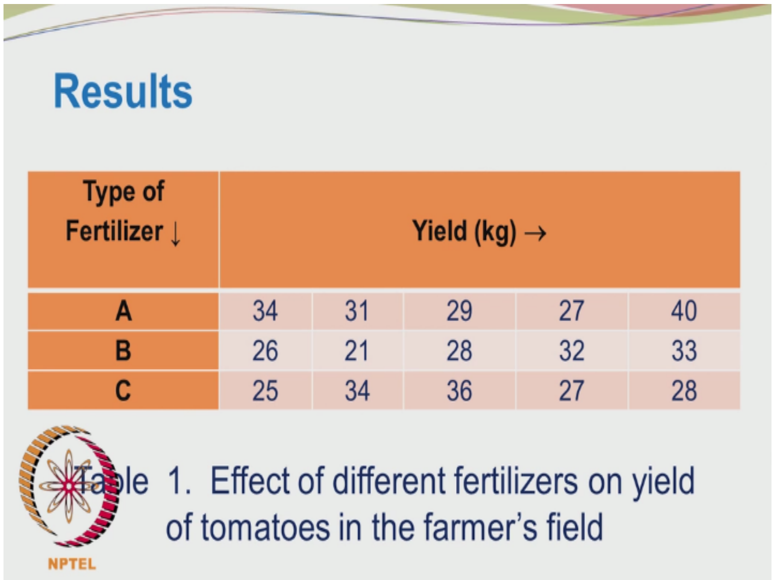


## ILLUSTRATION

A farmer wants to increase the yield of tomatoes from his land. He tries three fertilizers A, B and C. Rather than believing the results of a single comparison, he repeats his comparison trials 5 times. In other words, he tries the each of the three different fertilizers in 5 randomly chosen plots on his extensive land. The results of the tomato yield in kg are presented next.

We look at an interesting example. Many of the statistically concepts came from the field of agriculture. So we will also take an example from the agricultural area. Let us say that the farmer wants to increase the yield of tomatoes from his land and he tries 3 fertilizers A, B, and C and instead of doing just only once he repeats his comparison trials 5 times. What does it mean? So he takes 5 randomly chosen plots of land. He has 5 different fields and in each field he tries the 3 fertilizers.

**(Refer Slide Time: 31:53)**



## Results

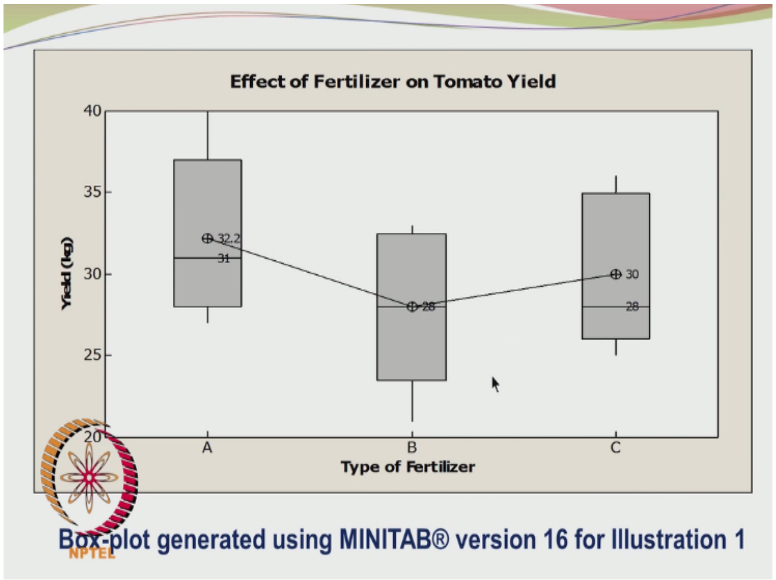
| Type of Fertilizer ↓ | Yield (kg) → | | | | |
|---|---|---|---|---|---|
| A | 34 | 31 | 29 | 27 | 40 |
| B | 26 | 21 | 28 | 32 | 33 |
| C | 25 | 34 | 36 | 27 | 28 |

Table 1. Effect of different fertilizers on yield of tomatoes in the farmer's field

So the results are shown here. So you can say that these are the results from the first field. The second column represents the results from the second field. The third column from the third field and so on and this is the effect of fertilizer A, fertilizer B, and fertilizer C in the different fields. Looking at the data we cannot infer much. In field 1, it looks like fertilizer A is the best. Now if you go to field 4, fertilizer B is the best.

If you look at field 3, fertilizer C is the best. So which is the better fertilizer or is there any difference between the fertilizers? This is the question we want to address.

**(Refer Slide Time: 32:53)**

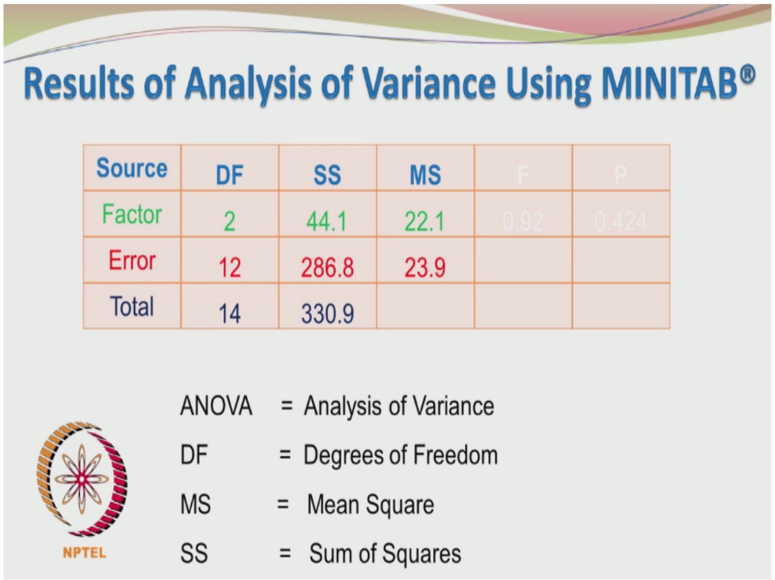


Box-plot generated using MINITAB® version 16 for Illustration 1

So we draw the box plot and we find the box plots are definitely showing some difference, but again we cannot strictly say whether fertilizer A is better than fertilizer B and whether it is better than fertilizer C because all of them have considerable scatter. The circular values in the box plot represent the average values and the second line as I said earlier represents the median.

In this case, the mean and median values are coinciding. So just by looking at the pictorial depiction we cannot convincingly conclude that the fertilizers are different in terms of influencing the yield of tomato from the land so we have to do a proper statistical analysis to check whether the fertilizers are different in influencing the yield or they are pretty much more or less the same.

**(Refer Slide Time: 34:23)**



## Results of Analysis of Variance Using MINITAB®

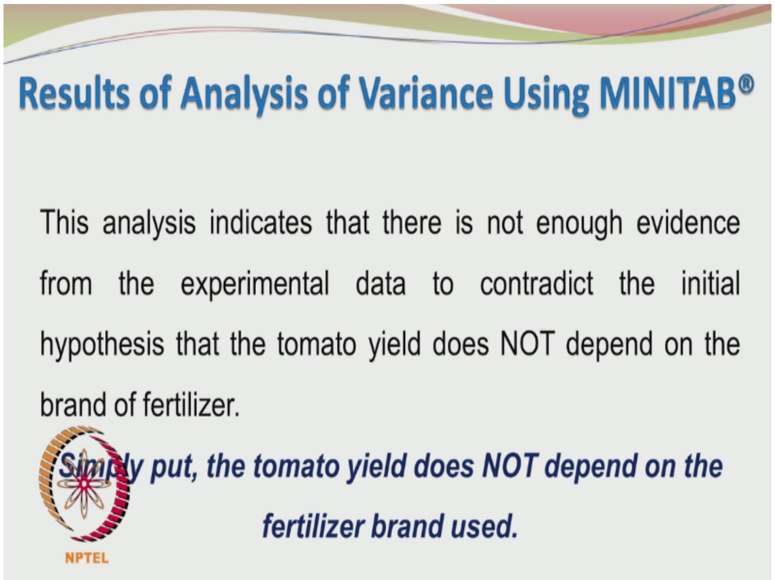
| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Factor | 2 | 44.1 | 22.1 | 0.92 | 0.424 |
| Error | 12 | 286.8 | 23.9 | | |
| Total | 14 | 330.9 | | | |

ANOVA = Analysis of Variance
DF = Degrees of Freedom
MS = Mean Square
SS = Sum of Squares

So we do analysis of variance and we have the degrees of freedom noted here. We had the 3 levels of fertilizers A, B, and C and there are 2 degrees of freedom. The degrees of freedom associated with the error is 12, and this is the sum of squares associated with the factor, or treatment and this is the error sum of squares. So we divide the sum of squares by the degrees of freedom and we get the mean square values.

You remember in the definition of variance we subtracted the average values from the data and then squared them and then divided by n - 1. Here also we are doing a similar kind of exercise. We get sum of squares and then we divide it by the degrees of freedom to get the mean square values. I am not going to all the details here just to illustrate that the variance is very important quantity and our decisions are based on the analysis of variance.

**(Refer Slide Time: 35:52)**



## Results of Analysis of Variance Using MINITAB®

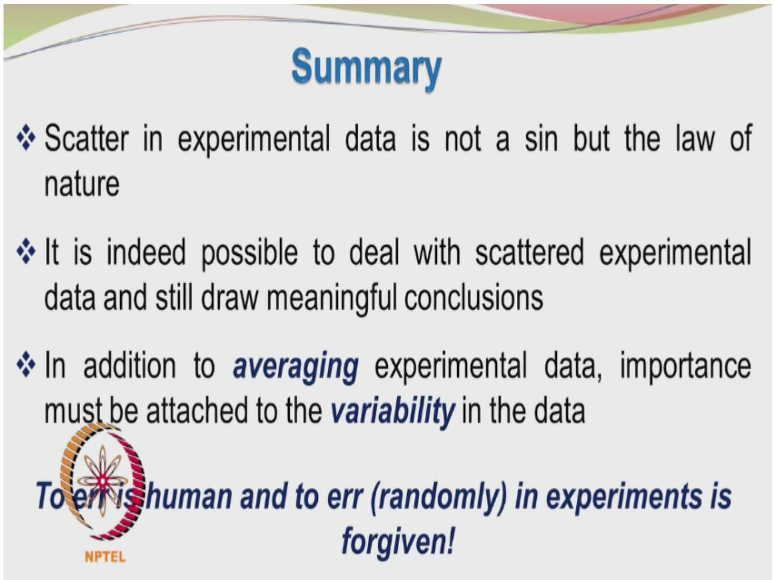
This analysis indicates that there is not enough evidence from the experimental data to contradict the initial hypothesis that the tomato yield does NOT depend on the brand of fertilizer.

***Simply put, the tomato yield does NOT depend on the fertilizer brand used.***

So after carrying out the exercise, this particular analysis indicated that there is not enough evidence from the experimental data to contradict the initial hypothesis that the tomato yield does not depend on the brand of fertilizer. So in simpler terms, the tomato yield does not depend on the brand of fertilizer used. So we have postulated the hypothesis and then we have carried out an analysis of variance and then we have drawn a suitable conclusion.

The details are not given at the present interactive stage, but we will be essentially doing the analysis of variance and then drawing suitable conclusions.

**(Refer Slide Time: 36:46)**



## Summary

- Scatter in experimental data is not a sin but the law of nature
- It is indeed possible to deal with scattered experimental data and still draw meaningful conclusions
- In addition to ***averaging*** experimental data, importance must be attached to the ***variability*** in the data

***To err is human and to err (randomly) in experiments is forgiven!***

To summarize, the scatter in experimental data is not a sin, but is the law of nature and you do not have to be ashamed that there is scatter in your experimental data. It is indeed possible to deal with scattered experimental data and still draw meaningful conclusions. In addition to averaging of the experimental data, importance must be attached to the variability is the main motivation and the reason for carrying out the statistical analysis.

Finally, to err is human and to err randomly in experiments is forgiven. So what we will do next is go to the first topic in our syllabus which would be random variables.