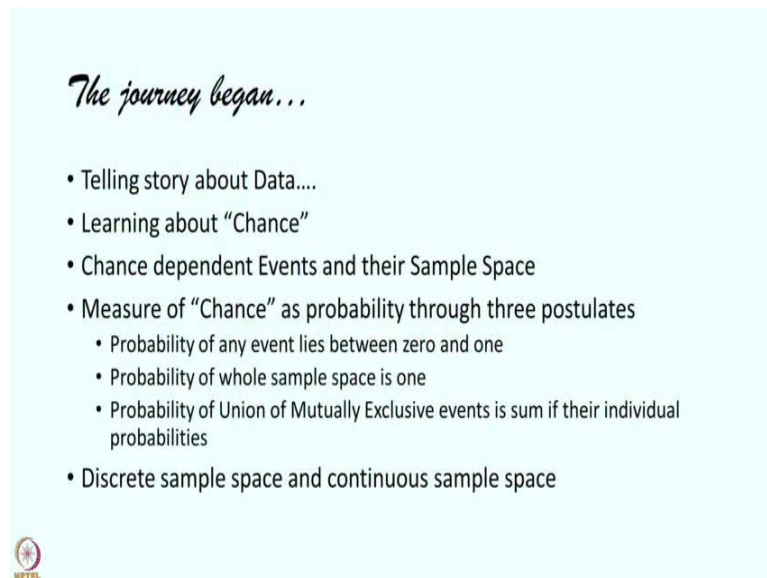


Dealing with Materials Data: Collection, Analysis and Interpretation
Professor M. P. Gururajan
Professor Hina. A. Gokhale
Department of Metallurgical Engineering and Materials Science
Indian Institute of Technology, Bombay
Lecture 91
Summary of the course


Hello and welcome back to the course on Dealing with Materials Data. We have come a long way. We had a 6 week of course on basic understanding, basic understanding of basic concepts of statistics. In this session, we are going to summarize whatever we have learned so far.

(Refer Slide Time: 0:41)



The journey began...

- Telling story about Data...
- Learning about "Chance"
- Chance dependent Events and their Sample Space
- Measure of "Chance" as probability through three postulates
 - Probability of any event lies between zero and one
 - Probability of whole sample space is one
 - Probability of Union of Mutually Exclusive events is sum if their individual probabilities
- Discrete sample space and continuous sample space

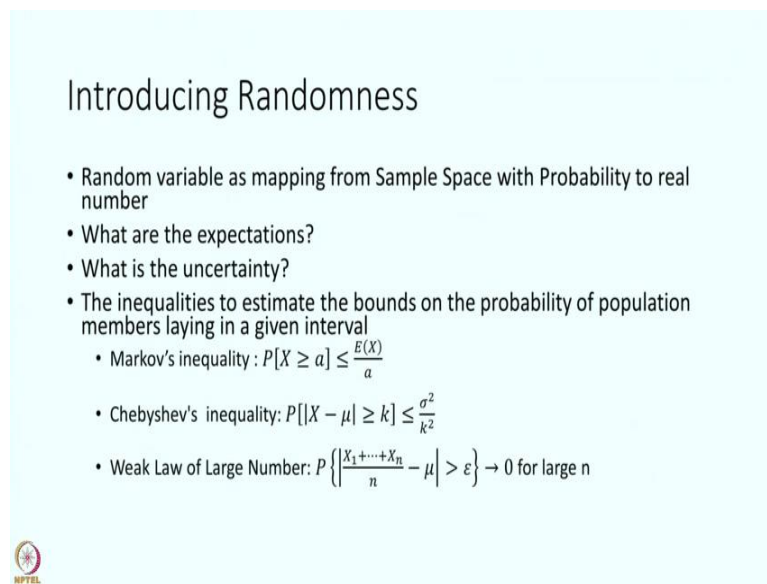


So, summing up, our journey began by telling the story about data, how to explain a data to someone who has not seen the data or the data is so large that cannot make out anything about it quickly. So, we what we called it, data description or descriptive statistics.

Then we started learning about chance and we defined the chance dependent events and their sample space. And then we came to the measure of chance as a probability with three postulates which said that probability of event, any event lies between 0 and 1. The probability of a whole space is 1 and if you have mutually exclusive events, then the sum of the union of probability of union of the event is sum of the individual probabilities.


And then we say that there are some sample spaces which are discrete, which tend to take discrete values and there are some sample spaces which are continuous because they tend to take continuous values.

(Refer Slide Time: 1:54)



Introducing Randomness

- Random variable as mapping from Sample Space with Probability to real number
- What are the expectations?
- What is the uncertainty?
- The inequalities to estimate the bounds on the probability of population members laying in a given interval
 - Markov's inequality : $P[X \geq a] \leq \frac{E(X)}{a}$
 - Chebyshev's inequality: $P[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}$
 - Weak Law of Large Number: $P\left\{\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \varepsilon\right\} \rightarrow 0$ for large n



Then, we introduced what is called randomness and here, friends I would like to tell you that after taking this course, we should be able to distinguish between a randomly occurring event and an arbitrary event.

You see, it is a very common habit to say that some random person came or some random thing happened. Sometimes, we actually mean arbitrary. So, let me take little bit of time to tell you, the random variable, the way we have define and that is why I call it a introduction to randomness. The random variable is always connected with a probability. It is a sample space with a probability mapped to a real number.

So, it is a number, but it is has a probability attached to it, while when you talk of arbitrary, arbitrariness has nothing to do with probability, so there is a distinction between the two. So, at least one takeaway from this course is that definitely that you know when to use a random word and when to use a word arbitrary.

How to define an event to be random and how when you say that it is arbitrary. Then we talked what are the observations of this random variables. What is the uncertainty, by that what I mean is that what is the variance? Because variance defines that it varies between this limits with certain probability as we have studied in the interval estimation. So, the variance as a takeaway, you can understand it that it is a measure of uncertainty of the random variable.

(Refer Slide Time: 3:40)

Introducing Randomness

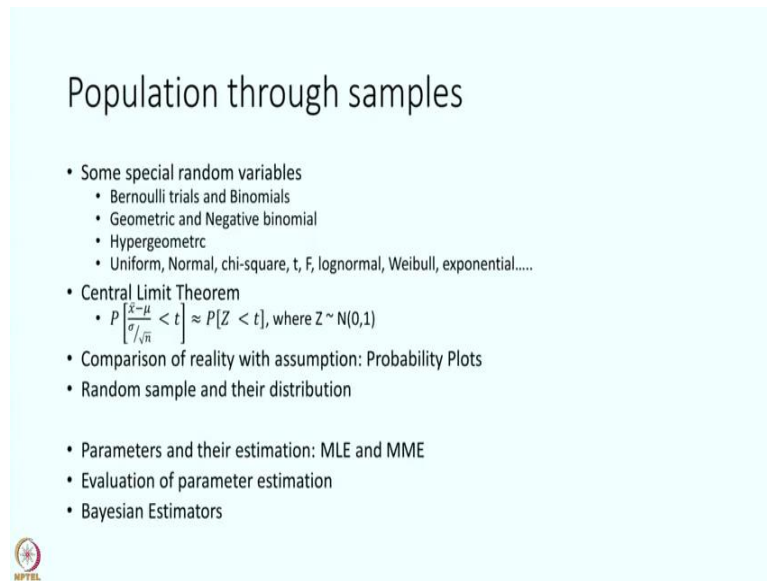
- Random variable as mapping from Sample Space with Probability to real number
- What are the expectations?
- What is the uncertainty?
- The inequalities to estimate the bounds on the probability of population members laying in a given interval
 - Markov's inequality : $P[X \geq a] \leq \frac{E(X)}{a}$
 - Chebyshev's inequality: $P[|X - \mu| \geq k] \leq \frac{\sigma^2}{k^2}$
 - Weak Law of Large Number: $P\left\{\left|\frac{X_1 + \dots + X_n}{n} - \mu\right| > \varepsilon\right\} \rightarrow 0$ for large n



Then, we understood certain inequalities, Markov's inequality, Chebyshev's inequality and weak law of large number. Here, what we have tried to say is we have to try to put a bound on the probability of population member line with a certain limits. So, for example Markov inequality says that if a, a random variable X is always positive, then probability that X takes a value larger than a predefined value, a predefined number a, it is smaller than expected value of X divided by a.

Chebyshev's inequality says that, the values that would lie between mu plus k and mu minus k interval of a random variable X that is random variable X would lie between mu minus k and mu plus k, where mu is its mean value and k is any real number greater than 0 then that is lesser or equal to the variance of the random variable divided by k square. The weak law of large number finally says that the sample mean comes very close to the actual mean of the population as the number of sample size gets larger and larger, this is called weak law of large number.


(Refer Slide Time: 5:10)



Population through samples

- Some special random variables
 - Bernoulli trials and Binomials
 - Geometric and Negative binomial
 - Hypergeometric
 - Uniform, Normal, chi-square, t, F, lognormal, Weibull, exponential....
- Central Limit Theorem
 - $P\left[\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < t\right] \approx P[Z < t]$, where $Z \sim N(0,1)$
- Comparison of reality with assumption: Probability Plots
- Random sample and their distribution

- Parameters and their estimation: MLE and MME
- Evaluation of parameter estimation
- Bayesian Estimators



Then, we started learning about population through samples. So, we said, first let us see what all kinds of different distribution forms that a population can take. So, we learned about some special random variables, Bernoulli trials, binomial distributions, geometric distribution, normal distribution, negative binomial distribution, chi square distribution, F distribution and so on and so forth.

And in this we came to a central limit theorem which said that for a very large value of N that is your large value of large sample size, the sample mean minus the population mean divided by its population variance by square root N behaves like a normal random variable with mean 0 and variance 1, this only helps in case of a large sample size. You know that the behaviour of mean is like a normal random variable and it is closed, its expected value is expected to be the mean of the population.

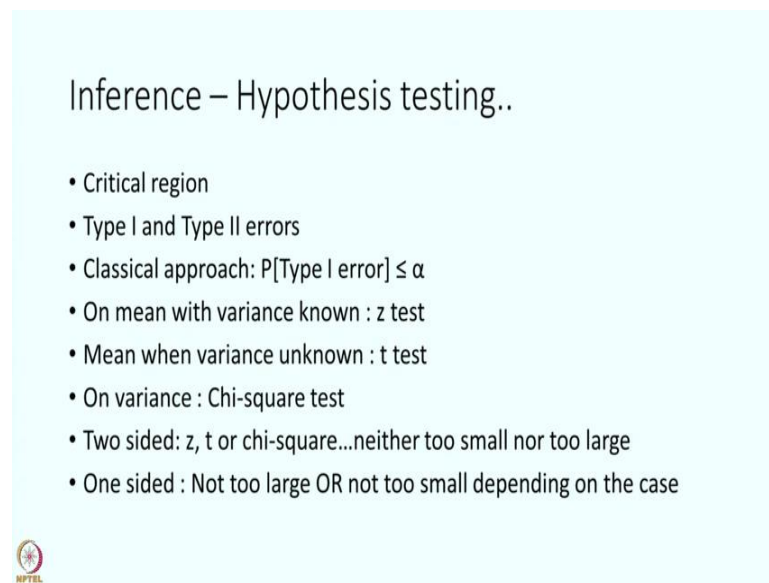
Then we said that suppose you have made this assumption on the population and your sample has certain values of its probability density function or probability distribution function, then we talked about probability plots. How to plot the hypothetical value that is what you have assumed value of the population along with that is assumed cumulative distribution function along with a sample cumulative distribution function and make a judgement whether it is, your assumption is true or close to the reality or not.

Then we came to parameter and their estimations, we learned about maximum likelihood estimator, we learned about the estimator that is method of moment's estimator. We also showed that there could be inconsistency if you use method of moment estimators. The

evaluation of parameter estimation also we learned, we said that there could be a, you can take a least squares distance, minimum biased distance and then if there is no bias then you call it an unbiased estimator.


And we also introduced very briefly, the Bayesian estimation, we did not really introduce Bayesian estimators per say, but we gave the conceptual design of how the Bayesian estimators are derived. We move forward and we started working on hypothesis testing.

(Refer Slide Time: 8:15)



Inference – Hypothesis testing..

- Critical region
- Type I and Type II errors
- Classical approach: $P[\text{Type I error}] \leq \alpha$
- On mean with variance known : z test
- Mean when variance unknown : t test
- On variance : Chi-square test
- Two sided: z, t or chi-square...neither too small nor too large
- One sided : Not too large OR not too small depending on the case



Now here, I would like to tell you that yes we did all the work by looking at or making an assumption that the population is normal. So, you might be wondering that if the population is all the time normal then what is a point because it was already said that the materials data tend to be skewed, they are not beautifully bell shaped curve, then what is the point of having it? But I would like to bring it to your notice that the concept of critical region, the concept of type 1 and type 2 errors and the approach to keep the type 1 error as small as possible by predefining a value alpha, this concept has nothing to do with any distributional assumption.

There are very general assumptions that have been made. So, these are the assumptions for any distribution, there is no assumption, there is no need to have a normal distribution assumption here. Also with respect to testing the mean equal to a given value or equality of two different means or when the means are equal when the variance are unknown to say that a variance is equal to a given value, you take any of this test which will be a Z test, a T test or chi square test. If you wish to have a look at the table and arrive at an alpha value or to arrive at a cut off value, then you need a normal distribution assumption.

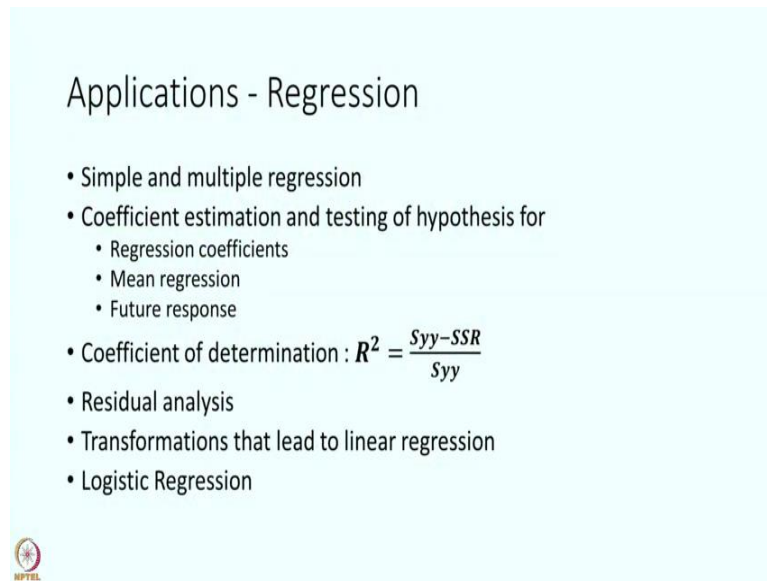
Otherwise frankly speaking, two sided test says that Z, T or chi square neither should be too small nor it should be too large. The value, you can always calculate the T or chi square and you just have to see that it should not be very very very large and it take neither should be very very small if you are looking at two sided test.

If you are looking at one sided test then depends on which side you are looking, it should not be very small or it should not be very large. And this we demonstrated by arriving at Weibull T and Weibull chi square test. So, here I would like to tell you that the rear takeaway for you is for you to understand that this statistics, the Z statistic, T statistic, chi square statistic and in future in ANOVA what comes as an F statistic.

Their value is comparatively large or small is all you are looking at. If you wish to have what we call a cut off value then as we have shown in the case of Weibull distribution, you can always do it by doing the simulation. Now, the computer is in your hands, the algorithm is there, you can simulate the random numbers and generate the cut off value you wish for, any given alpha value that you need to have.


So, the takeaway here is, please do not get carried away by the normality assumption, the takeaway is that these statistics which have been introduced here with the reference to the test, the hypothesis made are valid even beyond the normal distribution.

(Refer Slide Time: 12:05)



Applications - Regression

- Simple and multiple regression
- Coefficient estimation and testing of hypothesis for
 - Regression coefficients
 - Mean regression
 - Future response
- Coefficient of determination : $R^2 = \frac{S_{yy} - SSR}{S_{yy}}$
- Residual analysis
- Transformations that lead to linear regression
- Logistic Regression



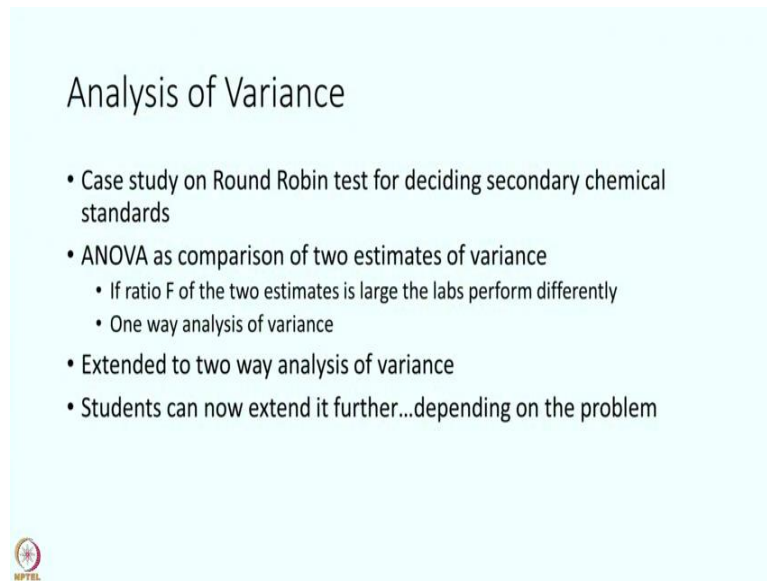
Then we started going into applications, we first consider regression application, we again looked into the coefficient estimation and then we did the hypothesis testing for regression coefficient, mean regression and future response.

Again, please remember in all this we have derived T statistic, and T statistic is a T statistic, it need not have the normal assumption, it is a unit less statistic and can be looked into it with reference to any distribution that you wish to assume or you can do it by simulating it and getting the cut off values if you are so keen to look into it, otherwise looking at the largeness and smallness is good enough to give you an answer.

Then, we talked about coefficient of determination to decide how good is the fit. We talked about residual analysis, very important. We have made some assumptions and this assumptions we must verify through the residual analysis. We also talked briefly about the transformations that can lead again back to regression, linear regression.

And I know that in the sessions doing your analysis with data analysis with R, you have been given variety of examples other than the what we discussed in the our regular statistics class for this kind of transformation leading to linear regression model. We very briefly talked about logistic regression.

(Refer Slide Time: 13:50)



The slide is titled "Analysis of Variance" and contains the following bullet points:

- Case study on Round Robin test for deciding secondary chemical standards
- ANOVA as comparison of two estimates of variance
 - If ratio F of the two estimates is large the labs perform differently
 - One way analysis of variance
- Extended to two way analysis of variance
- Students can now extend it further...depending on the problem

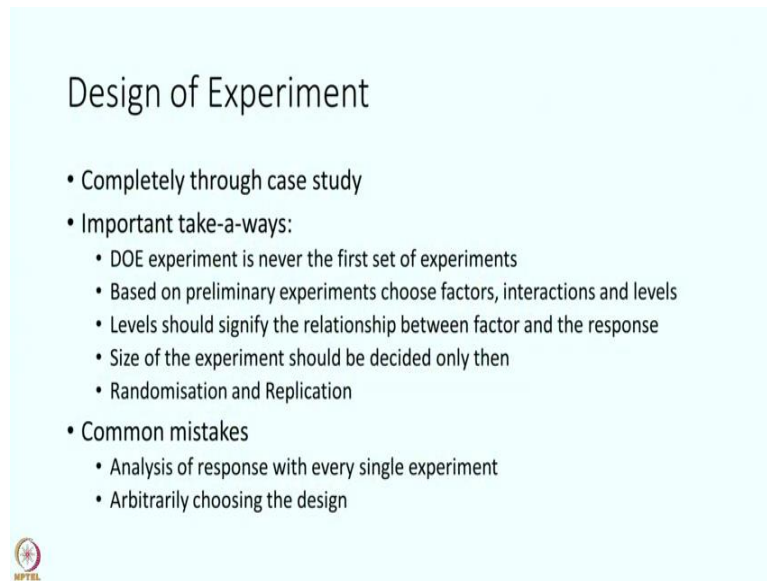
In the bottom left corner of the slide, there is a small circular logo with the text "NPTEL" below it.

Then we started going through the case study and understanding analysis of variance. We studied the analysis of variance again, we came across the F ratio because we realized that analysis of variance as it says, it is comparing the two variances which arrives under the hypothesis and due to the error.

And if you compare the two estimate of variances it leads to an F ratio. And larger the F ratio, more chances that you are deviating from your assumption of the rows or all the laboratories in our case of case study, all the laboratories are not different or the all the laboratories are same, so you are deviating from it.

Once again, F distribution is not important, F ratio is important. We extended it to a two way analysis of variance and is now you can all imagine that you can further extend it even to the three way analysis of variance and so on and so forth depending on the problem.

(Refer Slide Time: 15:06)



The slide is titled "Design of Experiment" and contains the following content:

- Completely through case study
- Important take-a-ways:
 - DOE experiment is never the first set of experiments
 - Based on preliminary experiments choose factors, interactions and levels
 - Levels should signify the relationship between factor and the response
 - Size of the experiment should be decided only then
 - Randomisation and Replication
- Common mistakes
 - Analysis of response with every single experiment
 - Arbitrarily choosing the design

At the bottom left of the slide is the NPTEL logo.

Finally, we came to design of experiments. This we understood completely through case study. There is one important point which I had forgotten which I would like to mention it here that when we chose the design matrix of size 16, please remember, had we not done this exercise, since we had 7 independent random variables varying, that is independent variables varying where 7, the size of the design matrix would have been 2 to the power 7, instead we worked out with 2 to the power 4.

So this is also called a fractional factorial experiment, this is confounded experiment, I totally agree, but these details as I said are the details when you go into more and more understanding and detailed or very intricate designed experiments, but to begin with because this says that dealing with materials data to begin with when you want to do the design of experiment, this case study gives you an idea that instead of having experiments number to of power 2 to the power 7, we are dealing with only 2 the power 4 and that makes the difference.

Once again, there are some important takeaways here which I had mentioned there. Design of experiment is never, the designed experiment is never the first set of experiments. You have to have some preliminary experiments done before in order to understand what are the factors effecting, what are the interactions that might be effecting, how the levels are playing a role.

Levels should signify the relationship between factor and the response. It should not be taken out of just curiosity, yes in scientific experimentation, curiosity plays a major role, but in design of experiment, the purpose is to have a systematic experiment as small number as possible and gain maximum out of it.

So, minimum number of experimentation, maximum gain requires that this level should truly signify the relationship between the factor and the response, some preliminary analysis or preliminary experimentation is very necessary for that. The size of the experiment should be decided only once these exercise are over. And please do not forget to randomize and replicate the estimation because randomization and replication takes away two other kinds of errors.

The randomization takes care of the nuisance factor, the errors or the or the random error occurring through the nuisance factor. And replication takes care of the random error occurring through the because of the experimentation itself. There are some common mistake I have seen when the young researcher comes to me to take a help on design of experiment, they try to analyse response after every single experiment. Please remember, we design a design matrix, we did not design a single experiment, we design a matrix.

So, the complete matrix of the experiment should be completed before that let us not get into the analysis, it actually diverts our attention from the experimentation. Second thing I have found out is that, without putting so much of thought as to what should be the levels, why should be three levels or two level experiment, how many replications you wish to make, what interactions are important, simply arbitrarily, now I am not saying randomly, I am saying, arbitrarily a design is chosen and experiment is done.

And when you come with such a experiment already over and your analysis is not very informative to you, well I have to remember R A Fisher, who used to say that if you approach a statistician after the experimentation is over, perhaps all the statistician can do is do the post-mortem to find out what the experiment died of. So, let us not make that mistake. So, these are the major important takeaways. And why am I spending time here? Actually, speaking this whole course was to make you understand regression, analysis of variance and design of experiment.

As we saw in design of experiment, it uses the principles of regression, it also uses the principles of analysis of variance. So, that is how it is in that order. But in order to understand the design of experiment, the analysis of variance and the regression, we had to understand the basic of statistics.

So, these techniques which a materials engineer would be using more frequently which is regression or analysis of variance of design of experiment, please remember, please understand

your basics clearly, what are you doing and then apply it. And these are the your major takeaways from the design of experiment.

I wish the journey continues and may the journey continue forever, you learn more about statistics, you learn more about materials data analysis. This is the age for material data science and I wish you all the very best. Thank you.