

Chemical Process Instrumentation
Prof. Debasis Sarkar
Department of Chemical Engineering
Indian Institute of Technology, Kharagpur

Lecture – 15
Performance Characteristics of Instruments and Data Analysis - II (Contd.)

Welcome to lecture 15. This is the last lecture of week 3 and also this is a last lecture of performance characteristics of instruments and data analysis part 2. So, we have talked about probability density functions in our previous lecture and we will talk about some more aspects of statistical data analysis in this lecture.

(Refer Slide Time: 00:42)

**Performance Characteristics of Instruments
and Data Analysis - 2**

Today's Topic:

- ✓ Statistical Analysis of Experimental Data:
 - Chi-square test of goodness of fit
 - Students' t-distribution
 - Regression analysis

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, today's topic is Chi-square test of goodness of fit student's distribution student's t distribution and regression analysis let us start with Chi-square test of goodness of fit, we have discussed that random experimental errors are expected to follow Gaussian distribution or normal distribution.

(Refer Slide Time: 01:18)

Chi-Square Test of Goodness of Fit

- Random experimental errors would be expected to follow *Gaussian distribution*. Can we determine if a set of experimental observation match some particular distribution?
- The chi-square test of goodness of fit can be used to answer this question. It is defined by

$$\chi^2 = \sum_{i=1}^n \frac{[(\text{Observed value})_i - (\text{Expected value})_i]^2}{(\text{Expected value})_i}$$

$F = n - k$ $n = \text{number of groups of observations}$, $F = \text{degree of freedom}$
 $k = \text{number of imposed condition}$

1. For the computed value of chi-square and degree of freedom (F), find the probability (P) from the Table.
2. If P lies between 0.1 and 0.9, the observed distribution follows assumed distribution.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now let us say you have a set of observations we have performed some experiments in laboratory let us say you have measured a pressure or you measured a temperature.

And obtained a set of data is it possible to say whether this experimental observation match some particular distribution. So, the question we ask is can you determine if a set of experimental observation match some particular distribution the Chi-square test of goodness of fit can be used to answer this question Chi-square is defined by this equation.

So, you have a set of data and you have expecting a value following some particular distribution for your data let us say you assume a normal distribution or you assume some other distribution. So, you have some expected value for the observation. So, you have each observation and you have corresponding expected value for the observation which follows for some distribution that you have assumed for your data.

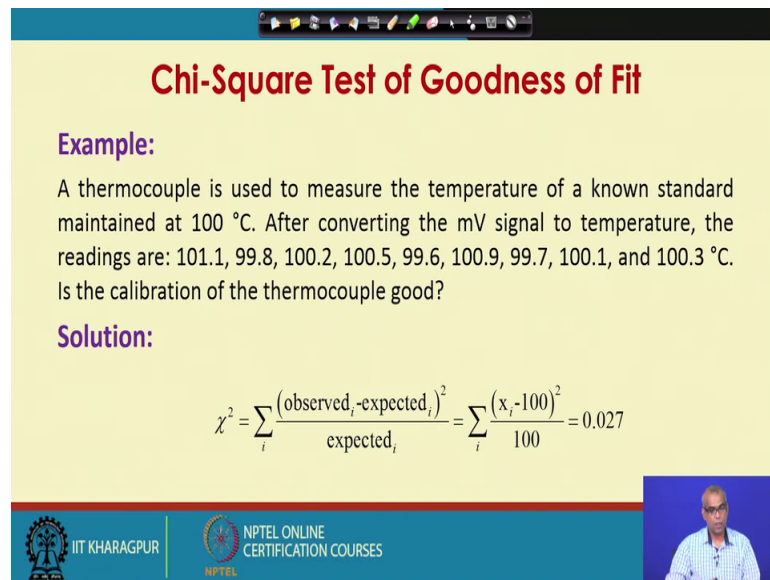
Now, you find out this quantity observed value minus expected value whole square divided by expected value this terms you sum for all the observations. So, you pick up one observation find the difference between observed value and expected value square it divide that by expected value sum such terms sum up such terms for all observations. So, that is defined as Chi-square let us define another term known as degrees of freedom which is n minus k where n is number of groups of observations is also known as number of sales or number of groups of observations and k equal to number of imposed

conditions for the computed value of Chi-square and degree of freedom you find the probability from the Chi-square distribution table.

So, once you have the set of data you find out the Chi-square value using the formula that you have talked about you can calculate the Chi-square value for the set of data you also calculate the degree of freedom then there is a table available known as Chi-square distribution Chi-square probability distribution table. Now look at this Chi-square distribution table and find the probability from the table corresponding to degrees of freedom if the probability lies between 0.1 and 0.9 the observed distribution follows assumed distribution. So, what is done is you find out the Chi-square value from the set of data in principle if Chi-square equal to 0; the distribution in the data is equal to the distribution that you have assumed large value of Chi-square means there is deviation or departure from the assumed distribution. So, the data does not follow exactly the distribution that you have assumed for large value of Chi-square.

Now, what you do is we again consider a level of confidence. So, imagine we let us assume that we consider 95 percent confidence level. So, now, we look at the Chi-square distribution table corresponding to degrees of freedom given degrees of freedom and under 95 percent confidence level we find out the Chi-square value from the table if our computed Chi-square is less than this critical Chi-square value obtained from the table then we say that our assumption about the distribution is good if our computed Chi-square is greater than the obtained critical Chi-square from the Chi-square distribution table then the assumption about the distribution in the data is not good actually we should look at both side a 95 percent confidence level also at the lower level; let us say 0.05 percent.

(Refer Slide Time: 08:36)



Chi-Square Test of Goodness of Fit

Example:
A thermocouple is used to measure the temperature of a known standard maintained at 100 °C. After converting the mV signal to temperature, the readings are: 101.1, 99.8, 100.2, 100.5, 99.6, 100.9, 99.7, 100.1, and 100.3 °C. Is the calibration of the thermocouple good?

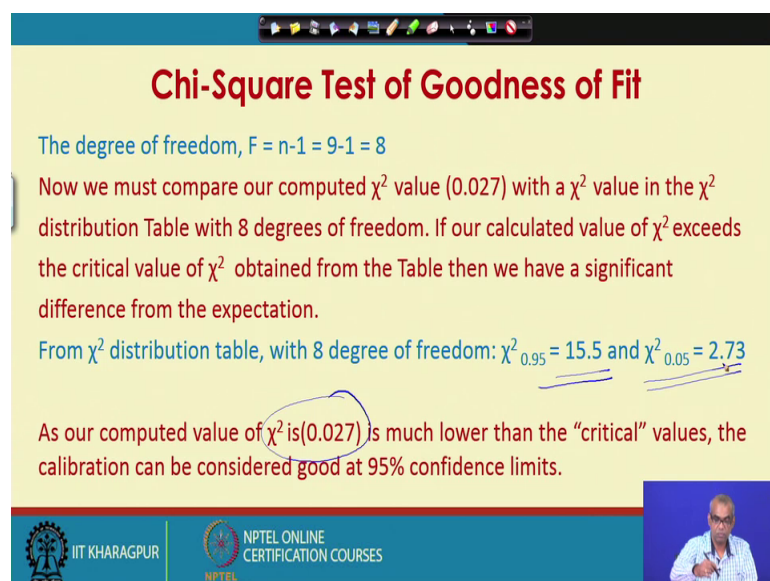
Solution:

$$\chi^2 = \sum_i \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i} = \sum_i \frac{(x_i - 100)^2}{100} = 0.027$$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, we will take an example to make it more clear a thermocouple is used to measure the temperature of a known standard maintained at 100 degree Celsius after converting the millivolts signal to temperature the readings are 101.1; 99.8; 100.2, 100.5, 99.6, 100.9, 99.7, 100.1 and 100.3 degree Celsius is the calibration of the thermocouple good or in other words are this values are expected to start with we must calculate the Chi-square value. So, the Chi-square value is computed using the formula and you go to get a value of 0.027. Now we will make use of the Chi-square distribution table, but before that.

(Refer Slide Time: 09:48)



Chi-Square Test of Goodness of Fit

The degree of freedom, $F = n - 1 = 9 - 1 = 8$

Now we must compare our computed χ^2 value (0.027) with a χ^2 value in the χ^2 distribution Table with 8 degrees of freedom. If our calculated value of χ^2 exceeds the critical value of χ^2 obtained from the Table then we have a significant difference from the expectation.

From χ^2 distribution table, with 8 degree of freedom: $\chi^2_{0.95} = 15.5$ and $\chi^2_{0.05} = 2.73$

As our computed value of χ^2 is (0.027) is much lower than the "critical" values, the calibration can be considered good at 95% confidence limits.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

I have to calculate the degrees of freedom the degrees of freedom here is number of goods of observation minus 1 which is 9 minus 1 equal to 8.

Now, you must compare our computed Chi-square value which is 0.027 with a Chi-square value in the Chi-square distribution table with 8 degrees of freedom if our calculated value of Chi-square exceeds the critical value of Chi-square obtained from the table, then we have a significant difference from the expectation if our Chi-square value our computed Chi-square value is less than the critical value of Chi-square obtain from the table, then the values you have obtained are; that means, data we have obtained are expected.

So, from Chi-square distribution table with 8 degree of freedom, let us calculate the Chi-square corresponding to 95 percent confidence level as well as 0.05 percent level we see that our computed Chi-square which is 0.027 is much less than Chi-square at 95 percent confidence level and Chi-square at five percent confidence level.

So, we say that the calibration can be considered good at 95 percent confidence limits.

(Refer Slide Time: 11:33)

Student's t-Distribution

To determine standard deviation of the mean in terms of standard deviation of the population we use

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

σ_m = stand deviation of the mean value
 σ = standard deviation of the set of measurements
 n = no. of measurements

For small samples $n < 10$ this relation is not very reliable. Student's developed a better way for determining confidence interval by introducing t

$$\Delta = \frac{t\sigma}{\sqrt{n}}$$

Where t is

$$t = \frac{\bar{x} - X}{\sigma} \sqrt{n}$$

\bar{x} = mean of n observations
 n = number of observations
 X = mean of normal population which the sample are taken from

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Now, there is another important aspects of distribution known as student's t distributions. So, students t distribution is another important aspect of statistical data analysis in our previous lecture, we have learned about standard error which is standard deviation divided by square root of number of observation sigma divided by square root of n now

when n is small this sigma divided by square root of n gives an unreliable estimate of uncertainty under such circumstances or where sample size is small students proposed an alternative and better way to calculate the uncertainty level. So, let us see; what exactly is this. So, standard deviation of the mean in terms of standard deviation of the population is sigma divided by square root of n for small samples n less than n this relation is not very reliable.

So, students developed a better way for determining confidence interval by introducing something called t which is defined as this. So, with t the confidence interval can be written as t sigma by square root of n where t is a mean of observations minus mean of normal population from which the sample are taken divided by standard deviation and multiply this whole quantity by square root of number of observation that is square root of n.

(Refer Slide Time: 14:29)

Student's t-Distribution

Example:
A thermocouple is used to measure the temperature of a known standard maintained at 100 °C. After converting the mV signal to temperature, the readings are: 101.1, 99.8, 100.2, 100.5, 99.6, 100.9, 99.7, 100.1, and 100.3 °C. Obtain the mean measurement and calculate the tolerance limit for 95 % confidence level.

Solution: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
 $= \frac{1}{9} (101.1 + 99.8 + 100.2 + 100.5 + 99.6 + 100.9 + 99.7 + 100.1 + 100.3)$
 $= 100.24 \text{ } ^\circ\text{C}$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Lets again take another example to make the point more clear a thermocouple is used to measure the temperature of a known standard maintained at 100 degree Celsius after converting the millivolts signal to temperature the readings are 101.1, 99.8, 100.2, 100.5, 99.6, 100.9, 99.7, 100.1.

And 100.3 degree Celsius obtain the mean measurement and calculate the tolerance limit for 95 percent confidence level. So, you want to find out the uncertainty associated at the level of 95 percent confidence. So, the mean can be computed in a straight forward

manner from this equation and it happens to be 104; 100.24 degree Celsius. So, this is the mean of the measurements. Now these measurement has to be expressed as this mean of this measurement plus minus some uncertainty sigma by square root of n may not be a good idea here because of sample size a small. So, we will make use of student's t distribution.

(Refer Slide Time: 16:17)

Student's t-Distribution

The sample standard deviation is $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = 0.5199 \text{ } ^\circ\text{C}$

The degree of freedom, $F = n-1 = 9-1 = 8$

Using the Student's distribution table, we obtain $t_{95} = 2.306$

$\Delta = \frac{t\sigma}{\sqrt{n}} = \frac{2.306(0.5199)}{\sqrt{9}} = 0.399 \text{ } ^\circ\text{C}$

$x = 100.24 \pm 0.39 \text{ } ^\circ\text{C}$

Value of Student's t as a function of degree of freedom and confidence level is available as a table in text books.

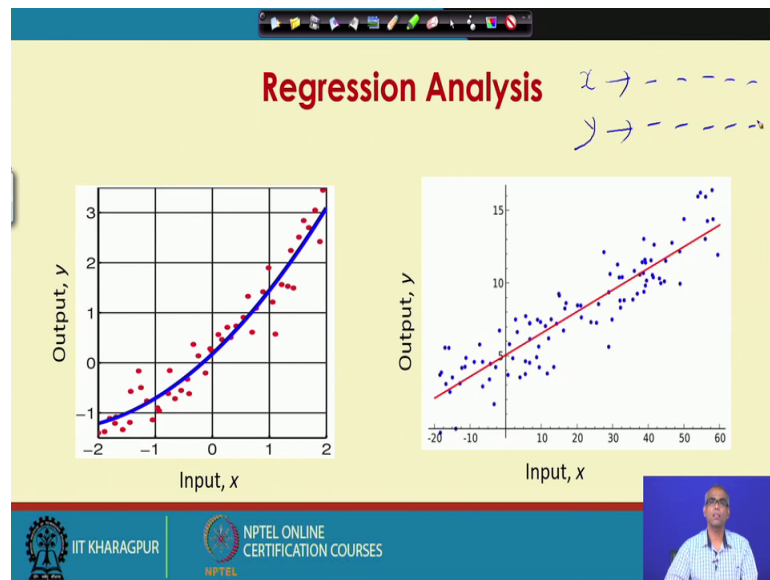
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, we compute the sample standard deviation using the standard formula for standard deviation the degree of freedom is number of groups of observations minus one which is 9 minus one equal to 8 now you have to look at student's t distribution table, which is available in many standard textbooks of statistical data analysis. If you look under students distribution table we will see the value of the t at 95 percent confidence level is 2.306.

Now let us make use of this formula to compute the level of uncertainty which is t into standard deviation divided by square root of n you put the values t is 2.306 standard deviation is 0.5199 and square root of n is square root of 9 and it happens to be 0.399. So, the temperature measurement is now expressed as the mean 100.24 plus minus 0.39 degree Celsius. So, we have seen sigma by square root of n as standard error in our previous lecture here we multiply this quantity with the term called t.

And this value of this t has to be obtained corresponding to a confidence level from student's t distribution table.

(Refer Slide Time: 19:24)



Now, let us talk about regression analysis you have input verses output data. So, you have input and corresponding output you have. So, this input output data will allow you to establish a relationship between this output and input which essentially is the calibration for the instrument now the relationship between input and output may be linear or it may be non-linear if the instrument is linear we are expected to get the linear relationship between input and output and if the relationship is non-linear we will get a non-linear relationship on non-linear behavior between output and input. Now if I have a set of data, I can express this relationship in terms of graph or I can also express in terms of equations.

Here we will take an example of linear regression analysis; that means, given a set of data will try to fit a linear equation for this data set. So, if I have x verses y data, I will try to fit an equation like $y = ax + b$ where x is input and y is output and a and b are the parameters of the equation. So, the question we ask is how do I estimate this parameters a and b.

(Refer Slide Time: 22:25)

Linear Regression Analysis

Method of least square:

- standard approach in regression analysis
- The most important application is in data fitting, finds the line of best fit for a dataset

We have data values as (x, y) . We want to fit: $y(x) = ax + b$
How to compute a and b?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Method of least square is a standard approach in regression analysis the most important application is in data fitting it finds the line of best fit for a data set we have a data values as x y and we want to fit y x equal to a x plus b. So, how do you compute the values of a and b.

(Refer Slide Time: 22:59)

Linear Regression Analysis

We want to fit a straight line such as $y(x) = ax + b$ to the data set (x, y) .
Thus we must minimize the following error:

$$\text{Minimize } \left\{ \text{Error} = \sum_{i=1}^N [y_i - (ax_i + b)]^2 \right\}$$

Take the derivative of the error with respect to a and b, and set each to zero.

$$\frac{\partial(\text{Error})}{\partial a} = -2 \sum_{i=1}^N x_i [y_i - (ax_i + b)] = 0 \quad \frac{\partial(\text{Error})}{\partial b} = -2 \sum_{i=1}^N [y_i - (ax_i + b)] = 0$$

Solve for a and b from these two equations.

Handwritten notes: $e_i \rightarrow y_i$, $ax_i + b$, $y_i - (ax_i + b)$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, you want to fit a straight line such as y x equal to a x plus b to the data set x y. So, we need to minimize the error and the error can be represented as the deviation from the deviation from the true value.

If the true value is y_i for a particular observation the equation predicted output is $a x_i + b$ because $y = a x + b$ is the equation that I am fitting in the data set x, y . So, for an input x_i the actual output is y_i the output predicted by the equation will be $a x_i + b$. So, the error will be $y_i - a x_i - b$. So, sum up all such squared errors must be minimized in other words, we have to find out the values of a and b such that these error is minimum under that situation $y = a x + b$ will be a best fit line that passes through the given data set. So, given a data set, I want to fit a straight line such as $y = a x + b$. So, I define the error which is $y_i - a x_i - b$ and sum all the squared errors.

So, the question now is find values of a and b such that sum up all this squared errors is minimum this problem is known as least square problem finding a and b that minimizes the sum of squared errors for all the observations. So, to find out a and b what we need to do is we have to take the derivative of this error with respect to parameter a and set it equal to 0 we have to do the same thing for parameter b ; that means, take the derivative of error with respect to parameter b and set it equal to 0. So, here you have 2 equations we solve this 2 equation simultaneously and get the values of a and b . So, we take the derivative of the error with respect to a we have these equation we take the derivative of error with respect to b , we have this equation we set this 2 equal to 0 and solve these 2 equations simultaneously to obtain a and b .


(Refer Slide Time: 27:19)

Linear Regression Analysis (Cont'd)


$$a \sum_{i=1}^N x_i^2 + b \sum_{i=1}^N x_i = \sum_{i=1}^N x_i y_i \quad \text{and} \quad a \sum_{i=1}^N x_i + bN = \sum_{i=1}^N y_i$$

We can write these in matrix form:


$$\begin{bmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N y_i \\ \sum_{i=1}^N x_i y_i \end{bmatrix}$$

$$a = \frac{N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{N \sum_{i=1}^N x_i^2 - \left[\sum_{i=1}^N x_i \right]^2} \quad b = \frac{\sum_{i=1}^N y_i \sum_{i=1}^N x_i^2 - \sum_{i=1}^N x_i \sum_{i=1}^N x_i y_i}{N \sum_{i=1}^N x_i^2 - \left[\sum_{i=1}^N x_i \right]^2}$$


IIT KHARAGPUR



NPTEL ONLINE
CERTIFICATION COURSES



So, this is one equation this is another equation we can write this in this matrix form the solution becomes a equal to this and b equal to this. So, the values of parameter say a and b can be obtained from these 2 expressions. So, this is about linear regression.

(Refer Slide Time: 28:13)

Regression Analysis

We can use an m^{th} degree polynomial for a nonlinear calibration:

$$y = a_0 + a_1x + a_2x^2 + \dots + a_mx^m$$

We must minimize the error:

$$\text{Minimize } \left\{ \text{Error} = \sum_{i=1}^n \left\{ y_i - [a_0 + a_1x_i + a_2x_i^2 + \dots + a_nx_i^n] \right\}^2 \right\}$$

The slide also features a small graph on the right showing a set of data points and a smooth curve fitted to them. At the bottom, there are logos for IIT KHARAGPUR and NPTEL ONLINE CERTIFICATION COURSES, along with a small video inset of the presenter.

But you cannot always set a linear; fit a linear equation through the data let us say your y verses x relationship is something like this we cannot fit a straight line can you try an mth degree polynomial it will be a non-linear calibration. So, you can try an mth order mth degree polynomial similar to linear regression.

In this case also we must minimize this error and again take the derivative of this error with respect to these parameters and set those derivatives equal to 0 and obtained equations will be solved simultaneously to find out the values of parameters. So, this is how we can fit an mth order polynomial through the data, but we will do this only if the linear data linear equation does not fit the data.

(Refer Slide Time: 30:01)

Regression Analysis: Goodness of Fit

- The **correlation coefficient**: conveys how good is the fit by either least square method or graphical curve fitting.
- **Correlation coefficient** is defined by

$$r = \left[1 - \frac{\sigma_{y,x}^2}{\sigma_y^2} \right]^{1/2}$$
$$\sigma_y = \left[\frac{\sum_{i=1}^n (y_i - y_m)^2}{n-1} \right]^{1/2}$$
$$\sigma_{y,x} = \left[\frac{\sum_{i=1}^n (y_i - y_{ic})^2}{n-2} \right]^{1/2}$$

y_i are actual value of y , and y_{ic} are value computed from the correlation coefficient equation for same value of x .

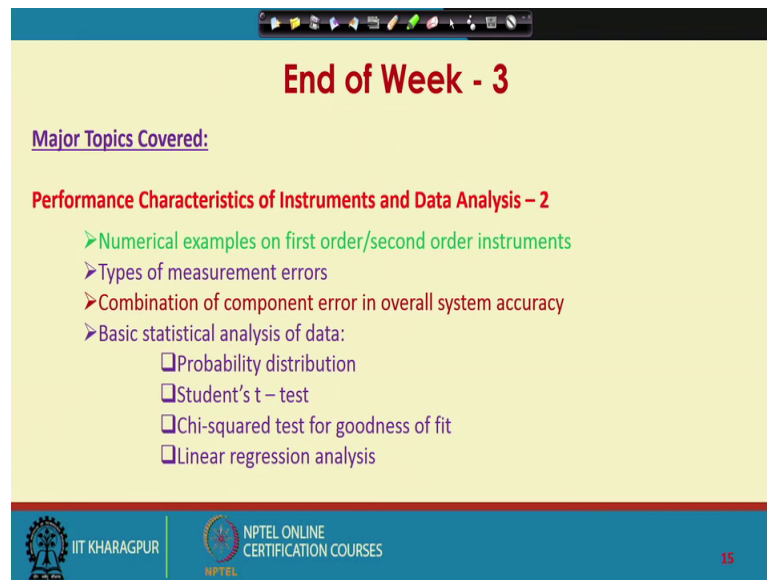
For a perfect fit, $r = 1.0$

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, how do I know that my fit is good. So, there should be a measure of goodness of fit there is something called correlation coefficient which conveys how good is the fit by either least square method or graphical curve fitting. So, correlation coefficient is a measure of goodness of fit correlation coefficient is defined by this equation note that you have terms like sigma y_x and sigma y .

So, sigma y is this and sigma y_x is this y_i are actual value of y and y_{ic} are value computed from the correlation coefficient equation for same value of x for a perfect fit the value of correlation coefficient r will be equal to 1, if the value is very close to 1 the goodness of fit is high. So, fit is good when the value of the correlation coefficient is close to one. So, correlation coefficient will tell you how good your fit is once you have fitted in linear equation or non-linear equation your correlation coefficient will tell you how good is the fit. So, you fit a linear equation and find out the correlation coefficient if it is close to 1; close to 1; let us say 0.995; the fitness is very good.

(Refer Slide Time: 32:06)



End of Week - 3

Major Topics Covered:

Performance Characteristics of Instruments and Data Analysis – 2

- Numerical examples on first order/second order instruments
- Types of measurement errors
- Combination of component error in overall system accuracy
- Basic statistical analysis of data:
 - ☐ Probability distribution
 - ☐ Student's t – test
 - ☐ Chi-squared test for goodness of fit
 - ☐ Linear regression analysis

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 15

So, this is end of week 3 we have covered performance characteristics of instruments and data analysis part 2 that was the major topic under which we have talked about numerical examples on first order and second order instruments.

We have taken examples of thermometer as first order system we have taken u tube manometer as second order instrument we have analyze them, we have talked about types of measurement errors we have seen how to combine compute component errors how to combine component errors to calculate overall system accuracy then we have discussed some basic statistical analysis of data in particular, we have talked about probability distribution functions, we have talked about student's t test, we have talked about Chi-squared test for goodness of fit and we have bit fit as upon linear regression analysis. So, this closes lecture of week 3 and you will be given assignments for these topics that we have covered over this week.