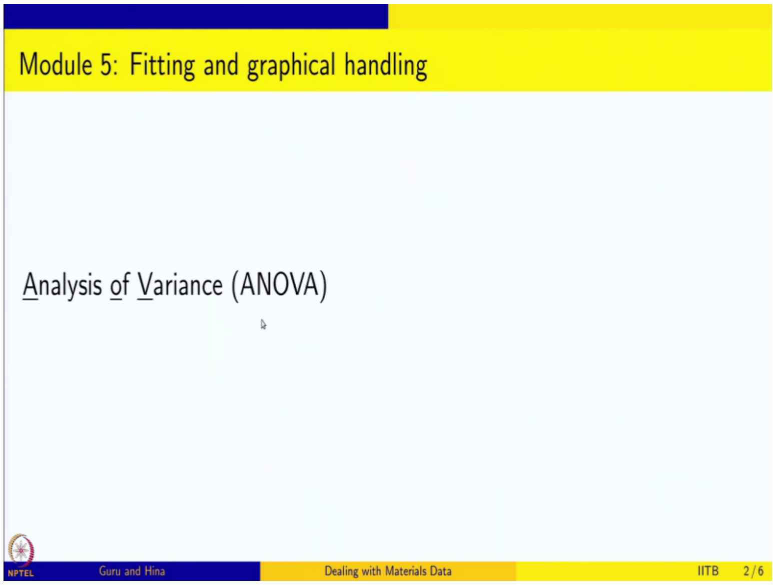
# Dealing with Material Data: Collection, Analysis and Interpretation

**Professor M. P. Gururajan**
**Professor Hina A. Gokhale**
**Department of Metallurgical Engineering and Materials Science**
**Indian Institute of Technology, Bombay**

## Lecture 79
## Analysis of Variance

(Refer Slide Time: 00:26)



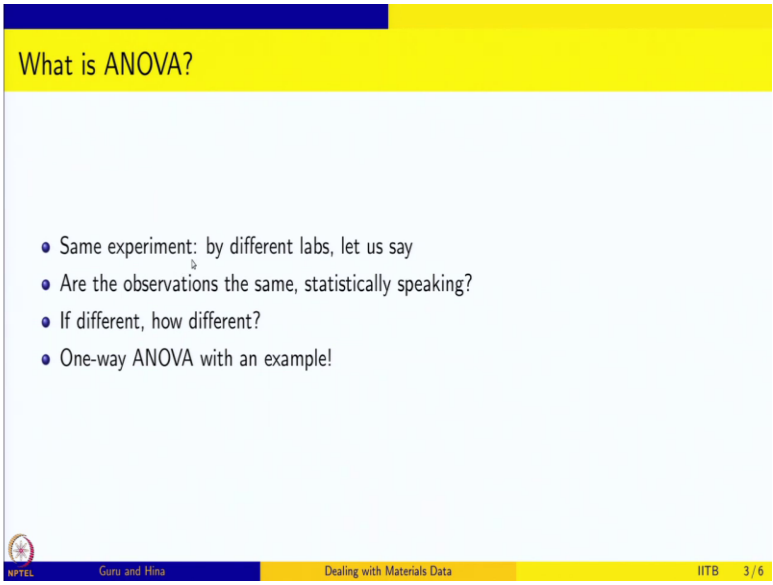
Module 5: Fitting and graphical handling

Analysis of Variance (ANOVA)

NPTEL Guru and Hina Dealing with Materials Data IITB 2/6

Welcome to Dealing with Materials Data. We are looking at the Collection Analysis and Interpretation of Data for Material Science and Engineering. We are in the module on fitting and the graphical handling of data. In this session, we are going to discuss analysis of variance ANOVA, we have already learnt about ANOVA from the other sessions in this course, but this session we will learn how to do ANOVA using the computer, using R.

(Refer Slide Time: 00:43)



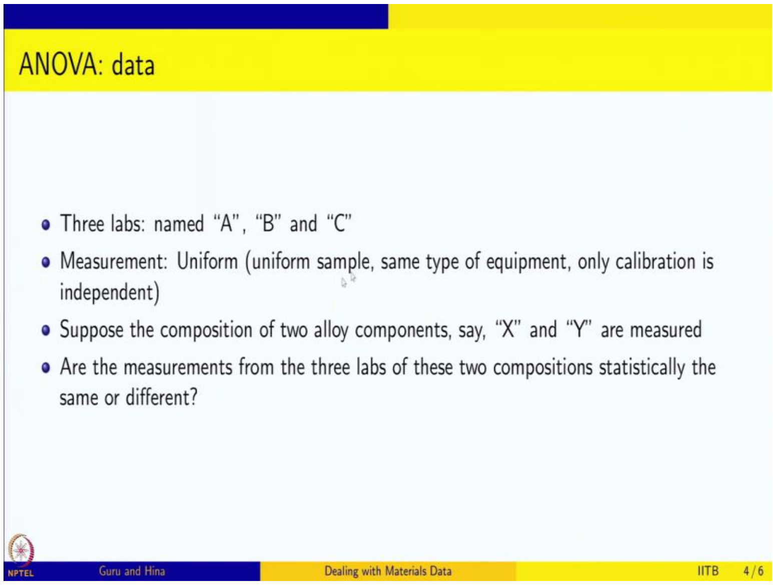
## What is ANOVA?

- Same experiment: by different labs, let us say
- Are the observations the same, statistically speaking?
- If different, how different?
- One-way ANOVA with an example!

NPTEL Guru and Hina Dealing with Materials Data IITB 3/6

What is ANOVA? You know that if you do some experiment, let us say the same experiment by three different labs and each lab does it more than one, so they will have their own mean value for the experiment and different labs will give you different means. What we are testing is that these observations are they the same, statistically speaking or if the differences between the different labs is statistically significant and if they are different, how different? So, this is done using one way ANOVA an you have seen an example and I will use the same example but I will just show you how to do it using R.

(Refer Slide Time: 01:25)



## ANOVA: data

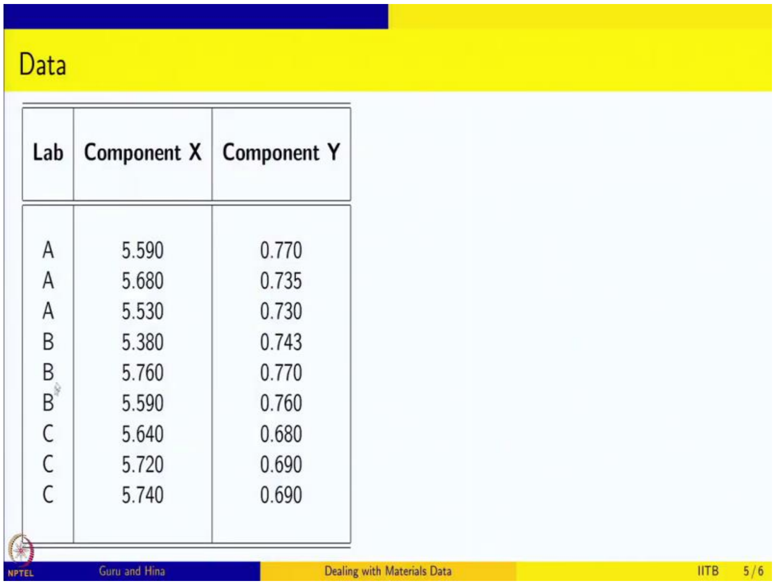
- Three labs: named "A", "B" and "C"
- Measurement: Uniform (uniform sample, same type of equipment, only calibration is independent)
- Suppose the composition of two alloy components, say, "X" and "Y" are measured
- Are the measurements from the three labs of these two compositions statistically the same or different?

NPTEL Guru and Hina Dealing with Materials Data IITB 4/6

The labs are named as A, B and C and the measurement is uniform that is the uniform sample and same type of equipment was use by all the three labs, but only their calibration was independent, each calibrated according to their own standards and then did the measurement. And what measurement they did is the measurement of composition of two alloy components, let us call them as X and Y.

So, you have three labs A, B, C and they measured two quantities X and Y. And we are trying to see if the measurements from the three labs of these two components X and Y are statistically same or different.

(Refer Slide Time: 02:04)



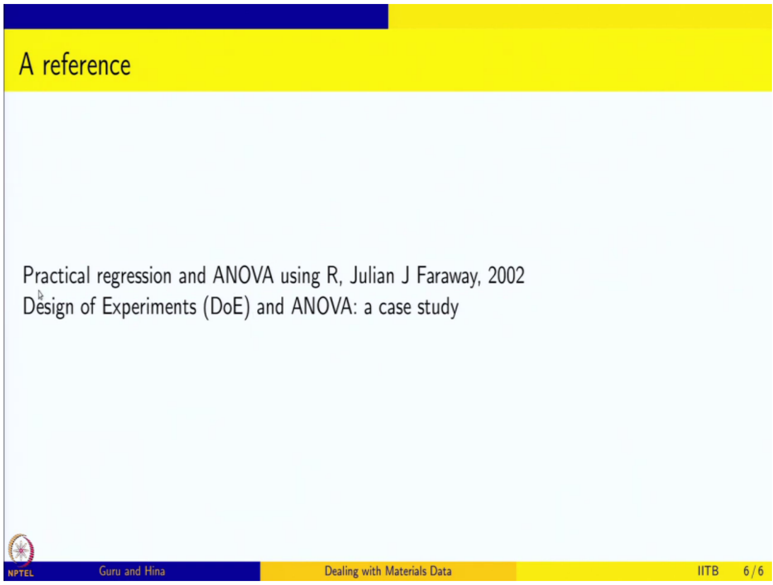
Data

| Lab | Component X | Component Y |
|---|---|---|
| A | 5.590 | 0.770 |
| A | 5.680 | 0.735 |
| A | 5.530 | 0.730 |
| B | 5.380 | 0.743 |
| B | 5.760 | 0.770 |
| B | 5.590 | 0.760 |
| C | 5.640 | 0.680 |
| C | 5.720 | 0.690 |
| C | 5.740 | 0.690 |

NPTEL Guru and Hina Dealing with Materials Data IITB 5/6

So, this is the data. So, the lab is A and then B and then C, and component X the measurements from A is here and here is B and here is C, similarly component Y there are three measurements from A, three measurements from B and three measurements from C. So, let us take this data and try to do the analysis.

(Refer Slide Time: 02:19)



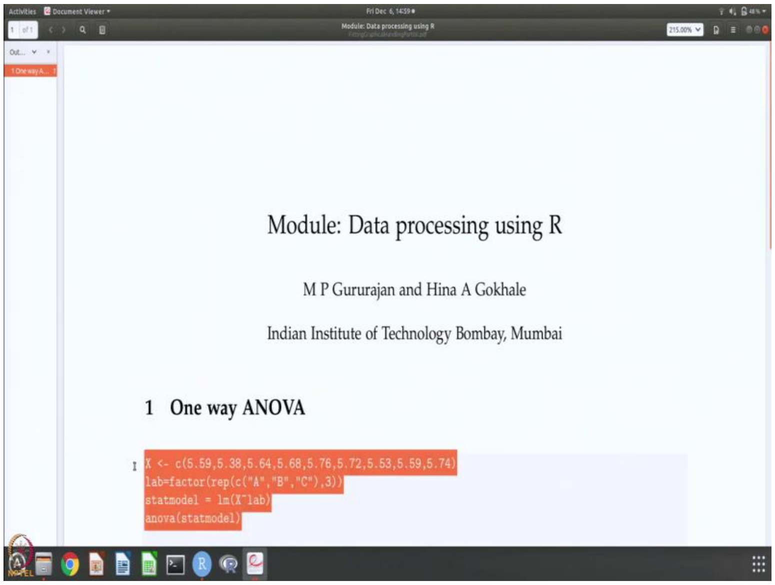
A reference

Practical regression and ANOVA using R, Julian J Faraway, 2002
Design of Experiments (DoE) and ANOVA: a case study

NPTEL
Guru and Hina
Dealing with Materials Data
IITB
6 / 6

We are going to do very simple one way ANOVA, but there is a good book called practical regression and ANOVA using R, by Julian Faraway and it is available for free at the R site, so you should download and take a look at this book. We will also go from here to design of experiments and ANOVA, but as a case study in the next module.

(Refer Slide Time: 03:03)



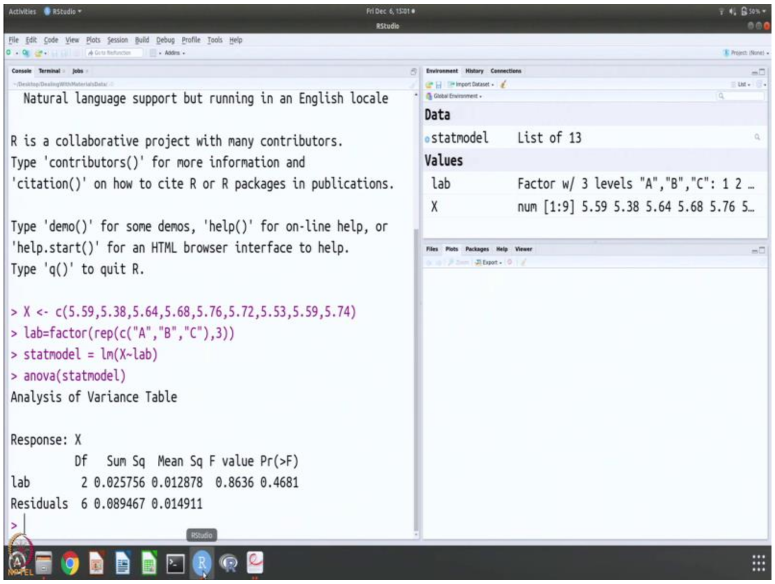
Module: Data processing using R

M P Gururajan and Hina A Gokhale

Indian Institute of Technology Bombay, Mumbai

1 One way ANOVA

```
X <- c(5.59,5.38,5.64,5.68,5.76,5.72,5.53,5.59,5.74)
lab=factor(rep(c("A","B","C"),3))
statmodel = lm(X~lab)
anova(statmodel)
```



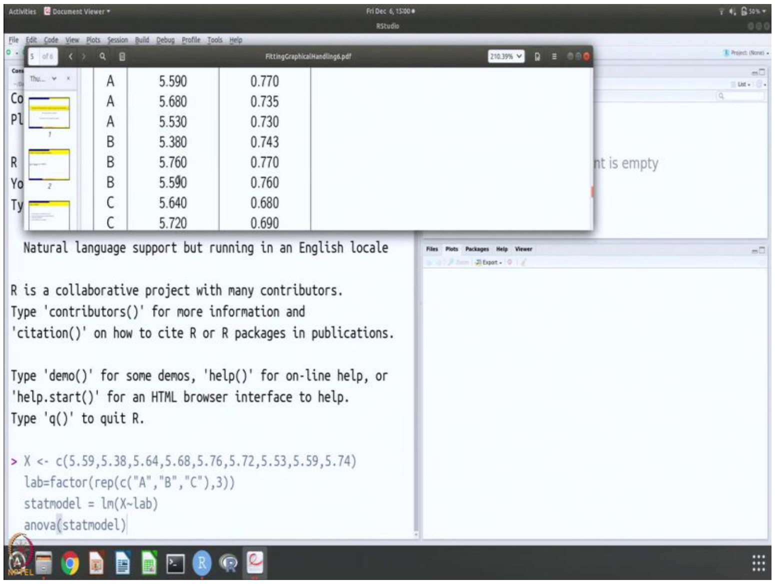
Activities
RStudio
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Console Terminal Jobs
Environment History Connections
Import Dataset
Global Environment
Files Plots Packages Help Viewer
Export

```
 Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> X <- c(5.59,5.38,5.64,5.68,5.76,5.72,5.53,5.59,5.74)
> lab=factor(rep(c("A","B","C"),3))
> statmodel = lm(X~lab)
> anova(statmodel)
Analysis of Variance Table

Response: X
          Df   Sum Sq  Mean Sq F value Pr(>F)
lab        2 0.025756 0.012878  0.8636 0.4681
Residuals  6 0.089467 0.014911
> 
```

```
Data
 statmodel      List of 13
Values
 lab            Factor w/ 3 levels "A","B","C": 1 2 …
 X              num [1:9] 5.59 5.38 5.64 5.68 5.76 5…
```

## Data

| Lab | Component X | Component Y |
|---|---|---|
| A | 5.590 | 0.770 |
| A | 5.680 | 0.735 |
| A | 5.530 | 0.730 |
| B | 5.380 | 0.743 |
| B | 5.760 | 0.770 |
| B | 5.590 | 0.760 |
| C | 5.640 | 0.680 |
| C | 5.720 | 0.690 |
| C | 5.740 | 0.690 |



Activities
Document Viewer
Fri Dec 6, 15:00
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
FittingGraphicalHandling6.pdf
210.39%
nt is empty
Files Plots Packages Help Viewer
Export

| | | |
|---|---|---|
| A | 5.590 | 0.770 |
| A | 5.680 | 0.735 |
| A | 5.530 | 0.730 |
| B | 5.380 | 0.743 |
| B | 5.760 | 0.770 |
| B | 5.590 | 0.760 |
| C | 5.640 | 0.680 |
| C | 5.720 | 0.690 |

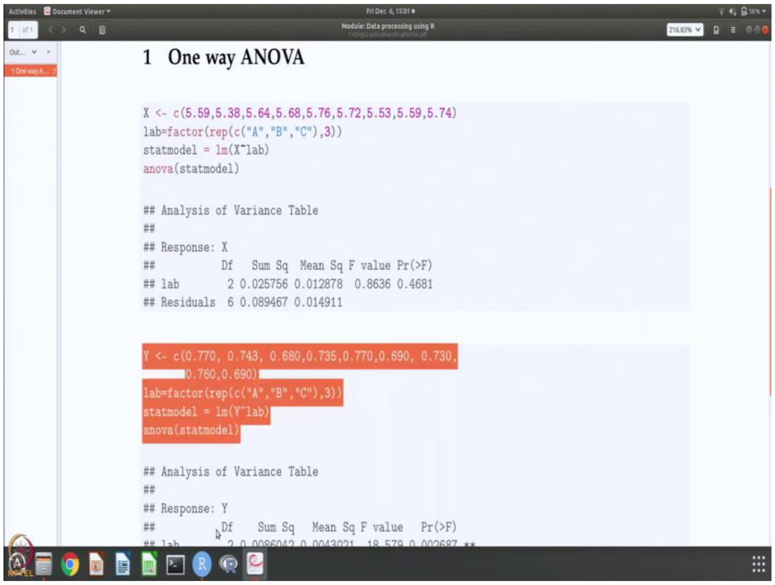
```
  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> X <- c(5.59,5.38,5.64,5.68,5.76,5.72,5.53,5.59,5.74)
  lab=factor(rep(c("A","B","C"),3))
  statmodel = lm(X~lab)
  anova(statmodel)
```

So, let us do this analysis and for doing that start R and let us get the so here is the analysis for the component X, the lab is three A, B, C and we have to give the data also in the corresponding order A, B, C, A, B, C because that is how the data will be read, if you look at the way, the data is represented in this table, it is all A first and all B next and all C third. But we are giving data as the A, B, C, so we should give 1 1 1 and 2 2 2 & 3 3 3.

So, this is important because if you do not give it in the right order, then you will get the wrong results, you can do this as a test also for yourself and you can see that the values are given 5.59 and then the 5.38 and then 5.64, so that is the first value for C, then 5.68, 5.76 and then the second value 5.64, 5.72 and then there is 5.53 and 5.59 and 5.74, so these are the values that we have read.

And the idea is very simple so develop a statistical model which connects the values to the labs and then do ANOVA on that model. So, you get this F value and basically the F value whether it is small or large tells you whether you can accept that within the statistical difference of these data as the same or that within statistical difference these data are significant.

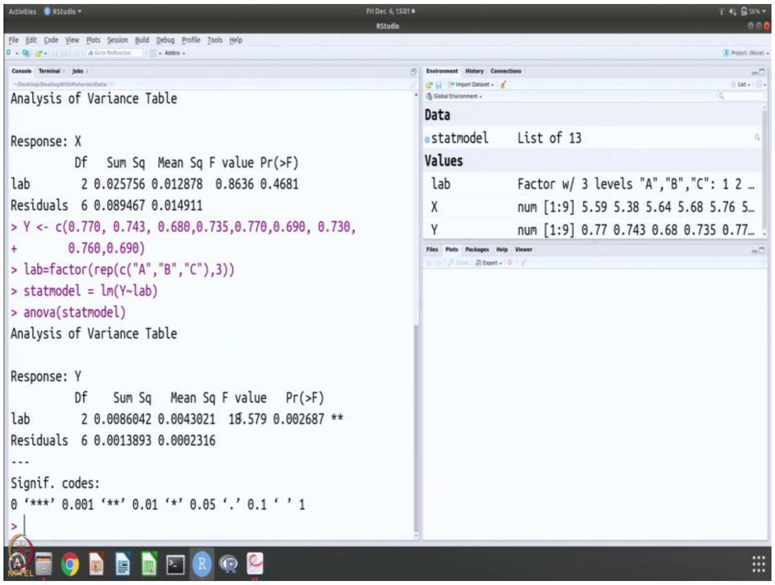
(Refer Slide Time: 05:02)



# 1 One way ANOVA

```
X <- c(5.59,5.38,5.64,5.68,5.76,5.72,5.53,5.59,5.74)
lab=factor(rep(c("A","B","C"),3))
statmodel = lm(X~lab)
anova(statmodel)

## Analysis of Variance Table
##
## Response: X
##           Df   Sum Sq  Mean Sq F value Pr(>F)
## lab        2 0.025756 0.012878  0.8636 0.4681
## Residuals  6 0.089467 0.014911
```

```
Y <- c(0.770, 0.743, 0.680,0.735,0.770,0.690, 0.730,
       0.760,0.690)
lab=factor(rep(c("A","B","C"),3))
statmodel = lm(Y~lab)
anova(statmodel)

## Analysis of Variance Table
##
## Response: Y
##           Df    Sum Sq   Mean Sq F value   Pr(>F)
## lab        2 0.0086042 0.0043021  18.579 0.002687 **
```



```
Analysis of Variance Table

Response: X
          Df   Sum Sq  Mean Sq F value Pr(>F)
lab        2 0.025756 0.012878  0.8636 0.4681
Residuals  6 0.089467 0.014911
> Y <- c(0.770, 0.743, 0.680,0.735,0.770,0.690, 0.730,
+        0.760,0.690)
> lab=factor(rep(c("A","B","C"),3))
> statmodel = lm(Y~lab)
> anova(statmodel)
Analysis of Variance Table

Response: Y
          Df    Sum Sq   Mean Sq F value   Pr(>F)
lab        2 0.0086042 0.0043021  18.579 0.002687 **
Residuals  6 0.0013893 0.0002316
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

```
Data
statmodel     List of 13
Values
lab           Factor w/ 3 levels "A","B","C": 1 2 …
X             num [1:9] 5.59 5.38 5.64 5.68 5.76 5…
Y             num [1:9] 0.77 0.743 0.68 0.735 0.77…
```

Let us, do it for Y, it is rather trivial, it is the same type of exercise, so the numbers have to be fed correctly and if you feed then you call for the statistical model and do one ANOVA. And you get the analysis of variance table, so you can see that F value is 0.86 here 18 here, so probably in this case the differences are significant and probably in this case the differences are not significant.

So, this is something that you have already seen in the other part of the course, this is just to show you that in R you can do and you can do it with just one line command and the book by Faraway gives you more information. So, to summarize fitting leads naturally to analysis of variance, because different labs can do the same experiment and do the fitting and give you the data.

Remember when we were looking at the data that was collected to by NIST from different sources, they did not do this, they mixed everything and then looked at the entire data and made an average out of it.

However, because different labs measure different ways and there could be small systematic differences between those measurements or statistically they could be really significantly different, that is not something that we did when we took the copper data, we just took all the data from all sources, mix them all up and did an analysis, that could be sometimes not quite correct and you have to do careful experiments and establish that statistically speaking the different labs or groups actually give you the same value and the way to do that is to use ANOVA.

In the case of NIST data of course the raw data is not available, so what was reported in the literature was what was taken and the analysis is done, so you can do that kind of analysis also and we will also see in our case study also that you can look at the literature take the numbers and do the analysis. But if you want to compare across different groups and labs, it is also essential or if you want to set up a standard that if you want to use the values from any lab and then decide whether it is acceptable or not and things like that, so it is very important to do this kind of analysis.

And with R of course you can do and you have learnt the idea behind this analysis in the other part of the course. So, we will, we have come to sort of end of this module, we will take up some case studies to better fix the ideas that we have learnt over these five different modules in the next module. Thank you.