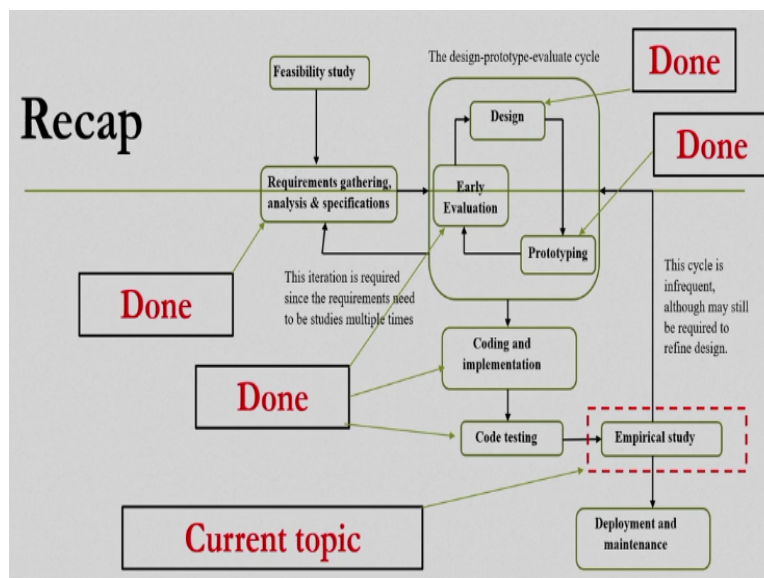


**Design and Implementation of Human-Computer Interfaces**  
**Dr. Samit Bhattacharya**  
**Department of Computer Science and Engineering**  
**Indian Institute of Technology – Guwahati**

**Lecture – 41**  
**Empirical Data Analysis**

Hello and welcome to the NPTEL MOOCS course on design and implementation of human-computer interfaces. We are going to start lecture number 35 where we will continue our discussion on the empirical study, the remaining portion of the discussion on empirical study. As is customary before we start let us quickly recap what we have learned so far and where we are now.

**(Refer Slide Time: 01:08)**



We are discussing interactive system development lifecycle, human-computer interfaces on interactive systems and when we try to develop them, we follow a particular stage wise approach which we are calling interactive system development lifecycle. It contains several stages. Each stage produces some outcome document. So, we have covered so far most of the stages that includes the requirement gathering, analysis and specification stage.

In this stage along with customer requirement, we also capture or gather and analyse the requirements of the end users from the point of view of usability, which is very important concern for interactive system development. Outcome of this stage is SRS or software requirement specification. Next, we have covered design stage, here primarily focused on interface and interaction design, outcome is a design document.

Then we covered prototyping, how to make prototypes out of the design document, what are the different types of prototypes, all these things we have covered, outcome is the prototype. Next, we covered quick evaluation of prototypes. Now, this evaluation is with the primary aim of unearthing usability issues. But here the evaluation is done by primarily quote unquote experts who are domain experts and here end users may not be part of the evaluation process, but it gives us a quick way of getting usability issues.

If usability issues are found, we can refine our design go for prototyping and evaluate again, so this cycle may continue till we arrive at a stable design. So, design prototype evaluate cycle we covered. Next, we cover design of the system where we have seen how we can come up with a modular system design using different approaches, we can follow either a function-oriented approach or object-oriented approach.

In the former case, we can use DFD to express the design system design in the latter case we can use UML to express the design. Outcome of these design stage is the system design document. Next, we covered coding and implementation stage. We learned about good coding practices and outcome of this stage as is obvious is the code, the system implementation which may include along with the code the documentation for the code also.

Next, we went through the details of code testing, several testing methods we have learned that includes review-based testing, then structural testing, functional testing. Functional Testing is also known as black box testing whereas structural testing is also known as white box testing. So, all these testing methods we have learned, outcome of this stage is testing document, testing reports.

Now, the testing purposes to find whether the code is executable. The next stage is empirical study where we try to ascertain whether the code is usable with the help of end users. So, currently we are discussing the stage empirical study.

**(Refer Slide Time: 04:48)**

## Empirical Study Stages

---

- Broadly, four stages
  - Identification of research question(s)
  - Determination of variables
  - Design of experiment
  - Analysis of empirical data

We have already covered several lectures on this topic, empirical study. So, the empirical study essentially refers to observation of user behaviour while they are asked to use our proposed system, in a very simple manner we can say this is an empirical study, we observe and then we analyse the observed data. Now, this is a very systematic approach. It is not to be done in a very random manner, ad hoc manner or casual manner.

In order to do it systematically, we divided into four stages. In the first stage, we try to identify the research questions. Second stage is determination of variables. In the third stage, we go for the design of the experiment and the fourth and final stages analysis of empirical data. So, these four stages together comprise the overall usability study process.

**(Refer Slide Time: 05:50)**

## Understanding the Stages

---

- We discussed first THREE stages
- In this lecture, we shall discuss last stage (data analysis)

In the previous lectures, we have covered the first three stages namely identification of research question, determinism of variables and design of experiment. In this lecture, we are going to cover the last topic or last stage that is analysis of data. Unless of course we analyse the data will not be able to come to any conclusion about the observation. So, analysis of data is a very important stage and we should be careful while analysing the data.

So, in this lecture we are going to learn about issues and challenges in data analysis as well as the methods that can be followed for analysis of empirical data. Let us begin our main discussion for this lecture. So, in order to understand the issues that are involved in data analysis, we will revert back to our earlier example.

**(Refer Slide Time: 06:42)**

**Basic Idea**

---

- Consider study to compare text entry speed of interfaces - (for RQ3)
- Let us decide to make use of twelve participants
- Also, instead of two, let's assume we compared text entry speed of our design with another 11 interfaces (total 12 interfaces)
- We performed a repeated-measure experiment and used Latin Square method for counterbalancing
- We have 144 data items ( $12 \times 12 = 144$ )

So, earlier we are talking of a text entry interface that we propose and we want to compare its performance with that of the existing systems. So, we framed a research question, research question number 3 that is whether our system is faster than MS Word. Now, earlier in the previous lectures, we mentioned about using 2 interfaces and 5 users to explain the concepts. For the sake of discussion here, we will expand that setup.

What we will do here, we will assume that there are 12 participants instead of 5 as we have seen earlier. Also instead of 2 let us assume for the sake of discussion that we compared the text entry speed of our design with another 11 interfaces. So, total there are then 12 interfaces, one is our design and the 11 are existing designs, earlier we talked about 2, one is our design and the other one was MS Word. Now, we are talking of 12 designs.

This is a hypothetical situation of course, so these can be any 11 interfaces, you can think of any 11 existing text entry interfaces as per your knowledge. In the last lecture, we learned about different experiment designs, we talked about within-subject or repeated-measure design as well as between-subject designs. So, let us further assume that we have conducted a repeated-measure experiment that is the within-subject experiment.

And in order to take into account the practice effect which we have discussed in details in the previous lecture, we used the Latin Square method approach for counterbalancing that is nullifying the effect of practice effect. To recap these concepts, you may refer to the previous lecture where we have discussed in details what is the Latin Square method and what is the idea of counterbalancing.

So, if that is the situation, we have 12 interfaces and 12 participants and it is a repeated measures design that means each participant perform text entry tasks on all the 12 interfaces. So, total we have 144 data items. Now, here each data item refers to the text entry speed. So, 12 into 12 or 144 text entry speeds we have collected.

**(Refer Slide Time: 09:10)**

**Basic Idea**

---

- A portion of data shown in Table

Participant	Interfaces											
	I <sub>1</sub>	I <sub>2</sub>	I <sub>3</sub>	...	I <sub>11</sub>	I <sub>12</sub>						
P <sub>1</sub>	3	1	2	...	5	4	2					
P <sub>2</sub>	4	2	3	...	4	4	3					
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮					
P <sub>12</sub>	2	2	2	...	3	5	2					

*empirical data*

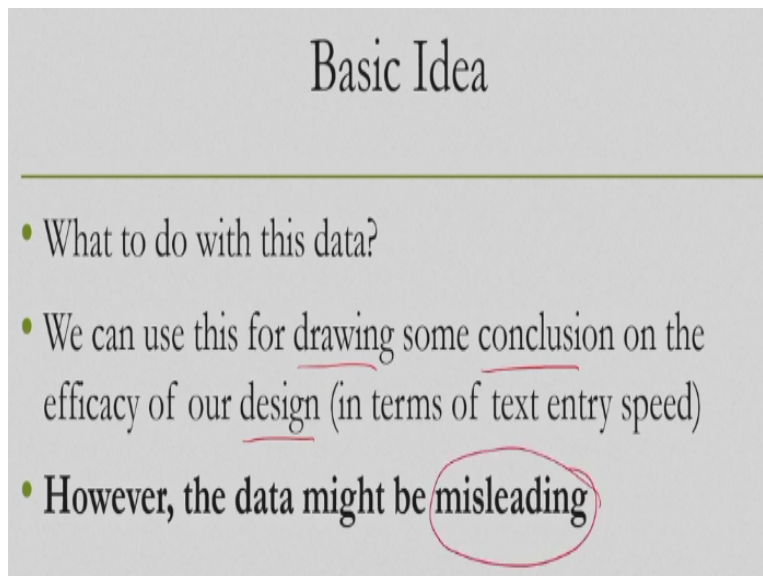
*n<sub>1</sub> n<sub>2</sub> ... n<sub>12</sub>*

Let us think of such a table. In the table on the one side we have interfaces, on the other side we have participants, participant 1, 2 to up to 12 and interfaces 1, 2 up to 12. So, each cell indicates the text entry speed that the participant achieved with the particular interface. So, for example 3 here in this cell indicates that the participant one performed text entry with interface 1 and achieved a speed of 3 characters per minute.

In that way we can interpret all the other values. So, this is the data we have collected. In other words, this is our so called empirical data. Now, the objective of any empirical study is to analyse this data to come to a conclusion. So, how we can analyse? Let us follow a very simple approach that is for each interface we can take an average value for these numbers that means for interface 1 average text entry speed obtained is say something  $n_1$ , this is  $n_2$  and in that way we get  $n_{12}$ .

Now, we can simply compare these  $n$  values, see the one with the maximum speed and then we can say that that particular interface provides maximum text entry speed. This sounds quite simple, fair and logical, but is it going to be the correct way to look at the data that is the question that we are going to understand in this overall lecture.

**(Refer Slide Time: 11:02)**



Basic Idea

- What to do with this data?
- We can use this for drawing some conclusion on the efficacy of our design (in terms of text entry speed)
- However, the data might be misleading

So, the question is what to do with this data? Definitely we can use this for drawing some conclusions on the efficacy of our design whether it is faster in comparison to other interfaces. As I said, we can simply take an average of each interface for all the 12 participants and draw a conclusion. But if we draw such a conclusion and conclude that particular interface is faster as compared to the other interfaces that need not be the correct conclusion, it may be actually misleading, why it is so? Why it would be misleading?

**(Refer Slide Time: 11:41)**

## Basic Idea

- We made use of twelve interfaces - represent only a fraction of all possible text input interfaces
- Twelve participants are also a very small fraction of all the potential users (even if we consider specific demographic profiles)

One thing is we made use of 12 interfaces. Now, these 12 interfaces represent only a fraction of all possible text input interfaces. So, when you draw a conclusion like our interface is faster than any other text input interfaces, ideally we should compare it with all possible text input interfaces that are available in the market, but 12 may be a very small number, in fact 11 because the other one is our own interface.

So, we have compared with 11 other existing interfaces, now that 11 may represent a very small fraction of all possible text input interfaces that are available in the market. We may not be aware of all such interfaces, so we cannot draw a general conclusion based on these 11 interfaces solely using a very simple method. Another concern is there are 12 participants who participated in our study.

Now, this is also a very small fraction of all the potential users. Suppose, we define the users to be the English speaking teenagers within the age group of 15 to 25, now if we employ say only 12 out of these population that represent a very tiny fraction, so conclusion drawn on the basis of these tiny fractions' behaviour did not be applicable to all the other members of the user group.

So, even if we are considering specific demographic profiles, still we may not be able to work with a large set of users belonging to that profile because of practical considerations. And can we really draw any generalizable conclusion based on that behaviour that we observed for that tiny fraction of the whole population, is that possible?

**(Refer Slide Time: 13:41)**

## Basic Idea

---

- We are dealing with samples rather than the actual population
- Where is the guarantee the observations were not due to chance

So, what it actually tells us is that we are dealing here with samples, samples of the interface, samples of all available text input interfaces, sample of the user group. So, we are dealing with samples rather than the actual population, actual population means all members of the particular groups, so all users or all interfaces represent the actual population. So, we are dealing with samples.

Now, the question is where is the guarantee that the observations were not due to chance, it may happen that whatever we have observed happened because of pure luck, pure chance instead of showing the actual behaviour of the user group.

**(Refer Slide Time: 14:36)**

## Basic Idea

---

- If we conduct another study with a completely different set of participants and interfaces, we may end up with a completely different data set leading to a different conclusion altogether

Why that is important? We need to ensure that suppose today we conducted one such experiment and observed the text entry speed of the participants, tomorrow we conduct



another study with a completely different set of participants and interfaces. The participants belonging to the same user group whereas the interface is belonging to the same group of text input interfaces. So, we want to ensure that tomorrow what we observe is not completely different from what we observe today.

That means tomorrow we are not going to end up with a completely different data set. If that is the case, then of course our conclusion will change and we are going to end up with a different conclusion than what we can conclude based on the data we have collected today. So, the issue is this. Today we are conducting a study, in this study we have employed some participants and used some interfaces and based on that study we got some data and we are analysing the data to conclude.

We want to ensure that the data that we have observed is not due to chance, it is because of the actual behaviour of the users, otherwise if we conduct a study tomorrow with a different group of users and different set of interfaces, then we may end up with another data set and analysing which may lead to a different conclusion, which of course is not desirable.

**(Refer Slide Time: 16:13)**

Basic Idea

---

- How likely is that possibility - we need to answer this question first
- Points to “statistical significance” of the data

So, how likely is that possibility that the data we got is not due to chance or rather the data that we got is purely by chance. So, we need to answer that question first. So, that is our primary concern in analysis of data. Whatever data is collected, whether that is reliable or whether that did not happen due to chance that we need to answer first. Now, to answer this question, we need to perform a statistical significance test.

And this particular question points to statistical significance of the data, in other words, the reliability of the data whether the data is statistically significant and for that we need to perform statistical significance test of the data. So, let us try to understand statistical significance.

**(Refer Slide Time: 17:03)**

Fundamental Concern

- We work with samples **BUT** wants to draw conclusion on larger population
- Thus, we wish to conclude if the result is applicable for any user and any text input interface
- Rather than applicable for only the twelve participants and interfaces

So, the fundamental concern here is we work with samples, small set of the actual population, but, a big but, what we want is we want to draw conclusion on the larger population or the entire population. So, we are working with samples, but we are trying to draw conclusion on the population, so that is our fundamental concern. Is that possible or what are the issues that we should answer or address before we can draw such a conclusion?

So, we wish to conclude if the result is applicable for any user and any text input interface in the specific context of our example rather than applicable for only the 12 participants and interfaces. So, in the example, we made use of 12 participants and 12 interfaces including ours the data we observed and collected.

If we analyse the data and come to a conclusion whether that conclusion is applicable to any interface, text input interface and any user belonging to that user group or it is specific to only the 12 participants and the 12 interfaces that we have considered. This is the fundamental constraint that we have.

**(Refer Slide Time: 18:29)**

## Fundamental Concern

- Necessary to determine *nature* of data
  - Is the data occurred *by chance*
  - Or due to *specially designed test conditions* (the *interfaces*) - often termed as *“treatment”*

So, in order to address this concern, it is necessary to determine the nature of the data. Is the data occurred by chance or the data occurred due to specially designed test conditions? That is in our case the interfaces, which is often termed as treatment conditions. This is a more popular term used that is the data occurred due to the treatment condition. So, in our case, specially designed test conditions where we made use of 12 interfaces. So, this nature of data, we first have to understand whether the data occurred by chance or because of the treatment condition.

**(Refer Slide Time: 19:04)**

## Basic Idea

- Statistical significance tests answer the question
- If we perform significance test on data and find the statistic is significant with  $p < 0.05$  (to learn soon), we can say with *confidence* that the data is due to the *“treatment”* and not by chance in *95%* of the times

In order to understand whether the data occurred by chance or due to the treatment condition, we need to go for statistical significance tests, special category of tests which answers this particular concern. If we perform significance test on data items and find that the statistic is

significant with  $p$  less than point 0.05, we can say with confidence that the data is due to the treatment condition and not by chance in 95% of the times.

This is a very simple interpretation of the statistical significance test outcome. So, we will learn about these terms what is  $p$ , what is this value 0.05. But if we find something like this, then we can conclude that the data happened due to the treatment condition and not due to chance or data is going to happen, the kind of data that we observed that event is going to happen in 95% of the cases and in 5% of the cases it may happen due to chance.

So, if we conduct the experiment 100 times, in 95% of the times we are going to get similar data whereas in 5% of the cases there may be deviation from the kind of data that we have observed which may happen due to chance. So, in this way we can actually conclude what is the confidence with which we can say that data is not by chance.

**(Refer Slide Time: 20:38)**

Basic Idea

- Our starting point is a statistic
- In the context of user-centric research, it is the mean (or average) of a group of data items (the sample)
- Typically, we are interested to test the significance of the difference between the means of several groups of data (in most situations)

So, here our starting point is a statistic. In the context of user centric research, it is the mean or average of a group of data items. Now, the data items are essentially samples or collected from samples and we are dealing with the mean of the data items. Typically, we are interested to test the significance of the difference between the means of several groups of data and this is the situation in most of the cases.

So, when we are analysing empirical data, in most of the cases we are interested to know about the difference between the means of multiple groups and whether those differences are statistically significant or not. So, note that we are not directly dealing with averages, we are

computing our ages are calculating mean or averages, but then the difference between averages is what we are more interested in than the absolute value of the averages. Let us try to understand this with example.

**(Refer Slide Time: 21:40)**

### Basic Idea

---

- Ex - let us assume only two interfaces (as originally done in RQ3)
- There are twelve participants as before
- We get two “groups” of text entry speed data

So, let us assume that only two interfaces are there as originally done in RQ3 instead of 12 that we started with. Now there are 12 participants as before. So, then we get two groups of text entry speed data, one group belongs to text entry speed for one interface and other group belongs to the other interface.

**(Refer Slide Time: 22:07)**

### Basic Idea

---

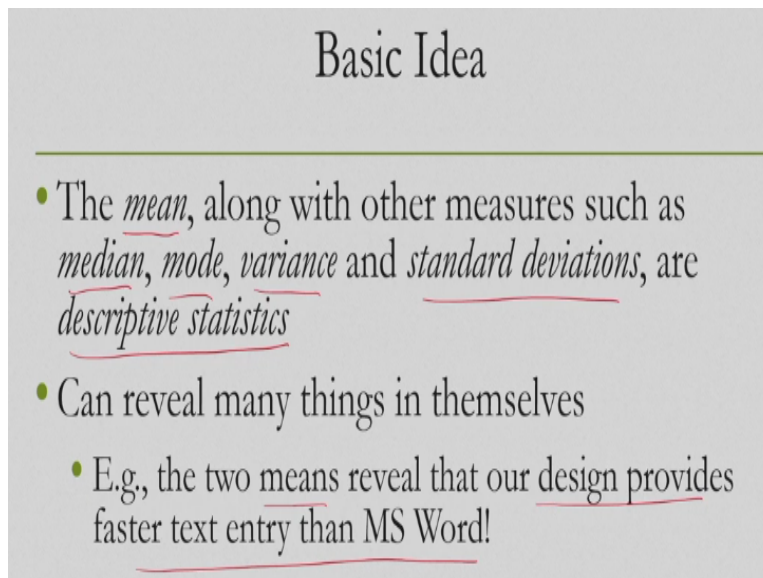
Participants	P <sub>1</sub>	P <sub>2</sub>	P <sub>3</sub>	P <sub>4</sub>	P <sub>5</sub>	P <sub>6</sub>	P <sub>7</sub>	P <sub>8</sub>	P <sub>9</sub>	P <sub>10</sub>	P <sub>11</sub>	P <sub>12</sub>	Group Mean
Our design	3	2	3	4	3	3	2	5	5	3	2	4	3.25
MS Word	2	1	1	3	3	4	2	2	1	1	3	1	2.00

So, we can represent the data in this way. So, there are 12 participants, P 1 to P 12. Now, there are two groups of data, one group belongs to our design where these participants produce some text entry speed based on the tasks that are given to them and then these are the

values. So, there are 12 values in this group and we computed the mean of this group which is 3.25. Similarly, for the other interface MS Word there are 12 values and this is the group mean 2.

Now, here you can see that there are two mean values, one is 3.25 for our design and 2 for MS Word. We may be tempted to immediately conclude that since our design produces a higher speed on average, so it is higher than the MS Word, this may be a very tempting way to conclude immediately based on the group means.

**(Refer Slide Time: 23:12)**



Basic Idea

- The *mean*, along with other measures such as *median*, *mode*, *variance* and *standard deviations*, are *descriptive statistics*
- Can reveal many things in themselves
  - E.g., the two *means* reveal that our *design provides faster text entry than MS Word!*

Now, the mean along with other measures which you may be already aware of such as median, mode, variance, standard deviations these are all called descriptive statistics. They of course on their own can reveal many things. For example, two means reveal that our design provides faster text entry than MS Word, but that is what apparently it seems. But can we really conclude like that?

**(Refer Slide Time: 23:44)**

## Basic Idea

---

- Can't have such general conclusion
- Same reason - where is the guarantee that the difference shall remain the way it is (i.e., group mean for our design is **greater** than other group mean)?

Problem is we cannot have such a general conclusion unless we perform some more statistical significance tests. Because of the same reason, where is the guarantee that the difference between the mean shall remain the way it is if we conduct the experiment with different setup that is different interfaces and different group of participants. Now here that we are talking of difference in group means.

So, group mean for our design is greater than other group mean that is what we observed in this study, but where is the guarantee that this relationship will hold in another study that we may conduct tomorrow? We have to first ensure that before drawing the conclusion that our design is faster than MS Word.

**(Refer Slide Time: 24:41)**

## Basic Idea

---

- We are **not concerned about absolute values** (may change)
- Rather about **relative difference** between them
- The only way to be **confidant** of the **reliability** of the observations is to go or the **significance tests**

So, one thing we should keep in mind is that we are not concerned about absolute values, those values may change. Rather what we are concerned about is relative difference between those values. Now, the only way to be confident of the reliability of the observations is to go for the significance test. So, the only way to know whether this difference between absolute value shall hold in any other experiment that we conduct tomorrow is to go through the significance test of the data that we have collected.

**(Refer Slide Time: 25:18)**

Basic Idea

- Significance tests are performed for hypothesis testing

Statistical significance { GM1 - 100  
GM2 - 5  
Diff maybe significant

A point to be noted here is that we perform significance tests. Here the term significance is not the general term significance as in English, but here we have to understand it as statistical significance. So, always remember that whenever we use the term significance, we are referring to the term statistical significance. Now, these two are different. For example, we may say that, suppose there is a group mean 1 with the value 100 and another group mean 2 with value 5.

This is obtained after some data collection process that we followed. So, the difference between them may look like significant. Difference may be significant it may appear, but it was so happened that they are not statistically significant that is although the difference between them is 95 which looks like a significant difference between these two numbers, statistically this difference may not be significant as we shall soon see.

So, what may look like significant may not be statistically significant and what may be statistically significant may not look like significant. So, both way you have to keep this in mind.



(Refer Slide Time: 27:04)

### Basic Idea

---

- Ex – consider the research question RQ3  
Does the new interface let me enter text “faster” than MS Word?
- Corresponding hypotheses (null and alternative)
  - H<sub>0</sub>: Our design is not faster than MS Word.
  - H<sub>1</sub>: Our design is faster than MS Word.

Now, as I said we perform statistical significance tests for hypothesis testing. Let us see how we can do that. Let us consider the research question RQ3 that is does the new interface let me enter text faster than MS Word that was our original research question. So, the corresponding hypothesis which is the null hypothesis as well as the alternative hypothesis, there can be two, recall our discussion on hypothesis.

The null hypothesis is denoted by H 0, our design is not faster than MS Word and the alternative hypothesis denoted by H 1 is our design is faster than MS Word, suppose these two are the two hypotheses we started with.

(Refer Slide Time: 27:54)

### Basic Idea

---

- With the significance test, we try to refute the null hypothesis
- Let us try to understand the process with one simple (probably the simplest) statistical significance test

With the significance test, we try to refute the null hypothesis that is the primary objective of significance test. Now, let us try to understand the process with one simple, probably the simplest statistical significance test, let us try to understand this.

**(Refer Slide Time: 28:21)**

### Paired Sample t-test

---

- If you perform the test, you are likely to get something like Table

✓ <u>The difference in group means = 1.25</u>
✓ <u>The two-tailed p value = 0.020583</u>
✓ <u>t = 2.702015</u>
✓ <u>Degrees of freedom (df) = 11</u>
✓ <u>The difference is statistically significant</u>

This is an example of a t-test which is one of many available significance tests. Suppose, we have performed the t-test, a paired sample t-test. Once you perform the t test, you are likely to get something like what is shown in this table. The difference in group means 1.25. The two-tailed p value 0.02058,  $t = 2.702015$  some number, degrees of freedom df 11. And the conclusion is the difference is statistically significant. So, what it tells?

**(Refer Slide Time: 28:55)**

### Paired Sample t-test

---

- Although everything in Table is important, we do not report all these information
- Difference of the means is found to be statistically significant - we simply report this fact

Now, in the table as you can see there are several entries. You will get this table if you perform the t-test with some automated tool. So, there are several entries in the table. First

one is of course the actual difference, then the p value, t value, the degrees of freedom value and the final conclusion. So, all of these need not be reported in this same manner, the same way that is shown in that table.

But all these entries are of course very important. Difference of the means is found to be statistically significant that is the final conclusion that is written in the table. We simply report these facts along with some other information that we shall see. Now reporting this conclusion at some form, so generally you will find the conclusion is reported with this type of format.

**(Refer Slide Time: 29:50)**

Paired Sample t-test

---

- With a specific format

$t(11) = 2.70, p < 0.05, \text{statistically significant}$

For a paired sample t-test we use the symbol t within parenthesis 11 that is the degrees of freedom df equal to the value of the t that is 2.70, then we report p less than 0.05 and statistically significant. So, this is the notion that typically is used to concisely report everything that is present in the table. So, this is about how to report what is the format in which we should report.

**(Refer Slide Time: 30:28)**

## Paired Sample t-test

- With a specific format

$$t(11) = 2.70, p < 0.05, \text{ statistically significant}$$

*df*      *.05*  
*0.05*

In the format, as you can see we use the lowercase t. After it, we put degrees of freedom df that is 11 in our case within parentheses and without any space in between this is important. This is followed by an equal to symbol followed by the t-statistic which in our case is 2.70. Then we put a comma and a space character finally we write the p value as p less than 0.05. This is the notation typically used.

Of course, there may be some other notations used, but this notation is more commonly used, so you are advised to use this notation that we have just discussed, namely in this format t 11 without any space and within parentheses equal to t-statistic value comma space p less than 0.05, do not write it .05, you should always write it 0.05 and then comma and then the conclusion statistically significant.

**(Refer Slide Time: 31:42)**

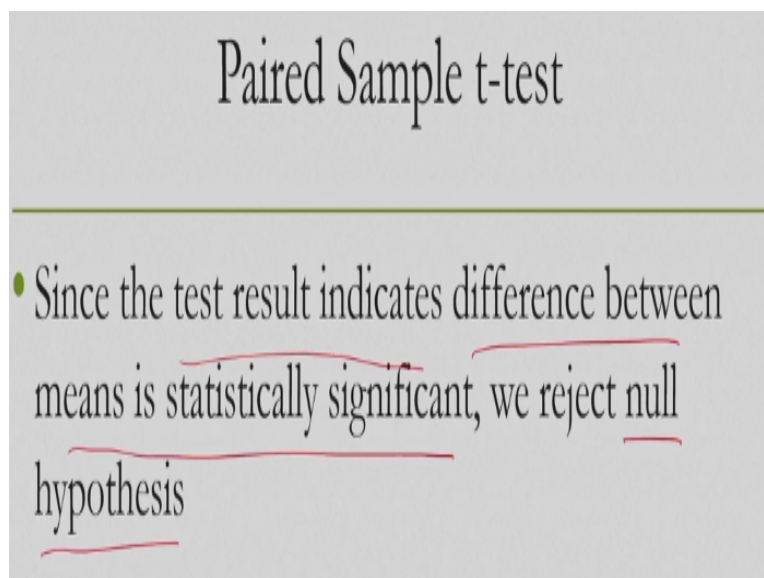
## Paired Sample t-test

- The 'p' value roughly indicates the probability that the data occurred by chance
- The value '0.05' is a pre-defined value
- It indicates that the probability of getting the results by chance is about 5%

Now, let us try to understand the significance of these notations. The p value roughly indicates the probability that the data occurred by chance. So, the value 0.05 is a predefined threshold value, generally we use this value to conclude whether the possibility of the data occurring by chance is less than 5% or not. So roughly we can interpret it in this way. So, p indicates roughly the probability that the data occurred by chance.

And 0.05 is the threshold value above which we generally do not consider the data to be statistically significant, generally. So, if we are using the value 0.05, then that indicates that the probability of getting the results by chance is about 5% or less that is the meaning of this threshold value and the significance of the notation p.

**(Refer Slide Time: 32:46)**



Since the test result indicates difference between means is statistically significant, we reject the null hypothesis. So, if we find that the result is statistically significant, then we can reject the null hypothesis and establish the alternative hypothesis.

**(Refer Slide Time: 33:05)**

## Paired Sample t-test

---

- It may be noted that the tests run the risk of **TWO** types of **errors**

Now, one issue that we have to keep in mind is that the tests run the risk of two types of errors. So, it is not that there is no issue with this type of testing, so there are maybe two types of errors and we have to be aware of it and we have to be careful about data analysis using statistical significance tests.

**(Refer Slide Time: 33:33)**

## Paired Sample t-test

---

- **Type I error** (also known as the  $\alpha$  error or “false positive”)
  - Occurs when we reject a null hypothesis, which is true and should not be rejected
  - To avoid Type I errors, we typically use a very low value of p (e.g,  $p < 0.05$ )

There is a type 1 error which is also known as the alpha error or false positive. Now, this type of error occurs when we reject a null hypothesis, which is true and should not be rejected that may happen. To avoid type 1 errors, typically use a very low value of p which is  $p = 0.05$ . So, if you set p to be higher, then you run the risk of having type 1 errors. So, what is the type 1 error? To recap type 1 error is the error when we reject a null hypothesis which should not have been rejected.

**(Refer Slide Time: 34:17)**

## Paired Sample t-test

- **Type II error** (also known as the  $\beta$  error or “false negative”)
  - Indicates a situation where we do not reject a null hypothesis although it is false and should have been rejected
  - To avoid, it is generally recommended to go for larger sample sizes

Then there is a type 2 error. It is also known as beta error or false negative. So, type 1 error is called alpha error or false positive and type 2 error is called beta error or false negative. So, this type 2 error indicates a situation where we do not reject a null hypothesis, although it is false and should have been rejected. So, this is just the opposite of type 1 error. In type 1 error, we rejected the null hypothesis which should not be rejected.

Here in type 2 error, we are not rejecting a null hypothesis we should have been rejected. Now to avoid type 2 errors, it is generally recommended to go for larger sample sizes so that amount of data items should be larger to avoid type 2 errors. So, we should be careful about these types of errors possibility of them occurring in our conclusion and accordingly take corrective actions.

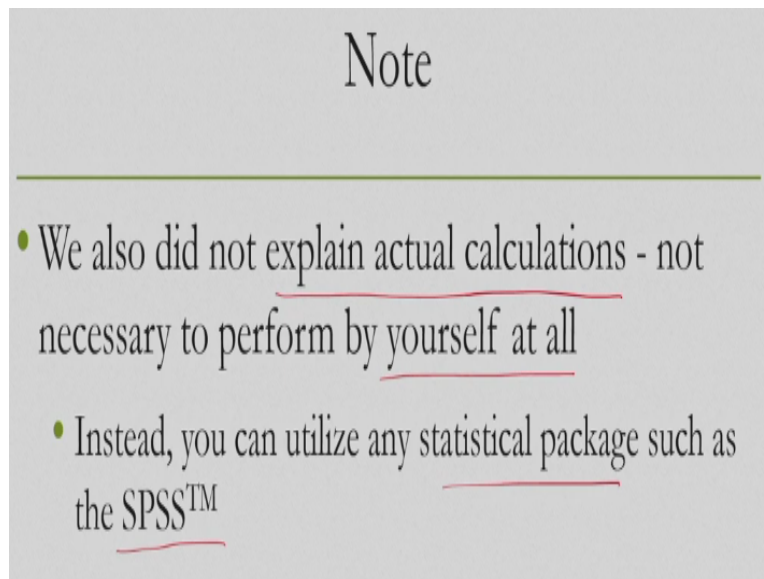
**(Refer Slide Time: 35:24)**

## Note

- Many important terminologies and issues (e.g., *p-value*, *two-tailed distribution*) we mentioned in the passing
- Our aim is to introduce the idea - for more details, reference may be consulted

So, many important terminologies and issues such as p value, two-tailed distribution, etc., we mentioned in passing, so rather casually. Our aim is to primarily introduce the idea rather than going for an in depth study of statistical significance test. So, if you want to know more about these things, more precise definitions of p values or these ideas of two-tailed distributions, what they mean, what they signify then you may refer to the reference books that are going to be mentioned at the end of this lecture.

**(Refer Slide Time: 36:05)**



Note

- We also did not explain actual calculations - not necessary to perform by yourself at all
- Instead, you can utilize any statistical package such as the SPSS™

Also, we did not explain the actual calculations here, we just reported the values that is to keep things simple not complicate with unnecessary details. In fact, the calculations you need not perform by yourself at all, there are already tools available which you can make use of such as SPSS tool or any other statistical packages. So, we can make use of any of these, just feed them your data and ask them to calculate the particular statistic.

It will automatically be done with all the details that we have just discussed. So, those are the two things that we explicitly did not cover, one is the precise definitions as well as more details about the specific terms that we used like the p value or the two-tailed distribution, these are some terms that we have used, but those details we have not covered here, you may refer to the references that are going to be mentioned at the end.

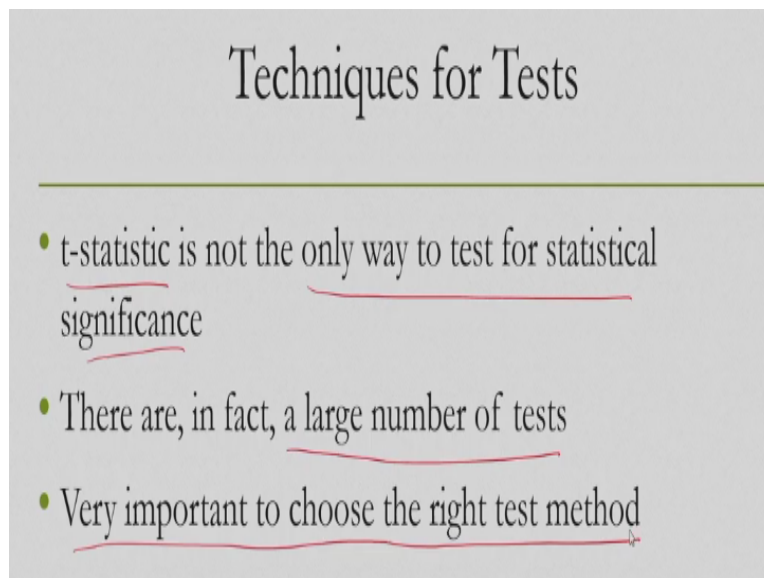
Also the actual calculations we did not discuss here in details; how to perform those calculations, how to get the t-statistic in particular or some other statistic that we are going to learn soon, for that statistical packages are available, you may choose any of the statistical packages and make use of the features provided to get those values automatically. Now, let us



move to our next topic of this lecture that is different techniques or t test that we have just seen is only one of many statistical significance tests that are possible.

Let us have a quick look at different techniques and which technique is applicable in which situation that also is very important to know, we are going to learn that also in brief.

**(Refer Slide Time: 37:54)**

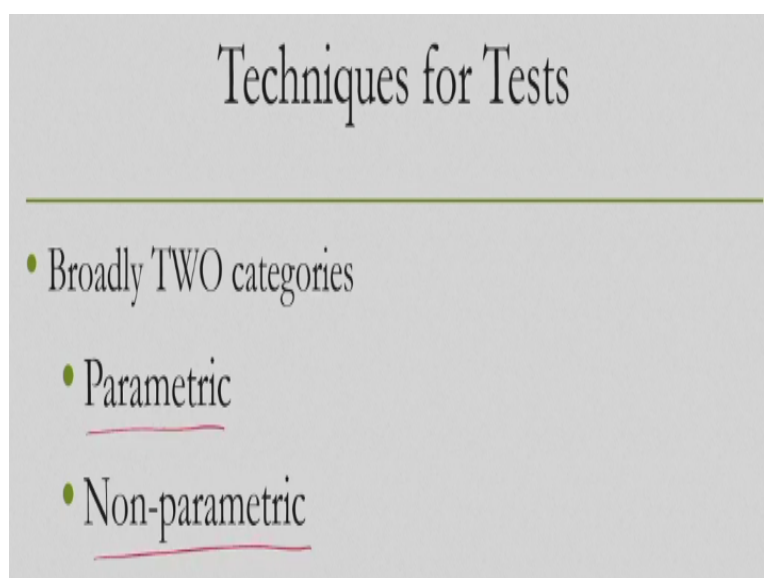


Techniques for Tests

- t-statistic is not the only way to test for statistical significance
- There are, in fact, a large number of tests
- Very important to choose the right test method

So, the t-statistic that we have just used in our example is not the only way to test for statistical significance as I just mentioned. There are in fact a large number of such tests available and very important to choose the right test method. So, given the data it is of course not advisable to randomly pick up a testing method and apply it on your data, you have to very carefully choose a testing method for analysis of a particular data set.

**(Refer Slide Time: 38:36)**



Techniques for Tests

- Broadly TWO categories
  - Parametric
  - Non-parametric

Now, the different techniques that are available for statistical significance test can be broadly categorized into these two groups, parametric tests and non-parametric tests.

**(Refer Slide Time: 38:50)**

## Parametric Tests

---

- t-test is one of the many methods collectively known as the parametric methods
- Applicable subject to the fulfilment of THREE conditions

Now, t-test is an example of the many methods that are collectively known as the parametric methods or parametric significance tests. Now, these tests are applicable subject to the fulfilment of three conditions. So, we should first check whether these three conditions are satisfied, then only we should go for one of the several parametric tests.

**(Refer Slide Time: 39:15)**

## Parametric Tests - Conditions

---

- Data should come from a “normally distributed” population
- We should use at least an interval scale (with equally-spaced intervals) of measurement for the dependent variable (a ratio scale is even better)
- Variance in the groups of data should be approximately equal

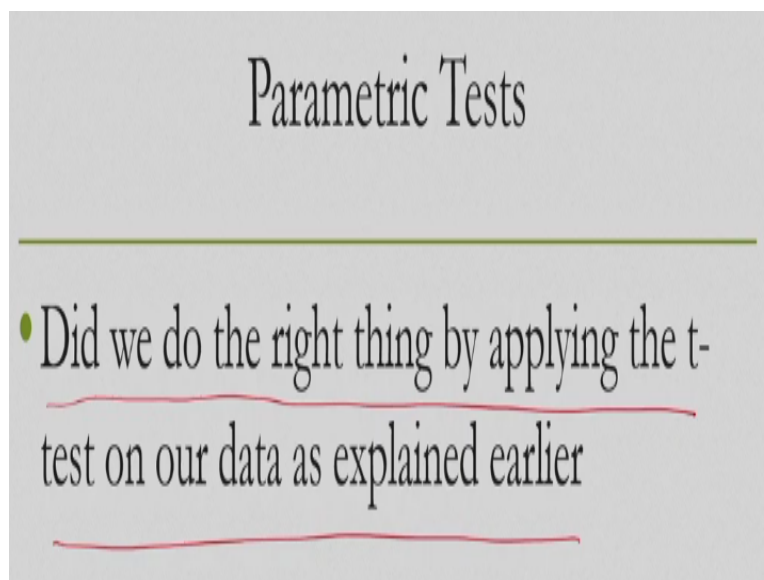
Now, what are the conditions for application of parametric tests, parametric significance test or other parametric statistical significance tests? Now, the data should come from a normally distributed population. So, whatever data we are dealing with should have come from a normally distributed population that is the first condition that has to be satisfied. Second is we

should use at least an interval scale of measurement to record data that means with equally spaced intervals to record data for that dependent variable.

A ratio scale is even better. So, we should preferably use ratio scale, if that is not possible, then at least interval scale should be used to measure and record the data. Otherwise, whatever data we record using the other scales nominal and ordinal are not useful for application of parametric testing methods. Variance in the groups of data should be approximately equal, this is another important condition.

So we are dealing with several groups of data and the variance in those groups should be approximately the same or equal. So, these three conditions must be satisfied if we are to use parametric methods for testing statistical significance of the data that we have collected.

**(Refer Slide Time: 40:51)**



So, let us see. Did we do the right thing by applying that t-test on our data as explained earlier. So, we simply used the t-test without actually bothering to check or apparently without bothering to check whether these three conditions were satisfied with our data. So, did we do the right thing? Let us see.

**(Refer Slide Time: 41:12)**

## Parametric Tests

- We used a ratio scale to record text entry speed (CPM)
- Thus, we are not violating the second condition for a parametric test – what about the other conditions
- Let's make an assumption

So, in our data collection method, we use the ratio scale that is the text entry speed. So, we are not violating the second condition for a parametric test that is the use of ratio scale. Now, what about the other two conditions that is whether the data came from a normally distributed population and whether the variance in the groups are almost equal, let us make some assumption to understand this.

**(Refer Slide Time: 41:41)**

## Assumption

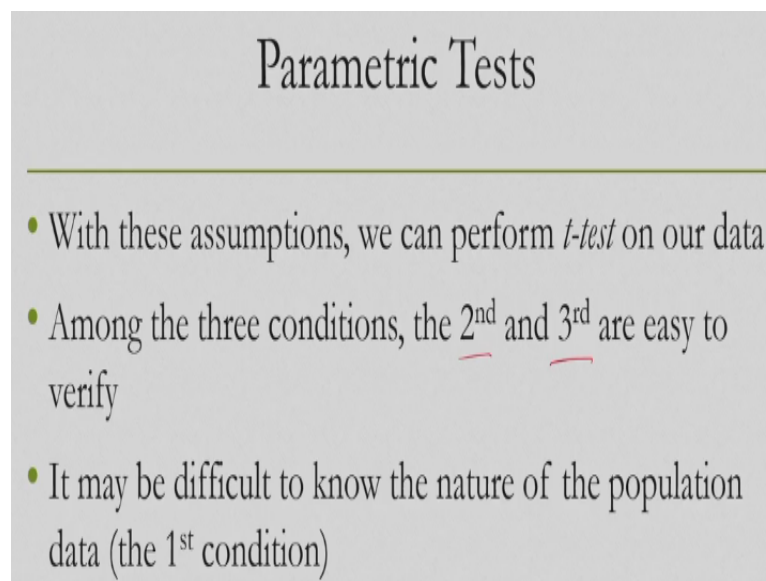
- Text entry speeds follow a “normal” (Gaussian) distribution
- Distribution refers to entire population – we plot the speed by all the users (if that is possible)
  - Our sample data (speeds of twelve participants) drawn from a normally distributed population
  - Sample data in itself need not be normally distributed

Text entry speeds follow a normal or Gaussian distribution that is one assumption we are making. Now, when we talk of distribution, the distribution is used to refer to the entire population. So, how we get this distribution? We plot the speed of all the users, all possible users if that is possible. Now, our sample data that is the speed of 12 participants drawn from a normally distributed population and sample data in itself need not be normally distributed.

So, we actually do not have in this case any way to know whether the distribution that we are likely to see for the entire user population that is their text entry speed is going to be normally distributed or not. Let us make that assumption that it is from normal distribution. Now, we are drawing the samples from that distribution.

So, even if this particular samples do not follow normal distribution that does not matter, it is drawn from the normal distribution as per our assumption, so we can say that it holds the first condition if we assume that to be correct, in general that kind of assumptions hold.

**(Refer Slide Time: 42:57)**



Parametric Tests

- With these assumptions, we can perform *t-test* on our data
- Among the three conditions, the 2<sup>nd</sup> and 3<sup>rd</sup> are easy to verify
- It may be difficult to know the nature of the population data (the 1<sup>st</sup> condition)

And of course, whether the variance is approximately equal or not we can simply calculate and in the kind of data that we have reported, you can check whether the variance between the two groups are approximately equal or not. So, it is of course easy as you can see from this example to check for conditions 2 and 3, second and third conditions are easier to check that is whether the variance are approximately well and whether we have used the at least an interval scale.

Here of course, we have used ratio scale which is anyway satisfying the condition and we can check the variance. So, second and third condition checking is generally easy. Problem is with the first condition that is whether the data that we have collected comes from a normally distributed population because it is not possible to capture data for all users belonging to that population, so it is generally difficult.

**(Refer Slide Time: 44:01)**

## Parametric Tests

- If you are not sure, you can go for additional tests such as the Shapiro-Wilk or the Kolmogorov-Smirnov test
- Can reveal if the sample data is taken from a normally distributed population

Now, if there is some element of uncertainty, then we have some way out. What we can do is we can go for additional tests, some additional statistical tests such as Shapiro-Wilk test or the Kolmogorov-Smirnov test. So, these tests again you did not do it yourself, the statistical packages can do it on your way up. These tests will reveal if the sample data is taken from a normally distributed population or not.

So, either we can assume based on our intuition or we can go for these additional tests to check whether the first condition holds. If all the three conditions hold, then only we can go for parametric tests on our data.

**(Refer Slide Time: 44:54)**

## Non-Parametric Tests

- If your data do not support any one of the above three condition, you should go for the non-parametric tests of significance

The other category of tests is called non-parametric tests. If your data do not support any one of the three conditions that we just mentioned, then you should not use parametric testing

methods, instead you should go for non-parametric statistical significance tests. So, any one of these three conditions if violated, then you should go for non-parametric testing methods.

**(Refer Slide Time: 45:21)**

Parametric test	Experiment design	Non-parametric test
<del>Independent-samples t-test</del>	Between-subject design with one factor having two levels.	Chi-square test
<del>Paired-sample t-test</del>	Nominal (categorical) scale of measurement	
<del>One-way ANOVA</del>	Within-subject design with one factor having two levels.	McNemar's test
<del>Factorial ANOVA</del>	Nominal (categorical) scale of measurement	
Independent-samples t-test	Between-subject design with one factor having two levels.	Man-Whitney U test
Paired-sample t-test	Within-subject design with one factor having two levels.	Wilcoxon signed ranks test
One-way ANOVA	Between-subject design with one factor having more than two levels.	Kruskal-Wallis test
Factorial ANOVA	Between-subject design with two or more factors, each having two or more levels.	
Repeated measure ANOVA	Within-subject design with one factor having three or more levels. Also applicable in within-subject design with two or more factors, each having two or more levels.	Friedman test

So, let us summarize then our discussion so far. So, which kind of test we are going to use depends on our experiment design. So, if we are designing an experiment following the between-subject design approach with one factor having to levels, remember factor means independent variable, levels mean the values that it can take. So, this is our experiment design set up between-subject design with one factor and two levels, then we can go for a non-parametric test that is Chi-square test.

There is no corresponding parametric test available, so this is not possible. If we are using nominal categorical scale of measurement with this kind of experiment design that is between-subject design with one factor having two levels and nominal or categorical scale of measurement, then we should go for Chi-square test this is one nonparametric test. So, there is no corresponding parametric test available for this type of design.

If we are using within-subject design with one factor having two levels like before and nominal or categorical scale of measurement, then we can go for McNemar's test, this is another non-parametric test. Again, there is no corresponding parametric test available for this experiment design. So, in the first case when we are talking about between-subject design, one factor two levels and nominal scale of measurement, we go for Chi-square test.

In the second case when we are having within-subject design one factor two levels and nominal scale of measurement we go for McNemar's test, both are non-parametric tests. When we are having a design that is between-subject with one factor having two levels, we can have independent samples t-test if the three conditions get satisfied as parametric test, and if not then Man-Whitney U test that is a non-parametric test.

Within-subject design with one factor having two levels, paired sample t-test that is a parametric test and in case of the conditions not getting satisfied we can have Wilcoxon signed ranks test. If we have between-subject design with one factor having more than two levels then we can go for one-way ANOVA as parametric test and Kruskal-Wallis test as non-parametric test.

So, parametric test is applicable when this general design is there plus the three conditions are satisfied and non-parametric test is applicable when either of the three conditions is not satisfied. When we have between-subject design with two or more factors each having two or more levels, then we can have factorial ANOVA, there is no corresponding non-parametric test available here.

Then so when we have between-subject design with two or more factors each having two or more levels for parametric case we can go for factorial ANOVA, for non-parametric case the same Kruskal-Wallis test is applicable. When we have within-subject design with one factor having three or more levels, then we can have repeated measure ANOVA under the parametric tests and Freidman test under the non-parametric test.

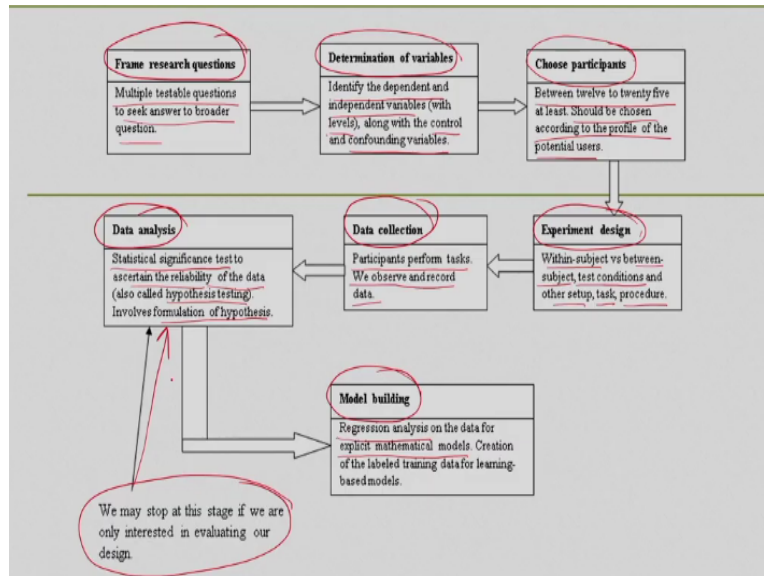
Now, this is also applicable in within-subject design with two or more factors each having two or more levels. So we have several parametric tests like independent sample t-test, paired sample t-test, one-way ANOVA, factorial ANOVA, repeated measure ANOVA; these are parametric tests. And under non-parametric we have Chi-square test, McNemar's test, Man-Whitney U test, Wilcoxon signed rank test, Kruskal-Wallis test, Freidman test.

And for each test we have to first ensure that the particular experimental design as mentioned in this table is satisfied and the three conditions that we have mentioned earlier, either they are satisfied or not satisfied. If they are satisfied, then we go for parametric test; if not then we go for non-parametric test. So, this is in summary different tests that are available and the



situations in which we are going to apply those tests. So, now let us summarize the entire workflow of empirical study.

**(Refer Slide Time: 50:21)**



So, first thing is we frame research question. So, generally we are supposed to frame multiple testable questions to seek answer to broader question as we have already discussed in details. This is followed by determination variables. What we do here? We identify the dependent and independent variables with their levels. Now, independent variables are also called factors and the values that they take are called levels along with the control and confounding variables.

So, I hope you remember the idea of control variables, the variables that we use as constants and confounding are those variables whose existence we are not aware of. Then we go for choosing our participants. Between 12 to 25 participants should be chosen to carry out a reliable test with reliable conclusion. Should be chosen according to the profile of the potential user, so they cannot be randomly taken.

First you have to identify the profile and from that group of users whose profile matches with the profile that you have identified, you can choose the participants Next you should go for experiment design. You have to decide whether you should go for within-subject or between-subject design depending on the availability of resources, volunteers or participants, etc. Also, you have to identify the test conditions and other setup, decide on the tasks and the procedure for data collection.

Next is the data collection, actual data collection where you ask the participants to perform the tasks and the data they produce you observe and record using either of the recording or measurement scales. The next is data analysis that is here you go for statistical significance test to ascertain the reliability of the data that is whether they happen by chance or due to the test condition. Now, this is also called hypothesis testing, which involves formulation of hypothesis.

So, although we discussed it during the research question framing, so actually it is required when we go for that statistical significance testing of the data. These are the steps that we have covered in the previous lectures. There may be one additional step which is not the focus of this course as well as this lecture, but I will just mention it here that is model building. So, once you are satisfied that the data is not due to chance, what you can do?

You can perform some regression analysis on the data between the dependent and independent variables to come up with explicit mathematical models or you can use the data for training in some machine learning or deep learning based approach to learn and model user behaviour. So, from the point of view of our interest in this course, we can stop at this stage here as soon where we conclude based on the analysis of the data.

But if required we may go one stage ahead and come up with a model of user behaviour which can be used to automate certain aspects of the system as well as the design process. So, that is in summary what we can do in empirical study. As you can see, the study is not a simple thing. Initially, when we started it may have seemed simple, we just asked friends to give feedback on our design.

But it is not as simple as you may be thinking and probably by now it is clear to you after going through the previous few lectures that it is a very rigorous process, involves lots of carefully considered stages and carefully considered activities to complete the process.

**(Refer Slide Time: 54:31)**

# Book

- **Bhattacharya, S.** (2019). Human-Computer Interaction: User-Centric Computing for Design, McGraw-Hill India
  - Print Edition: ISBN-13: 978-93-5316-804-9; ISBN-10: 93-5316-804-X
  - E-book Edition: ISBN-13: 978-93-5316-805-6; ISBN-10: 93-5316-805-8

## Chapter 7

So, whatever we have discussed so far, you can find in this book. In fact, several concepts we have mentioned in the passing as mentioned during the lecture, So, in this book some of those concepts are explained in detail. Also in this book, you will get reference to further study materials where you can learn in more details about those concepts. So, you are advised to refer to chapter 7 of this book to know in more details about those concepts.

So, with that I would like to end this lecture. I hope you have understood the concepts and enjoyed the lecture, looking forward to meet you all in the next lecture. Thank you and goodbye.