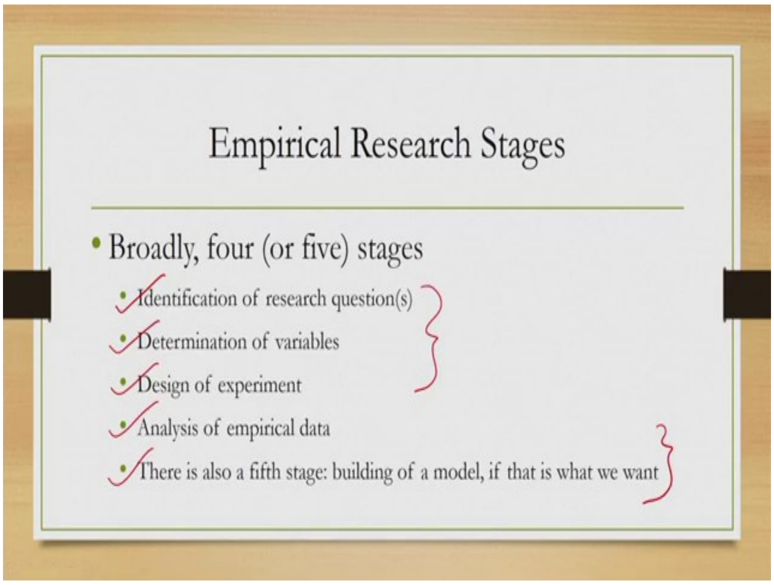
User-Centric Computing for Human-Computer Interaction

Prof. Samit Bhattacharya
Department of Computer Science and Engineering
Indian Institute of Technology, Guwahati

Lecture - 28
Data analysis including model building

Hello and welcome to lecture number 28 in the course User Centric Computing for Human Computer Interaction. So, if you recollect we have started our discussion on the topic of empirical research. Now, in this topic we mentioned that there are five stages, five stages involved in the process of empirical research.

(Refer Slide Time: 01:11)

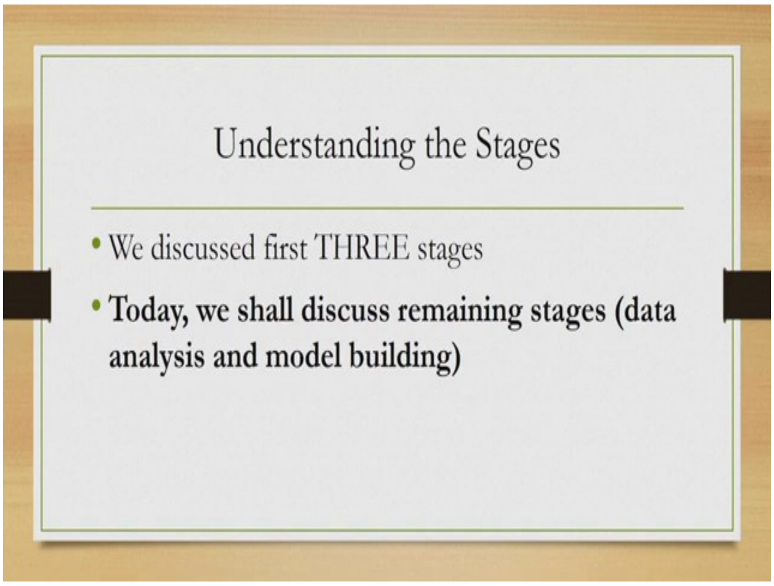


Empirical Research Stages

- Broadly, four (or five) stages
  - Identification of research question(s)
  - Determination of variables
  - Design of experiment
  - Analysis of empirical data
  - There is also a fifth stage: building of a model, if that is what we want

What are those stages? Let us just have a quick relook: first stage is identification of research questions, second is determination of variables, third is design of experiment, fourth is analysis of data. Sometimes we may stop at this stage or if we are interested in building a model then there is a fifth stage that is model building. Among these we have already discussed the first three stages in the previous two lectures; namely how to formulate research questions, how to identify variables and how to design an experiment. Today we are going to cover the remaining two stages namely analysis of data and model building.

(Refer Slide Time: 01:57)



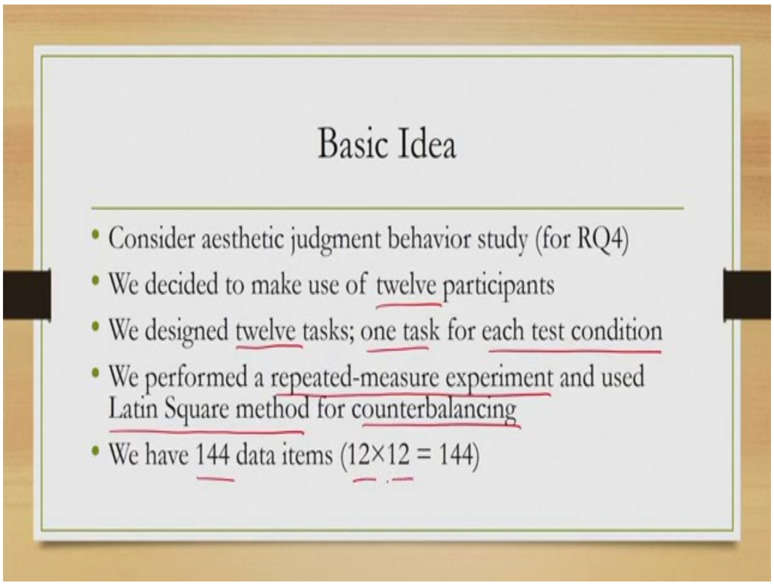
Understanding the Stages

- We discussed first THREE stages
- **Today, we shall discuss remaining stages (data analysis and model building)**

So, let us start with data analysis, in order to explain the idea of data analysis and its significance we will start with an example. The example that we are discussing throughout the lectures namely building of a relationship between the aesthetic judgment behavior of a user and the interfaces. In the previous lecture, if you may recollect we have formulated few research questions to perform an empirical study; so, that we can build the relationship.

Now, one question we formulated which we termed as research question 4 or RQ 4 for the benefit of recollection let me just quickly restate it. The question was on how the aesthetic judgment behavior is measured, how the aesthetic score in a scale of 1 to 10 is related to the number of objects, the type of the objects and the layout of an interface.

(Refer Slide Time: 03:14)



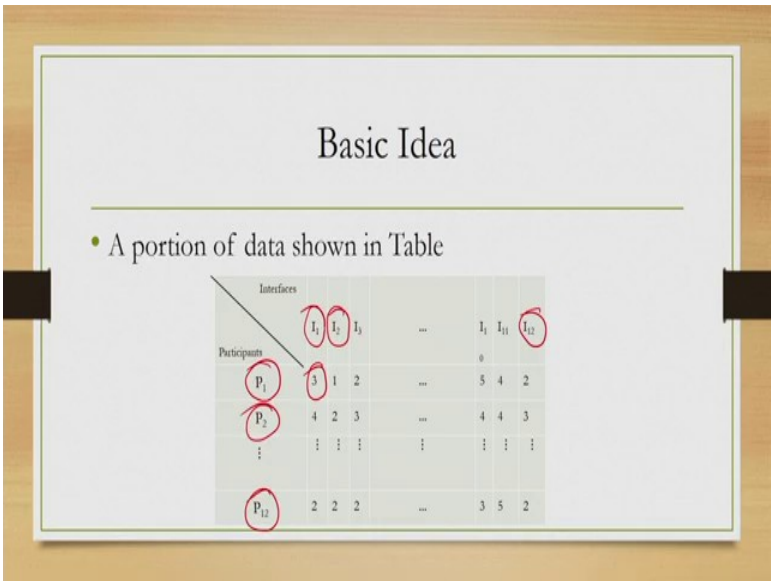
Basic Idea

- Consider aesthetic judgment behavior study (for RQ4)
- We decided to make use of twelve participants
- We designed twelve tasks; one task for each test condition
- We performed a repeated-measure experiment and used Latin Square method for counterbalancing
- We have 144 data items (12×12 = 144)

So, for this question suppose we want to conduct an experiment and we decided to make use of twelve participants and there were twelve tasks. There are twelve tasks and we have designed one task for each of twelve test conditions. Furthermore, we have decided to go for a repeated measure experiment; that means, each participant is asked to perform all the twelve tasks.

And, in order to avoid the practice effect we designed the task following a Latin Square method for counterbalancing. So, all these are fine and after we perform the study we end up with 144 data items. So, 12 participants each participant created 12 ratings for 12 interfaces and total 144 ratings, each rating is a data item.

(Refer Slide Time: 04:05)



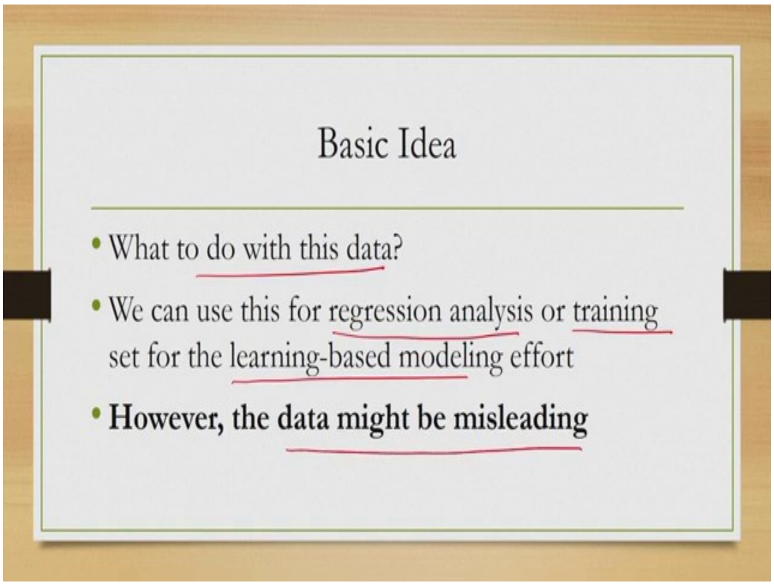
Basic Idea

- A portion of data shown in Table

Interfaces Participants	I_1	I_2	I_3	...	I_1 0	I_{11}	I_{12}
P_1	3	1	2	...	5	4	2
P_2	4	2	3	...	4	4	3
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
P_{12}	2	2	2	...	3	5	2

Let us see one sample of these data items, this is a hypothetical sample. So, here as you can see on one side we have mentioned participants 1, 2 up to 12 and on the other side we have mentioned interfaces 1, 2 up to 12; each sale in the matrix represent a rating. So, participant 1 rated 3 for interface 1 and so on.

(Refer Slide Time: 04:37)

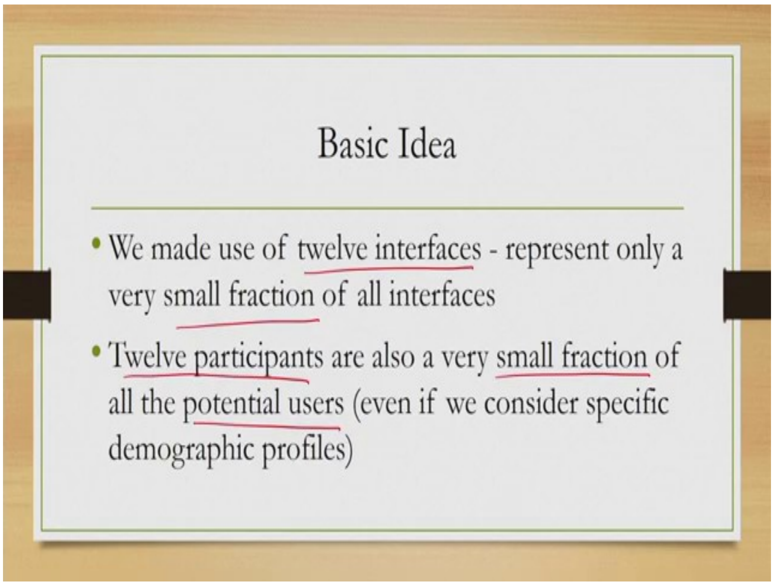


Basic Idea

- What to do with this data?
- We can use this for regression analysis or training set for the learning-based modeling effort
- **However, the data might be misleading**

Now, this is the data that we have collected. So, what to do with this data? As we have mentioned in one of our previous lectures, we can use it for regression analysis to build a model or use the data for training a learning based model. But, the problem is before we can go for this regression analysis or learning based model, we need to be very sure about the reliability of the data in fact, the data may be misleading.

(Refer Slide Time: 05:14)

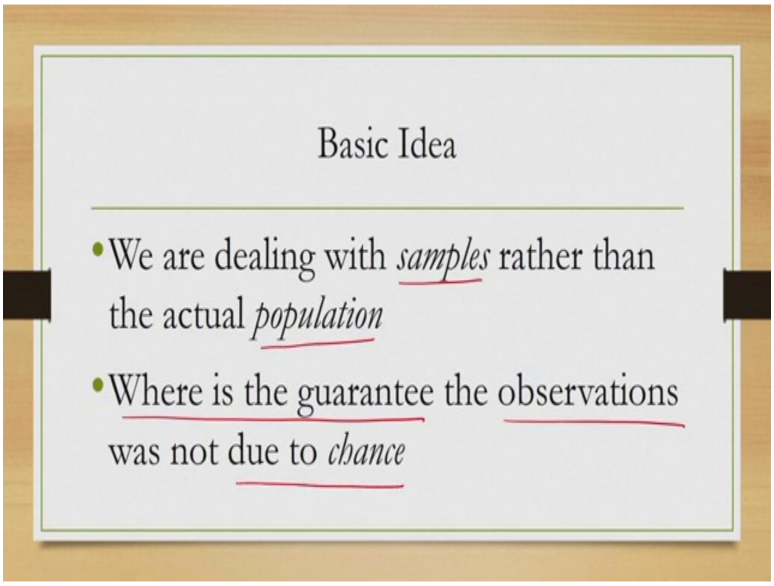


Basic Idea

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

So, whatever we observed may not reflect the true nature of the judgment behavior. Why? Because, in our experiment we have used twelve interfaces and twelve participants; now these twelve interfaces represent a very small fraction of all possible interfaces that are available. Similarly, the twelve participants represent a very small fraction of all the potential users, even if we consider specific user profiles.

(Refer Slide Time: 05:54)

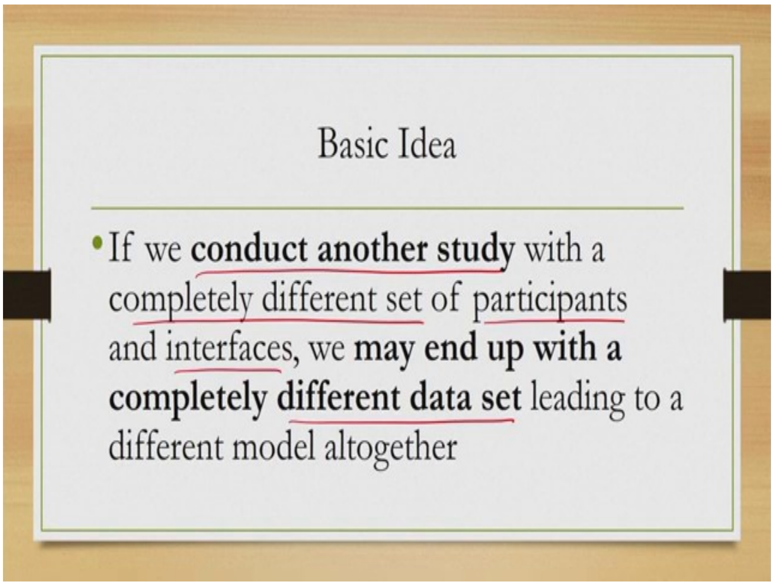


Basic Idea

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

What these indicate? It indicates that we are dealing with samples rather than the actual population. So, here the actual population consists of all possible interfaces and all potential users, that is impractical we cannot perform an experiment with the entire population. So, what we do is select samples, but the question is where is the guarantee that the observations that we make are not due to chance? This is a very important question and unless we are able to answer this question satisfactorily, we cannot do anything with the data that we have collected.

(Refer Slide Time: 06:39)



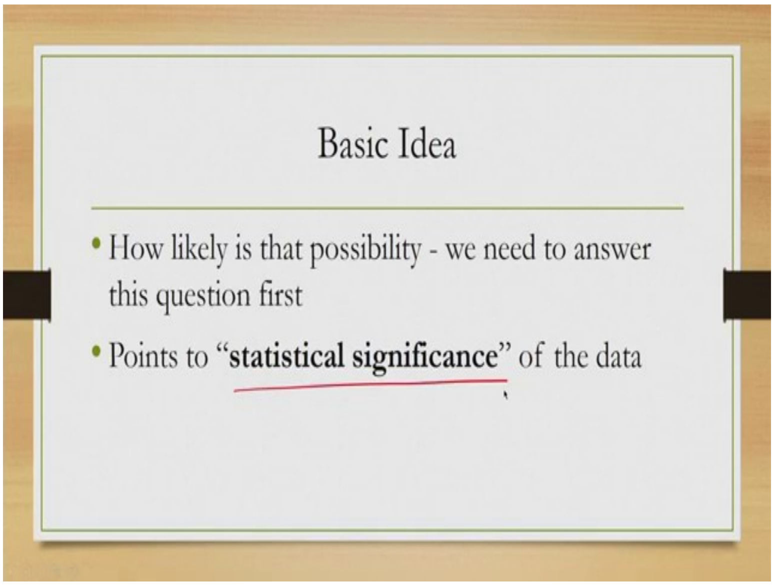
Basic Idea

- 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 model altogether

Now, let us try to understand what it means. Suppose, we conducted another study with a completely different set of participants and interfaces, then we may end up with different dataset. And, if we use this data set we can get a different model, if we are using it for regression analysis or training a learning based model.

So, one day we conduct a study to build a model with a particular set of interfaces and users, another day we conduct another study with a different set of users and interfaces. And, the models that we get out of the data for each day maybe different because the data on one day we may have got by chance and the other day we may have got not by chance, but because of the design of the interfaces.

(Refer Slide Time: 07:36)

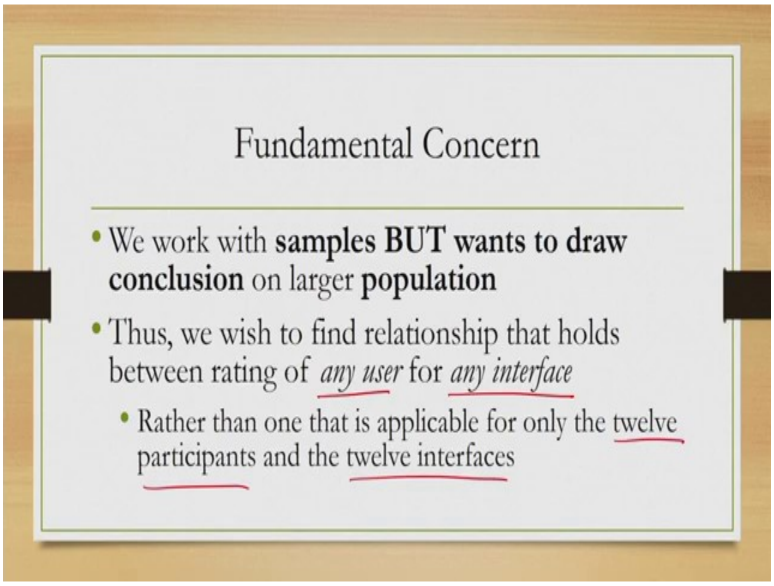


Basic Idea

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

So, our objective is to find out the possibility, the possibility of getting the data by chance or by design and to know that we need to go for a particular analysis of data that is known as statistical significance test. So, let us first try to understand the significance test, what it means and how to conduct this test.

(Refer Slide Time: 08:10)



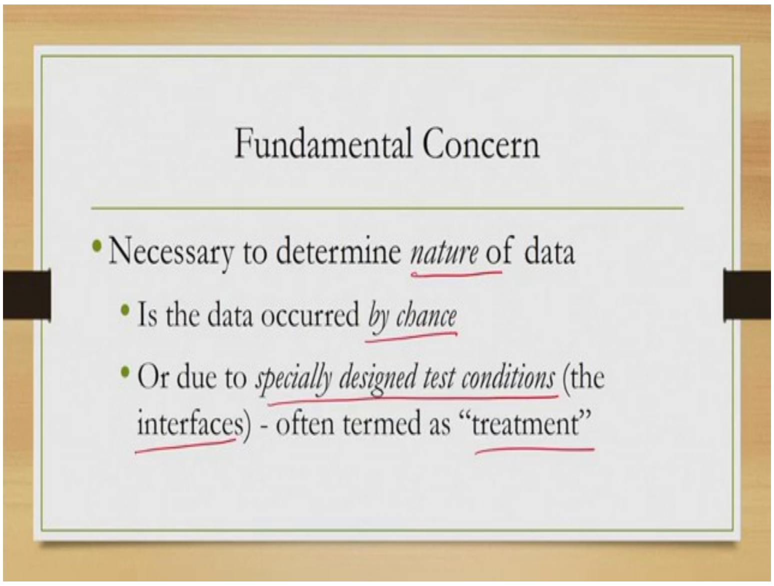
Fundamental Concern

- We work with **samples BUT wants to draw conclusion** on larger **population**
- Thus, we wish to find relationship that holds between rating of *any user* for *any interface*
  - Rather than one that is applicable for only the twelve participants and the twelve interfaces

So, what is our objective in empirical study? We are working with samples, as we have already mentioned that it is not possible to work with the entire population; instead what we do is decide on a sample and try to observe data for that sample. However, our overall objective is to come to a conclusion that is not applicable only to the samples, but to the entire population. So, we want to collect data for samples and one to generalize the findings to the entire population.

In other words in the context of our example on modeling aesthetic judgement behavior; so, we made some observations that we have shown before and those observations are for twelve interfaces and twelve participants. However, the relationship that we want to find out should be applicable to any user and any interface, not only specific to the twelve interfaces and twelve participants that we have used in our study, that is our overall objective. How to ensure that data that we get is actually because of the design of the interfaces rather than by chance?

(Refer Slide Time: 09:39)

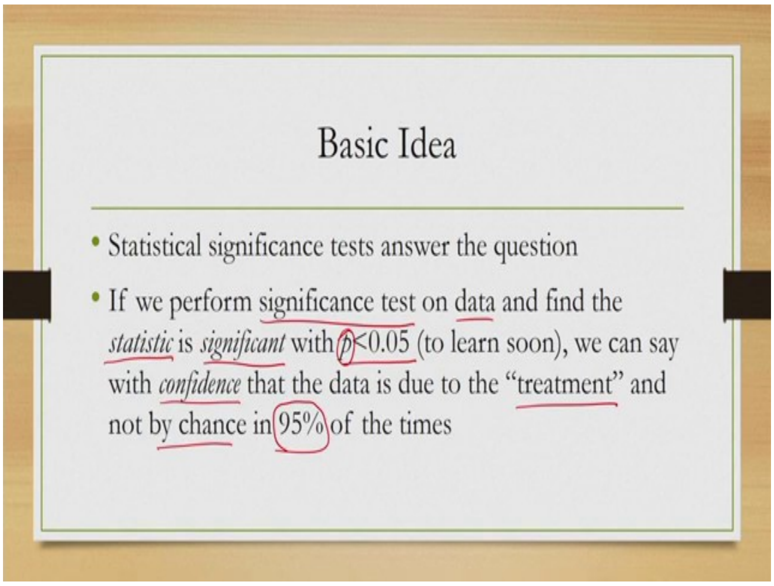


Fundamental Concern

- 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, what we need is to determine the nature of the data whether it occurred by chance or due to the specially designed test conditions often called treatment conditions. So, in our example this specially designed test conditions are the particular interfaces that we have used. And, we are assuming that whatever data we collect are because of the design of the interfaces rather than by chance, but there should be some way to ensure that, that is the case whatever data we got is because of the interfaces not because of chance.

(Refer Slide Time: 10:16)



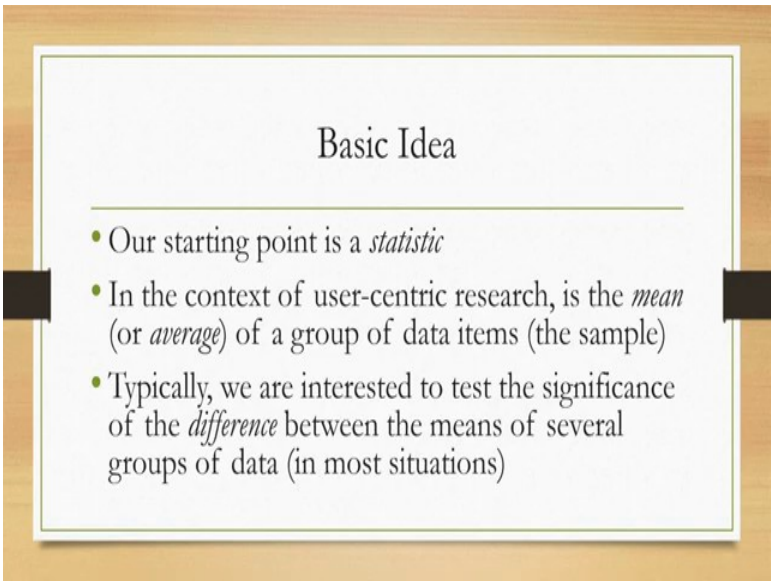
Basic Idea

- 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

So, its statistical significance test we can get that assurance. So, if we perform a test statistical significance test on the data and find that the statistic is significant with the specific value of p, then we can say that the data is due to the treatment rather than due to chance in a certain percentage of cases. So, if we set p to be 0.05 then we can say that in 95 percent of the cases the data that we have observed is due to the treatment condition and not by chance and the remaining 5 percent cases it may occur by chance.

Now, what is the meaning of statistic or the term confidence? These things we will learn soon, but the idea is something like this that we conduct a test with the set value of a particular concept called p. And, based on the result of the test we conclude that whether the data occurred by chance or by treatment with a confidence value.

(Refer Slide Time: 11:31)

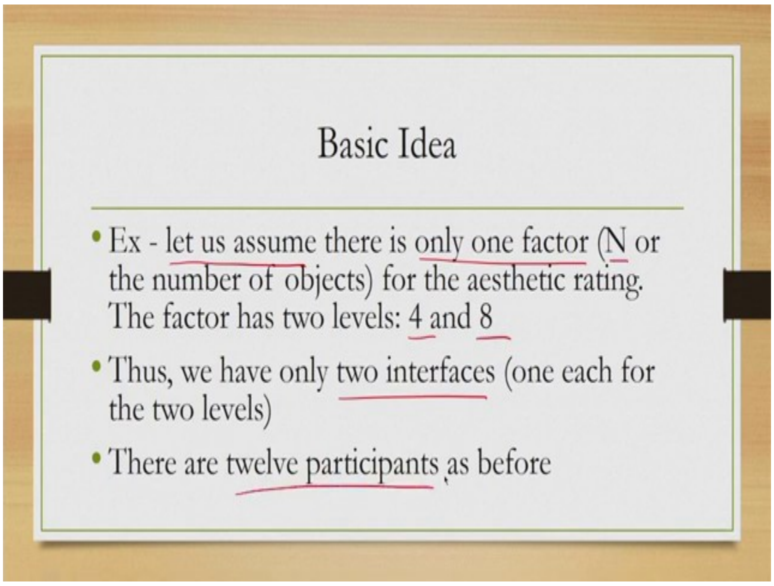


Basic Idea

- Our starting point is a *statistic*
- In the context of user-centric research, 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, our starting point is a statistic. Now, mostly in the context of user centric research these statistic is typically the mean or average of a group of data items where these data items represent the data items collected over the samples. And, what we are interested in? We are typically interested in testing the significance of the difference between the means of several groups of data items, that is our overall objective to test the significance of the difference between several group of data items.

(Refer Slide Time: 12:22)

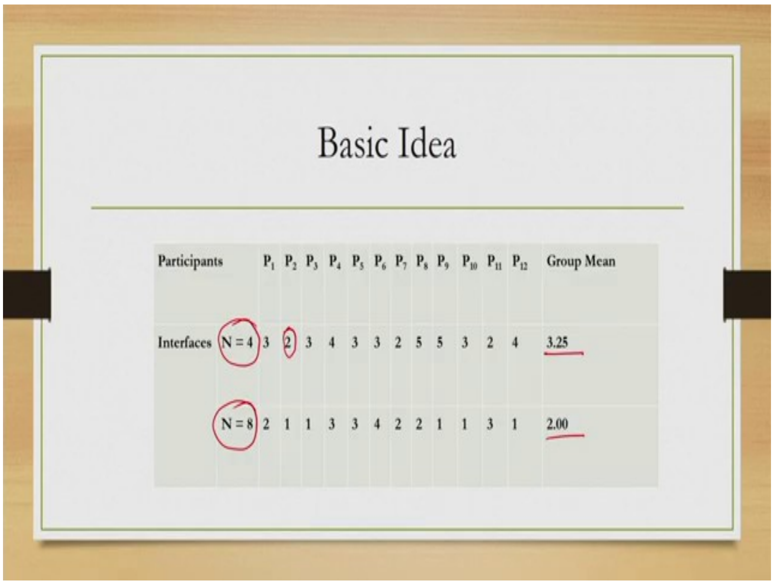


Basic Idea

- Ex - let us assume there is only one factor (N or the number of objects) for the aesthetic rating. The factor has two levels: 4 and 8
- Thus, we have only two interfaces (one each for the two levels)
- There are twelve participants as before

Now, let us try to understand in terms of an example. Let us assume that there is only one factor in our study on aesthetic judgment behavior. The factor is N or the number of objects on the interface and we have chosen two levels for this factor 4 and 8. So, in our study we have used only two test conditions or in other words two interfaces; one with 4 objects and other one with 8 objects and there are twelve participants as before. Now, with this setup we have collected data.

(Refer Slide Time: 13:01)

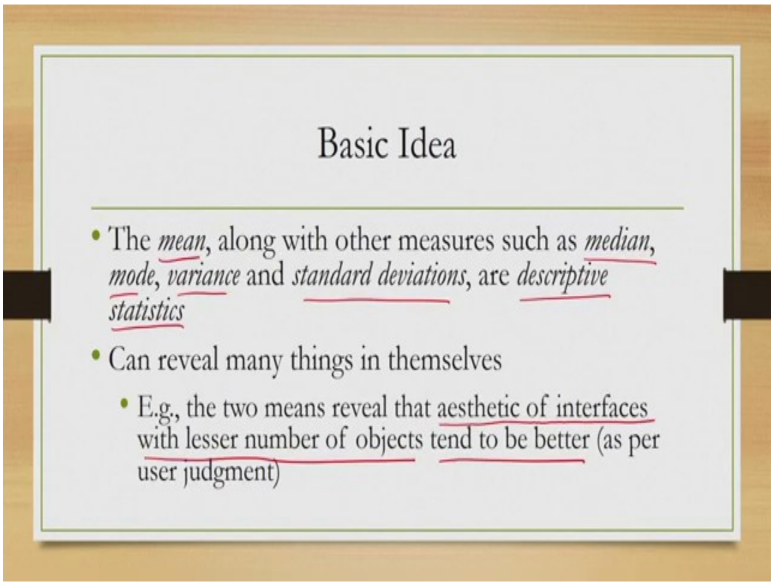


Basic Idea

Participants		P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	Group Mean
Interfaces	N = 4	3	2	3	4	3	3	2	5	5	3	2	4	3.25
	N = 8	2	1	1	3	3	4	2	2	1	1	3	1	2.00

And, this table shows a sample of the data that we have collected and we have also reported the mean for each group. So, there is one group for the interface with N equal to 4 and there is another group of data items for interface with N equal to 8. Now, each data item represents a rating provided by the participants and we have also reported the mean of those ratings for each group. So, there are two group means one is 3.25 for N equal to 4 and one is 2 for N equal to 8. Now, from this means what we can infer?

(Refer Slide Time: 13:48)

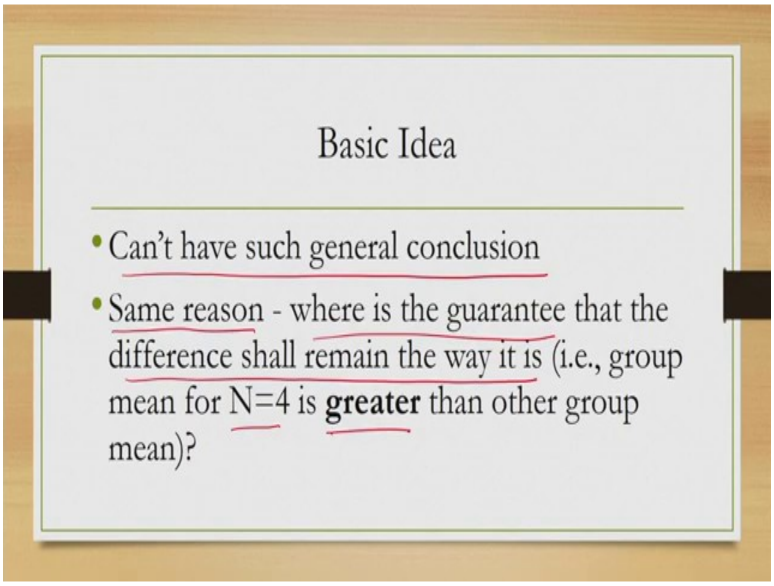


Basic Idea

- 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 aesthetic of interfaces with lesser number of objects tend to be better (as per user judgment)

So, as you know or you probably know mean along with some other measures such as median, mode, variance, standard deviations are collectively called descriptive statistics and these descriptive statistics can reveal many things in themselves. So, you have seen the two means, based on these two means we can simply conclude that the aesthetic judgment of interfaces with less number of objects tend to be better because the group means say so.

(Refer Slide Time: 14:42)



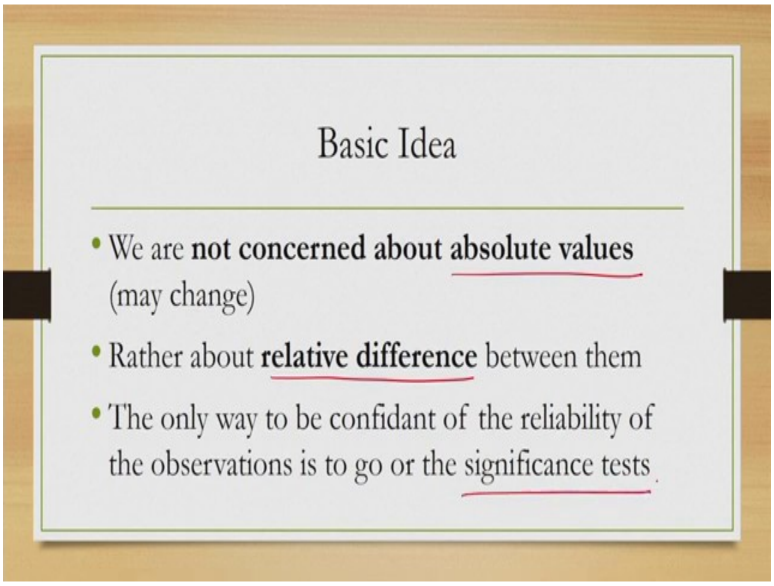
Basic Idea

- 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 N=4 is **greater** than other group mean)?

But, can we really conclude like this? That is the question and the answer is that we cannot have such general conclusion based on these two group means, because we have already mentioned this reason before the same reason. Because, of the same reason that is where is the guarantee that the difference shall remain the way it is.

And, what is this difference? What is the nature of this difference? That the group mean for N equal to 4 is greater than the group mean for N equal to 8. So, tomorrow if we conduct another test with another set of participants, where is the guarantee that this relative differences between the two groups shall remain the same?

(Refer Slide Time: 15:28)



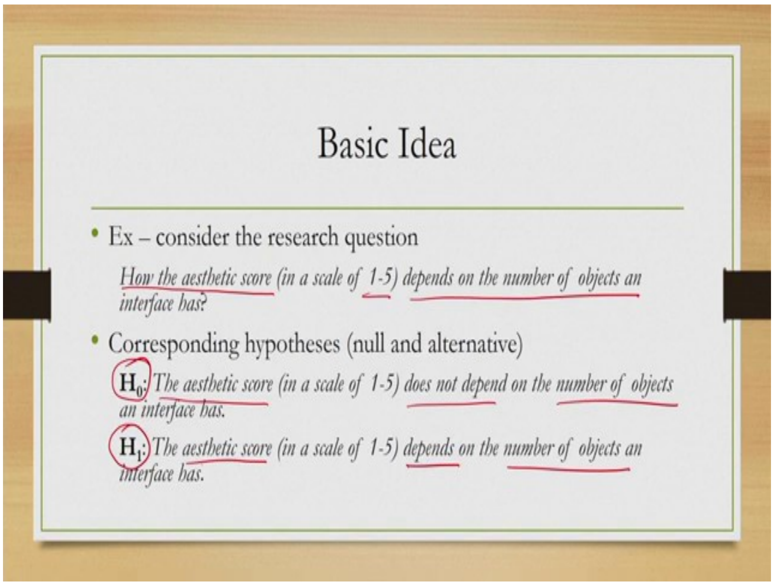
Basic Idea

- 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

Note here that we are not concerned about absolute values rather we are concerned about relative differences between the groups. So, absolute values may change; however, the relative differences between these values that is one is greater than the other or one is equal to the other or one is lower than the other, this relative differences are the main concern for us. Our objective is to ensure that these differences that we have observed in our study have not occurred by chance and the way to do it is to go for significance tests.

Now, there is another point to be noted here that is we are talking of significance tests, but the significance tests are not performed on research questions instead we need hypothesis. So, earlier we have seen how to generate the null and alternative hypothesis from a research question. So, that same thing we have to do here, if we conduct an experiment for a research question; in order to analyze the data collected in the experiment to perform significance tests, we have to create the two hypothesis: null and alternative and based on that hypothesis only we can go for a significance test.

(Refer Slide Time: 17:03)



Basic Idea

- Ex – consider the research question

  How the aesthetic score (in a scale of 1-5) depends on the number of objects an interface has?

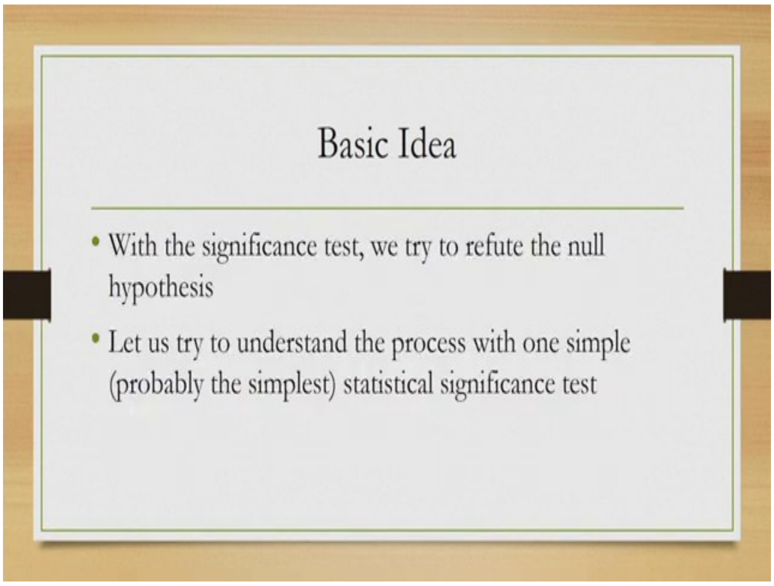
- Corresponding hypotheses (null and alternative)

  H_0: *The aesthetic score (in a scale of 1-5) does not depend on the number of objects an interface has.*

  H_1: *The aesthetic score (in a scale of 1-5) depends on the number of objects an interface has.*

As an example consider the research question how the aesthetic score in a scale of 1 to 5 depends on the number of objects or N. Now, the corresponding hypothesis from this research question can be framed in the following way. The null hypothesis is H naught that is the aesthetic score does not depend on the number of objects. And, the alternative hypothesis is the score depends on the number of objects and the data we use to refute the null hypothesis and establish, that the alternative hypothesis is true.

(Refer Slide Time: 17:52)

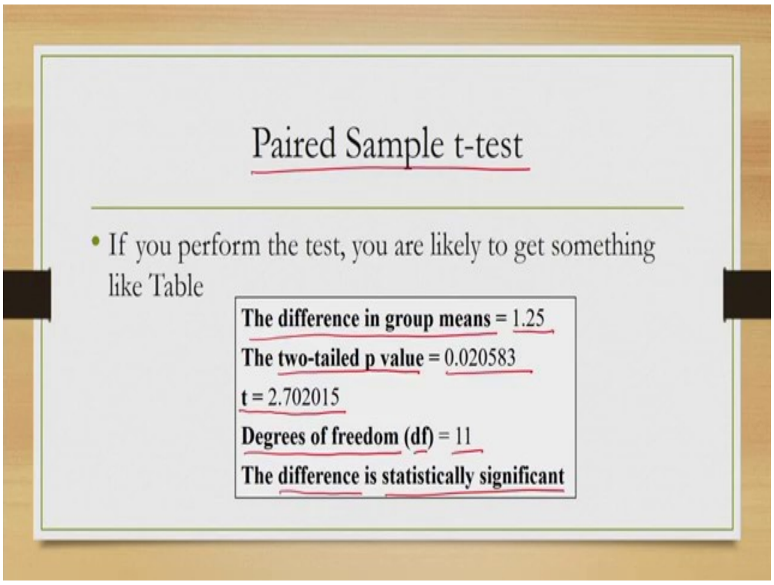


Basic Idea

- 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

And how we do that? Let us try to understand in terms of a simple and probably the simplest example of significance test that is called the t test. So, we will try to explain the idea of significance test in terms of t test.

(Refer Slide Time: 18:13)



Paired Sample t-test

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

The difference in group means = 1.25

The two-tailed p value = 0.020583

t = 2.702015

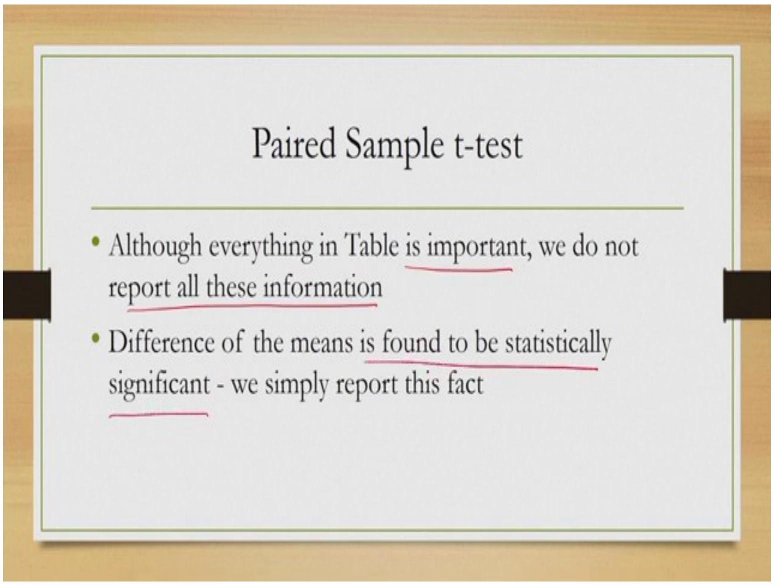
Degrees of freedom (df) = 11

The difference is statistically significant

Now, if we perform a paired sample t test, that is one type of statistical significance test then what we are likely to observe? Few things we are likely to observe that is one quantity which is the absolute difference between the group means that is 1.25, one statistic two tailed p value something like this 0.020583, the t statistic which is having a value 2.70, another quantity degrees of freedom typically represented with the symbol df which is 11.

And, based on these values the conclusion is that the difference is statistically significant; that means, whatever we have observed is not by chance in at least 95 percent of the cases, if we have set a value of p to be 0.05.

(Refer Slide Time: 19:23)

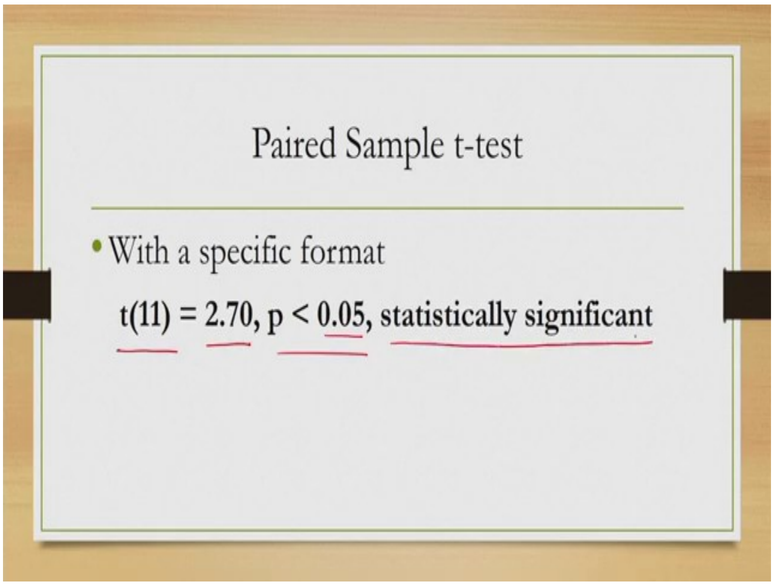


Paired Sample t-test

- 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

All these things that are mentioned in the table are important. However, it is not necessary to report all these information, instead we report only the final outcome that is the difference of the means its found to be statistically significant, but in a slightly different way.

(Refer Slide Time: 19:45)



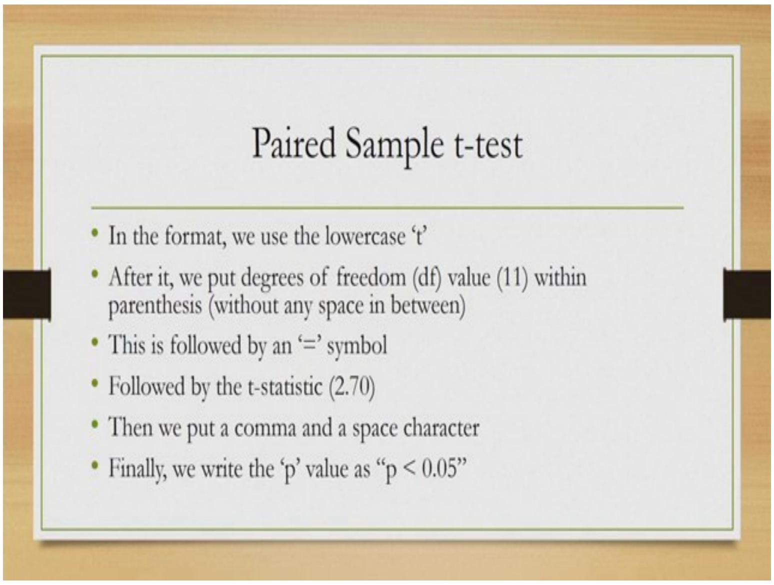
Paired Sample t-test

- With a specific format

t(11) = 2.70, p < 0.05, statistically significant

So, we use a specific format which is shown here. So, this term the t represents the t test within parenthesis we represent the degrees of freedom, then there is a equal to sign followed by the t statistic value followed by comma followed by the p value; in this notation p less than the set value 0.05 and then the final outcome that it is statistically significant.

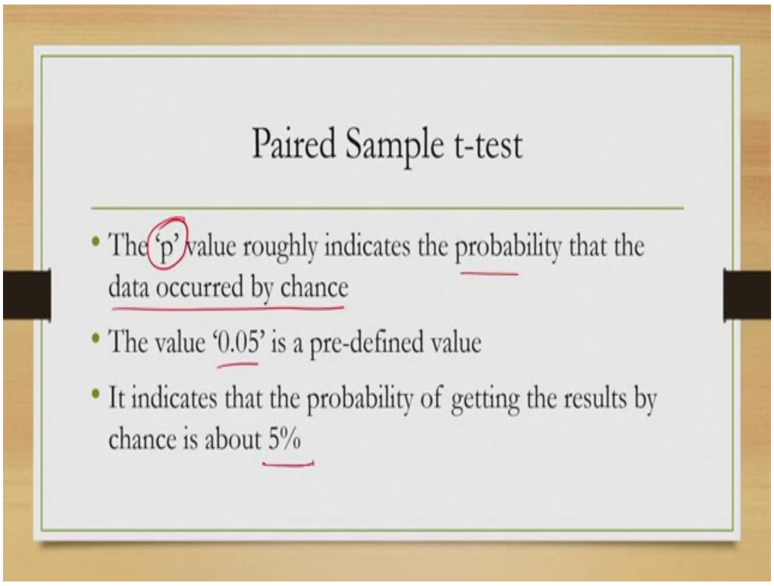
(Refer Slide Time: 20:23)



Paired Sample t-test

- In the format, we use the lowercase 't'
- After it, we put degrees of freedom (df) value (11) within parenthesis (without any space in between)
- This is followed by an '=' symbol
- Followed by the t-statistic (2.70)
- Then we put a comma and a space character
- Finally, we write the 'p' value as "p < 0.05"

(Refer Slide Time: 20:27)

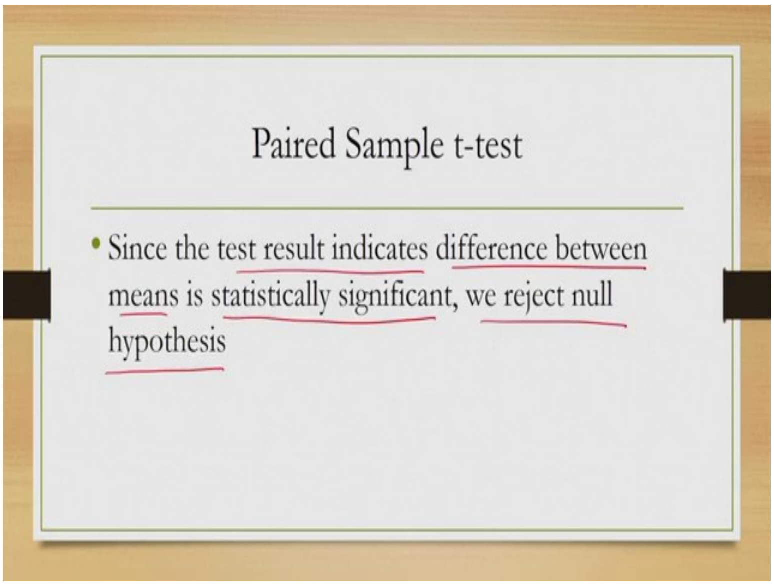


Paired Sample t-test

- 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, this notation is very important. What this p value indicates? It roughly indicates the probability that the data occurred by chance. So, if we set a value of 0.05 we are indicating that we are interested to know that the probability of getting the results by chance is about 5 percent. So, depending on the value of p, the conclusion that we draw may change. So, if we set a higher value then definitely it will indicate that the result is significant subject to the condition that the probability of occurrence by chance is of that particular value.

(Refer Slide Time: 21:07)



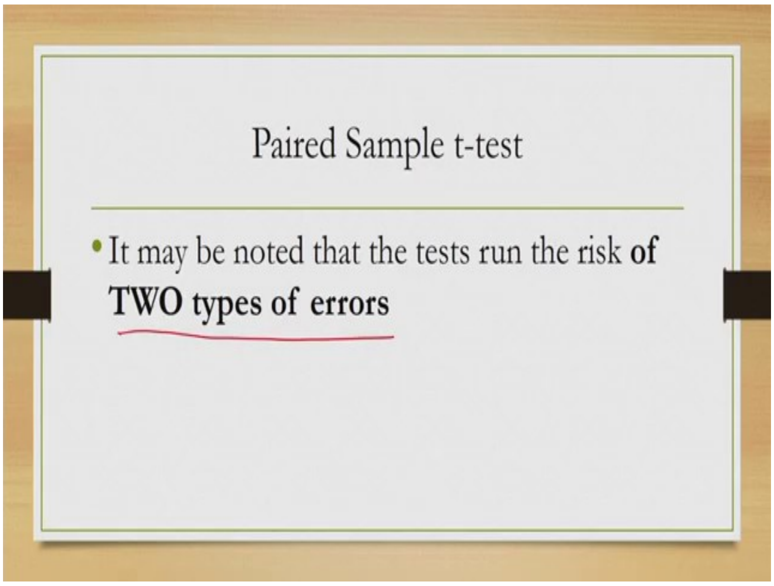
Paired Sample t-test

- Since the test result indicates difference between means is statistically significant, we reject null hypothesis

Now, since the test results have indicated that the difference between the means is statistically significant, it indicates that the judgment is actually dependent on the number of objects. So, we reject the null hypothesis which is our overall objective, in other words the alternative hypothesis turns out to be true; so, there exists some relation. So, that is the overall conclusion that we can draw from this test.

So, to summarize so, we have performed some calculations, found out values for some statistic and based on those values we conclude that the alternative hypothesis is true. Now, note that this alternative hypothesis is a generalized concept. So, it is not applicable only to the samples rather the entire population. So, based on the results of the test we can conclude on the population which is our main concern.

(Refer Slide Time: 22:15)

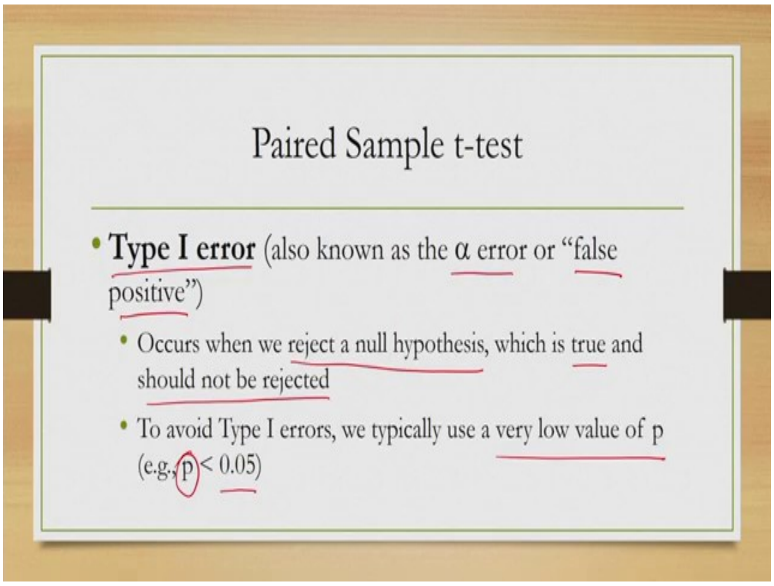


Paired Sample t-test

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

However, when you are performing a test, you should be aware of two types of errors that may occur.

(Refer Slide Time: 22:26)

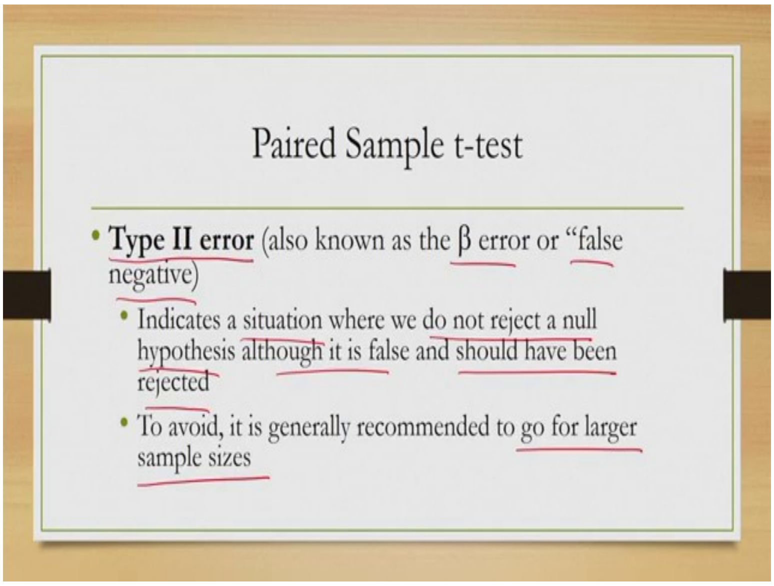


Paired Sample t-test

- **Type I error** (also known as the α 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., $p < 0.05$)

First one is called type I error, often called alpha error or false positive errors. Now, it occurs when we reject null hypothesis, although the hypothesis is true and should not be rejected. And, when it happens? Typically, when we use higher value of p or the probability values. So, to avoid type I errors, we typically use a very low value of p, typical value used for p is 0.05, it can even be 0.01, but in most of the cases we use p as 0.05.

(Refer Slide Time: 23:11)

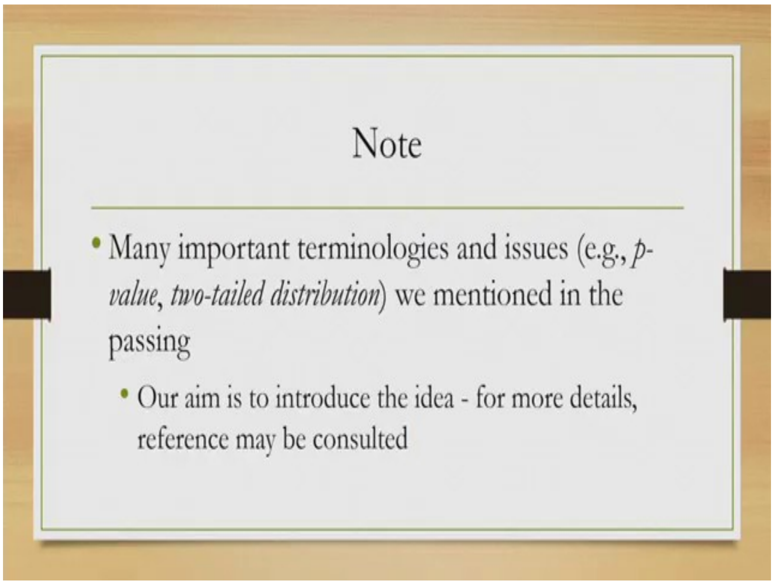


Paired Sample t-test

- **Type II error** (also known as the β 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

Now, there is another type of error called type II error which is also known as the beta error or false negative. What it indicates? It indicates a situation where we do not reject a null hypothesis although that is false and should have been rejected and the only way to avoid these type of errors in our tests is to generally go for larger sample sizes. So, if we increase sample size, if we work with more participants and more interfaces then it is likely that these type II errors can be avoided.

(Refer Slide Time: 24:04)

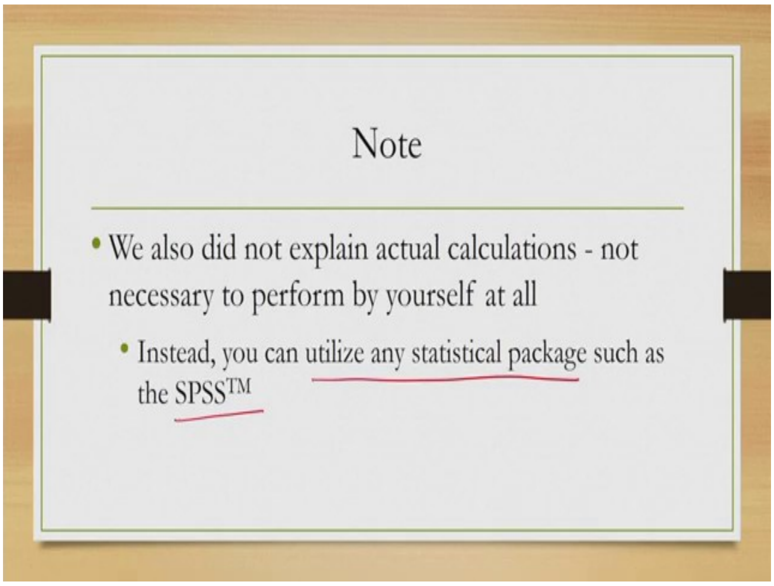


Note

- 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

Now, in this discussion we have used many terms like the degrees of freedom, the p value and we have given some rough ideas of what these mean. However, the exact definitions are not given and we will not try to give an exact definition here, our overall objective is to explain the idea rather than talking about the exact things.

(Refer Slide Time: 24:36)



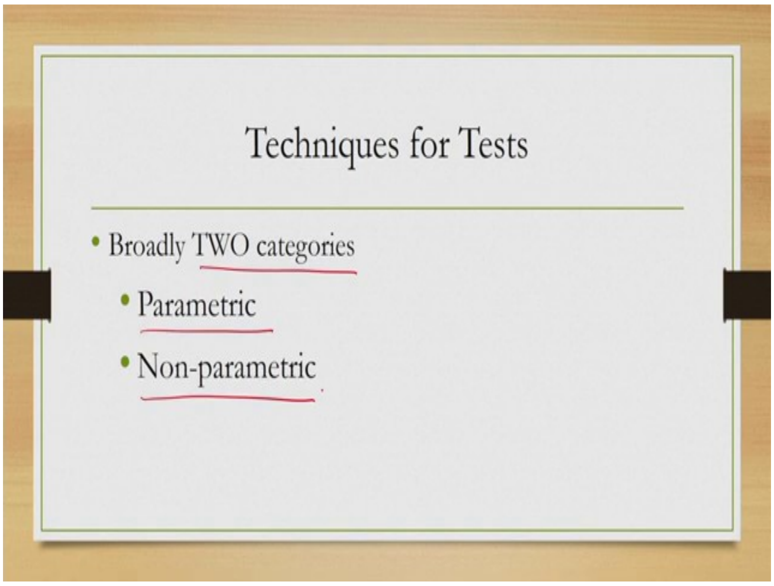
Note

- 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™

For more details you may consult the references that will be mentioned at the end of this lecture. Another thing is how to make those calculations? We mentioned some values for the t statistic or for degrees of freedom or for mean and other things. So, how to do that? You can actually utilize any specialized statistical package, any statistical package such as SPSS where all these tools are available to perform all the calculations.

So, you do not need to actually do it manually, you can use any statistical package for performing the calculations. Now, other thing is the type of tests we should go for; as an example we have explained to some extent the idea of the paired t test. However, this is not the only test available and many other tests are there which you can use. It is very important to identify the right test for the right data items. So, let us try to have some idea on how to choose a right test.

(Refer Slide Time: 25:55)

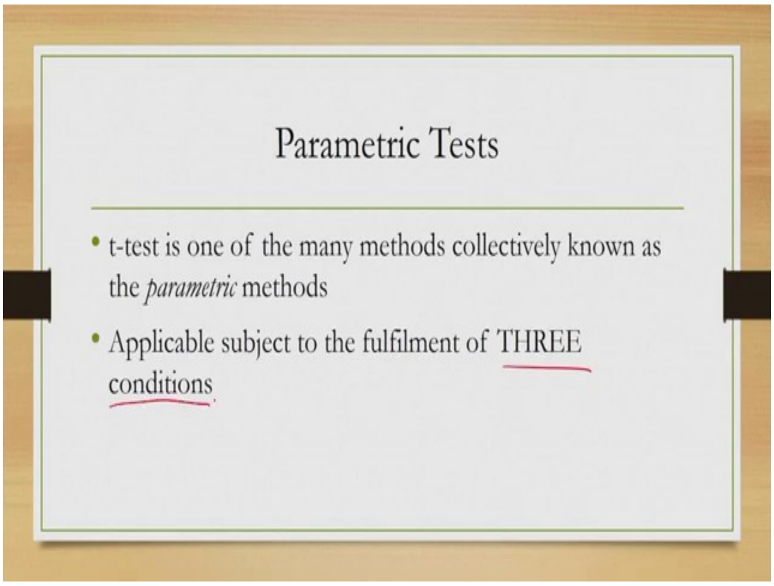


Techniques for Tests

- Broadly TWO categories
  - Parametric
  - Non-parametric

There are broadly two categories of tests possible: one is called parametric and one is called non-parametric.

(Refer Slide Time: 26:07)

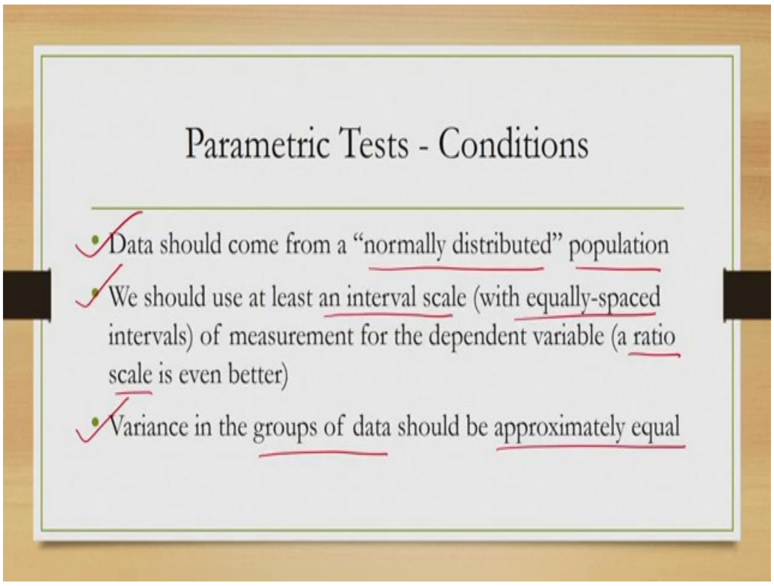


Parametric Tests

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

Now, in parametric tests there are three conditions that should be fulfilled. What are these three conditions?

(Refer Slide Time: 26:13)



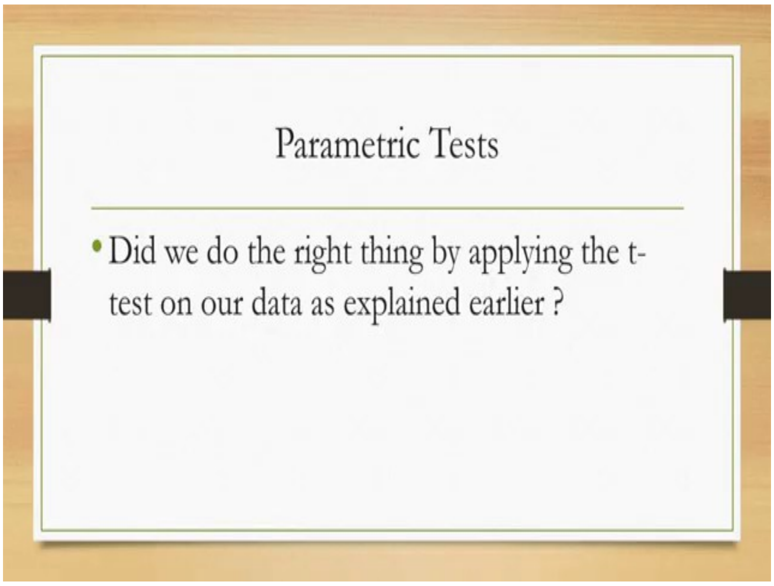
Parametric Tests - Conditions

- ✓ 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

First of all the data should come from a normally distributed population. Secondly, we should use at least an interval scale with equal spacing or a ratio scale and thirdly the variance in the groups of data should be approximately equal. So, there are these three conditions which should be fulfilled in order to be able to use a parametric test to test for statistical significance of your empirical data.

The first condition is that the data should come from a normally distributed population. The second condition is that the measurement skills that we have used to record the data should be at least an interval scale with equal spacing, a ratio scale is even better.

(Refer Slide Time: 27:11)

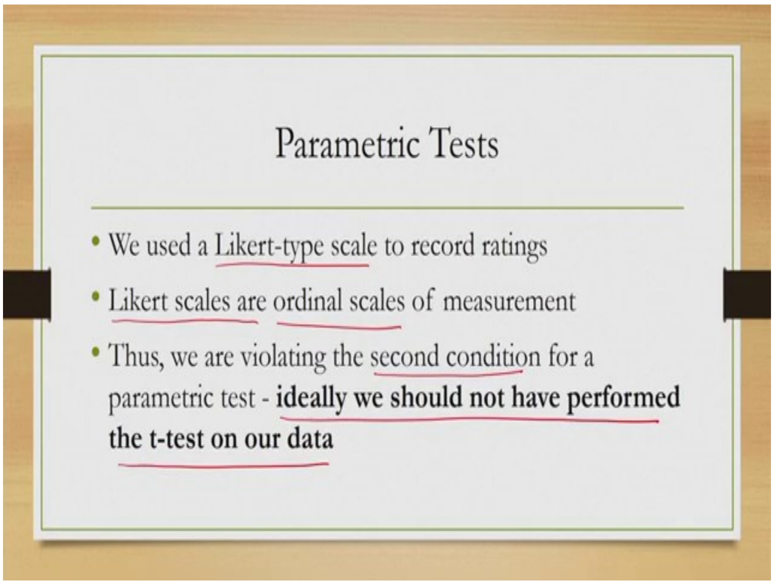


Parametric Tests

- Did we do the right thing by applying the t-test on our data as explained earlier ?

And the third condition is that the variance in the groups of data item should be approximately equal. So, once we are aware of these conditions, we can go for choosing a right test. Now, note that in our earlier example we have applied the t test. Now, t test is a parametric test. So, the question is: did we do the right thing by choosing the t test to analyze our data? Let us try to answer this question.

(Refer Slide Time: 27:49)



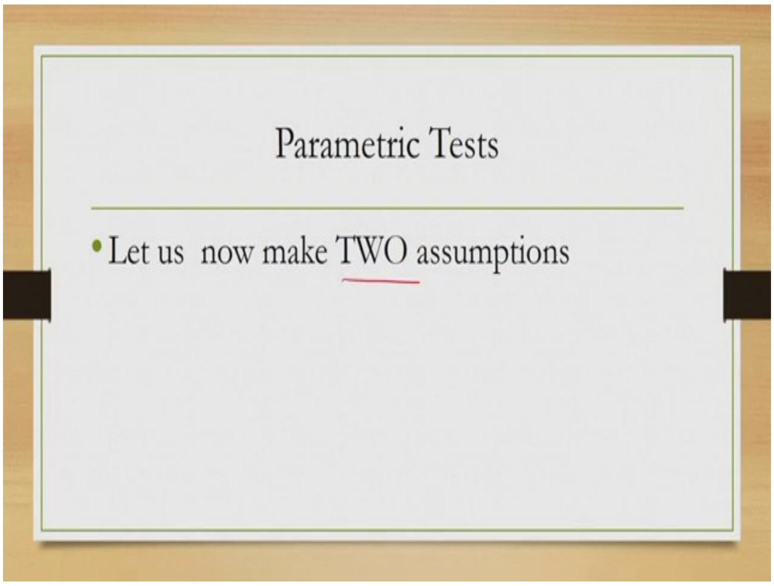
Parametric Tests

- We used a Likert-type scale to record ratings
- Likert scales are ordinal scales of measurement
- Thus, we are violating the second condition for a parametric test - **ideally we should not have performed the t-test on our data**

Note that we talked about representing the aesthetic judgment behavior in terms of a rating. And, we used a Likert type rating scale where 1 means something, 2 mean something else with no relation to 1 and so on. Now, Likert scales as we have mentioned earlier are ordinal scales of measurement. However, for application of parametric tests we require the use of at least an interval scale or even better is to use a ratio scale.

So, the data items on which we applied the t test violated the second condition that is the data items were not recorded using either an interval scale or a ratio scale. So, ideally we should not have perform the t test on our data. Now, to make the t test applicable on our data, what we can do?

(Refer Slide Time: 28:56)

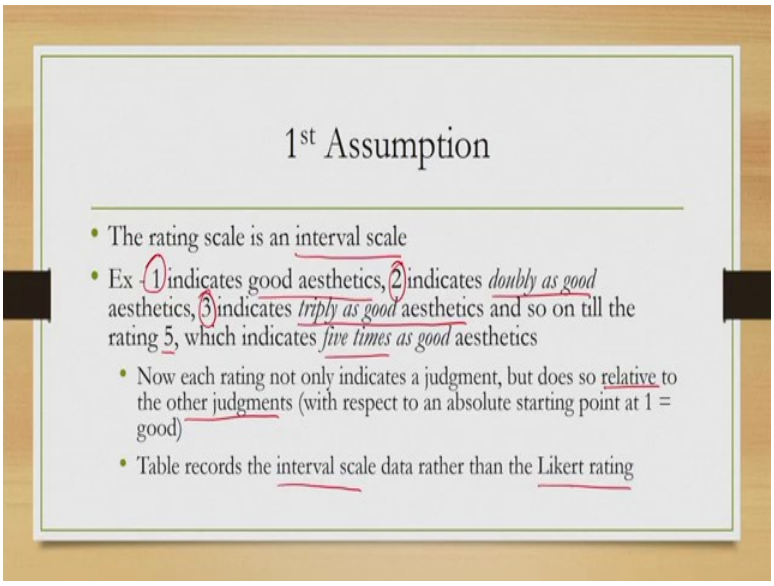


Parametric Tests

- Let us now make TWO assumptions

Let us now make two assumptions about the data which we did not do earlier.

(Refer Slide Time: 29:04)



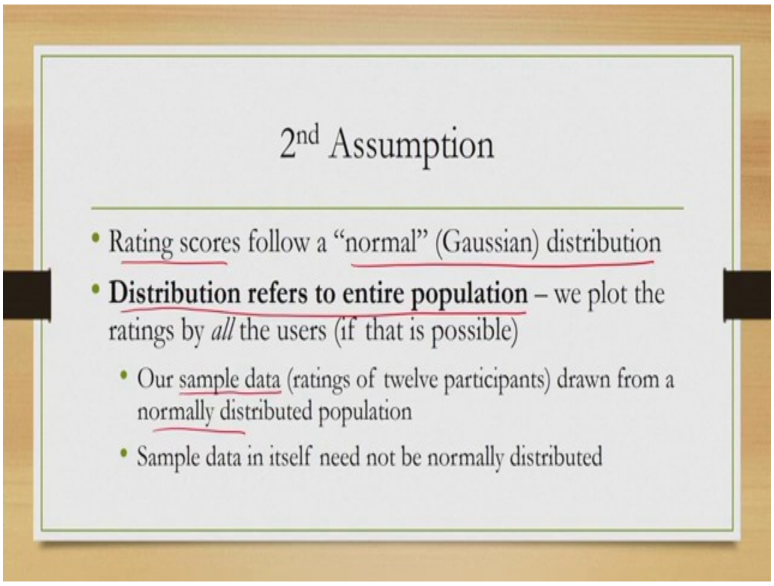
1st Assumption

- The rating scale is an interval scale
- Ex - 1 indicates good aesthetics, 2 indicates *doubly as good* aesthetics, 3 indicates *triply as good* aesthetics and so on till the rating 5, which indicates *five times as good* aesthetics
  - Now each rating not only indicates a judgment, but does so relative to the other judgments (with respect to an absolute starting point at 1 = good)
  - Table records the interval scale data rather than the Likert rating

First of all the rating scale now is redefined as an interval scale rather than a Likert scale. So, we can define it in a different way; for example, now 1 indicates a good aesthetic, 2 indicates doubly good aesthetic rather than something else, 3 indicates triply as good aesthetic and so on till rating 5 which indicates five times as good aesthetic.

So now, each rating is not only representing a judgment behavior, but also it does so, relative to the other judgments with respect to an absolute starting point at 1 which indicates good and the table records this interval scale rather than the Likert ratings. Now, if we make these assumptions, then definitely we adhere to the second condition for application of parametric tests.

(Refer Slide Time: 30:07)

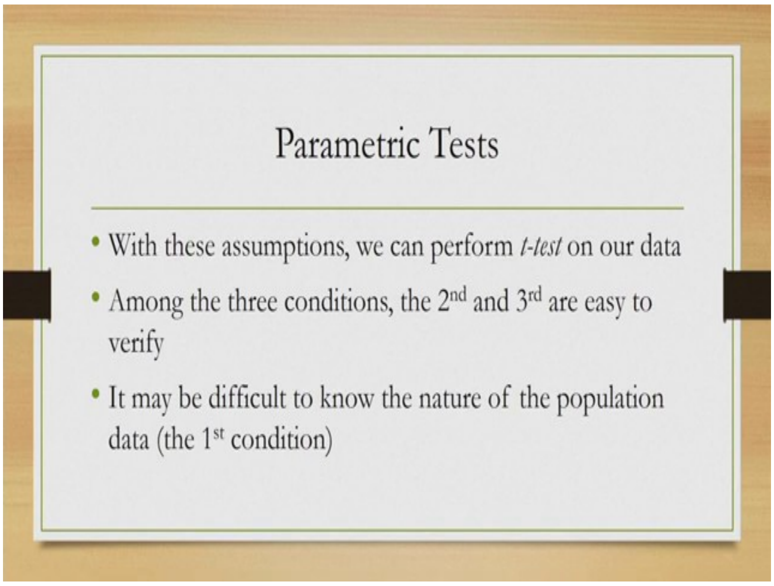


2nd Assumption

- Rating scores follow a "normal" (Gaussian) distribution
- **Distribution refers to entire population** – we plot the ratings by *all* the users (if that is possible)
  - Our sample data (ratings of twelve participants) drawn from a normally distributed population
  - Sample data in itself need not be normally distributed

And, there is a second assumption which is that the rating scores follow a normal or a Gaussian distribution. Note here that this distribution refers to the distribution for the entire population, not for the sample. So, it is not necessary that the sample data follows a normal distribution instead the sample data can follow any distribution. But, we are assuming that the population data from where the sample data is drawn follows a normal distribution or a Gaussian distribution.

(Refer Slide Time: 30:45)

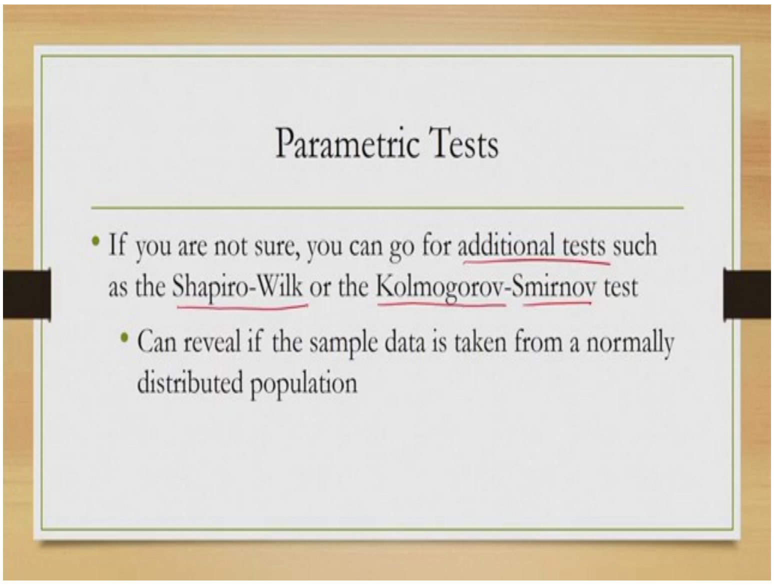


Parametric Tests

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

So, if we make these two assumptions then we can perform the t test, otherwise we cannot. Now, among the three conditions that we have mentioned to check if parametric tests are applicable for a given data set or not, the 2nd and 3rd conditions namely the measurement scale used and the variance between the groups of data items are easier to determine. However, the first condition that is whether the sample data comes from a normally distributed population data is not that easy, it is somewhat difficult.

(Refer Slide Time: 31:36)

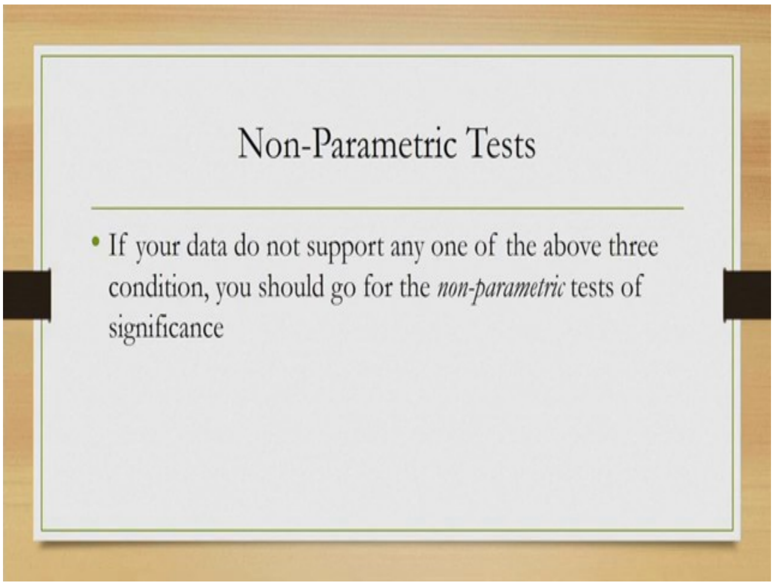


Parametric Tests

- 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 we are not sure then we can go for additional tests such as Shapiro-Wilk test or Kolmogorov-Smirnov test. Now, these tests can reveal if the sample data is taken from a normally distributed population data.

(Refer Slide Time: 32:01)

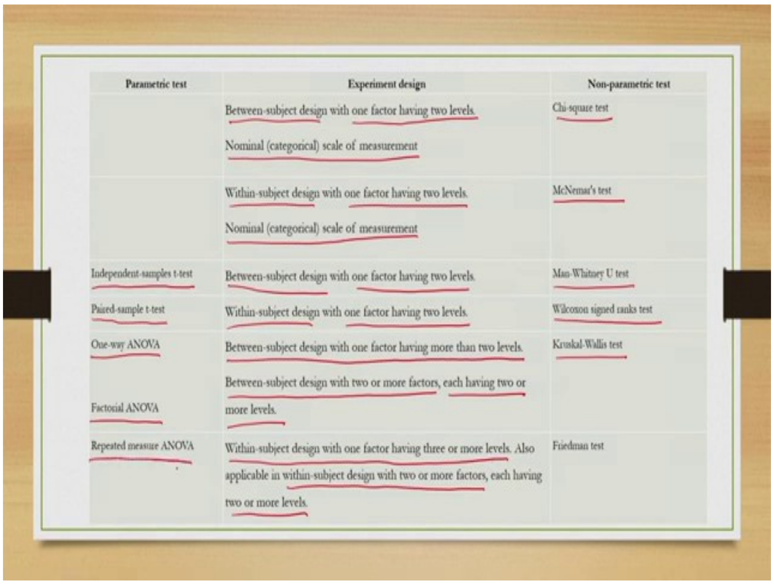


Non-Parametric Tests

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

Now, if after these tests we find that our sample data does not come from a normally distributed population data, then we should not apply parametric tests; in that case we should go for non-parametric tests. Now, there are many parametric tests as well as non-parametric tests and it is very important to identify the right situation for application of a particular test.

(Refer Slide Time: 32:37)



Parametric test	Experiment design	Non-parametric test
	Between-subject design with one factor having two levels. Nominal (categorical) scale of measurement	Chi-square test
	Within-subject design with one factor having two levels. Nominal (categorical) scale of measurement	McNemar's test
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, in this table we have listed the popular tests that are used to analyze data. So, when we have a between subject experiment design with one factor having two levels and a nominal or categorical scale of measurement, then we can go for a chi square test which is a non-parametric test. When we have within subject design with one factor with two levels and again the same nominal scale of measurement, we can go for McNemar's test which is again a non-parametric test. For these two situations we do not have any parametric test.

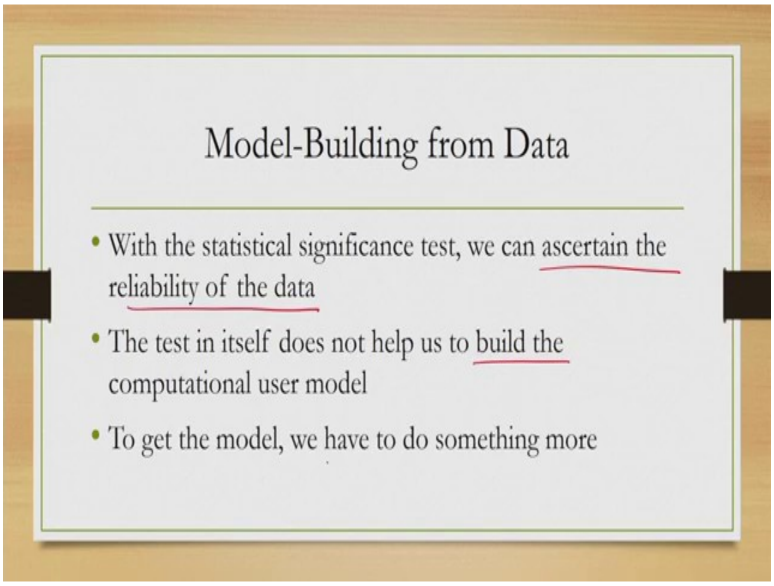
However, when we have a between subject experiment design with one factor and two levels, then if the data comes from a normal population and other conditions are satisfied then we have a parametric test which is independent samples t test. And, if those conditions are not satisfied, then we have a non-parametric test which is Man-Whitney U test. If we have a within subject design with one factor having two levels and the conditions for parametric tests are satisfied, then we can apply a paired sample t test, otherwise we can apply Wilcoxon signed ranks test.

If we have a between subject design with one factor and more than two levels, then we can apply one way ANOVA, if the parametric test conditions are satisfied; otherwise we can apply Kruskal-Wallis test. And, if we have a between subject designed with two or more factors each having two or more levels and the conditions for parametric tests are satisfied, then we can apply factorial ANOVA. And, in both the cases we can go for Kruskal-Wallis test, if the conditions for parametric tests are not satisfied.

When we have within subject design with one factor having three or more level or within subject design with two or more factors, each having two or more levels we can go for repeated measure ANOVA; if the data satisfies the condition for parametric tests or we can go for Freidman test, if that is not the case. So, this table summarizes different situations in which we can apply specific test and depending on the situation you should choose your test carefully. Now, that is about application of significance tests for data analysis.

This is primarily the analysis task that we do on our empirical data to determine if the observations that we got or observations that we have made are due to the experimental condition rather than by chance. So, once we do that, we can go for building a model from the data. What we can do?

(Refer Slide Time: 36:06)

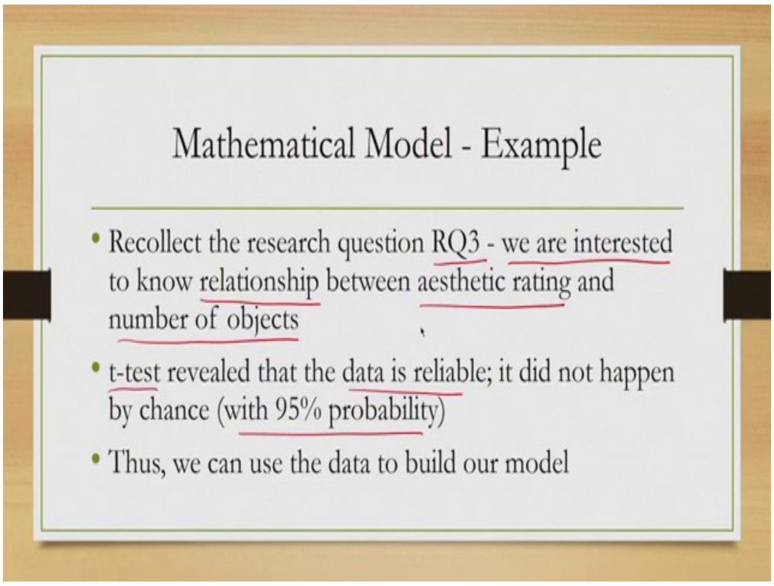


Model-Building from Data

- With the statistical significance test, we can ascertain the reliability of the data
- The test in itself does not help us to build the computational user model
- To get the model, we have to do something more

So, with the statistical significance test we can ascertain the reliability of the data, but from the test we cannot build a model, for that we need to do something more.

(Refer Slide Time: 36:25)



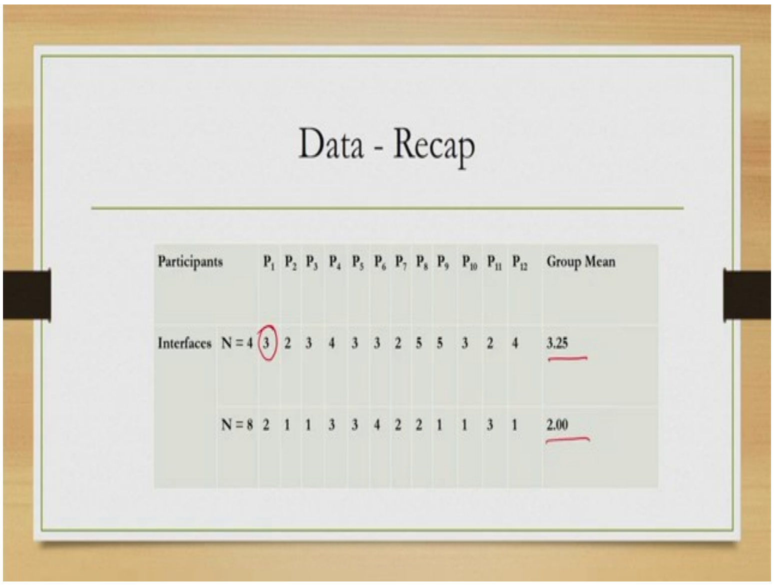
Mathematical Model - Example

- Recollect the research question RQ3 - we are interested to know relationship between aesthetic rating and number of objects
- t-test revealed that the data is reliable; it did not happen by chance (with 95% probability)
- Thus, we can use the data to build our model

Let us try to understand this in terms of an example, the research question RQ3 which we have mentioned in a previous lecture. So, what this research question is all about? It talks about a relationship between aesthetic rating and the number of objects. So, there is only one independent variable number of objects and one dependent variable aesthetic rating and a research question 3 actually talks about the relationship between these two.

Now, from the research question we have formulated two hypothesis and performed a t test to find that the data is reliable. So, it did not happen by chance with 95 percent probability. Once we know that then we can use the data to build a model.

(Refer Slide Time: 37:20)

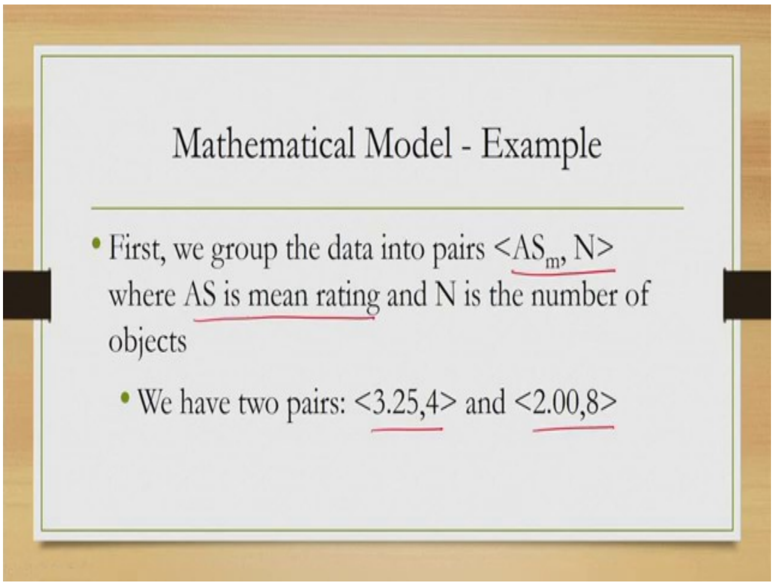


Data - Recap

Participants		P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	Group Mean
Interfaces	N = 4	3	2	3	4	3	3	2	5	5	3	2	4	3.25
	N = 8	2	1	1	3	3	4	2	2	1	1	3	1	2.00

So, for convenience the table of data items is reproduced here, as you can see each cell indicates a rating for a particular value of N. We have also indicated group mean at the end of the table.

(Refer Slide Time: 37:48)

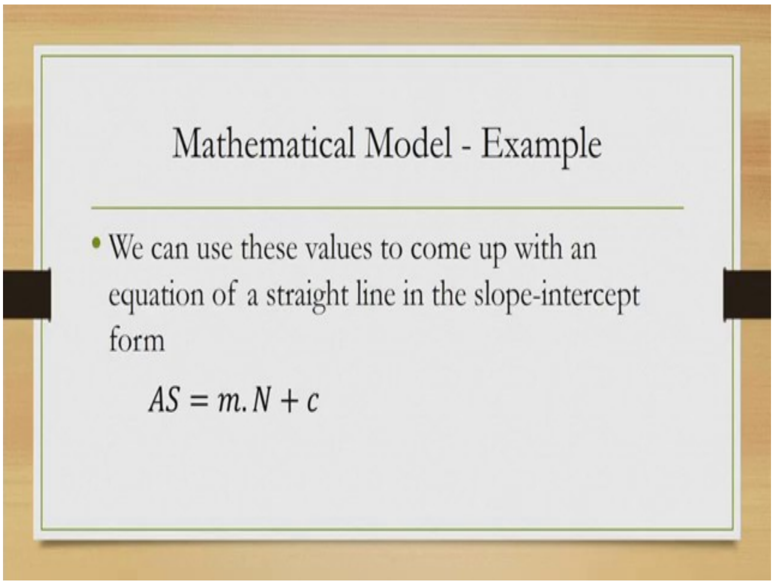


Mathematical Model - Example

- First, we group the data into pairs <AS_m, N> where AS is mean rating and N is the number of objects
    - We have two pairs: <3.25,4> and <2.00,8>

Now, with this data we can actually come up with a model. So, in order to do that we first group the data in to pairs of rating and number of objects, where this rating is the mean rating. So, then we have two pairs: one is the group mean for 4 items on the interface and other one is the group mean for 8 items on the interface.

(Refer Slide Time: 38:22)



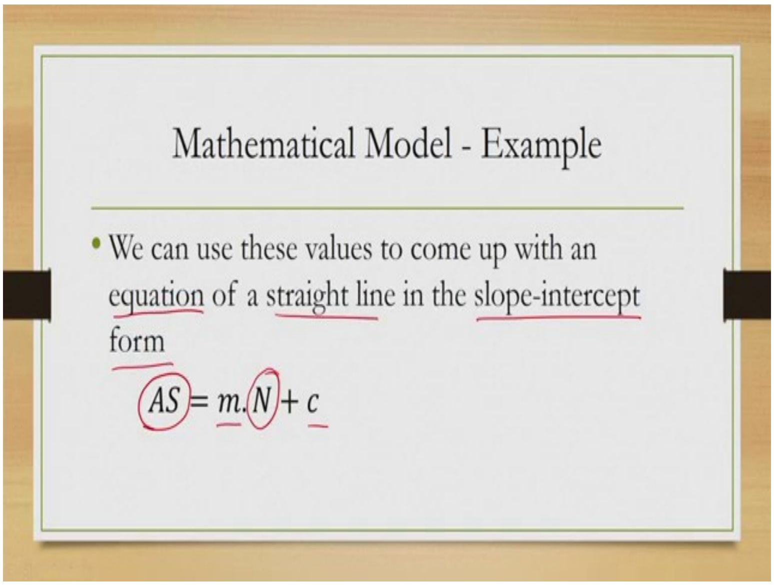
Mathematical Model - Example

- We can use these values to come up with an equation of a straight line in the slope-intercept form

$$AS = m.N + c$$

Now, using these two pairs we can come up with an equation of a straight line using the slope intercept form. So, this form is something like this AS which is the aesthetic score is linearly dependent on N, the number of objects on the interface, m is the slope and c is the intercept of the line.

(Refer Slide Time: 38:53)



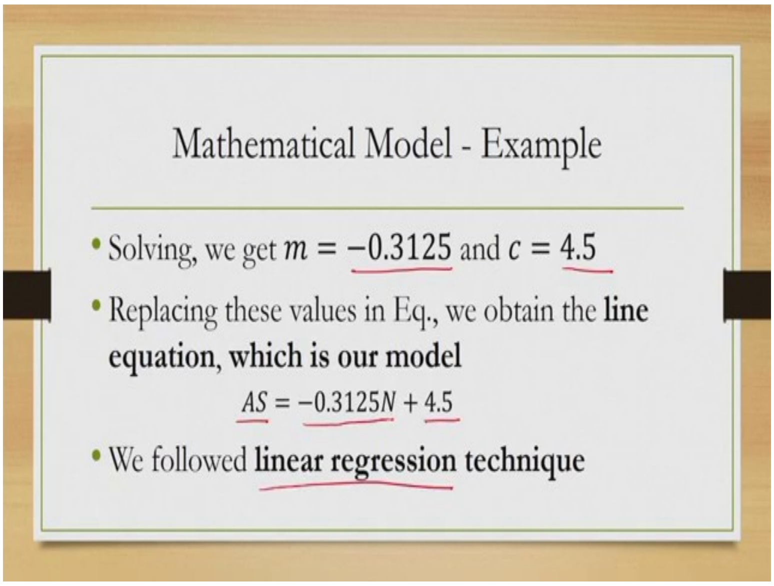
Mathematical Model - Example

- We can use these values to come up with an equation of a straight line in the slope-intercept form

$$AS = m.N + c$$

Now, we replace the two values that we have in this equation to get the values of m and c. So, we set up this system of equations, solve it to get the values of m and c.

(Refer Slide Time: 39:14)



Mathematical Model - Example

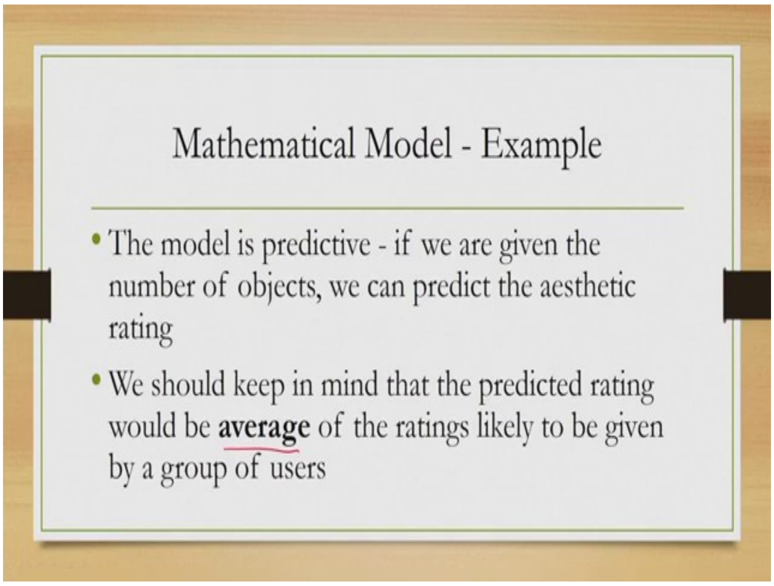
- Solving, we get $m = -0.3125$ and $c = 4.5$
- Replacing these values in Eq., we obtain the **line equation, which is our model**

$$AS = -0.3125N + 4.5$$

- We followed **linear regression technique**

And, if you do this you will find that m is minus 0.3125 and c is 4.5. So, we use these values in our equation to get the final model which is of the form; aesthetic score is equal to minus 0.3125 N plus 4.5. Now, this equation we obtain through linear regression and this equation is our predictive model. So, if we know the value of N, we use the value in this equation to get a likely rating for that particular interface.

(Refer Slide Time: 39:51)

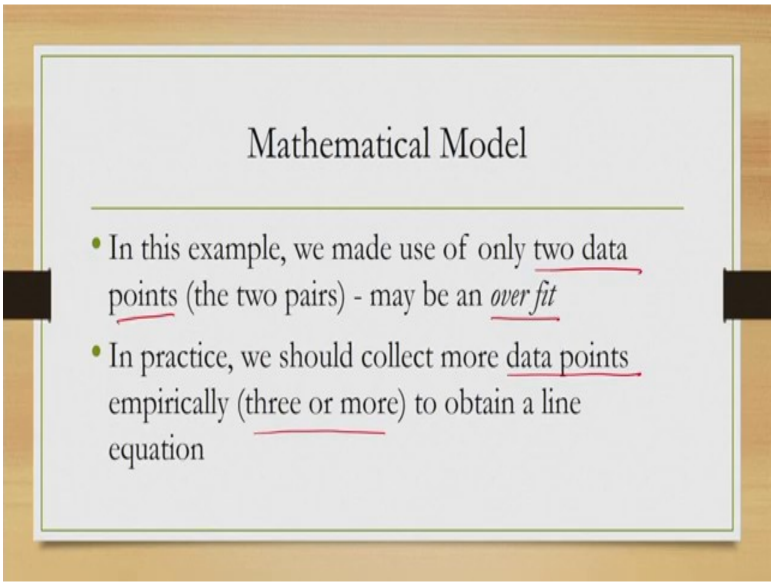


Mathematical Model - Example

- The model is predictive - if we are given the number of objects, we can predict the aesthetic rating
- We should keep in mind that the predicted rating would be **average** of the ratings likely to be given by a group of users

Something we should keep in mind always that is the model that we got in this case, the model for predicting a rating always gives us some average behavior not individual behavior. So, the rating that we may get using this model would be an average of the ratings that we are likely to get from a group of users. So, for each individual user, it may not be the same rating. So, what we get is a group behavior rather than individual behavior.

(Refer Slide Time: 40:29)

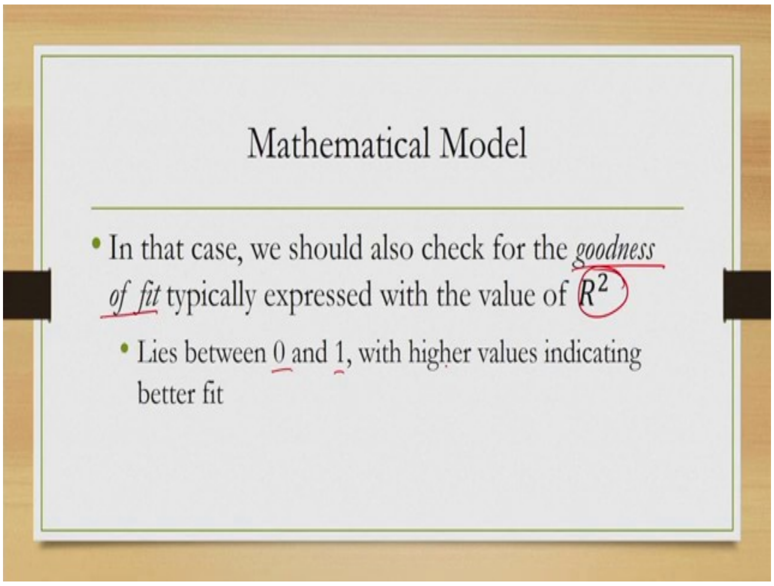


Mathematical Model

- In this example, we made use of only two data points (the two pairs) - may be an *over fit*
- In practice, we should collect more data points empirically (three or more) to obtain a line equation

Another thing is that we have come up with this model by using only two data points which as you know maybe an overfit. Ideally, we should collect more data points typically three or more to get a line equation.

(Refer Slide Time: 40:54)

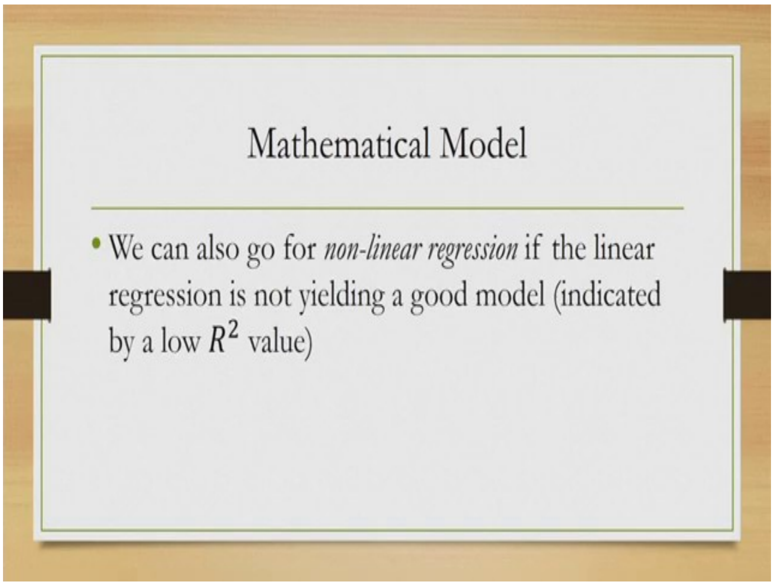


Mathematical Model

- In that case, we should also check for the *goodness of fit* typically expressed with the value of R^2
  - Lies between 0 and 1, with higher values indicating better fit

And, once we are establishing a line equation using regression technique, we should also check for goodness of fit using the value of R square. The R square value lies between 0 and 1 and the higher the value the better is the model fit. So, if you are getting a line equation with R square to be close to 0; that means, it is a overfit and it actually is not modeling the behavior.

(Refer Slide Time: 41:34)

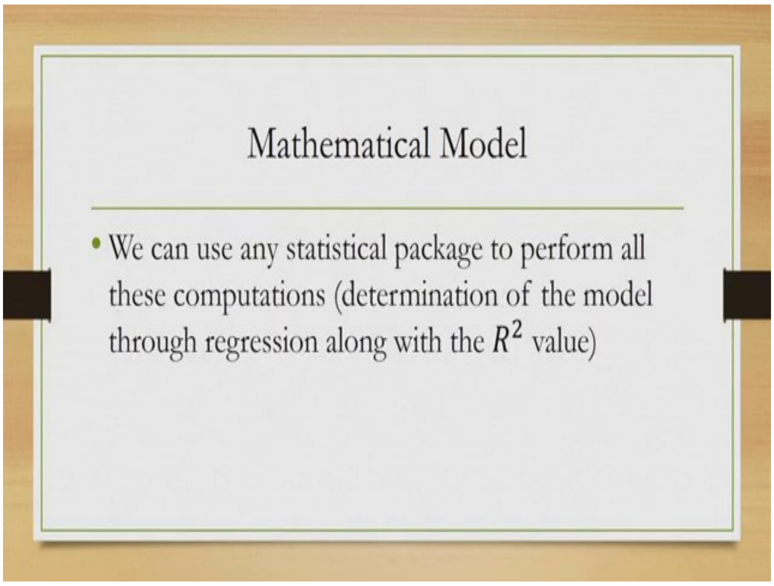


Mathematical Model

- We can also go for *non-linear regression* if the linear regression is not yielding a good model (indicated by a low R^2 value)

Now, linear regression is of course, one way of going for regression technique, it only gives us line equation. And, as you know you can always go for non-linear regression as well to get some non-linear equations and there also we need to check the goodness of fit using R square values.

(Refer Slide Time: 41:57)

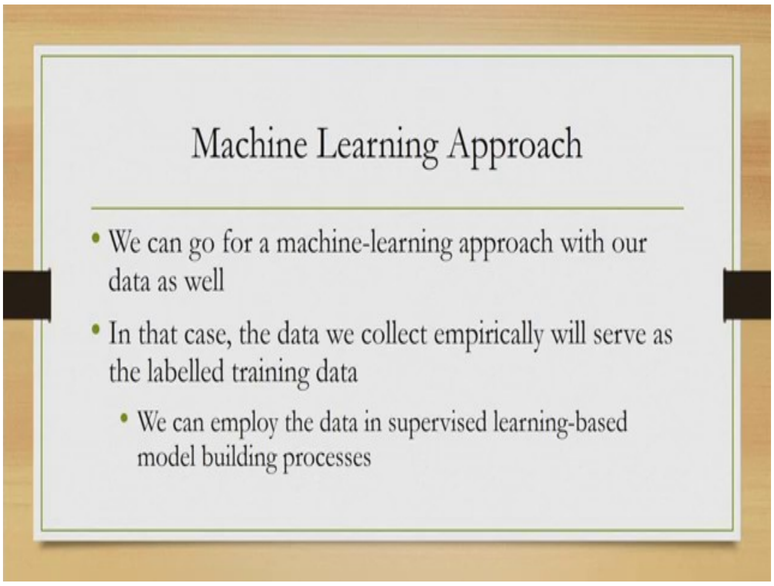


Mathematical Model

- We can use any statistical package to perform all these computations (determination of the model through regression along with the R^2 value)

And, these regression techniques or checking with R square values, all these things can be done automatically using any statistical package. You do not need to do the computation yourself, this is similar to using the package for significance tests. So, already available tools are there and you can use those for the purpose.

(Refer Slide Time: 42:21)

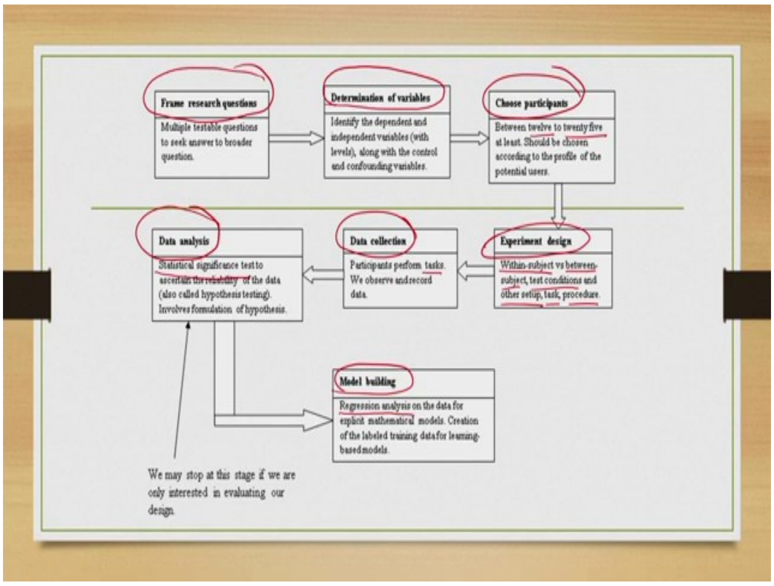


Machine Learning Approach

- We can go for a machine-learning approach with our data as well
- In that case, the data we collect empirically will serve as the labelled training data
    - We can employ the data in supervised learning-based model building processes

Now, getting an equation using regression technique is one way of building a model other way is of course, to build a learning based model where we can use the data to train our model which is typically a classifier and using that trained model we can go for predicting the behavior in real life. So, the data can be used both for training as well as testing. So, that is in summary what we can do with the data, how we can build a model either through regression technique or through supervised learning technique.

(Refer Slide Time: 43:03)



Frame research questions

Multiple testable questions to seek answer to broader question.

Determination of variables

Identify the dependent and independent variables (with levels), along with the control and confounding variables.

Choose participants

Between twelve to twenty five at least. Should be chosen according to the profile of the potential users.

Experiment design

Within-subject vs between-subject, test conditions and other setup, task, procedure.

Data collection

Participants perform tasks. We observe and record data.

Data analysis

Statistical significance test to ascertain the reliability of the data (also called hypothesis testing). Involves formulation of hypothesis.

Model building

Regression analysis on the data for explicit mathematical models. Creation of the labeled training data for learning-based models.

We may stop at this stage if we are only interested in evaluating our design.

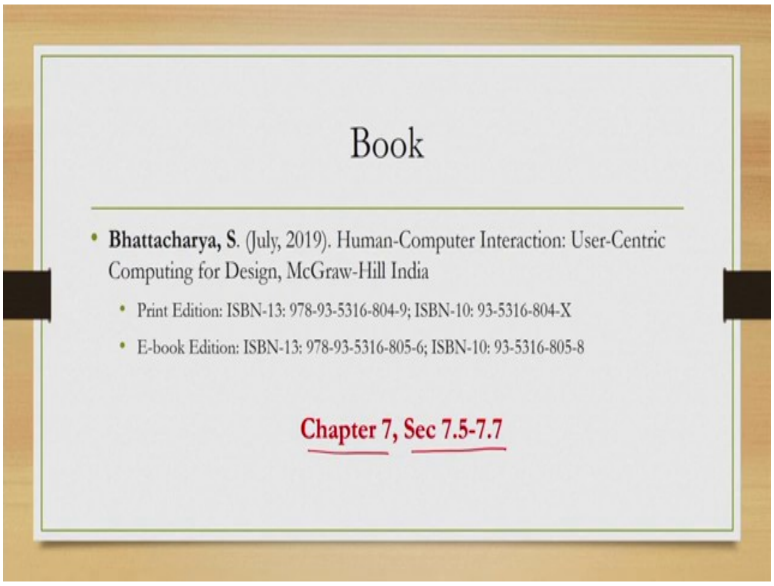
So, at the end I would like to show the entire discussion that we have made in terms of a flow chart. So, this flow chart actually indicates the entire process of empirical research. So, first we start with framing a research question and typically our objective is to frame multiple testable questions to seek answer to a broader question. Once, the questions are framed we go for identification of variables. We identify dependent, independent, control and confounding variables.

The third task is to choose participants, for a pilot study we can use 5 participants, but for actual study between twelve to twenty five participants should be chosen; according to the profile of the users, potential users. This is followed by design of the experiment, where we decide whether to go for a within subject design or between subject design, what should be the test conditions and how to assign the tasks to the participants, what should be the procedure followed and what setup is required and so on.

And, in the actual data collection phase participants are asked to perform tasks and we observe and record data. This is followed by data analysis phase, where we perform statistical significance test to ascertain the reliability of the data. Now, sometimes this phase is also called hypothesis testing which involves formulation of hypothesis from the research questions.

Now, if we are interested in empirical evaluation of our system or the quality of our system, then we can stop at this stage. But, sometimes we can extend it further, sometimes our objective is to build a model. So, we can go to the next stage, that is perform a regression analysis or supervised learning to build a predictive model. So, this is in summary what we do in empirical research.

(Refer Slide Time: 45:32)



Book

- **Bhattacharya, S.** (July, 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, Sec 7.5-7.7

So, whatever we have discussed today can be found in this book, you are advised to refer to chapter 7, section 7.5 to 7.7 to get more details on the topics that we have covered in this lecture.

Thank you and good bye.