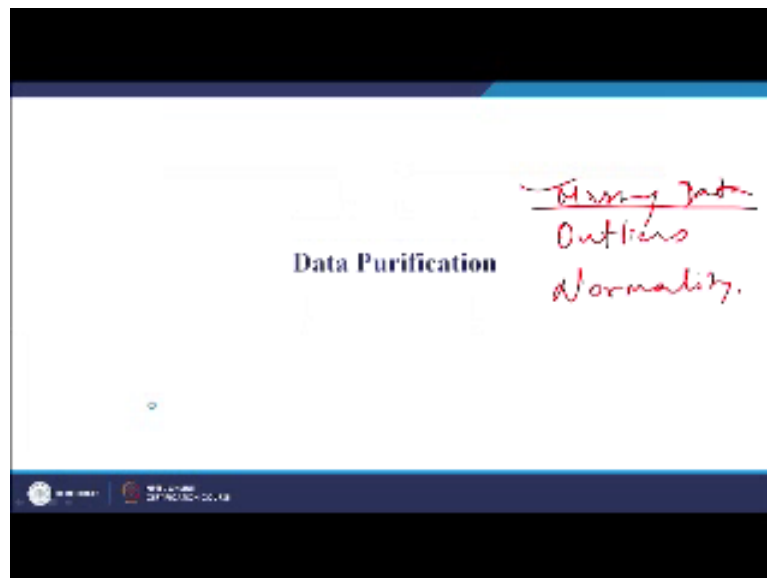


**Marketing Research and Analysis-II**  
**(Application Oriented)**  
**Prof. Jogendra Kumar Nayak**  
**Department of Management Studies**  
**Indian Institute of Technology – Roorkee**

**Lecture - 18**  
**Data Purification and handling – II**

A very warm welcomes to everyone; so, welcome to the lecture series of the course Marketing Research and Analysis. In the last lecture, we were discussing about the importance of data purification, why researchers need to be very, very careful, this is one of the areas which researchers tend to ignore, in the haste for you know doing a publishing paper or completing a research work, but, then they land up into more difficult trouble and then they have to regress back and maybe do the entire research again. So, that consumes in the totality more time, had they been slightly more careful and handle the data better in the beginning itself.

**(Refer Slide Time: 01:25)**



So, data purification becomes a very important subject of discussion for any research course. In the last class, we were talking about the three things, for example, one was the case of missing data. So missing data, then we talked about outliers, we have just begun we have not gone deep into outliers and then the third case was of normality, in which we have started about the first case this is what we were discussing and we will may be carried forward with this only today, in this lecture. So missing data, as I said there is what will you do when there is missing data, with you in your case, in your data set.

(Refer Slide Time: 01:46)

TABLE 1 Hypothetical Example of Missing Data

Case ID	V <sub>1</sub>	V <sub>2</sub>	V <sub>3</sub>	V <sub>4</sub>	V <sub>5</sub>	Missing Data by Case	
						Number	Percent
1	1.2	3.9	6.7	7.0	2.6	0	0
2	4.1	5.7			2.9	2	40
3		9.9		7.0		3	60
4	9	8.6		2.1	1.8	1	20
5	4	8.3		1.2	1.7	1	20
6	7.5	6.7	4.8		2.5	1	20
7	2	8.8	4.5	1.0	2.8	0	0
8	2.7	8.0	1.0	1.8	1.8	0	0
9	1.8	7.6		1.2	2.5	1	20
10	4.5	8.0		1.1	2.2	1	20
11	2.5	8.2		1.1	3.9	1	20
12	4.5	6.4	5.9	1.0	2.5	0	0
13					2.7	5	100
14	2.8	5.7	6.4		3.8	1	20
15	3.7			1.0		5	100
16	7.8	6.4	5.0		2.1	1	20
17	5	3.2		3.3	2.8	1	20
18	2.8	3.2	5.0		2.7	1	20
19	2.2	5.7		2.6	2.9	1	20
20	1.8	8.0	5.0	2.2	3.0	0	0
Missing Data by Variable						Total Missing Values	
Number	2	2	11	5	2	Number: 23	
Percent	10	10	55	25	10	Percent: 23	

For example, this is a hypothetical example you seen there are certain variables V1 V2 V3 V4 V5. Now each variable has got there are certain missing data for example here there is missing data, here there is missing data, and here. So if you can see variable wise, is also you can calculate the missing data's and case wise you can calculate the missing data.

For example from in case if you seen the first case has got no missing data but others there are some missing data, for example two missing data here, ah there is 3 missing data here, 11 so it is distributed. In case of variable you can also you see that there are some missing data here in this variable and maybe this variable has got the highest missing data and in this case this value, this case has the highest missing data.

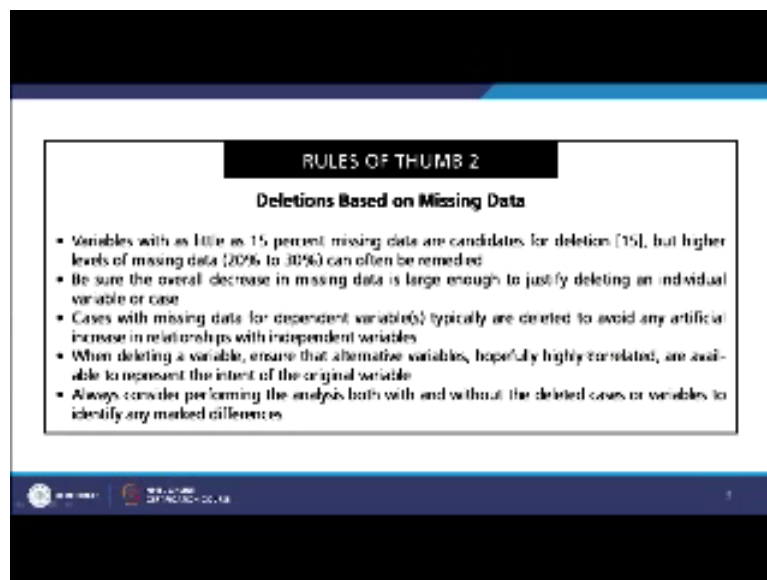
So the question is what will you do when you have such kind of missing data, will you ignore? Is it possible to ignore? Then you can ignore no issues if it is not creating much of a trouble to you. If it is too less and it is not that important. But suppose you cannot ignore, then what? Then you have to may be replace the missing data with something, more realistic, more correct data ok fine, so let us see how do we approach this problem of handling the missing data.

I am sure every researcher must be, will face the problem of missing data but they tend to know you know, general approach, I will tell you honestly speaking is have seen people filling it up with some numbers just some numbers. Because they realise that the computer does not understand and it takes any number given to it but then they do not realise that they

are doing more harm to themselves instead of doing good. Why I am saying that they are doing harm to themselves, because once you have filled in the data, it then how it works, it works through a, large permutation combination across the variables. So once you have done it, the entire data the mechanism changes of how it functions changes.

Now suppose you even yourself might have forgotten which data you have replaced or you have changed. Now when you do that, it could be possible that the entire analysis is coming, off track, it is coming something wrong, weird and wrong. And then you are in a loss, because you do not remember even, what you have done. So, why you should do that? When there is a scientific approach to do handle your missing data, then why should you be doing all this things? Ok. So let us see.

**(Refer Slide Time: 04:27)**



So this is the Rule of Thumb. Deletions based on missing data. The question is if you have less missing data, you may delete it, or let us see, that is, you find that there are few variables which have large number of missing data you can delete them. But the question is, here you have to use your own logic. How much deletion is possible? If I delete one variable out of maybe 20 maybe it is ok, still. But if there are only 5 variables, I am deleting one; maybe I am deleting 20%. So, that you have to decide. Variables with as little as 15% missing data are candidates for deletion.

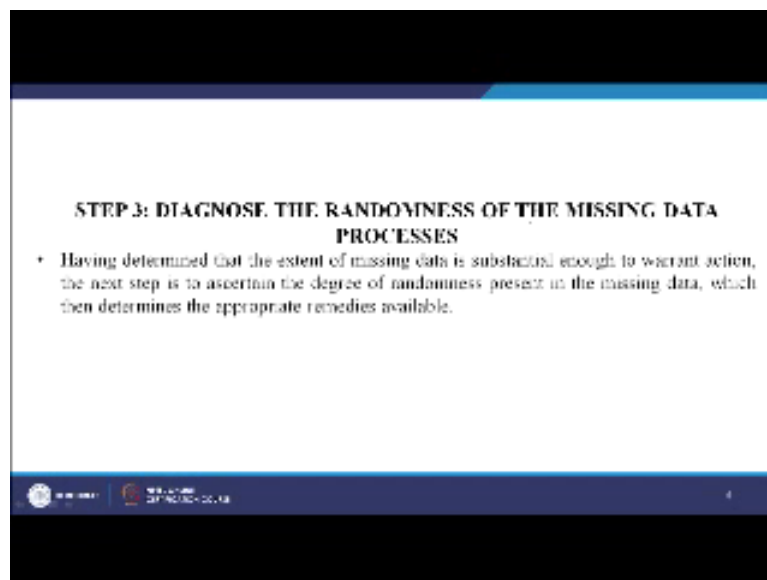
See books could always tell you very nicely that we should be very clear and we should delete it, but the point is, always deletion does not serve our purpose. Because sometimes you know, we may land up in a lesser number of variables or cases. Higher levels of missing data

up to 20% to 30% can often be remitted. So it says if it is less than 15% you may delete it, if it is more than 20 to 30%, you may remit it.

My question is why not even treat the 15%? Why should you delete the first thing? So if it is 15% if you can have a remedy for 20 to 30%, then why not handle it even with 15%? Ok. Be sure the overall decrease in missing data is large enough to justify deleting an individual variable data. So it is, there are sufficient data is there with you that means. Cases with missing data for dependent variables typically are deleted, to avoid any artificial increase in relationship with independent variable.

So, if a dependent variable has lot of missing data, in such a condition they need to be carefully handled. While deleting a variable, ensure that alternative variables hopefully highly correlated means they are more or less behaving the same time to explain the same thing,, highly correlated, are available to represent the intent of the original variable, that means, there is another variable which can take care of the variable that you might be possibly deleting. Always consider performing the analysis both with and without the deleted cases or variable to identify the marked differences.

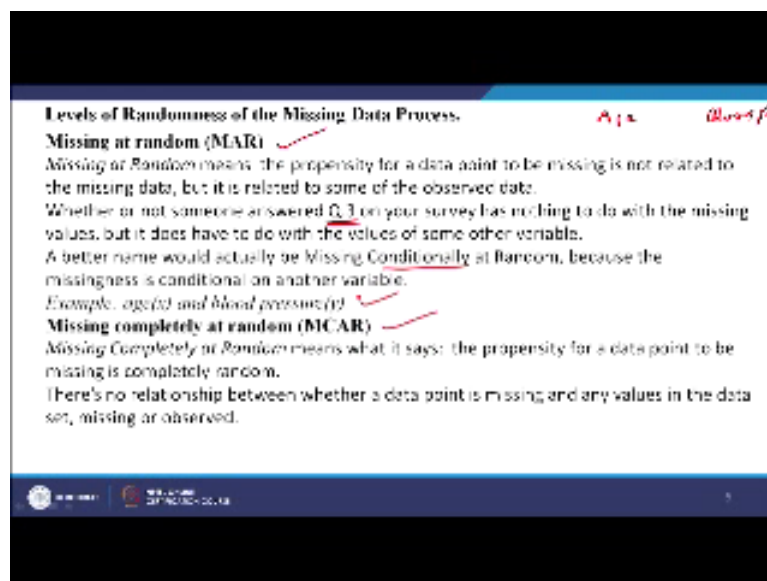
**(Refer Slide Time: 06:41)**



So there is a statistical technique we will see how to deal with such problem diagnose the randomness of the missing data, is it how what is the pattern of the missing data, what is it random, highly random, completely random or it is just at random. So having determined the extent which you have done is substantial enough to one and action, the next is to find the degree of randomness present in it.

Now why in that, why it is important? what will happen if I do not consider the, I have sufficiently large number of data so what happens if I do not study the randomness, it is possible that you may you might be missing a very, very important observation that why are people only missing, there is missing data in a particular data set or a particular variable or why a particular kind of respondent for of a particular place maybe are having missing data. So those are very important observations for a researcher and they need to be carefully handled and observed and understood also.

(Refer Slide Time: 07:50)



**Levels of Randomness of the Missing Data Process.**

**Missing at random (MAR)**  
Missing at Random means the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data. Whether or not someone answered Q. 3 on your survey has nothing to do with the missing values, but it does have to do with the values of some other variable. A better name would actually be Missing Conditionally at Random, because the missingness is conditional on another variable.  
Example: age and blood pressure.

**Missing completely at random (MCAR)**  
Missing Completely at Random means what it says: the propensity for a data point to be missing is completely random. There's no relationship between whether a data point is missing and any values in the data set, missing or observed.

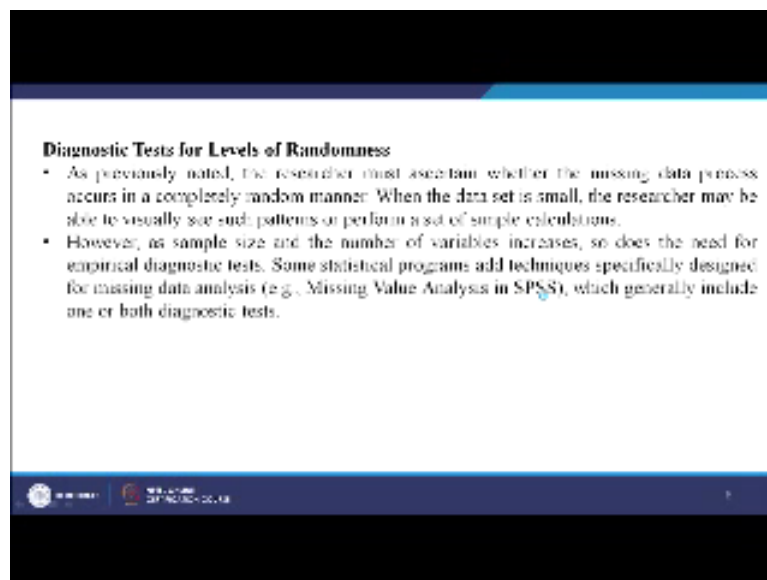
So, the randomness also missing data sometimes give few extra information, ok. So, two things. Missing at random, Missing completely at random. Missing at random means the propensity for a data point to be missing, is not related to the missing data, but it is related to some of the observed data. Now let us see this. Now whether or not someone answer question number 3 let us say you have certain questions, I have just given example question number 3 out of 20 questions question third.

On a survey, has nothing to do with the missing values but it has to do with the values of some other variable. In the last lecture I had explained you about blood pressure and age relation, age and blood pressure. So when I said is there any relationship here, it was found that young people were generally not reporting their blood pressure values, ok. So this is the case, ok. So a better name would actually be missing conditionally.

So the condition, there is a condition, if it is young they are not reporting so conditionally at random, because the missingness is conditional on another value. That means the missing value of one variable is dependent because of another variable. That is why this case is called a missing at random case and it is more complicated in comparison to the other one which is missing completely at random. What it says?

The propensity for a data point to be missing is completely random. There is no pattern. It is that, people have not given reply. Why they have not given? There is no why for it. Yes the difference is, in the missing at random case, there is maybe a why? Why they have not given there is a reason. But here there is no reason, this relationship between whether a data point is missing and any values in a data set missing out of that. So this is very easy to handle, ok.

**(Refer Slide Time: 09:35)**

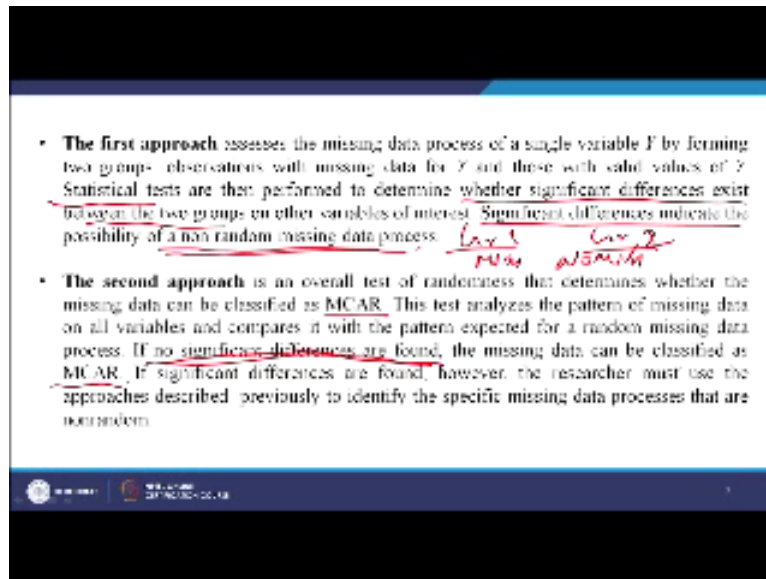


So testing how to test the level of randomness previously noted the researcher must ascertain, whether the missing data occurs in a completely random manner, when the data set is small the researcher may be able to visualise it. Suppose it is a small dataset you visualise the same as the patterns and perform some calculations however as sample size increases then our value increases, the need for test also goes on increasing some statistical programs and technique specifically designed for missing value for example in SPSS which I will conduct and show you, in just in this lecture.

So, when you have a large data sets, suppose you have a data set of 10,000 people, can you and the hundred variables can you visually do anything with that? Impossible so in such a

condition, you can use some techniques available in the software and take care of the missing data.

(Refer Slide Time: 10:32)



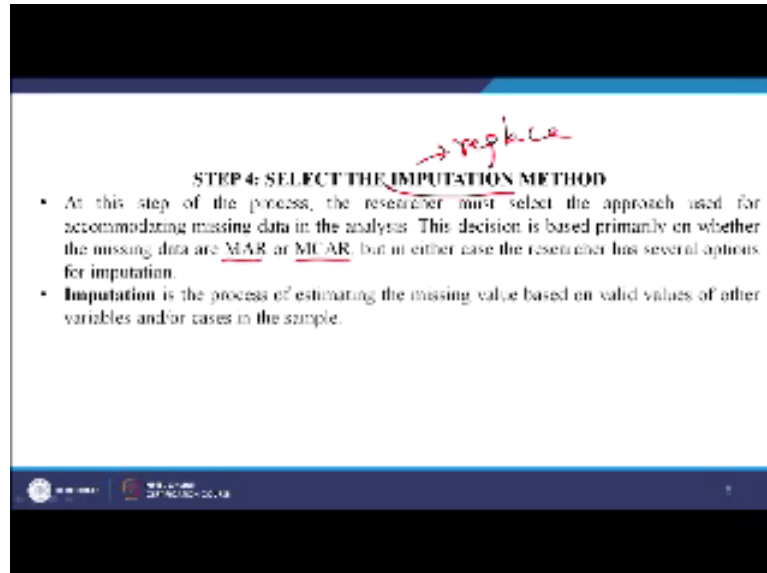
The first approach accesses the missing data process for single variable  $Y$  while forming two. Now what is done handle, to check whether the missing data is actually having impact or not? What is done? Two, the data is divided into 2 groups, one group where some missing data is available, the other group where no missing data is available. And then the simple t test is a statistical t test is performed to determine whether significant differences exist between the two, so you understood the group one let us say group two.

So group one let us say has got missing values? it has got no missing values. No missing values. Now when I compare the data set, these two datasets, suppose significant differences exist, it indicates the possibility of a non random missing data process. So what you say, if a significant differences exist it is a case of non random missing data. The second is called the overall test of randomness.

It determines whether the missing data can be classified as MCAR, how you see. This test analyse all the pattern of missing data on all variables and compares it with the pattern expected for a random missing data. If no significant differences are found in this case significant differences were found. Here no significant difference are found, then the missing data can be classified as I hope you have understood what happens take two groups and compare these two groups.

If significant differences exist between group 1 and group 2 then it is a case of a missing at random data if there is no significant difference then, it is just a randomness and is a such a worry some factor. So here we will not be so much worried, ok.

**(Refer Slide Time: 12:32)**

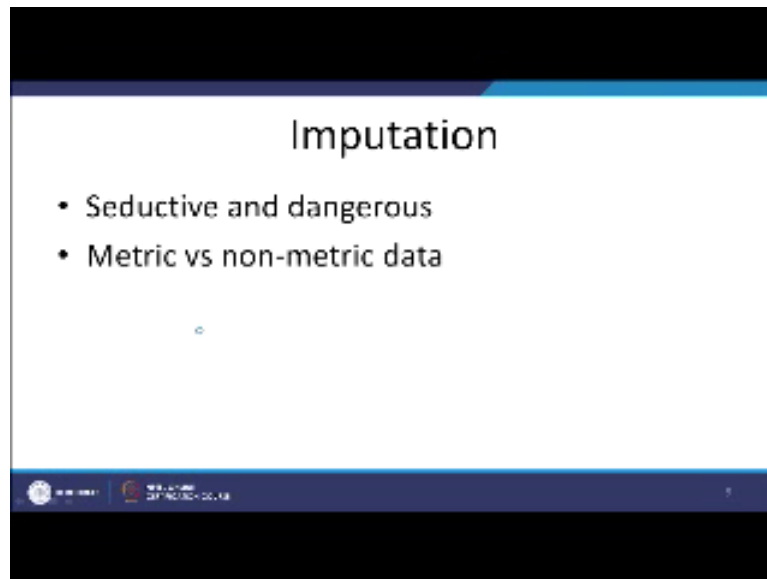


Now once you are done with these things then we come to the fourth and the most important part, what is this? Here we would try to impute or replace, impute or replace, replace the missing value with some logically right value, ok. The researcher must select the approach used for accommodating missing data analysis. The decision is based primarily on whether the missing data are missing at random or missing completely at random.

But either case the result has several options for application. So let see, lets not get into this random and non random rather see, how you can impute, replace the data, because some of you might be very early, young to research, So I am not trying to make it more complicated let us understand, you have understood the meaning, that is very important you have, you should understand, what is it somebody ask you should be able to answer, but just how do I impute, how do I correct my data? Let us go to that.

**(Refer Slide Time: 13:29)**





Imputation is the process of estimating the missing value based on values based on some other variables and cases. Let us see, either you can calculate or there are some other methods also. But remember one thing this is a caution. Imputation can be sometimes very seductive and dangerous why? it is seductive, because you are getting a full complete data, so there is no missing data, so you are very happy, why it is dangerous, because you did not understand a reason for the missing data.

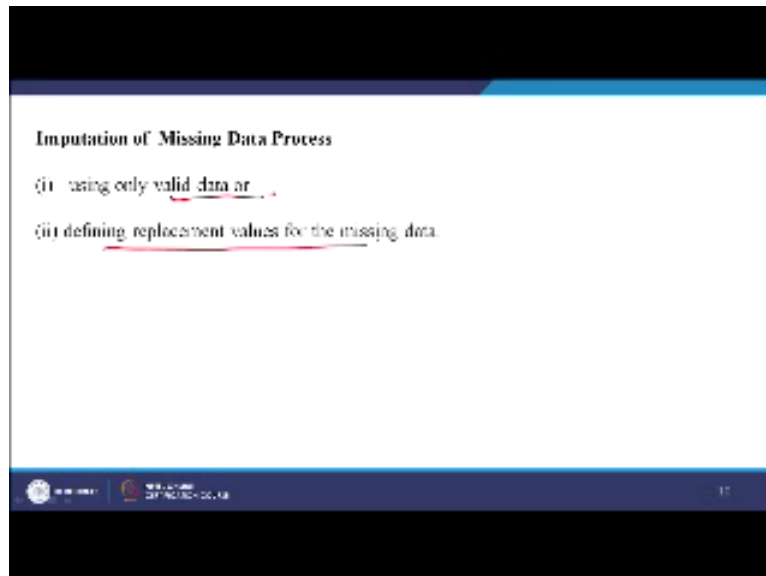
Maybe you are, did not understand why there was a missing data, was it a logical reason is there a logical reason for having missing data and you ignored it and you just filled up the division data so now that thing is still lying you why he is not the respondent did not fill up the data that thing you have just ignored without any logical reason and you tried to just go across to your analysis, so now it is very dangerous. So you do not know why it had happened but you are filled it up. So you are forcing the data now, ok.

The second thing is metric versus non metric data, see naturally when the data is metric, that means LSA interval scale or ratio scale I continue scale for that. So when the data is metric scale, metric data then ok, fine, you can go for a missing data approach. What is the data for example, non-metric, for example, gender male and female, now, suppose there is a person has not filled up whether is a male or female, should you try to impute the data, no. you should not be doing it.

I would advice never ever go for imputation of non metric data, because it makes no logical sense. Had the data being a metric data for example in a scale of 1 to 5, let us say, how much

is the score the respondent is giving. Now suppose people have given, different scores somebody has given 2, somebody has given 4, somebody has given 3, somebody has given five, somebody has again given 4 this is just fantastic. somebody is not given listen here he is not given we can determine by using some substitution method, but if it is a non metric data it is 10, or it is just a nominal scale, it is only a representation, what is the meaning of a train to fill it up? it is a very big mistake people do, so be very careful with it.

(Refer Slide Time: 15:40)

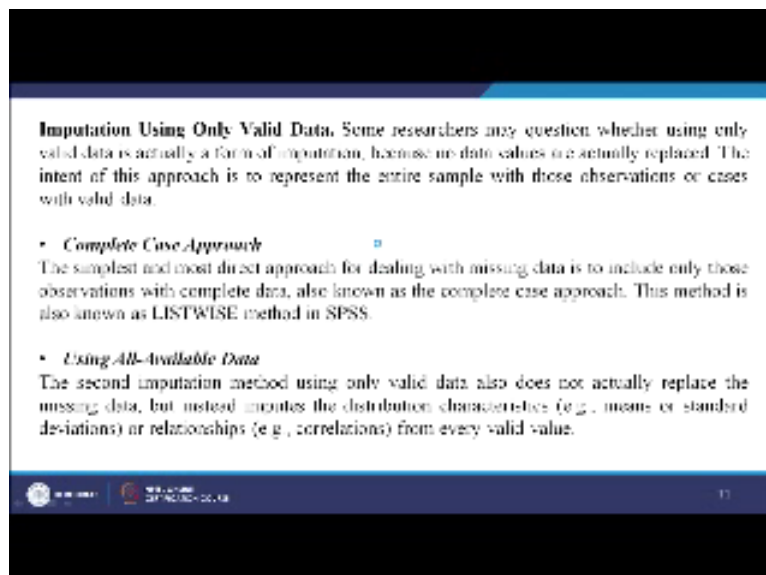


**Imputation of Missing Data Process**

- (i) using only valid data or ...
- (ii) defining replacement values for the missing data.

17

(Refer Slide Time: 15:47)



**Imputation Using Only Valid Data.** Some researchers may question whether using only valid data is actually a form of imputation, because no data values are actually replaced. The intent of this approach is to represent the entire sample with those observations or cases with valid data.

- **Complete Case Approach**  
The simplest and most direct approach for dealing with missing data is to include only those observations with complete data, also known as the complete case approach. This method is also known as LISTWISE method in SPSS.
- **Using All-Available Data**  
The second imputation method using only valid data also does not actually replace the missing data, but instead imputes the distribution characteristics (e.g., means or standard deviations) or relationships (e.g., correlations) from every valid value.

17

Imputation of missing data process use only the valid data or replace values for the missing data. Now the different methods let us see each one of them. Imputation using only valid data some researchers they question whether using only valid data, is actually a form of

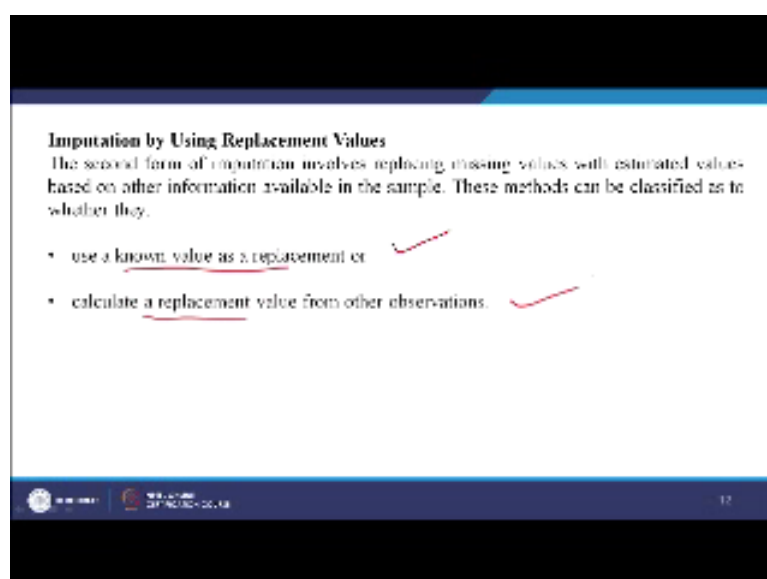
imputation because no data values that are actually represents you just deleted maybe you are not replaced anything.

So, the intent of the approaches to represent the entire sample with those observations which have a valid data. it ok fine no issues good but suppose in the process you have deleted a large number of cases, then their opinion or their observation their person, you know perception, you have ignored. By doing this may be a large size of the population you have missed or a large sample you have missed so you do not understand what these people were even thinking what was that feeling, so that is dangerous. So deleting too much can also be a dangerous.

Now coming to second, the methods complete case approach. The best the simplest and the most direct approach for dealing with missing data to include only those observations with complete data also known as the complete case. This method is also known as list wise method in SPSS, ok forget it, this is complete so anyway a complete case is always good whereupon again I am saying is complete case, if you delete too much, then also it is not advisable.

The second imputation method is used as only validator that does not actually replace the missing data but instead imputes the distribution characteristics. Now what is the distribution characteristic? Through the mean, the standard deviation, or the correlation, so it uses all the available data and tries to use these parameters to fill up the value, the missing data.

**(Refer Slide Time: 17:40)**

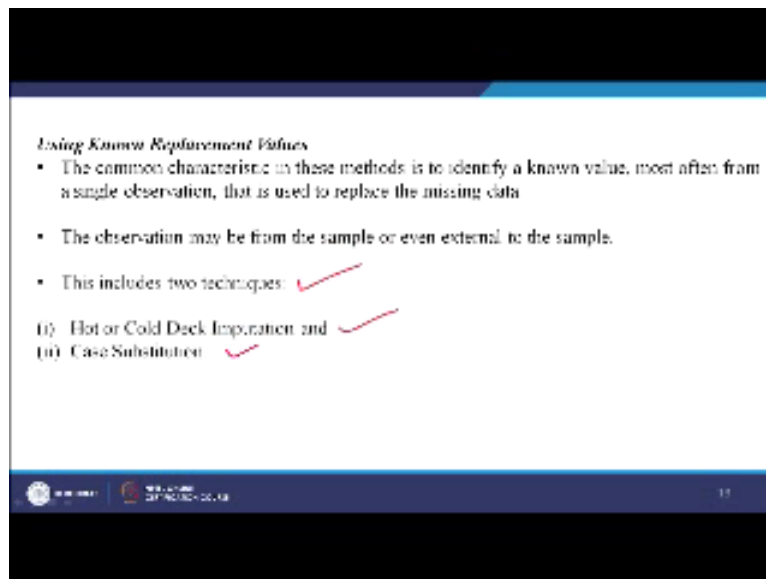


**Imputation by Using Replacement Values**  
The second form of imputation involves replacing missing values with estimated values based on other information available in the sample. These methods can be classified as to whether they:

- use a known value as a replacement or ✓
- calculate a replacement value from other observations. ✓

At the bottom of the slide, there is a logo on the left, the text 'SPSS 2019-2020' in the center, and the number '12' on the right.

(Refer Slide Time: 17:42)

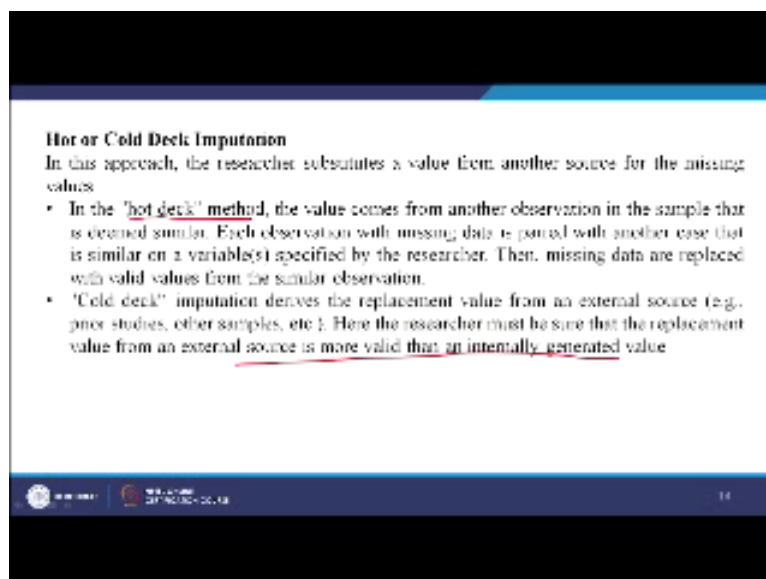


*Using Known Replacement Values*

- The common characteristic in these methods is to identify a known value, most often from a single observation, that is used to replace the missing data.
- The observation may be from the sample or even external to the sample.
- This includes two techniques: ✓
  - (i) Hot or Cold Deck Imputation and ✓
  - (ii) Case Substitution ✓

17

(Refer Slide Time: 17:55)



**Hot or Cold Deck Imputation**

In this approach, the researcher substitutes a value from another source for the missing values.

- In the "hot deck" method, the value comes from another observation in the sample that is deemed similar. Each observation with missing data is paired with another case that is similar on a variable(s) specified by the researcher. Then, missing data are replaced with valid values from the similar observation.
- "Cold deck" imputation derives the replacement value from an external source (e.g., prior studies, other samples, etc.). Here the researcher must be sure that the replacement value from an external source is more valid than an internally generated value.

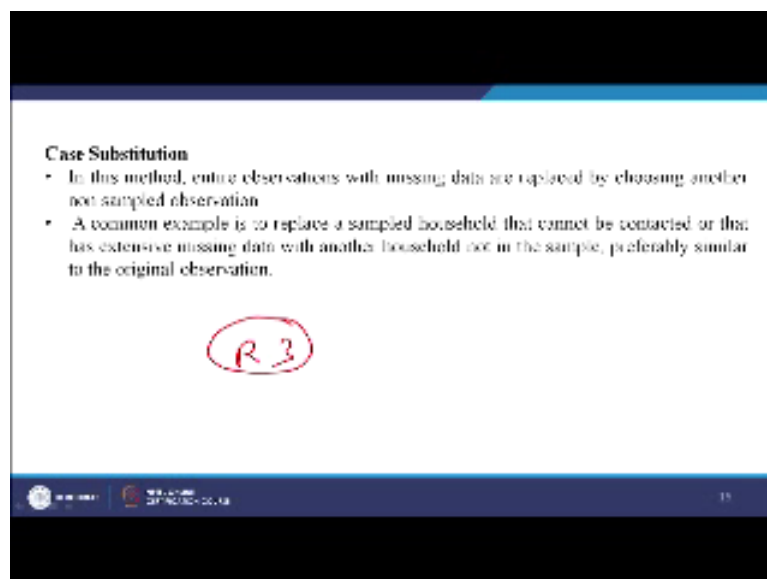
18

Now coming to imputation by using replacement values so use a known value, or you calculate the value. Now when you do this using the known replacement value, the observation techniques, first is called the hot and cold deck imputation and the other is called case substitution method. I will explain each one. Now what is hot and cold deck imputation, in the hot and cold deck imputation method, the value comes from another observation in the sample that is deemed similar.

So another observation which is very similar looking, each observation with missing data is paired with another case, that is similar on a variable specified by the researcher. So any similar looking variable if it is available, you can use those variables data and fill up the missing data. So, then the missing that replaced with valid values from the similar observation this is a hot approach ok.

The cold approach says, the replacement is from the external source, not from the existing data set from external source for example, prior studies or other samples. So you have seen somebody has done a similar study and there is a data set you feel it is very similar to mine and you can and there is a pattern so you can use those data and fill it up. This is the cold data approach. Here the researcher must be sure the replacement value is more valid than an internally generated value. So if he, if the researcher is so sure, you are free to do it no issues.

**(Refer Slide Time: 19:05)**



**Case Substitution**

- In this method, entire observations with missing data are replaced by choosing another non-sampled observation
- A common example is to replace a sampled household that cannot be contacted or that has extensive missing data with another household not in the sample, preferably similar to the original observation.

(R 3)

19

Next, comes the Case substitution. In this method entire observations is missing data are replaced by using another non sample of solution for example I have I can fill up a case you know, a person and respondent 3 is missing, let us say respondent 3 is missing. Whatever responses, profile is I can select similar another person and take data from him. So fine so and the case is substituted. A common example is replace a sample household that cannot be contacted or that has extensive missing data with another household not in the sample preferably similar to the original observations, fantastic, clear.

**(Refer Slide Time: 19:43)**

**Calculating Replacement Values**

- The second basic approach involves calculating a replacement value from a set of observations with valid data in the sample
- This includes two techniques
  - (i) Mean Substitution and
  - (ii) Regression Imputation

18

Now coming to the calculation of the replacement values so the second approach includes two methods, one the substitution mean substitution and the regression.

**(Refer Slide Time: 19:56)**

**Mean Substitution**  
One of the most widely used methods, mean substitution replaces the missing values for a variable with the mean value of that variable calculated from all valid responses.

**Regression Imputation**  
In this method, regression analysis is used to predict the missing values of a variable based on its relationship to other variables in the data set.

19

Now what is the difference, let us see, Mean substitution you replace the missing values for a variable with the mean value. So the mean is most the most standard method and it is one of the best methods. In the regression method, the regression analysis is used to predict the missing values of a variable based on its relationship to other variables of that data set. It is also possible, you can use the normal equation and find out and you can replace it or you can use the mean substitution, in the mean.

**(Refer Slide Time: 20:26)**

**RULES OF THUMB 3**

**Imputation of Missing Data**

- Under 10% Any of the imputation methods can be applied when missing data are this low, although the complete case method has been shown to be the least preferred
- 10% to 20% The increased presence of missing data makes the all-available, hot deck case substitution, and regression methods most preferred for MCAR data, whereas model-based methods are necessary with MAR missing data processes
- Over 20% If it is deemed necessary to impute missing data when the level is over 20 percent, the preferred methods are:
  - The regression method for MCAR situations
  - Model-based methods when MAR missing data occur

17

What is the Rule of Thumb? The Rule of Thumb says, under 10% of the missing data is there, any of the imputation methods can be applied when missing data are this low, although, the complete case method has been shown to be the least preferred. When it is 10 to 20%, the increased presence of missing data, makes the all available hot deck substitution method, and regression method most preferred for missing completely at random, whereas model based approaches are necessary with the missing at random missing data processes. So I said about the EM method I think in the last lecture, the Estimation and the measures of parameter.

Over 20% if is there, so the method is regression and model based method. So this is basically for the MAR. So these are some of the; what would you do when the missing data's. So this tells you some of the advantages when it is a complete data for example it is simple, but the disadvantage is, it is most affected by the non random processes, best used when the large sample size is large, strong relationship among variables exist, ok.

**(Refer Slide Time: 21:15)**

**Imputation Using Known Replacement Values**

<p><b>Case Substitution</b></p> <ul style="list-style-type: none"> <li>Provides realistic replacement values (i.e., another actual observation) rather than calculated values</li> </ul>	<ul style="list-style-type: none"> <li>Must have additional cases not in the original sample</li> <li>Must define similarity measure to identify replacement case</li> </ul>	<ul style="list-style-type: none"> <li>Additional cases are available</li> <li>Able to identify appropriate replacement cases</li> </ul>
<p><b>Hot and Cold Deck Imputation</b></p> <ul style="list-style-type: none"> <li>Replaces missing data with actual values from the most similar case or best known value</li> </ul>	<ul style="list-style-type: none"> <li>Must define suitably similar cases or appropriate external values</li> </ul>	<ul style="list-style-type: none"> <li>Established replacement values are known, or</li> <li>Missing data process indicates variables upon which to base similarity</li> </ul>

17

Low levels of missing data, here are some of the thing, which you can.

(Refer Slide Time: 21:38)

**Imputation by Calculating Replacement Values**

<p><b>Mean Substitution</b></p> <ul style="list-style-type: none"> <li>Easily implemented</li> <li>Provides all cases with complete information</li> </ul>	<ul style="list-style-type: none"> <li>Ignores variance of the distribution</li> <li>Distorts distribution of the data</li> <li>Depresses observed correlations</li> </ul>	<ul style="list-style-type: none"> <li>Relatively low levels of missing data</li> <li>Relatively strong relationships among variables</li> </ul>
<p><b>Regression Imputation</b></p> <ul style="list-style-type: none"> <li>Employs actual relationships among the variables</li> <li>Replacement values calculated based on an observation's own values on other variables</li> <li>Unique set of predictors can be used for each variable with missing data</li> </ul>	<ul style="list-style-type: none"> <li>Removes existing relationships and reduces generalizability</li> <li>Must have sufficient relationships among variables to generate valid predicted values</li> <li>Understates variance unless error term added to replacement value</li> <li>Replacement values may be "out of range"</li> </ul>	<ul style="list-style-type: none"> <li>Moderate to high levels of missing data</li> <li>Relationships sufficiently established so as to not impact generalizability</li> <li>Software availability</li> </ul>

17

Case substitution, for example is, it provides a realistic replacement value you have replaced the case itself, must have but the problem is this and disadvantage is that you have to have additional inventory additional cases, which is not there in the original sample. So, the best used when additional case are available. And you are able to identify appropriate cases ok. Hot and deck similarly, replace the missing data with actual values of the most similar cases so this is good thing advantage, but disadvantage is must define suitably similar cases, your logic of understanding, which is the most similar is also a questionable, sometime it can be wrong also.



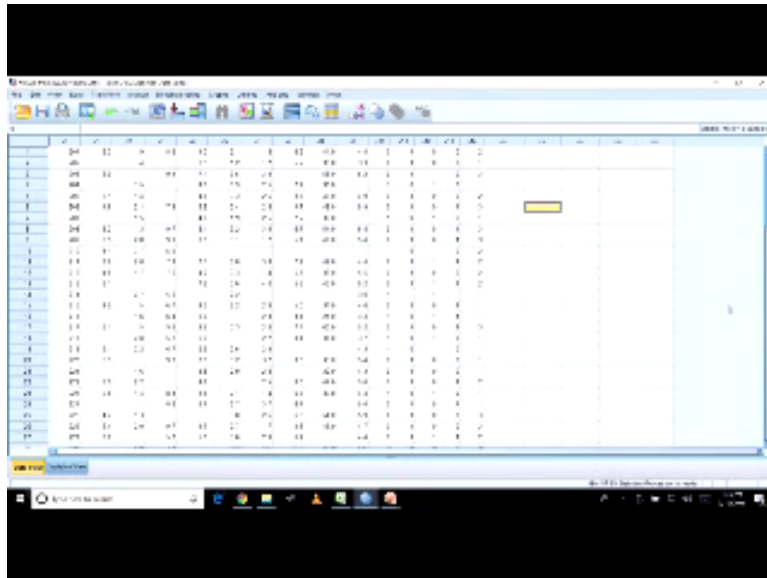
So, establishment replacement values are known then it is wonderful then you can easily go for it. Mean substitution is easily implemented, very simple. It reduces the variants of the distribution. So, the unexplained component is reduced. And when it is used? When the low levels of missing data are there relatively speaking, strong relationship among variable exists.

Regression is used when employees actual relationship among the variables replacement values calculate based on the observations own values and other values, variables. That is the let us say, the number of independent variables available. Unique set of predictors can be used for each variable with missing data. What is the disadvantage? It reinforces the existing relationship; after all you are trying to use the same relationship.

So, it reinforces the existing relationships and reduces the generalizability, so now you cannot generalize, what is generalized ability it means external validity is reduced. Now you cannot say that this study can be applied everywhere because the question is by again and again using the same existing relationships and trying to utilise the values for example in a bootstrapping method also we do that, which is questionable, you should not be doing more.

And of repeating the exercise again and again must have sufficient relationship among variables to generate value predicted value. So, this is some of the thing. So, where it is used? Moderate to high level missing data is there. Relation sufficiently established who is not to impact general. So this is very important. So these are some of the things that you need to be very careful. Now let me explain how you should do it. So this is the case which I have brought a data set. Now this data set as I have shown you there are large numbers of missing data.

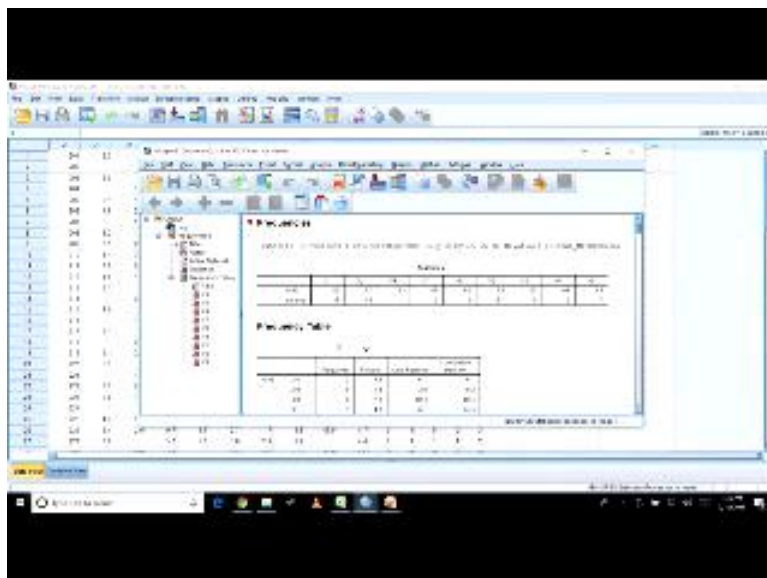
**(Refer Slide Time: 24:05)**



At the initial stage, let me find out how much of missing data's are there, ok. So what I am done is, I have just taken, some missing you know so I have taken some of the I can even use this but since I am not going to correct them, so I am not interested much. So what I am doing is trying to find out how many missing values are there. In V1 to V9, if you see the highest number of missing values are in V1, and the lowest one is in I think 6, V6.

So, variable one has got the highest number of missing data, the question is passed as a researcher, you should start thinking, why is it that when the questionnaire is the same and is used at the same respondent why is variable one having so many missing data? Is there any specific reason? People are not willing to answer it or they do not understand it or something is wrong is happening, so you need to understand this ok.

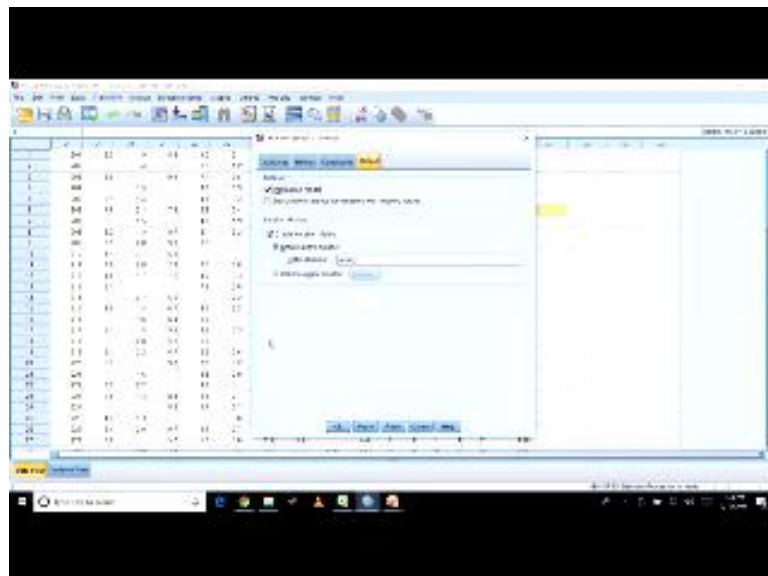
**(Refer Slide Time: 25:12)**



So this is something like the free from the frequency table what I have done is a just find out found out how much of missing data is there, ok. Now how do you impute the missing data or replace. Now let us go, 7 methods are there. Let me explain you the one of the methods here for example go to analyse see if you want to be multiple imputation you can do it or if you just want to, you know, you will go to transform and see here replace missing values so what you can do it you can use this method also, there are several options given you mean of nearby point median of nearby point linear interposition linear trend at point.

Whatever, the simplest some serious mean I am taking, for example, I can just try it out on this one. So I can use this method and what I will get is, I will get a new one variable. So, I can have it on my table and I can make the study for example, let us see. So what I have done, now if you minimise this, now V1 has been created can you see this, now this V1 has got almost no there is no missing data. So you can do this, but it is one by one doing is very cumbersome. So what I am doing is, I will delete this, so kindly delete this if it is possible, ok. Let us related with their otherwise we are we are not bothered at the moment

**(Refer Slide Time: 27:03)**

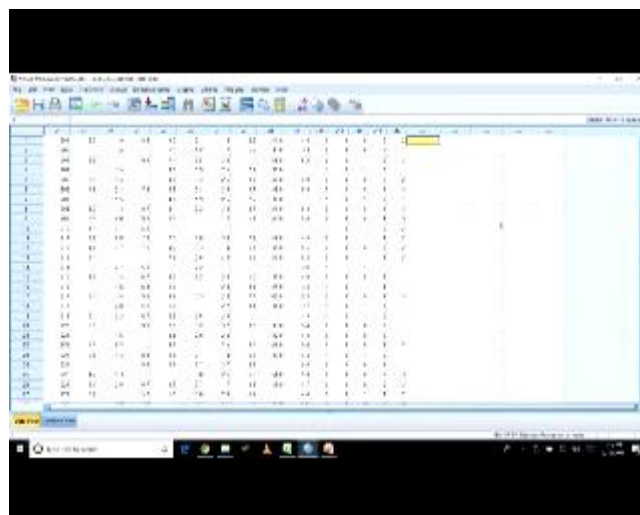


So what I am doing is I am running exercise again, how do you do this? So go to analyse there is something you can see multiple imputation, can you see this, so it is multiple imputation it will ask you to handle the missing data, so what I have done is I am creating a data set, for example, so I will create the name is JK and so I am saying jkn is my name, so jkn I am giving it a name let us say 3, ok. So the new data sets name is taken 3 ok so I have done it and I am saying ok.

So, once I have done this, so what I can do it now I can go to the jkn 3 data set, ok. So the jkn 3 data set you can see now if you look at it properly the missing data I don't think you can see any missing data out here. So all the missing data has been handled properly, so V1 V2 V3 ok so that the question is, let me, let me read it again, so what I am doing is I am taking all the variables 1 2 3 4 5 6 7 8 9 ok, now, you can see it has given you some imputations, correct, the number of imputation.

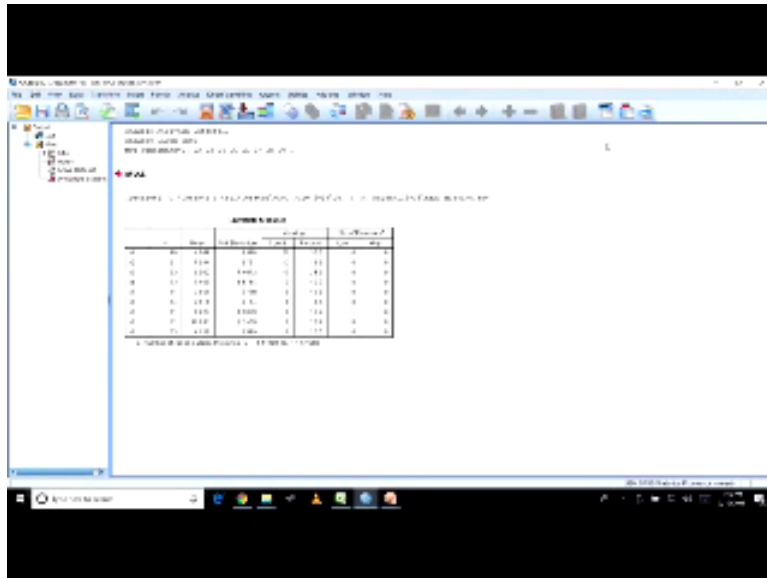
Now what is the number of imputation, imputation is but nothing but the number of iterations, it is a allowing you, generally this software is allowing for 5, you can increase it further also. So what it does is takes five imputed values and then after taking this five imputed values, it takes a pooled value. The pooled value is nothing but a representation of the five imputed values.

**(Refer Slide Time: 28:32)**



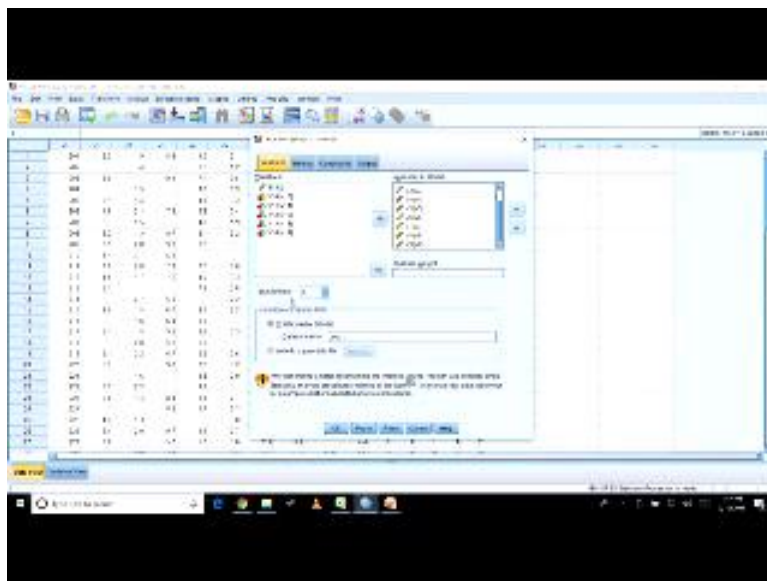
Let me show you how to do that, now let me show you how to handle the missing data in a data set. This is a case of the data set, which I had always showed you also. Now there are lots of missing data's as you can see. So first you can do one thing you can just check how much of missing data is there, to do that, you go to analyse, missing value analysis, ok. So, V1 V2 V3 V4 V5 V6 V7 V8 V9 so I want to see the patterns for example, so you can check the them also.

**(Refer Slide Time: 29:19)**



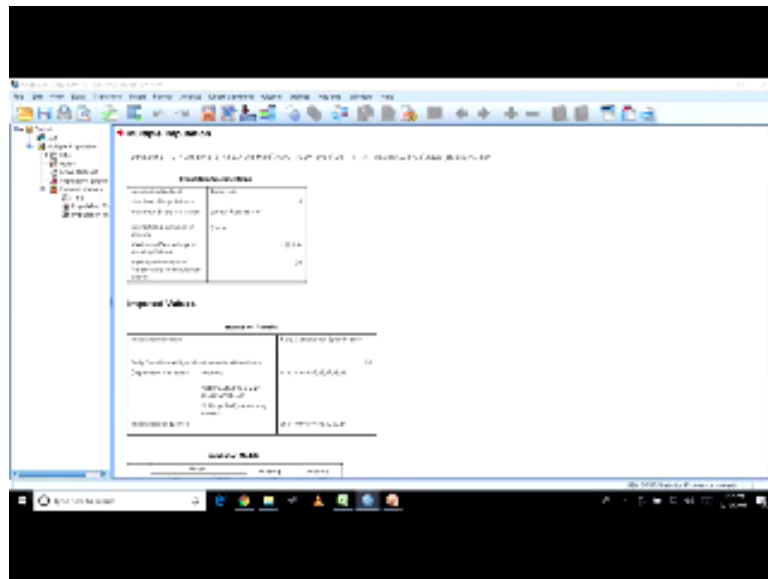
Or I want to check the descriptive you can check that also so what I am doing is an just trying to see you what is the kind of; so this is what it says, that this is basically the missing data is given to you in V1 there are this kind of 21 13 17 7 9 7 all this are given to you, correct, and the percentage is given to you.

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



So once you have understood ok this is the number of missing data's we are having, now do what do I do here again. Now again go to this, and go to the multiple imputation, and input the missing values. Now the variables you have taken earlier I have taken this variables, so it is already there. Now you also give a data set a name. So I have given here jkn in my name one so the method with automatic method constraints are coming to output the imputation model, So create a name of data set and create a hydration history has to given a new name so I give this one too ok.

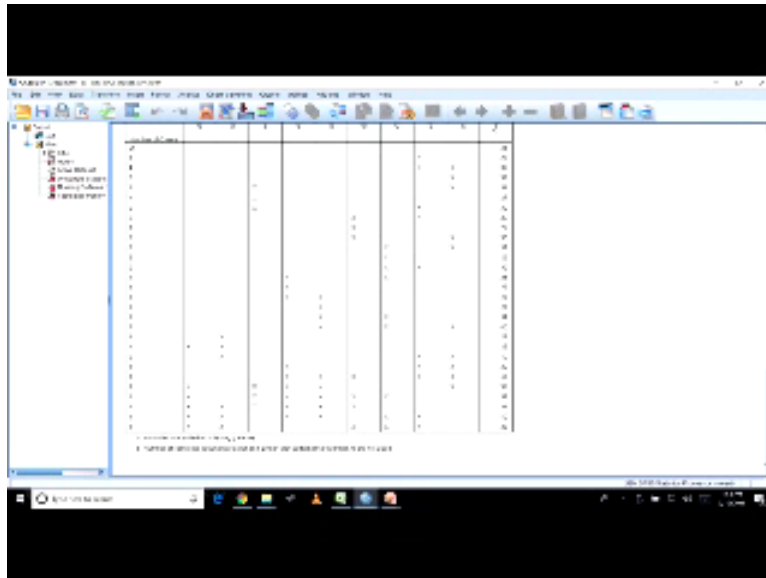
**(Refer Slide Time: 30:14)**



So now I have got this multiple imputation table with me. So it is done now let us leave this one, and go to the main file. So the main file a new file has been created this new file has been created jkn and 1 you can see jkn 1 is kind of matrix kind. Now what does it give you, now go to analyse ok. Suppose I want to compare the mean, ok let us look at how to impute the missing value. So you can go to analyse and check how much of missing value is there in your study.

So I have taken all the variables, I have generally ignored the nominal variables, the nominal scale, and has at check. For example, I can go to here also, tabulated cases, cases of missing value pattern. So now let us see how much of missing data is there. This will bring you a pattern, so now ok. First we have understood there is some missing data. So this table has already told us so again if you want to see this at least this table is very important for you 21 13 and percentage given.

**(Refer Slide Time: 31:41)**

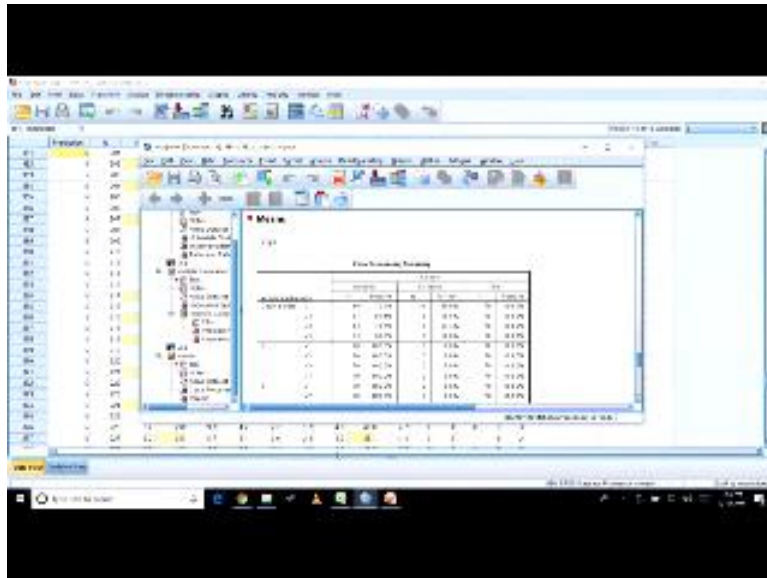


Now I have done with it, now how do I replace the value so I can I go to this multiple imputation. Now what I am doing is, I have already given a name we can either keep that or you can give a fresh name or whatever it is, so just I am going ok so while doing this, what I have done, let me see, what is variable, ok so there are 2 files that has come up, so this is the one which is if you can see here the imputation 0 is given, now if I change it to 1, you can see some yellow coloured marks so this yellow coloured one if you go back to 0, let us go back to the original is 0. So 0 is the original data.

You had this missing value kindly remember for example the second one the fourth one so now what it has done it has replaced one, the missing values within 5 imputations as you remember that given the five imputations, each imputation the values are given to you. So there are 5 imputations. So the question is we have got five kind of data sets now for example 5 imputed data sets, but the one which what will leave you, which one will you use for the question here, when it comes, you can use.

You will use the pooled estimate, now what is this, how do you do that, for example before I want to analyse this is the means of the study for example, sample means, so what I am doing is am taking for example, let us say, I am taking some of these values, and I am just checking the V1 V2 V3 V4 I am taking and I want to see them mean. So if you look at it, if you look at the output table, the output table I think, here it should be, ok.

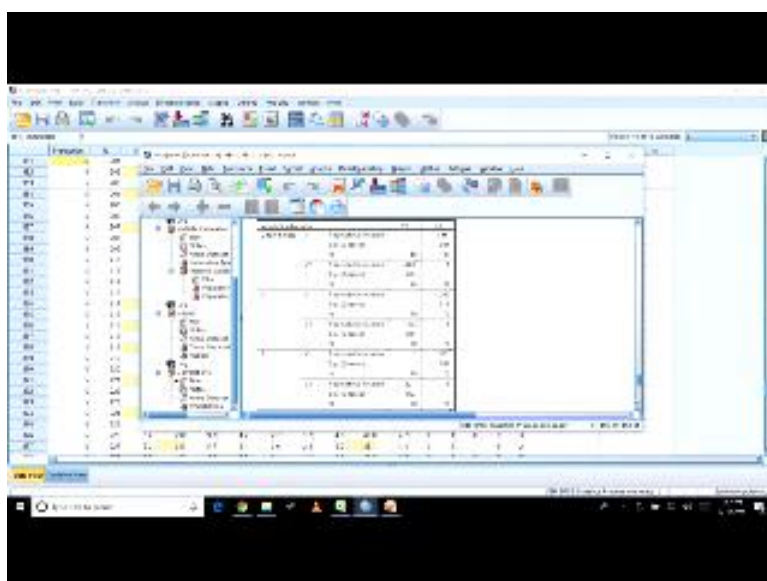
**(Refer Slide Time: 33:37)**



So you can see, you can see here, so when I took taken you know 5, 4 variables. So what it has given me some imputation number, the first imputation the V1 V2 V3 V4 and 5 imputations so first imputation after first imputation, the variable V1 its mean was 3.684 and possibility standard deviation was 1.13 similarly for V2 V3 V4 but the question is, these are all individual imputation.

So what will you take while you write in your research report or something? you will take the one which is the old sorry is the pooled one, so the pooled estimate tells you that the mean is 3.70 to 2.10 to 8.013 and 5.160 so the pooled estimate is basically representing all this 5 imputed values so you can do some other studies also.

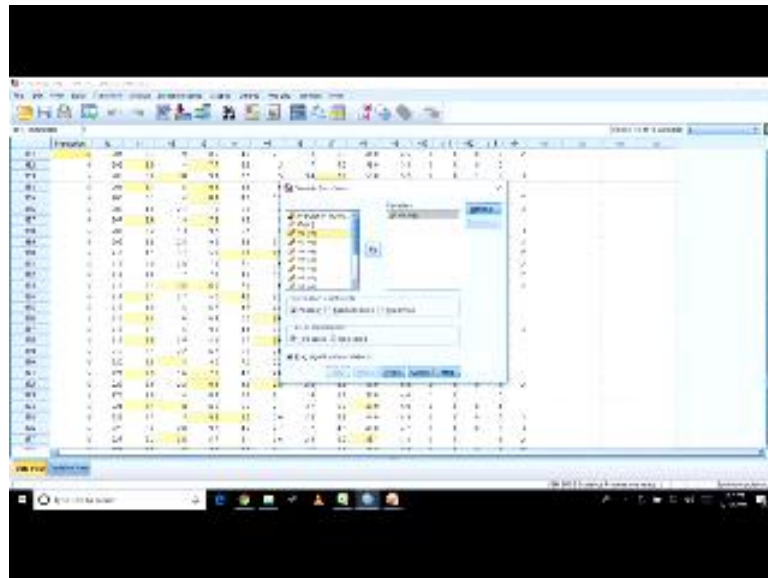
**(Refer Slide Time: 34:48)**





For example you want to check for example let us say anything if even see this for example I want buy with correlation, so I am taking a buy with correlation between lesser V1 and V2, now again if you see if you go back V1 V2 V4 and the pool so you did not write in your resource report about the V1 12345 instead you will get the pooled value. So this is what is a final outcome?

**(Refer Slide Time: 35:13)**



So, this is how you impute the values in a scientific manner and once you do it you can use it for your further analysis. I think I hope this is clear to you and in case you have doubt you practice and you can ask me this questions, whenever you are appearing for your in a doubt clearing sessions are done you can ask me. You have a way of writing your question and I will reply to them. Thank you so much.