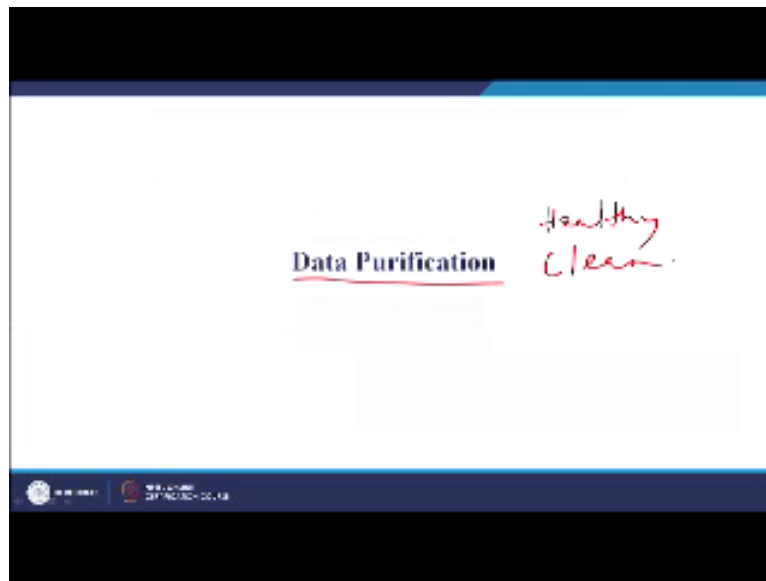


Marketing Research and Analysis-II
(Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture - 17
Data Purification and handling – I

Welcome everyone to the elected series of, Marketing Research and Analysis. So, in today's lecture we will talk about in this lecture and the next one will be mostly discussing about data. What is data and how to manage the data? So the term that have given to this lecture is you can see data purification, why I am saying purification, the simple reason being, whenever a researcher does any study he collects data and on basis of these data he makes an analysis. The research and its analysis, but what if your data is not a healthy data or not in a good condition? What happens then?

(Refer Slide Time: 01:36)



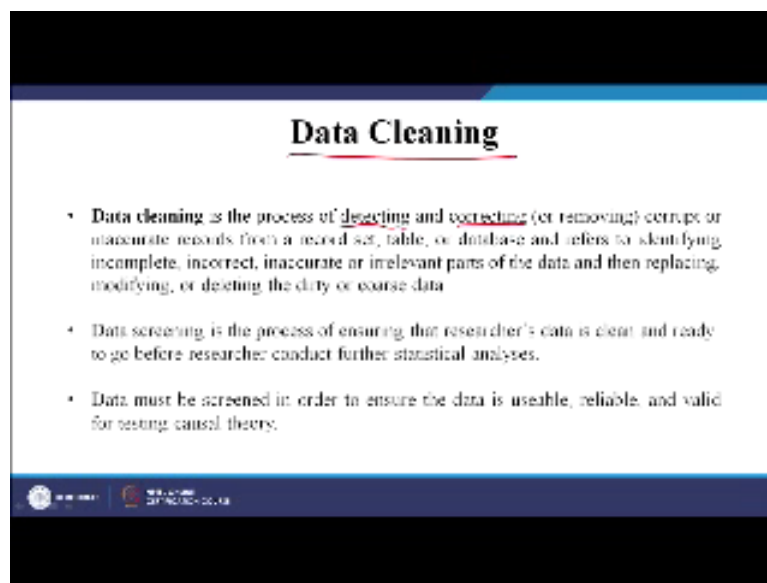
It is like you know the doctor treats you without understanding your problem, similarly the researcher analyses the data or treats the data or tries to make a analysis on the data but without understanding whether the data itself is in a healthy condition or not. So before you try to do anything you should first be very sure, to the; your data is a healthy and clean data. So it should be healthy and clean, why I mean if it is healthy it is a clean data generally. So how do you do that?

And what happens if you do not do this? So if you do not, do this that means you would be you know interpreting on output which might be very different from what actually should

have been happening. That means in a test of significance, or a test of hypothesis, you might be trying to say that the hypothesis is insignificant, which on the contrary should have been significant, just because you handled your data wrongly, you have not corrected them, that is why you have faced this problem.

So this is very important and one should not go further, any researcher should not go further, be it in the field of any field, because, every field request data, be it in the social science management, may medical, professional, engineer, there is data everywhere.

(Refer Slide Time: 02:42)



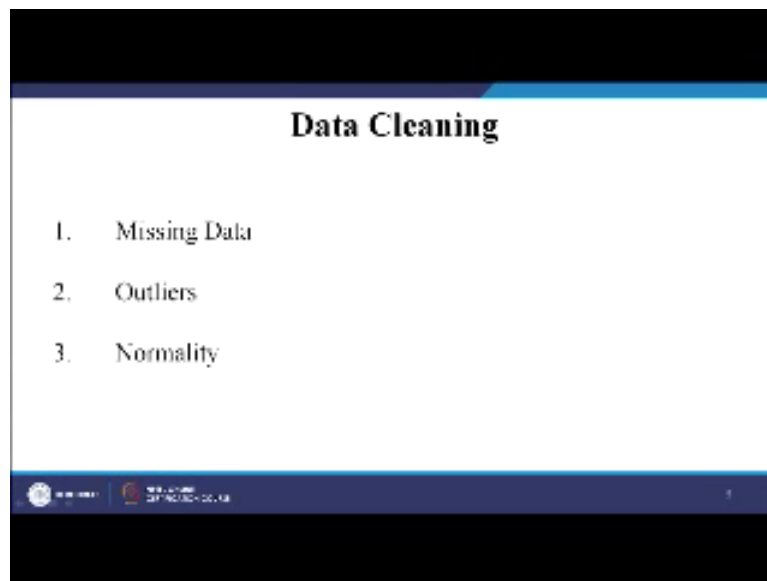
So this data has to be corrected, so what do I say here in the first scene if you see it is called data cleaning, cleaning. So your data has come to you, now you need to clean the data. So what is it the saying? Data cleaning is the process of detecting, you detect, and correct, correction involves sometimes even removing, in an extreme case, corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and replacing it with more you know clear data clean data.

So, this is, the first part, cleaning itself that tells you that you need to identify, first see where the problem is, you want the doctor and the doctor does not even test you properly and says if you have a pain in abdomen it is a kidney problem, it is very, very incorrect to say that so it could be just a gastric problem or an acidity problem of your body. So understanding first of all where, is the problem lying, is very important.

The second thing that says is data screening, that means, when screen the data, it is the process of ensuring the researchers data is clean and ready, it is finally say, yes the data is clean, if it is not clean it is not ready then you need to correct it further, before you conduct any further statistical analysis. Data must be screened, in order to ensure the data is usable, reliable, and valid for testing theory, that means, finally that means what, in order to do any enough for example we have seen, you are going to appear in interview and you are extremely nervous, and the interviewer without making you comfortable.

If he starts asking you volume of questions, you get more and more nervous, and your may be, your output becomes very, very poor. In such a condition, the true skill of the candidate is not known. So, if you want to know the true skill of the candidate, it is better that first you should be made, he should be made relaxed and then he should be asked questions, so that, he can give his best in the condition.

(Refer Slide Time: 05:09)



So data analysis can only begin, when the data is clean and ready, ok. How do I clean my data? So there are basically three problems we will talk about, the first problem is ne of the most discussed problem, is missing data. What is the reason if my data is missing or there is a large portion of the data which is not available to me? The question would further come to why it is not available? What is wrong with that? Is it deliberate or not deliberate? Anything so first is Missing Data.

(Refer Slide Time: 05:37)

The screenshot shows an SPSS data editor window with a grid of data. The columns are labeled V1 through V14. The rows contain numerical values, with several cells being empty, representing missing data. The interface includes a menu bar at the top and a taskbar at the bottom.

Second is Outliers, now let me just show you a missing data for a case, you can see this file this is an SPSS file, where if you see, there are different variables, so V1 V2 up to V14 and if you see in V1 there are certain datas in V2 there are certain datas, in everywhere you can see some blank spots, some blank one, so these blanks and in this case if you can see for the respondent 10 may be, there is a large number of blank spaces, correct.

So it is, you can go down, I think if you can scroll, this way, so you can see that there are this case, this study, has got some missing data, now what do I do, and the missing data is distributed both in the metric scale which is you can see from V1 to V10, V9, this data is mostly in a metric scale and V10 to 14 is basically if I am not wrong is a non metric scale and yeah it is a nominal scale, you can see here.

(Refer Slide Time: 06:48)

Data Cleaning

1. Missing Data ✓
2. Outliers ✓
3. Normality

The slide has a blue header and footer. The footer contains a logo and some text that is difficult to read.

so what happens when the scale is metric and non metric also that also I will tell you. The second case is the case of outlier, now what is this outlier and why it is important? I think we have discussed, in the past also, outliers are very important and the need to be identified. Now what if suddenly you are trying to find the income of people or let's say the speed at which you can run and we are trying to find the average speed of a class and suddenly we find somebody like a Usain Bolt coming into the class you know, becomes one of the candidates, you see by adding Usain Bolt, the mean and standard deviation is suddenly changing.

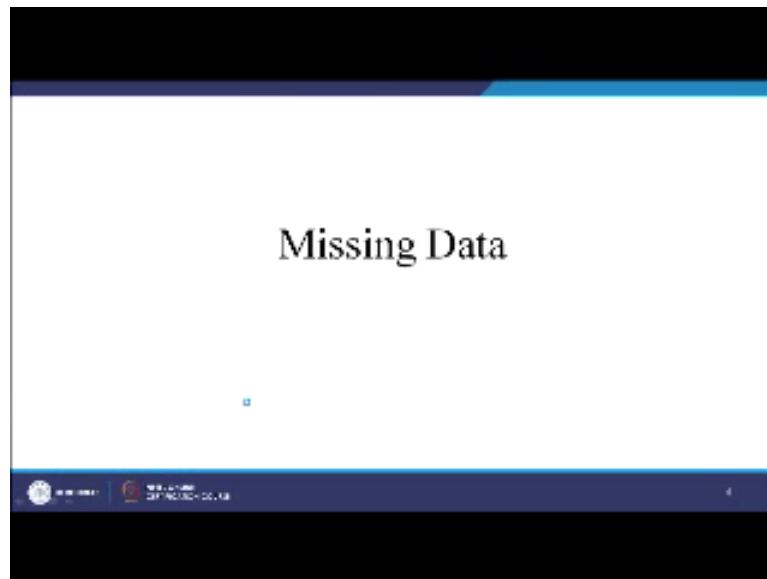
The parameters of estimation is suddenly changing. So this will give you a not true figure, it can be it will be rather very, you can say a concocted figure. So outliers are very important in a data set, for example let say, this is how the data is spread up and suddenly you find there is a data point here. So this data point, look like more or less outliers and the presence of outliers can again create problems for the data analysis.

The third thing is the normality. Now what is normality, before going to any statistical analysis, one needs to understand what is the distribution pattern of the sample. Now how is the sample distributed, if the sample is distributed in a uniform manner, in a unified way, and it is following a normal distribution, you see that most of the statistical tools and techniques can be used to utilise. However if the data is not normally distributed then either it is a special case we may think of some other distribution.

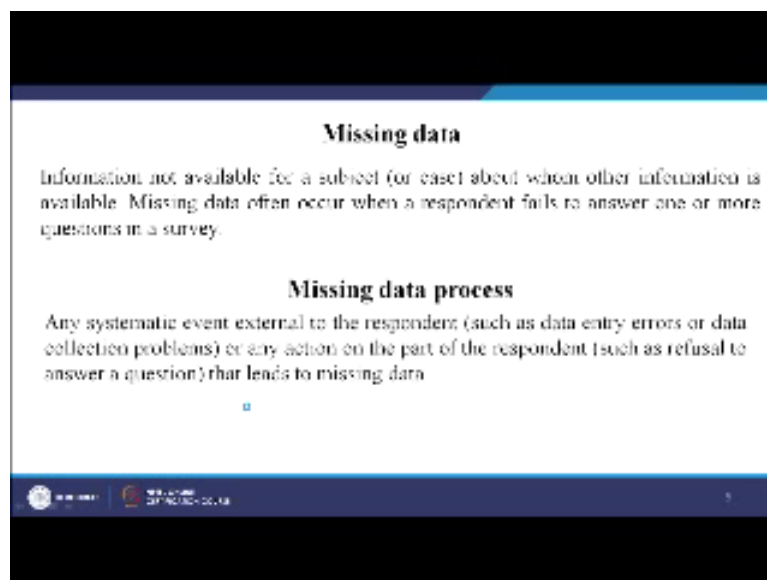
For example Binomial Distribution or Poisson distribution, you know or some other distribution. But the point is, our analysis is dependent that in statistics we are saying, it has if the data is not normal, then most of our analysis will not be correct, especially in case of social sciences. So the question is, if my data is not normal, what should I do? Should I go back to the field, should I collect the data again or should I, can I do something with it?

So these three cases are very important to be checked, there is something, sometimes people also say linearity is also equal important but I will tell you, if your data is mostly normal then linearity is automatically taken care of generally. So if the data is linear or nonlinear what happens? For example you must have heard about a linear regression or a non linear data, so if both condition, how does it matter, how it changes, the effect on the study.

(Refer Slide Time: 09:43)



(Refer Slide Time: 09:45)

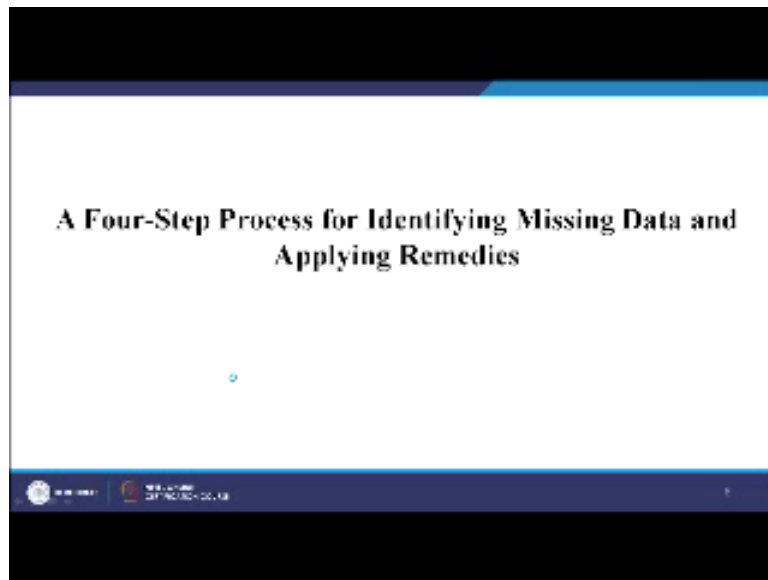


So next, let us start the first one, missing data. Information not available for a subject, or case, that means, or a particular variable or a particular respondent, there is a problem, about whom, other information is available but some of the information is not available. So missing data are often, occur a respond fails to answer, either he is failing to answer, one or more questions in a survey. So it is not necessary that is only failing to survey, no I am sorry feeling to answer but it could be also that, he is not interested in filling your questionnaire, he is not willing to answer. So there could be several reasons behind it.

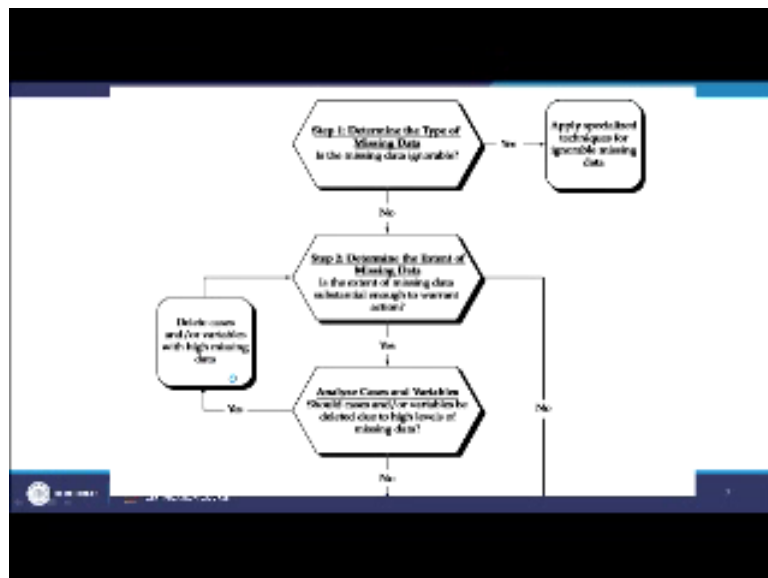
In a systematic event external to the respondent such as data entry errors or data collection problems or any action on the part of the respondent that leads to the missing data. Now, what is this missing data process? A missing data process is saying any systematic event,

external, suggest, data entry, or let say collection problem data collection or any action which that leads to the missing data. Now when you have for example by chance, you have filled in the data and you missed some data, by fast you missed 2-3 data's or put in some wrong data's or by putting some questions which are not very clear to the may be the respondent, he has not filled it up. So this could be the problem that can occur ok.

(Refer Slide Time: 11:15)



(Refer Slide Time: 11:20)



This is the 4 step process for identifying the missing data and applying the remedies. So let us see one by one. So this is how it looks like, the first step is determine the type of missing data, now the question is what do you mean by this type of missing data. Now the other types of missing data, now are there any types of missing data? Yes there are. The first is the say, is the missing data ignorable? That means, what happens if I ignore the missing data? I am not

worried about missing data because the patient has come and the doctor is ignoring the symptom.

So sometimes if we ignore the symptoms it is fine, it does not matter but suppose in some condition the patient is seriously serious patient of kidney, you know then there is problem. Then if he is ignoring the symptoms, it could be extremely dangerous. But suppose it is a simple case of let us a slight acidity in my chest, a burning sensation and I know it is not but big of a serious problem, and if I missed it, ignore it, it can be still manageable, it is ok, because I know naturally I will get correct in sometime, but if it is a problem related to my kidney or something, then missing it or ignoring it, could be extremely fatal.

Same thing you can bring the analogy here, determine the type of, is the missing data ignorable? No, if it is yes, it is ignorable, you can ignore. So apply specialise techniques for ignorable missing data, when it is ignorable, sometimes we just ignore and carry forward the process of you know, analysing the data. Second, if it is no, let us say you cannot ignore, that means, it is a sufficiently large problem, then you go to the second step.

Determine the extent, how much is the extent of the missing data, is the extent of missing data large enough to warrant action? First was, whether the problem was important or not? second is how much is the problem, suppose I am having a burning sensation for may be a day, which was ignorable, fine but suppose it is happening me to me if it out of 30 days in a month may be 15 days or 20 days, then it could lead to ulcer.

So is it sufficient large enough if I am having 15 days or if I am having let say 3 days in the month of two days in a month, is it ignorable? The question lies there. That is where you need to use your own logic. So if it here, is the extent of missing data substantial enough to warrant action? Yes then comes the point, analyse the cases and variables. So cases or variables be deleted due to high levels are missing data? Yes, then you delete the cases or the variables with high missings.

Suppose a particular respondent out of 10 questions he has not filled up 8, so that means we should it is better to ignore the respondent, the case, correct. Now the question is if it is a no, should cases and will be deleted due to high level data, no because if you start deleting every

Now that means if I am trying to make a study on age and you know blood pressure, there I can see that age, at a particular age, if there is a missing data could be possible that most of the people who are not reporting or missing, they must be young people. So in such a condition it is a missing at random, so there is there is a relationship, there is a relationship and this is why this becomes more complicated and little you know risky to handle it, had it been completely random case, then you need not worry much.

Since once you are done it, so if you are having a missing at random approach, the best thing is modelling based approaches. Now modelling based approaches, basically we use a maximum likelihood approach, where we use the; you know Estimation and the EM approach, we say that, EM approach is estimated by Estimation and this is the parameter values. So ok, so the EM approach, so this is the this is basically the modelling based approaches you use.

Suppose this is a completely at random approach then what happens, there are you use that different types of imputation, imputation is nothing but replacement method. Do you want to replace the missing database values, yes or no? Suppose no then do you want to use only case with complete data and use all possible valid data, fine, go ahead. Suppose you do not want to impute, no issues.

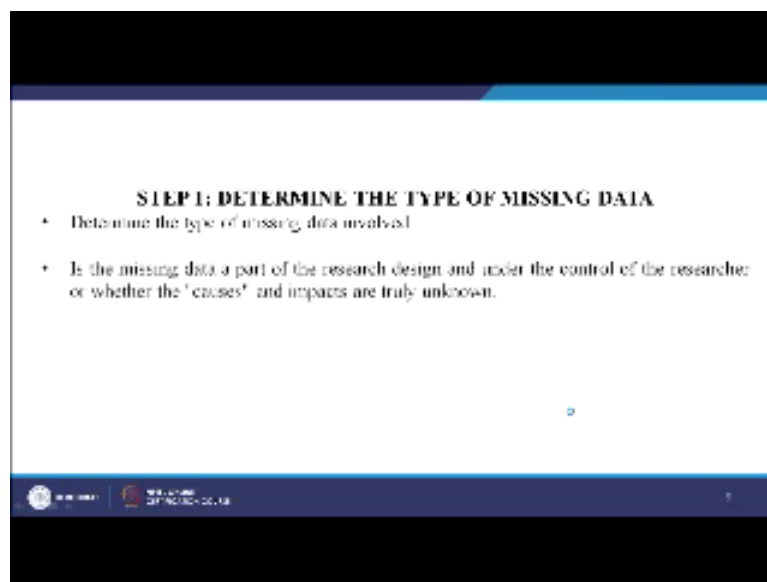
Looking at the extent of the data if it is not much, then you need not worry, but suppose you feel no it is large, the problem is large, I need to impute the data replace the data, the data applications, which method do you want to use? do you want to use known values or calculate replacement value from the valid data, now what does it mean, that means, I can if I want to replace the value I can do one thing I can use some earlier data.

Let say, respondent one, has a large number of listed let say respondent 5 has a large number of, so respondent 1,2 3 4 5, respondent 1 and respondent 4 are very similar in nature, let us see. Lets try and here, there are some missing data, but here, I found he has got a X1 X2 X3. Now what I can do is, in some conditions, I may use the values of the respondent 4 for filling up respondent one's data, so this is one possibility.

There are several methods which you can see, for example, complete data only, complete case approach, a method whereas it is the best method obviously. when use any for complete case then it is always the best thing or all available subset approach case substitution, so how do you substitute through a means of subdivision regression method, many methods are there.

This is the known value; I will explain here. Then you can either you can tab to the known values or you can use substitution case, case substitution. In this, case has been substituted here, or hot and cold deck imputation, this is something like a similar kind of a situation is seen, and then they try to replace the value, I will explain that. Then is the, you calculate the values, and you upload this method, regression basic method and the means substitution method. This method has been found to be the most significant; the best approach the best method.

(Refer Slide Time: 19:40)



So, let us go each one of them. Determine the type of missing data. So either this is the missing data or part of the Research Design and under the control of the researcher whether the causes and impacts I truly known or not, not ok.

(Refer Slide Time: 20:02)

Ignorable Missing Data.

- Many times, missing data are expected and part of the research design. In these instances, the missing data are termed ignorable missing data, meaning that specific remedies for missing data are not needed because the allowances for missing data are inherent in the technique used.
- The justification for designating missing data as ignorable is that the missing data process is operating at random (i.e., the observed values are a random sample of the total set of values, observed and missing) or explicitly accommodated in the technique used.
- For example, Missing data resulting from taking a sample of the population rather than gathering data from the entire population. In these instances, the missing data are those observations in a population that are not included when taking a sample.

WUOLAH UNIVERSITY
WUOLAH UNIVERSITY
2019/2020
17

So you need to understand what is the type of the missing data how important it is for me, right in my studies, ignorable many times data expected and part of research design, in this instances the missing data term ignorable, ignorable missing data meaning that specific remedies for missing data's not needed, it is ignorable, you can ignore. Because along certain missing data are not in the technique used, so you need to worry about it.

The justification for designating missing data as ignorable when you say it is ignorable is that, missing data process, is operating at random, that is the observed values are a random sample of the total set of values, or explicit so in case of a, for example missing completely at random, you need not worry about it much. For example, missing data, resulting from taking a sample of the population, rather than the gathering data from the entire population, in these instances, the missing data, are those observations in the population, that are not included when taking a sample.

(Refer Slide Time: 21:17)

Missing Data Processes That Are Not Ignorable.

- Missing data that are not ignorable occur for many reasons and in many situations. In general, these missing data fall into two classes based on their source: **known** versus **unknown** processes.
- Many missing data processes are known to the researcher in that they can be identified due to procedural factors, such as errors in data entry that create invalid codes, disclosure restrictions (e.g., small counts in Census data), failure to complete the entire questionnaire, or even the poor health of the respondent.
- In these situations, the researcher has little control over the missing data processes, but some remedies may be applicable if the missing data are found to be random.

11

So such conditions can happen. So, whether you can ignore the data or you cannot ignore. That is where the researchers has to use his or her own logic ok, sometimes it is not ignorable so how do you know that? Let us see. Missing data that are not ignorable occur from many reasons and in many situations. In general this fall into two categories known and unknown processes. Sometimes it is a missing data that are not ignorable is you know about it, they are not ignorable and you know about it sometimes is an unknown case.

So let see what is, now, known case, it is known to the researcher and they can be identified due to procedural factors such as, errors in data. I know that there was a mistake in my data entry. That creates an invalid course. Disclosure restrictions, sometimes some data are not available to us, failure to complete the entire questionnaire, the person is not interested, respondent is not interested, even the poor health of the respondent.

So, he is not willing to fill it up, in this situation the researcher has little control over the missing data process but some remedies may be applicable, if the missing data are found to be random. No issues. If suppose this is, conditions are there which you will generally come across in such a condition, the researcher can handle such problems by utilising certain statistical methods which I will show you.

(Refer Slide Time: 22:29)

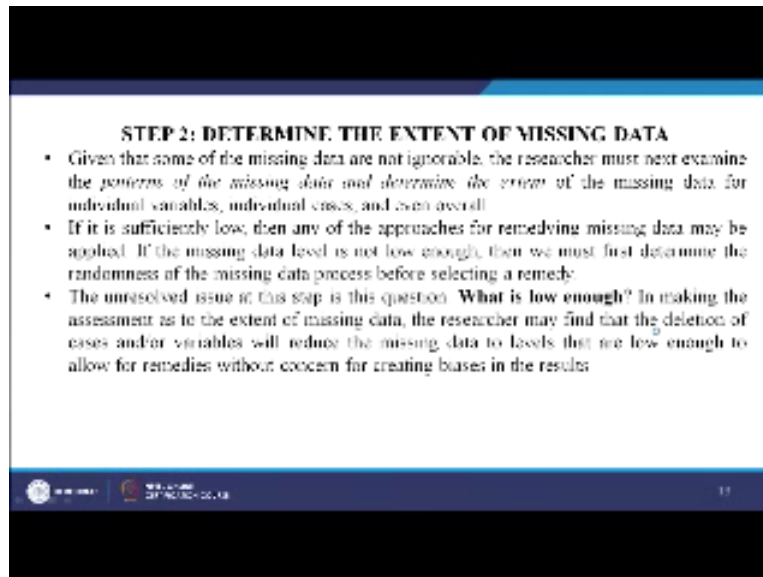
- Unknown missing data processes are less easily identified and accommodated. Most often these instances are related directly to the respondent.
- One example is the refusal to respond to certain questions, which is common in questions of a sensitive nature (e.g., income or controversial issues) or when the respondent has no opinion or insufficient knowledge to answer the question.
- The researcher should anticipate these problems and attempt to minimize them in the research design and data collection stages of the research.
- However, they still may occur, and the researcher must now deal with the resulting missing data. But all is not lost. When the missing data occur in a random pattern, remedies may be available to mitigate their effect.

Unknown missing data, now this is a, I do not know why it is happened, and less easily identified and accommodated. Most often these are related directly to the respondent, ok, example, the refusal to respond to certain questions. In the last case, it was a different thing. Now here the respondent is refusing, he is refusing to answer your question, which is common in questions of sensitive nature, for example, you know, income or any controversial issues your political affiliation, for example, or when the respondent has no opinion or insufficient knowledge you are asking maybe a rural person poor chap, about let say Newton's law of gravitation.

He might not be in a position to answer you, and then there is a missing data. The researcher should anticipate these problems, now, how do you handle this, now you should be, your first job as a researcher is to anticipate and attempt to minimise them in the Research Design and data collection stages. However they still may occur, the researcher must now deal, if suppose after doing everything still there is a problem, you must not deal with the resulting missing data, but all is not lost.

So when the missing data occur in a random pattern, remedies may be available to be that. So the question is, as I said if there are certain questions which people would not answer, they are more dangerous, because then you should not have, first of all, in the first hand, you should not had it should have had such questions in your questionnaire. And if it is the it is mandatory then you should may be had, educated the respondent before you have started filling up the questionnaire, anyway now the job is done the data is in your hand now you need to handle it.

(Refer Slide Time: 24:14)



STEP 2: DETERMINE THE EXTENT OF MISSING DATA

- Given that some of the missing data are not ignorable, the researcher must next examine the *patterns of the missing data and determine the extent* of the missing data for individual variables, individual cases, and even overall.
- If it is sufficiently low, then any of the approaches for remedying missing data may be applied. If the missing data level is not low enough, then we must first determine the randomness of the missing data process before selecting a remedy.
- The unresolved issue at this step is this question: **What is low enough?** In making the assessment as to the extent of missing data, the researcher may find that the deletion of cases and/or variables will reduce the missing data to levels that are low enough to allow for remedies without concern for creating biases in the results.

So, how you will go for it? Second thing is determine the extent which we said. Now you see some of the missing data are not ignorable, if it is ignorable, go ahead no worries. If it is not ignorable, the researcher must next examine the pattern of the missing data and the extent. What is a pattern is there any pattern of you know, the data getting missed or what is the amount of missing it? So, for he can check for individual variable, variable by V1 V2 V3 V4 individual cases respondent 1, respondent 2 or total, in total.

If it is sufficiently low, then any of the approaches for remedying missing data may be applied, even I can tell you if it is very low, less than 5%, then you can just ignore the missing data. Until unless your study is a very sensitive study. We say in social science at least less than 5% we are not worried, because it hardly makes any impact, on the final you know final check. A final analysis interpretation of the data, the data analysis, the unresolved issue at this step is what is low, what is low enough, in making the assessment to the extent of missing data, the researcher may find that the deletion of variables is reducing the missing data, now for example, let us go to this case.

(Refer Slide Time: 25:37)

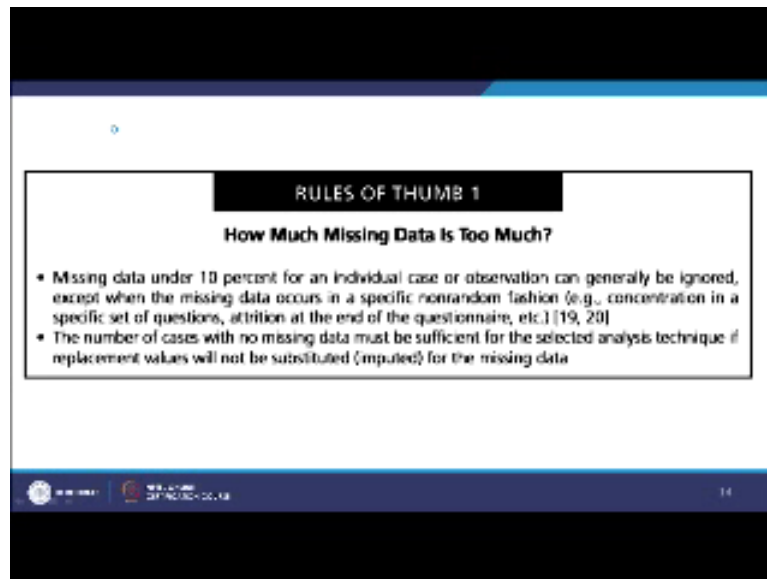
The image shows a screenshot of a data analysis software interface, likely SPSS, displaying a large table of data. The table has many columns and rows, with some cells highlighted in yellow. The interface includes a menu bar at the top and a taskbar at the bottom. The data appears to be organized into columns, with some cells containing numerical values and others being empty, indicating missing data.

Now suppose I see that lets say 210 response, case number 10 here, case number, let us say, suppose, some other, let us say some few correspondents, there are lot of missing data's. In in typical, if you see, what I can do it just go for a analyse a frequency, frequency and I will check, lets check this only, now what I am saying is in V1, there are 49 valid respondents, a valid case data and 21 is missing, in V2 13, V4 7 missing, V7 so it is very clear that V1 has got a large number of missing data.

So this has got a large number of missing data. So in this case, should you delete the entire V1 or would you do something to handle it? If you delete it, then in this case it is ok, because, you have less such cases. But some studies which is randomly occurring data and you found that there are out of 10, 3 4 variables are showing very poor data or high level of missing data.

Will you delete all of them or will you delete, let us say there are 100 respondents, 20 of them are already showing you poor data. So lot of missing data is there, will you remove them? So if you remove maybe there will be a problem sample size adequacy, so sampling adequacy or sample size adequacy, become an issue of concern. So the researcher has to understand that we have to deal with it.

(Refer Slide Time: 27:20)



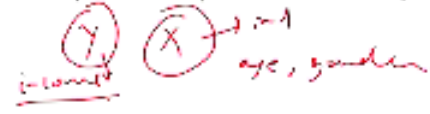
So Rules of Thumb, how much missing data is too much? Missing data under 10%, so I was saying 5% up to 10% or observation can generally be ignored. This is mostly true for at least social sciences. See, I am what I am saying please take it from case to case basis because you are if you are studying a very sensitive issue, let us say Aids patient or Rocket launch, then, that case would be different.




I am talking about general social science case. When the missing data that occurs in a specific random fashion, example, concentration in a specific set of questions, attrition at the end of the questionnaire, the number of cases with no missing data must be sufficient, the number of cases with the no missing data must be sufficient, for the selected analysis technique. If replacement values will not be substituted suppose I do not replace and still have sufficient amount of data, then I can ignore. No issue is there. What I will do is I will wind up here.

(Refer Slide Time: 28:19)

Assessing the Extent and Patterns of Missing Data.

- The most direct means of assessing the extent of missing data is by tabulating (1) the percentage of variables with missing data for each case and (2) the number of cases with missing data for each variable.
- This simple process identifies not only the extent of missing data, but any exceptionally high levels of missing data that occur for individual cases or observations. The researcher should look for any non-random patterns in the data, such as concentration of missing data in a specific set of questions, attrition in not completing the questionnaire, and so on.
- Finally, the researcher should determine the number of cases with no missing data on any of the variables, which will provide the sample size available for analysis if remedies are not applied.



We have understood now the extended pattern of missing data as I was saying just lets finish with this light and we will wind up. So what happens accessing the extent and pattern of missing data, the most direct means of assessing is by tabulating. You tabulate, take a frequency, may be the percentage variables with missing data for each case, which we just did, the simple process is to identify, not only the extent of missing data but any exceptionally high level of missing data that occurs in individual case observations.

So the researcher should look for, any non random, see if random is there, randomness is there is ok but if there is someone unsystematic pattern, or you know some non randomness is there, then you need to be very careful. The researcher should determine the number of cases with no missing data on any of the variables and which will provide the sample size to the end when I will wind up here. So the question is, when you are dealing with missing data you have to see you whether it is ignorable or not.

How do you justify that I have set 10%, 5% whatever you take, the question is, is there any relationship between let say, the X and Y variable, let say the X and Y, suppose there is missing data in Y due to X, now what is x, x is my independent variable that say as I said that case age or gender suppose and this is let us say my income, if my income has got lot of missing data, due to my gender being, because of my the male or female you know the senior or the young people or something.

If there is some kind of problem, ok then this is a missing at random case that means you should be little bit careful about it. Ok, these are also handled by the modelling approaches,

anyway so I will wind up here, this lecture, in the next lecture will continue with the same data preparation, thank you so much.