

Exploring Survey Data on Health Care
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 14
Deal with Missing Values

Welcome students to the NPTEL MOOC module on handling healthcare survey data. After discussing the previous class on normalizing data sets, I have mentioned the possibility of missing values. Missing values have a vital role in analyzing big data or in large scale data sets. There is a huge possibility of generating missing values because of many reasons.




If missing values are there, that may create some inconsistency in our result, and those have to be dealt with appropriately instead of avoided. So, let us get started with the discussion of missing values in statistics. Missing data or missing values often occur when no data value is stored for the variable in an observation.

(Refer Slide Time: 01:29)

Introduction

- ❑ In statistics, missing data, or missing values, occur when **no data value is stored** for the variable in an observation.
- ❑ This situation mostly occur as a result of **manual data entry procedures, equipment errors and incorrect measurements.**

	Gender	Glucose	AST	Age
Patient 1	?	120	?	?
Patient 2	M	105	?	68
Patient 3	F	203	45	63
Patient 4	M	145	?	42
Patient 5	M	89	?	80

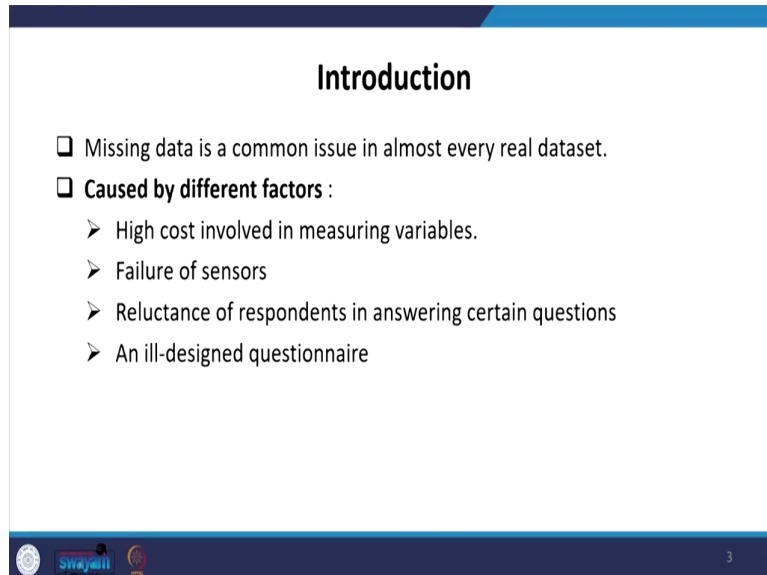
2

This situation mainly occurs as a result of manual data entry procedures, equipment errors and incorrect measurements. So, there are three important reasons by which we get the missing information or missing values like equipment errors, or incorrect measurements are most important. Sometimes in the schedule, we have followed wrong procedures or the question sequencings are not followed, or the questions are asked to the wrong person, or

even if it is asked to the person, the person might be reluctant to answer. So, there are possibilities of missing entries or missing values.

In this sample data set, there are question marks given; this indicates some missing points.

(Refer Slide Time: 02:21)



The slide is titled "Introduction" and contains the following text:

- ❑ Missing data is a common issue in almost every real dataset.
- ❑ **Caused by different factors :**
 - High cost involved in measuring variables.
 - Failure of sensors
 - Reluctance of respondents in answering certain questions
 - An ill-designed questionnaire

At the bottom of the slide, there are logos for "swayam" and "swayam" on the left, and the number "3" on the right.

As I mentioned that missing data is quite common in everyday research. This is caused by different factors such as the high cost involved in measuring variables, failure of sensors, respondents' reluctance to answer certain questions, and an ill-designed questionnaire.

So, the problems with missing data are it reduces statistical power. So, statistical power or power of the test, you might have heard about it accepting or non-accepting a null hypothesis. These are various discussions we usually deal with in hypothesis testing.

(Refer Slide Time: 03:17)

Problems of Missing Data

- Reduces statistical power
 - Statistical power : probability that the test will reject the null hypothesis when it is false.
- Lost data can cause bias in the estimation of parameters.
- Reduce the representativeness of the samples.
- Complications in handling and analyzing the data.
- Loss of efficiency.

swayamii 4

Statistical power means the probability that the test will reject the null hypothesis when it is false. The problem with missing data is that lost data can cause bias in estimating parameters.

The missing data may result in less representative of your sample, and these missing data are generated due to complications in handling and analyzing the data. This has a huge connection with the loss of efficiency in your results or in your data.

(Refer Slide Time: 04:15)

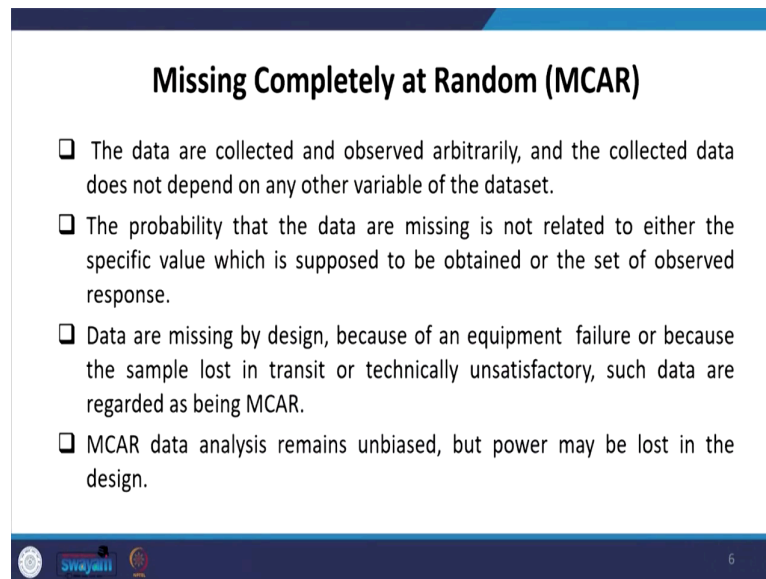
Types of Missing Data

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR)

swayamii 5

So, coming to the missing data again, we are addressing 3 important types of missing data. The one is called missing completely at random in short; we write it as MCAR, then missing at random (MAR) and missing not at random (MNR). So, the first one is missing completely at random; that means the data are collected and observed arbitrarily. The data does not depend on any other variable of the data set.

(Refer Slide Time: 04:56)



Missing Completely at Random (MCAR)

- ❑ The data are collected and observed arbitrarily, and the collected data does not depend on any other variable of the dataset.
- ❑ The probability that the data are missing is not related to either the specific value which is supposed to be obtained or the set of observed response.
- ❑ Data are missing by design, because of an equipment failure or because the sample lost in transit or technically unsatisfactory, such data are regarded as being MCAR.
- ❑ MCAR data analysis remains unbiased, but power may be lost in the design.

swayam 6

If simply arbitrarily is collected, it is not dependent on any other data set variable. The probability that the data are missing is not related to either the specific value that is supposed to be obtained or the set of observed responses required for the analysis. Data are missing by design. We already said there might be various possible designs in the data set in the survey.

So, through the design, it is possible that there will be complete randomness in the analysis or the collection of the data. The data are missing by design because of an equipment failure or because the sample was lost in transit or technically unsatisfactory; such data are regarded as missing completely at random.

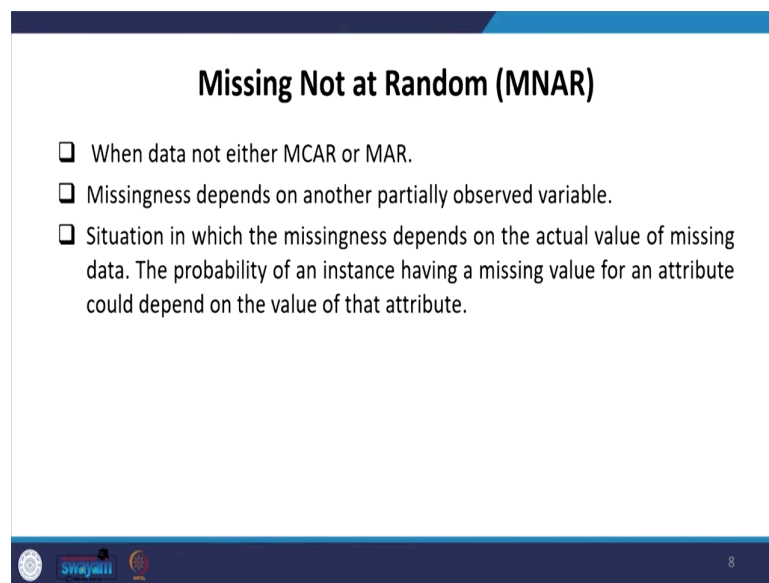
Missing completed data at random (MCAR) data analysis remains unbiased, but power may be lost in the design. So, data may be unbiased. It might give the mean value of the estimates, but the power may be lost in the design.

So, missing at random is another type of missing value. The probability of an instance having a missing value for an attribute may depend on the known values but not on the values of the

missing data itself. So, missingness can only be explained by fully observed variables, whereas partially observed variables cannot be responsible for missingness in others.

For example, women in the population are more likely not to reveal their age; therefore, the percentage of missing data among female individuals will be very high, which is also possible at random.

(Refer Slide Time: 07:20)



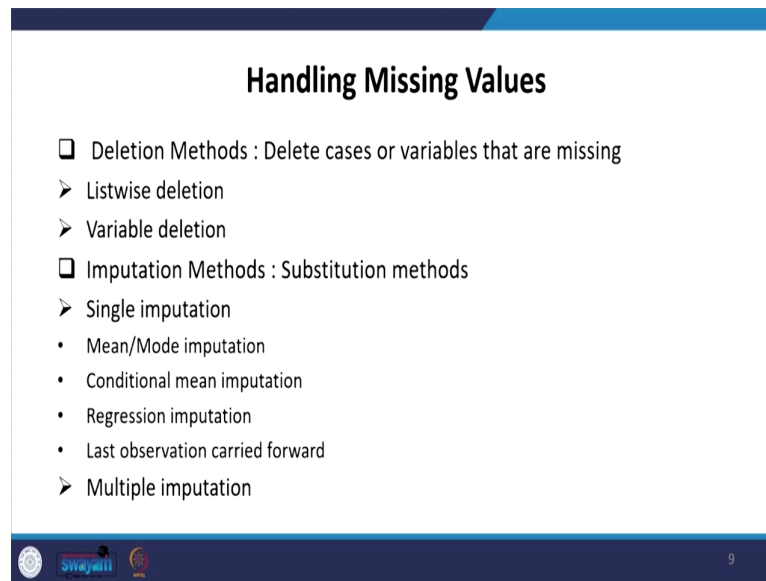
Missing Not at Random (MNAR)

- ❑ When data not either MCAR or MAR.
- ❑ Missingness depends on another partially observed variable.
- ❑ Situation in which the missingness depends on the actual value of missing data. The probability of an instance having a missing value for an attribute could depend on the value of that attribute.

swayam 8

The third type is missing not at random. The missingness depends on another partially observed variable. The situation in which the missingness depends on the actual value of missing data. The probability of an instance having a missing value for an attribute could depend on the value of that attribute.

(Refer Slide Time: 07:53)



Handling Missing Values

- ❑ Deletion Methods : Delete cases or variables that are missing
 - Listwise deletion
 - Variable deletion
- ❑ Imputation Methods : Substitution methods
 - Single imputation
 - Mean/Mode imputation
 - Conditional mean imputation
 - Regression imputation
 - Last observation carried forward
 - Multiple imputation

swayamii 9

The approach to handling missing values is quite important in research; one is called the deletion method: once your cases or your variables are missing, you can simply delete them. You should have been huge data on your pool that may not create much difference. We are going to deal with that in detail now.

Another most important and frequently used technique is the imputation technique and imputation method, which might be a single imputation; that might be multiple imputations. So, most frequently in detailed research analysis, multiple imputation is applied. In single imputation you can go with mean or mode imputations or conditional mean imputations, then regression imputations and last observation carried forward imputation.


So, these are now discussed the first case is called listwise deletion when these lists have certain missing values. Like in the example data set, the serial numbers 3, 4 and 8 have some missing values. The missing value in any of the entries might lead to some possibility of biased estimates. So, better to delete the complete list.

(Refer Slide Time: 09:50)

List-wise Deletion

- A good method when the proportion of missing data is less than 15%.
- **Advantages:**
 - It can be used for any type of statistical analysis.
 - No special computation are required
 - The parameters estimations are unbiased.
 - The standard errors are appropriate compared to original data.
- **Disadvantages:**
 - May remove a considerable fraction of data.

SN	Age	Gender	Health Expenditure
1	30	M	40000
2	40	M	36000
3	61	M	Missing
4	23	Missing	18000
5	41	M	36000
6	33	F	45000
7	22	F	22000
8	Missing	F	54000
9	35	F	31000
10	47	F	28000

 10

Whereas in the case of pair, you need not delete all the lists together. You can try to check which 2 pairs are missing. So, it depends upon how you are handling it and how you are looking at it. The list wise deletion method is good when the proportion of missing data is less than 15 percent.

There are advantages it can be used for any type of statistical analysis; no special computations are required. The parameter estimations are unbiased, and the standard errors are appropriate compared to the original data. So, these are the advantages we just discussed. So, that is why this is the most used technique. One disadvantage is that this may remove a considerable fraction of data since we are actually deleting the entire observation.

(Refer Slide Time: 11:16)

Variable Deletion

- Variable deletion involves dropping variables with missing values on an case by case basis.
- **Advantages:**
 - Makes sense when lot of missing values in a variable and if the variable is of relatively less importance.
- **Disadvantages:**
 - Loss of information regarding the variable.

SN	Age	Gender	Health Expenditure
1	30	M	40000
2	40	M	36000
3	61	M	Missing
4	23	Missing	18000
5	41	M	36000
6	33	F	45000
7	22	F	22000
8	Missing	F	54000
9	35	F	31000
10	47	F	28000

swayamii

11

The second one is called variable deletion: variable deletion involves variables with missing values on a case by case basis. What once that one variable is not that important to you and has a missing value. Out of 3 variables, the gender variable has a missing value. When I know that this gender variable is not that important in my analysis, that variable could be deleted or dropped.

So, there are advantages like this makes sense when a lot of missing values in a variable and if the variable is relatively less important and the disadvantage is loss of information regarding the variable.

(Refer Slide Time: 12:40)

Application in Stata

- ❑ Example – NSS 75th health round data.
- Stata reads missing (.) as a value greater than any number.

	PersonID	Expenditure	Level_of
58	5	1120	4
59	6	6300	3
60	1	.	.
61	2	3500	3
62	1	920	4
63	5	450	4
64	2	50	.
65	3	200	1
66	2	.	3
67	4	400	4
68	5	.	.
69	1	950	4
70	3	170	1
71	1	70	.
72	1	10750	4
73	2	350	4
74	2	160	.

Missing value

13

We are highlighting missing values with the application of stata. This is very easy and will be guiding all those details in our regression analysis, but at this moment, we are just giving you the snapshot from our own Stata operation.

Here the dot represents missing data. So, Stata reads the missing (.) as a value greater than any number. So, if you are giving a command in Stata, your data should be caveated within a range and suppose in expenditure you wanted to highlight the expenditure greater than 1000 in the data. So, below 1000 whichever will be displayed with a dot. So, it depends upon how you give the command, but in general, Stata gives a dot as the missing value information.

(Refer Slide Time: 14:12)

Application in Stata

Any analysis including multiple variables automatically applies list wise deletion.

```
. sum PersonID Expenditure Level_of_care
```

Variable	Obs	Mean	Std. Dev.	Min	Max
PersonID	43,240	2.756846	2.834012	1	92
Expenditure	41,062	914.2377	4125.129	0	501000
Level_of_care	39,363	2.714885	1.296413	1	5

```
. mean PersonID Expenditure Level_of_care
```

Mean estimation Number of obs = 37,880

	Mean	Std. Err.	[95% Conf. Interval]
PersonID	2.721621	.0134951	2.69517 2.748072
Expenditure	961.6973	22.00358	918.5697 1004.825
Level_of_care	2.757656	.0066125	2.744695 2.770616

So, now I am explaining here about including multiple variables. If there are so many variables, then analysis automatically comes with a fine observation to avoid missing values.

For example, person ID, expenditure and level of care have different observations 43240, 41062 and 39363 but in the simple model or mean estimation, it is not considering the entire one. It is only considering the one which has all the information it might have deleted. So, finally, you are left with the observation in a net where the mean could be derived.

So, this is the power of the missing value. It seems as if there are some missing values out of these 3 variables, and if you have corrected these numbers, then observations could have been increased.

(Refer Slide Time: 15:38)

Mean/Mode Substitution

- Replace missing value with sample mean/mode
- **Advantages:**
 - Can use complete case analysis methods
- **Disadvantages:**
 - Reduce variability
 - Weakens covariance and correlation estimates in the data (because ignores relationship between variables)

SN	X1	X2	X3
1	9	8	8
2	7.44	7	6
3	8	5	6
4	7	4	5
5	9	5	7
6	8	8	9
7	6	7	6
8	5	9	7
9	7	8	6.77
10	8	8	7

swayamii 15

Now, coming to the mean or mode substitution based approach to deal with the missing values. You can replace missing values with sample mean or mode. In the X1, X2 and X3, we either replace this missing value with its mean value or mode value, depending on which type of variable you are considering.

If it is a scaled variable, then it is suggested to go with the replacement with the mean value. Suppose it is a categorical variable or qualitative variable. In that case, it is suggested to go with a mode replacement in that the missing values are the complete case analysis method in this approach. Disadvantages like this will reduce variability and weaken covariance and correlation estimates in the data. It ignores relationships between variables. So, of course, since you are replacing with something, there is the possibility of weakening its correlation or covariance.

(Refer Slide Time: 17:07)

Conditional Mean Substitution

➤ Replace missing values with values of the variable mean for a relevant subgroup.

SN	Gender	X1	X2
1	F	8	8
2	F	7	6
3	F	5	6
4	F	6.25	5
5	F	5	7
6	M	8	9
7	M	7	6
8	M	9	7
9	M	8	7.25
10	M	8	7

There are certain missing entries in variables X1 and X2. So, instead of entire averaging these, you could have averaged out with its females and averaged out the value with this male information.

So, replacing missing values with values of the relevant mean for a relevant subgroup is very important. So, this is carrying these averages or even as per the requirement; it could also be the mode value. So, instead of the entire, you are following its subset and exactly replacing it.

(Refer Slide Time: 18:08)

Application in Stata

```
. use "C:\Users\admin\Desktop\nsa health\nsa75t"
. summ Level_of_care
```

Notes:

1. Unicode is supported; see [help unicond](#)
2. Maximum number of variables is set to

Now it is very interesting that you can take a snapshot or follow the video very carefully. You can open the stata window and accordingly click on this data; then you click on out of all those options and click on this create or change data, and there will be a further drop down like out of all those things you click this.

The other variable transformation commands and then select this change missing values to numeric. If you have missing values with the command only, you can replace those with numeric values, which is very important, and it can be dealt with stata

(Refer Slide Time: 19:04)

The screenshot shows a Stata window titled "Application in Stata". At the top left, the command `. summ Level_of_care` is entered. Below it is a summary table:

Variable	Obs	Mean	Std. Dev.	Min	Max
Level_of_care	39,363	2.714885	1.296413	1	5

The mean value, 2.714885, is circled in red. A red checkmark and the word "Sum" are written next to the command. A blue callout box labeled "mean" points to the mean value. Below the table is a dialog box titled "mvencode - Change missing values to numeric v...". The "Variables:" field contains "Level_of_care". The "Conversion rules:" field contains "2.71". A blue callout box labeled "Put mean value here" points to the conversion rules field. The "Submit" button is highlighted with a blue callout box labeled "Submit".

Once you have understood your distribution, you can easily replace the missing value. Suppose you are replacing by mean then how to know that you have to go with summary statistics either with summ or simply as with sum. Sum with that variable will be displayed with mean, standard deviation, minimum value and maximum value.

Once you click on the previous window, it will be displayed with this type of window, and you will simply enter the value that is 2.71 or the entire 2.714885 that will be replacing all those missing values. Usually, most of the database comes up with missing entries like 9 or 999, or they will instruct it what exactly they have entered.

The next one is your regression imputation. I suggest everyone, please take a note of it instead of just randomly getting a result without taking care of missing values. It is essential.

So, regression, imputation, what does this mean. This means you are replacing missing values with the predicted score.

(Refer Slide Time: 20:45)

Regression Imputation

- Replaces missing values with predicted score from a regression equation.
 $(X1)' = 4.621 - (0.734 * X2) + (1.139 * X3)$
- Advantages:
 - Use information from observed data
- Disadvantages:
 - Overestimates model fit and correlation estimates.
 - Weakens variance

SN	X1	X2	X3
1	9	8	8
2	6.32	7	6
3	8	5	6
4	7	4	5
5	9	5	7
6	8	8	9
7	6	7	6
8	5	9	7
9	7	8	8
10	8	8	7

19

Regression gives predicted values and as we all know that simple linear regression equation gives coefficients, whether it is a standardized coefficient or a non-standardized coefficient.

Once you have filtered the data set, I have already guided that and then normalized the data set. After that, if you are running your regression, your coefficient becomes a standardized coefficient instead of a non-standardized one. What is the advantage of this regression imputation? This is basically the expected value we are replacing.

The disadvantage of this method is that sometimes it overestimates the model fit and the correlation estimates, which weakens the variance. So this will reduce the variance, which may not be good for interpreting results.

(Refer Slide Time: 22:34)

Application in Stata

Level_of_care	Expenditure	Nature_of_ailment
17	2500	43
18	43	4
19	180	40
20	1900	24
21	990	14
22	.	43
23	300	44
24	380	16

$(\text{Level_of_care})(\text{col18}) = 2.87$
 $+ 5.39\text{e-}06(45) - 0.005(4)$
 $= 2.85$

```

regress Level_of_care Expenditure Nature_of_ailment
    
```

Source	SS	df	MS	Number of obs	F(2, 37877)	Prob > F
Model	226.913334	2	113.456667	= 37,880	= 68.74	= 0.0000
Residual	62512.3665	37,877	1.65040437			= 0.0036
Total	62739.2798	37,879	1.65630771			= 0.0036

Level_of_care	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
Expenditure	5.39e-06	1.54e-06	3.49	0.000	2.37e-06 8.41e-06
nature_of_ailment	-.0047665	.0004214	-11.31	0.000	-.0055923 -.0039406
_cons	2.871597	.012471	230.26	0.000	2.847153 2.896604

We have to run the regression of those 3 variables simply, 3 variables let it be as “regress level of care expenditure nature of element” as per our 75th round sample estimates we have given here in the data.

You can also check some of that basic information like observation, say F statistics, and P-value. Your model is perfectly fit though the R square value is not that good since what we have shown here is a sample data set.

I wanted to mention that once you run the regression, you get with your coefficients and, most importantly, your constant value and the other 2 coefficients. So, your constant value and beta 1 and beta 2 are given here.

(Refer Slide Time: 24:44)

Last Observation Carried Forward

- Imputes the missing value as a value on the same outcome the most recent time it was observed.
- Take average of X1 and X2

SN	X1	X2	X3
1	9	8	8
2	6	7	6
3	8	5	6
4	7	4	5
5	9	5	7
6	8	8	9
7	6	7	6
8	5	9	7
9	7	8	8
10	8	8	7

21

So, last observation carried forward is another method; once your last observation has a value that is going to be very relevant for your analysis, that one can be carried out. It imputes the missing values as a value on the same outcome the most recent time it was observed.

So, you can also take the average of X1 and X2, and once that is close to the value, that value could be entered as well. The last couple of slides is there to guide you and give you a complete catch of understanding missing values.

We have called multiple imputations like it consider impute, analyze, pooling and simultaneously, we are leading with a missing a value that will be filled in the missing data.

So, data is filled in with imputed values using a specified regression model. This is what we have already said earlier. This step is repeated m times because of the number of missing values, resulting in a separate data set each time.

(Refer Slide Time: 26:11)

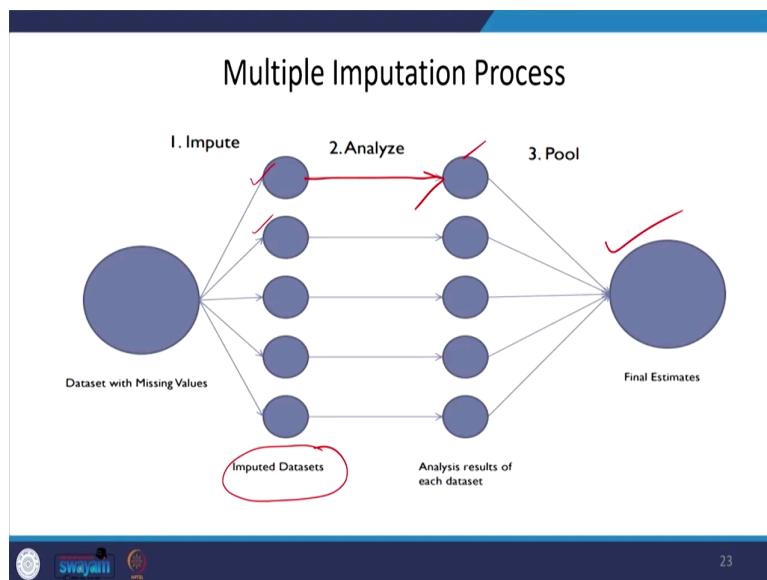
Multiple Imputation

- ❑ **Impute** : Data is filled in with imputed values using specified regression model.
 - This step is repeated m times, resulting in a separate dataset each time.
- ❑ **Analyze**: Analyses performed within each dataset.
- ❑ **Pool**: Results pooled into one estimate.
- ❑ **Advantages** :
 - Variability more accurate with multiple imputations for each missing value.
 - Considers variability due to sampling and variability due to imputation.
- ❑ **Disadvantages**:
 - Cumbersome coding.
 - Room for error when specifying models.

22

Then you will analyze and perform within each data set, and based on that, we will again pool those new findings or new predicted values based on that we can pool into one estimate.

(Refer Slide Time: 26:28)



In the chart, you can easily see if the data set has many missing values we analyze categorically one by one, then we get the number of imputed data sets. We will analyze it through regression and this coefficient will be derived and that will be considered. We will pool or append the extra information derived based on our new data.

Once we pool each corresponding data entry, we have generated a new data set and based on that; we will get our final estimate. So, this is called multiple imputations since we are actually having our analysis turn and one advantage is that the variability is well captured. It is more accurate than the simple imputation; simple regression imputation entries will be different instead of a single regression.

We are going with different models every time, different regression every time and getting a pool of data and based on that either you can again moderate it and get the missing values. This considers variables due to sampling and variability due to its multiple steps of imputation.

There are disadvantages like it's very tough to deal with because it has a number of steps involved. It is, in fact, quite cumbersome in terms of coding and its entering. So, room for error is higher since you are dealing with the approach multiple times.

So, these are all for today. So, you can combine the last lecture and today's lecture while you are following. I am you will fill the gap so far as the existing research is concerned, and these are all fine-tuning or beautifying your research.

Therefore, I suggest that you please follow, and accordingly, you go for the handling missing values for final analysis. I hope it is understood and I am sure you will have certain difficulties that could be dealt with in our doubt class or live session. So, I must thank you because of your presence you may also get your query solved through my TA involved in this program: Mr.Kamal and Mr. Milind.

So, their details will be also available through the NPTEL office you may connect to them as well. So, let us stop here.

Thank you.