## Handling Large-Scale Unit Level Data Using STATA Professor. Pratap C. Mohanty Department of Humanities and Social Sciences Indian Institute of Technology, Roorkee Lecture No. 05 Understanding Unit Level Data: NFHS

Welcome to the MOOC module on handling large scale unit level data with STATA; myself Dr. Pratap Mohanty, faculty member in the department of Humanities and Social Sciences. We have already started the discussion on this particular module by making you familiarize with the basics of unit level data. We are trying to make you understood what type of different unit level data are available for research at a larger context in India. And so far we have already completed 4 lectures, this is our fifth lecture; and with this lecture we can able to finish the first week lectures.

And in the last two lectures particularly, I unfolded the discussion on two important databases on unit level. Those are national sample survey as well as India Human Development Survey; so that is called IHDS. So, those two we have already completed and there are number of variables also dealt; but still it requires further attention when we introduce STATA in our next week. Next week we will steadily introduce once again with the databases for working; but majorly we will be working with NSS (National Sample Survey) database.

The last lecture on the very first week of this module is on understanding national family health survey data. This is very unique in the sense that this is one of the largest dataset in the world, so far as health and its different measures are concerned. So, this is this unique platform on observing various qualities of health especially women's health and child health care.

And those who are interested in linking their work or wish to have a complete devotion on working with the nutrition related issues, and challenges related to poverty, and challenges with the women, women empowerment. This is probably the best dataset I have seen among the Indian datasets. So, national family health survey therefore is very important for discussion. So, let us have a look and again why this is important?

(Refer Slide Time: 03:22)



Because this is very much comparable and this has been developed under the umbrella called DHS (Demographic and Health Survey). And it collects data of various countries and sticking to more or less the same format. The pattern of questions in all the rounds though they are differing; but, the idea or the approach by which the questions are dealt in other countries are more or less similar. So, the data extracted or derived from this platform is comparable to other countries in the world. And that too this is also nationally representative; we have also discussed why this is so.

So, NFHS stands for National Family Health Survey, and in India the leading organization in this database is IIPS (International Institute of Population Sciences). And IIPS follows again number of approaches to conduct the surveys; and this is very authentic. Let me understand the data through different perspectives.

(Refer Slide Time: 04:43)

- □ It is a part of "The Demographic and Health Surveys (DHS)" which is an international project designed to collect comparable survey data across countries.
- □ DHS is conducted across the globe.
- □ It is also called the Indian Demographic and Health Survey (IDHS).
- The National Family Health Survey (NFHS) is a large-scale, multi-round survey conducted in a representative sample of households throughout India.



As I already mentioned it is part of the DHS which is an international project designed to act to come to collect data and those can be comparable across the countries. I have already mentioned and this is where the first point stands. So, you may prepare one objective question like why is it the case that NFHS is comparable across the globe or across the countries, where DHS already conducted surveys on healthcare. So, it is across the globe because many countries have been covered; it is also called the Indian Demographic and Health Survey (IDHS) data.

The National Family Health Survey is also called a large-scale multi round survey conducted in a representative sample of household throughout India. We are going by the fifth episode of NFHS; though the fifth episode data is not yet available for the public, but it is at the verge of finalization.

(Refer Slide Time: 05:54)

Each successive round of the NFHS has had two specific goals:

 To provide essential data on health and family welfare needed by the Ministry of Health and Family Welfare and other agencies for policy and programme purposes, and
 To provide information on important emerging health and family welfare issues.

 The Ministry of Health and Family Welfare (MOHFW), Government of India, designated the International Institute for Population Sciences (IIPS) Mumbai, as the nodal agency, responsible for providing coordination and technical guidance for the survey.

Each successive round of NFHS has had two specific goals. One is to provide essential data on health and family welfare needed by the Ministry of Health and Family Welfare. Other agencies for policy and programme related purposes. And to provide the information on the second aspect is to give information on important emerging health and family welfare issues; or you may say family welfare challenges.

So, these two objective are based across the globe for entire surveys in the world, so far as DHS is concerned. The Ministry of Health and Family Welfare (MOHFW) located in Delhi under the government of India designated the IIPS that is International Institute of Population Sciences, Mumbai as the nodal agencies. I already mentioned. responsible for providing coordination and technical guidance for the survey.

(Refer Slide Time: 07:06)

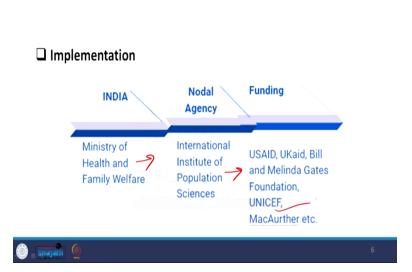
- Technical assistance for the NFHS was provided mainly by ORC Macro (USA) and other organizations on specific issues.
- □ The funding for different rounds of NFHS has been provided by USAID, DFID, the Bill and Melinda Gates Foundation, UNICEF, UNFPA, and MOHFW, GOI.



The technical assistance for the NFHS was provided mainly by the ORC Macro that is USA based and other organizations on specific issues; so far the technical assistance and its authentication is concerned. The funding for different rounds of NFHS has been provided jointly by USAID, DFID, the Bill and Melinda Gates Foundation, UNICEF, UNFPA, MOHFW that is Government of India as Ministry of Health and Family Welfare. So, there are so many top most organizations who give funds for this purpose, and they do expect number of directions while granting the fund to the countries.

So, eradicate number of challenges, number of goals; again these goals are also linked to the SDG's where sustainable development goals; we can work related to healthcare. There are 3 important goals can be linked and will certainly deal with those aspects later.

(Refer Slide Time: 08:20)



But, at this moment I am just clarifying what is called NFHS and how it is important, and what are the legacies behind NFHS. So far as the implementation is concerned of NFHS, as I mentioned it is MOHFW and the nodal agencies IIPS; then the funding agencies are all those. Even MacAurther is also important, MacAurther is also important for source of funding, you can take a note.

(Refer Slide Time: 08:49)

Apart from socio-economic information of respondent the Survey provides information on many qualitative aspects of respondent/ individual.

- Gamily planning and welfare.
- □ Healthcare services and utilization
- Maternal and child health
- Nutritional status and anemia among women and children
- Sexual Behavior
- U Women empowerment
- Domestic violence etc.



Apart from the socio-economic information of respondent; the survey provides information on many qualitative aspects of respondent or individual. Those qualitative aspects are under family planning or welfare, health services and utilization, maternal and child health, nutritional status or anemia among women and children both can be tapped, sexual behavior, women empowerment, domestic violence. Even substance use, even intimate pattern violence as part of the domestic violence content; and we can extract number of information out of it.

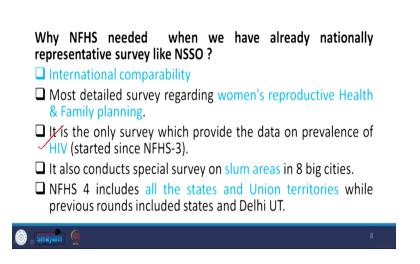
First of all, when a detailed information on maternal and child health care is given, we can have the relationship between mother and child, and that too in different years. Though this is not a panel data but some pseudo panel can be developed with certain techniques at a different cohort level. The cohort maybe your state level maybe at the district level; but you need to be very careful enough in making the cohort quite rational in the sense it has to be comparable across other data sets and that too across previous rounds. District level information are not available in all the rounds; I will discuss in nutshell I mean in my successive slides.

What I just wanted to develop the curiosity here for you as a researcher, because many challenges are there. India is still ranked among the countries, ranked quite below in terms of comparison among the country in the world; it is still comparable to African countries. So, far as the hunger industry is concerned, India is ranked at a very poor level. if you are interested in comparing with other health indicators of other countries; you can refer to other countries African countries and their DHS database.

And you can compare the nutritional status, the anemia level. Even though nutritional status you can calculate their stunting level, their wasting level or mal-nutritional related aspects; even over nutrition, under nutrition. There are possibility of research related to double nutritional challenges, multiple nutritional challenges within the household or within or across the individuals. So, there are various possibilities, I am saying it is a rich database.

Let us go ahead with certain other direction why NFHS needed, when we have already nationally representative survey like NSSO. It is because of the case that NSSO may not be comparable with other countries in the world; whereas NFHS is strictly comparable.

(Refer Slide Time: 12:22)



Because the approach considered in each countries are more or less similar, and I will show you those datasets how you can go through and how you can find the comparability. So, most detailed survey regarding women reproductive health and family planning; there is no such database in NSS which gives detailed observation on women's reproductive health and their family planning. Similarly, it is the only survey which provide the data on prevalence of HIV that is very very interesting; and controlling HIV is one of the trivial question across different forums.

And especially since NFHS-3, this has been included in the questionnaire, though it is very sensitive. If I am an individual researcher, I wanted to study at a ground level; but in the ground level the persons or the women or even men may not cooperate. And so in order to make them convinced and unbiased, you require lots of interventions; so, lots of spending is also required at the individual level. Without spending, probably they are not in a position to respond. Some forms of incentivization are required or some kind of that is I mean we may mention as a form of intervention.

So, interventions require funds, so at an individual level it is very difficult to capture many issues. Whereas, since this is a larger platform and the nodal agency as IIPS; IIPS is famously known for population related research. So, the information derived from this platform is quite authentic and comparable; and lots of research paper can be derived, if you are sincere enough in understanding the database carefully. So, this is interesting, HIV is considered.

It also conducts special survey on slum areas in 8 big cities. So, slum related information you can also get it from NSS (National Sample Survey); but, this coverage and the NSS coverage are not comparable. Let me make it very clear because the sampling frame considered by NSS and NFHS are entirely different. So, just one result from using STATA of NFHS and another result of NSS are not comparable; so be careful about these particular databases. Where, NFHS-4 includes all the states union territories, while previous rounds included states and Delhi as the union territories.

(Refer Slide Time: 15:37)

Data Information, Sources and Utilization □ Five rounds of this survey has been conducted since its inception. Given First round- 1991-92 Second round- 1998-99 □ Third round- 2005-06 Generation Found- 2015-16 Fifth round- 2018-19 (Only questionnaire is available) 🌀 \_ swayam 🎯

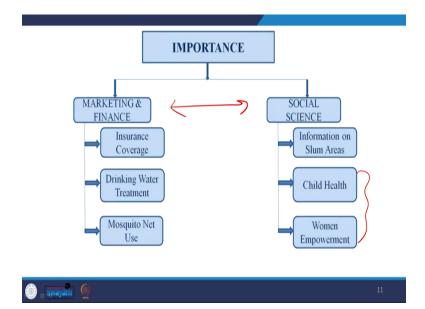
Data information and sources and utilization; so what are the sources? Which kind of databases? In which year those are studied? 5 rounds of this survey has been conducted since its inception. It started with 1991-92 that was the first round of data on health and family welfare related information; and the second round was published or reported in 1998-99. Third one is 2005-06; fourth there is a gap of 10 years. The fourth occurred in 2015-16 and the fifth is in the process and about to be published soon as reported in their website. But so far only questionnaire is available in this regard and this is in fact the original source; where I know you can understand which are the round.

(Refer Slide Time: 16:41)



This is NFHS-1 mentioned, this is NFHS-2, 3 and 4 and 5; and 5 here we have highlighted questionnaires. And in the questionnaires biomarkers are also interesting, and it is important to note, we will also discuss in our presentation in my discussion. And this is from the IIPS source International Institute of Population Sciences, and also NFHS also National Family Health Survey is mentioned. And so download data if you click on the download here; it will link you to the DHS website. In the demographic health survey website, you can able to download the data in your right format. Maybe in STATA also in the successive slides.

(Refer Slide Time: 17:33)

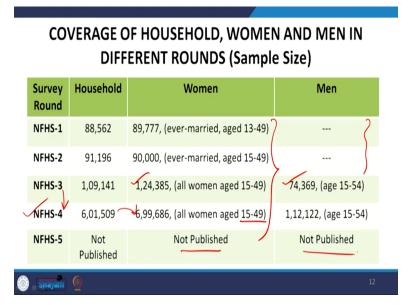


Why is this important? What are the importance of these, if anybody is interested in having research and there are marketing and finance? This data can also be useful like insurance coverage is also given. Like insurance coverage how many are the person insured and which are the sources of insurance. It is private, public or even interestingly the community health insurance is also given; which is very difficult to tap. And its respective definition is clearly given in the report; those who are interested you may go through.

Drinking water and treatment and so far as the individual finance is concerned, how much out of pocket is borne by the individual in this regard can also be calculated. Mosquito net so far as marketing is concerned, how many are using mosquito net; and what kind of mosquito related diseases are there in the country. So, even the morbidity profiles are clearly given in in the database. So, it is quite important. So far as social science is concerned, I think you should not look at in isolation. These are quid-pro-quo and relationships are there between these two domain.

Broadly, if you categorize the social science and their importance, we can derive information on slum areas, child health, women empowerment. Even the sociologists, the anthropologists are also highly interested in observing these two information carefully.

(Refer Slide Time: 19:18)



Let us have a look further on the coverage of this database like in terms of men and women composition; or their sample size in different rounds. It is very interesting to note I will start with the last one; that is NFHS-4 dataset, which is downloadable and freely available to the public. Only you are supposed to register at the designated place that is through the online platform. So, I will show you the link also. Here NFHS-4 covers 6,01,509 households, which is much higher than that of any other unit level database in India is presently having.

No database addressing on a single round covered such a large number of households. Out of this 6,99,686 women are covered. These are number of women individual, women are covered and 1,12,122 men are covered. It is interesting to note what do you mean by women here? Who are the women covered? though all individuals are covered, but more most interestingly all individual men and women; but specifically NFHS is defined eligible women. That too with the age group of 15 to 49.

And 15 to 49 is interesting to note because this is as per the WHO classification called reproductive age or the reproductive span. So, in the reproductive age this number are covered around 7 lakhs individuals are covered; those are women and more than 1 lakh are covered as men. And men age groups are covered as 15 to 54. if you look at combinely men and women together; this leads to a complete household or if you combine together probably will get a

couple file. The couple file is also important, so for as some of the specific answers are required, specific research is required.

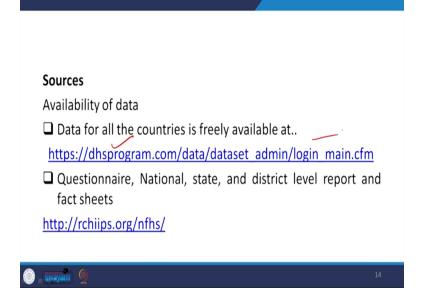
Whereas, NFHS-3 covered 1,09,141, compare NFHS-3 to NFHS-4; the number multiplied or more than around 6 times, and similarly for individual women and men. NFHS-2, 91,196 and other figures you can read and in all the rounds women information are given; whereas in the first two men information are not covered. And so it is also NFHS-5 since these are not published, we are not mentioning these number at this moment.

(Refer Slide Time: 22:37)



So, questionnaires or the schedule used for all the rounds are very important for analysis; because at the household level household information in all the rounds are covered. Yes is given all the rounds, you will get household information. Women are in all the round, men only in lasts three rounds, if you are considering the NFHS-5; and the biomarkers started from NFHS-4. So, NFHS-4 onwards the database is getting rich and there is proper calculation weight, height information is given; with the professionals, health professionals and the biomarkers are carefully considered. The village information are not presently recorded; so information are not given.

(Refer Slide Time: 23:38)



These are the link; you can have a click on those link in our PPT these PPTs will be uploaded for your reference. If you just click here, you will be redirected to the DHS website for login; and you just need to enter if you are a student or researcher. You simply give your basic details; you will be allowed to get the data; otherwise you can also click here to directly get further information on NFHS.

(Refer Slide Time: 24:11)

	DOWNLOADING THE DATA				
Demographic and Health					
The DHS Prooram > Data > Di	taset Account Hang = logn_main				
ABOUT THE DATA	Login and Download Instructions				
Getting Started	New Users Register				
Data Collection Data Processing	Before you can download detaents, you must reporter as a DoS data user. Dataset access is only granted for legitimate research purposes, Learn more about data restrictions and <u>how to restrict access</u> or view a <u>list of available datasets</u> . <b>Register for Dataset Access</b>				
UNDERSTANDING SURVEY STATISTICS					
Data Quality and Use	Model Datasets Available				
Data Tabulation Plan Survey Indicators	If you would like to try out some data files we have provided some example model datasets for which you do not need to register. Click here: to download model dataset files.				
Data Tools and Manuals					
Online Guide to DHS Statistics	Registered Users Please Login Here				
WORKING WITH DATASETS	Email: Password: NEW (passwords are new case sensitive)				
	Password: NEW! (passwords are now case sensitive)				

This is the website I am referring to and this is the user id you need to register here. It calls for password, you need to enter a new password and you get the data.



(Refer Slide Time: 24:22)

And how it looks like? It looks like this India 2015-16; that is the NFHS-4 data. This is NFHS-3 and so on, this is 2 and 1; and so it is written as in which phase those are covered. And it is always advisable to the readers or to the researcher to get the latest series of the data. The DHS covered the 74 series, and I will show you why it is 74. There are some other series because there are some calculation added, some recode made by the institution; and they help the individual researcher to get the authentic source and the correct data.

So, earlier series may not be referred, so I will show you. Recode is given. Recode is the column is given because they carefully provide the latest episodes of the data. And GPS indicator that is another interesting indicator, so far as India's dataset is concerned. From NFHS-4 onwards GPS information are available. In the previous round the geographical positioning information is not available; I think if you can look at Dr. William Joe paper published in EPW. Basically, he tried to look at the parliamentary constitutions and their impact on development; how a politician lead with the policy changes. How carefully or sensitive the politician is in developing their constitution.

So, you can get the constitution data by geographical positioning system and its availability. So, there is so this is little rich and HIV and all other biomarkers data available since NFHS-3 onwards; I have already mentioned in the previous slides.

Data	Survey Dataset Files					
ABOUT THE DATA	Te dia standard Dus a			_		
	India: Standard DHS, 2	India: Standard DHS, 2015-16 This page shows a listing of all dataset files available for the selected survey.				
Data Collection						
Data Processing	If you are a registered user, <u>please login here</u> to gain access to these files. If you are not a registered user, <u>please go here to register</u> .					
	n pos are not a report to ano,	READER OF LEASE AN AD DESIGNAL				
Recommended Citations	Survey Datasets					
UNDERSTANDING SURVEY	File Name	File Size	File Format			
STATISTICS	All-Ipera					
Data Quality and Use	Births Recode					
	IAB 74 DT.ZIP	202 MB	💙 Stata dataset (.dta)			
	LABR74FL_ZIP	206 MB	Flat ASCII data (.dat)			
Survey Indicators	IABR74SD.ZIP	396 MB	SAS dataset (.sas7bdat)			
Data Tools and Manuals	IABR245V.ZIP	326 MB	SPSS dataset (.sav)	_		
Online Guide to DHS Statistics	Couples' Recode					
	LACR74OT.ZIP	42.2 MB 44.4 MB	Stata dataset (.dta)			
WORKING WITH DATASETS	IACR74FL.ZIP	44.4 MB 66.5 MB	Flat ASCII data (.dat) SAS dataset (.sas7bdat)			
	IACR74SD 21P	47.4 M8	SPSS dataset (.sav)			
Using Datasets for Analysis	Heusehold Recode		as an ansatz (Tant)	-		
Model Datasets	JAHR74DT.ZIP	309 M8	Stata dataset (.dta)			
Data Variables & Definitions	IAHR74FL.ZIP	283 MB	Flat ASCII data (.dat)			
Dataset Types	LAHR7450.ZIP	444 MB	SAS dataset (.sas7bdat)			
File Formats	LAMB 4SV.ZIP	295 MB	SPSS dataset (.sav)			
File Types & Names	Individual Recode					
Merging Datasets	IAIR74.ZIP	450 MB	Hierarchical ASCII data (.dat)	_		
Herging Datasets	IAIR24DT.ZIP	326 MB	Stata dataset (.dta)			

(Refer Slide Time: 26:40)

This is the DHS dataset where you can able to download; so like as I told you 74 is the series, 74 is written here. And this is couple's information, birth information, household information, individual information. Individual information with the code is IR given; that is called individual recode. And recode data is important because they have already recoded some of the variable for our easy use; and couple information with the code CR. Birth recode with the code BR; but this 74 series is important and it is available. , what else is important in this page you need to check carefully? Which dataset you are going to download?

So, since we are using dot dta format that is STATA data; our module is on STATA. So, our title of the MOOC module is on handling large-scale unit level data with STATA. So, STATA dataset I will suggest you to download; it is user friendly and faster. So, there are even though if you are comfortable enough in SPSS also you can download SPSS data here; SAS dataset is also there. Even in the notepad file dataset is there, ASCII format dataset also there. So, I will suggest you download the STATA dataset and start looking at the instruction we are giving to you. Let me proceed for the next information.

(Refer Slide Time: 28:23)

Number of Questions	Women's Questionnaire	NFHS-1	NFHS-2	NFHS-3	NFHS-4
Number of Questions	Respondent's background	56	24	18	16
Dealt in Four Rounds	Reproduction	32	34	42	59
	Marriage and cohabitation	NA	NA	18	17
of NFHS, 1992–93 –	Contraception	54	50	39	51
	Contacts with health personnel	NA	NA	18	12
2015–16	Quality of care	NA	19	NA	NA
	Antenatal and postnatal care	47	52	82	95
	Immunisation, health and women's nutrition	51	47	83	89
	Utilisation of ICDS	NA	NA	15	15
	Sexual life	NA	NA	20	22
	Fertility preference	18	10	23	23
	Husband's background and work	21	22	29	39
	Household relations (status of women)	NA	7	29	40
	HIV/AIDS	NA	6	38	-55
	Total	279	271	454	517
	Number of questions in household questionnaire	79	64	68	27
	Number of questions in men's questionnaire	NA	NA	216	231
Source: K Srinivasan, Rakesh Mishra, EPW	Biomarker questionnaire	10	9	18 <sup>.</sup>	156
2020, V.6	Grand total	368	344	756	981
	NA= Questions related to the sections have not been adr	ninistere	d.		18

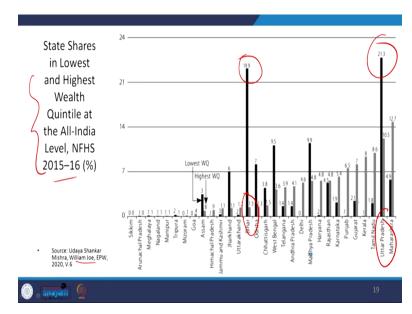
Number of questions dealt in 4 rounds so far of NFHS from 1992-93 till 2015-16; so let me read between the lines of the table, where respondents background the number of information are given is mentioned here. We are referring to the case innovation and Rakesh Mishra, EPW paper that was published in 2020 very recently in the volume number 6. So, reproduction information is also there and so number of questions dealt this; this is important because how many questions on this particular issue is there. Look at there are in the NFHS-4, 59 questions on reproductions are there, which is too interesting.

So, far as contraception use is there and family planning and contraception how matured our society is in using contraception; so, it is a healthy symbol of managing population. So, look at there are 51 variety of questions are covered; so qualitative in nature and looking at utilization of ICDS. Those who are very interested in program related analysis, ICDS program is internationally famous. This is considered to be one of the very successful programs so far as healthcare is concerned in India. And ICDS utilization the coverage of it started from NFHS-3; so there are 15 questions covered.

But in this particular question I will suggest as an individual researcher, you should do a parallel field study; because the information 15 variety of questions are though covered but not exhaustive. Still many challenges are there for better coverage. Similarly, sexual life, fertility preferences, household background in work, household relations, HIV. HIV I already mentioned

from the very latest two episodes we are covering HIV AIDS information; and in total 517 information questions are covered in the latest round of NFHS.

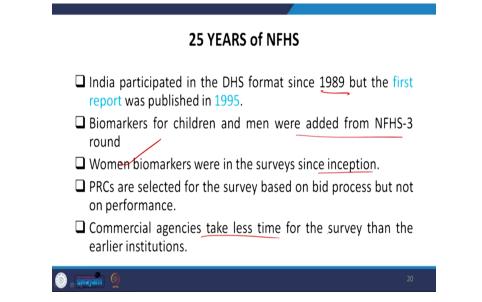
So far number of questions in the household questionnaire is concerned; it is 77. In the latest round number of question related to main questionnaire is 231; and biomarker question interestingly there are 156 variety of questions biomarkers. So, you can able to get a scientific observation on it, and in total there are 981 questions covered.



(Refer Slide Time: 31:28)

So, you can just follow, this is also the source office given William Joe, the U. S. Mishra CDS; professor William Joe of institute of economic growth published in the paper EPW is quite interesting. So, we are referring for your analysis and look at they derive some interesting result out of the NFHS-4. Here compare the top quantile and bottom quantile population; look at the gap and how the differences are there in different states. I am not analyzing these details at this moment; you can follow it up and this is very indicative.

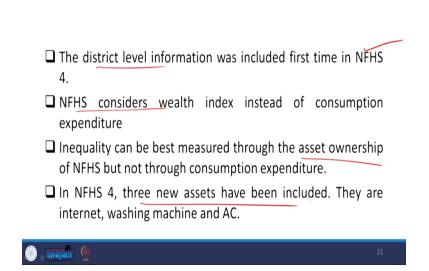
(Refer Slide Time: 32:17)



The legacy of 25 years of NFHS is important because of the fact that India participated in the NFHS or the DHS format since 1989; but the first report was published in 1995. The biomarkers as I said for children and men were added from NFHS-3 rounds onwards. Whereas, women biomarkers have been since its inception that is from the first round onwards; women information is given and the biomarkers are given. So, PRCs (Population Research Centers) are selected for the survey based on the bidding process; but not on their performance that is one of the criticism given in the recent paper, so far as NFHS is concerned.

Commercial agencies take very less time in managing the survey; nowadays they are best doing upon the commercial agencies instead of the PRCs. Only 3 PRCs are considered for this purpose; whereas they are simply giving to the commercial agencies. Which is consuming less time but they are skipping very qualitative questions. They cover the questions at a very less time which is raising huge question, huge doubts on their survey process.

(Refer Slide Time: 33:33)



The district level information was included for the first time in NFHS-4 as I already mentioned. So, district level comparison in the previous round is not possible, and NFHS considers wealth index instead of consumption expenditure. So, they give the wealth information not the income as such. Inquality can be measured through asset ownership since wealth index is given. So, in NFHS-4, 3 new assets have been added and they are internet, washing machine and AC. So, please follow those details in the dataset; you can able to get it. I think I have already given enough information; we will experiment with the STATA in our respective lecture. With this let me stop here. Thank you.