

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

**Lecture - 12**  
**Weights and Representative Sampling**

Welcome friends to this NPTEL module on handling Health Care Survey Data. We are in the third week. My name is Dr. Pratap Mohanty attached with IIT Roorkee in the Department of Humanities and Social Sciences. This lecture we have targeted to make you clarify with certain important directions in research.

The weights and representative sampling are important in writing research papers. The reviewers are demanding certain contents of the analysis, such as weights and representative sampling. In any analysis you do, the reviewers are expecting that your results must be representative enough.

Your results from the sampling or from the data must tell stories to a larger segment. So, any sort of policy implementation could be made based on your paper, based on your results. But representative sampling, as we have already read earlier, that it requires more sample size.

It requires the sample size to be reasonably higher. Without that in a small sample case, it is very difficult to represent the population. So, there are certain techniques to represent our population through weights.

Why should I give weights; why should I give overemphasis on a certain segment as we know that the structure of India is highly divided. There are huge hierarchies in the Indian context so far as development is concerned.

Certain sections have huge inequalities and there are resources not well distributed. That is the reason why your data itself is not normal. It is skewed and with all our statistical techniques, we know that the data has to be presented in a normal distribution. Then we can only estimate the number of inferences out of it.

But if your data is skewed, then averaging the outcome is not the right indicator. So, just based on data like income, for example, in society, only 10 percent of the population holds 90 percent of the income. In that case, if you are simply taking per capita income, I think it is

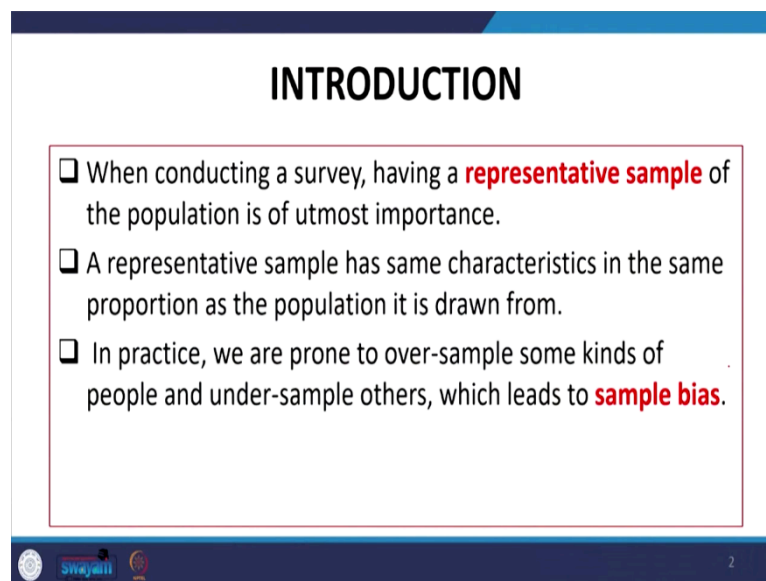
going to be misleading. So, when you are estimating any average income, it should be represent people.

But at this moment I am just trying to say that certain weights are required, certain valuations are required to a particular variable that would actually lead your interpretation in a better way and increase the acceptability of your paper to a larger segment. So, this is all about the background regarding weight.

Now, let us understand the introduction to it. My team in this particular segment has been consistently helping out. They are Mr. Milind and Mr. Kamal, and they are part of the Humanity and Social Science Department. They have been working on this NPTEL project.

Milind and Kamal have developed all those important connections in the PPT, and you will learn all those things very clearly. So, let us go ahead and understand.

(Refer Slide Time: 04:51)



The slide is titled "INTRODUCTION" in bold black text. Below the title is a red-bordered box containing three bullet points, each preceded by a square icon. The first bullet point states that having a representative sample is important. The second states that a representative sample has the same characteristics in the same proportion as the population. The third states that in practice, over-sampling and under-sampling lead to sample bias. The slide footer includes the Swajani logo and the number 2.

## INTRODUCTION

- ❑ When conducting a survey, having a **representative sample** of the population is of utmost importance.
- ❑ A representative sample has same characteristics in the same proportion as the population it is drawn from.
- ❑ In practice, we are prone to over-sample some kinds of people and under-sample others, which leads to **sample bias**.

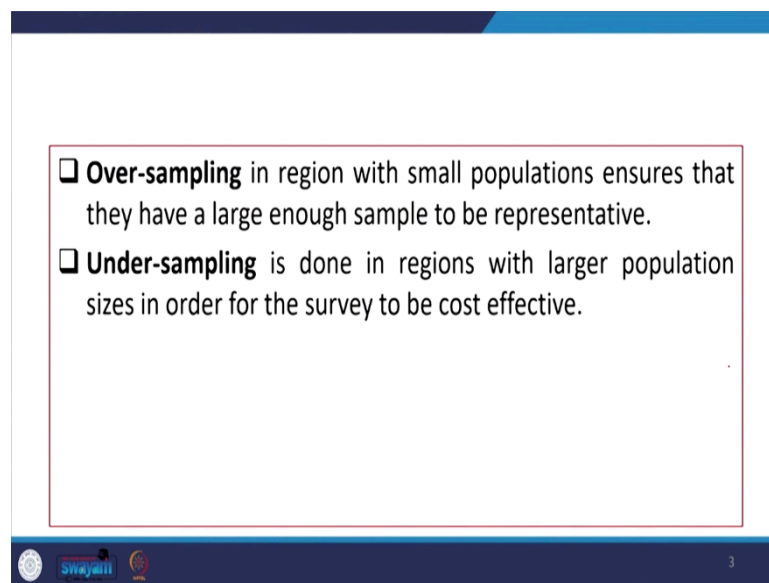
So, first of all, when conducting a survey, having a representative sample of the population in any kind of study is of the utmost importance. A representative sample has some characteristics in the same proportion as the population it is drawn from.

So, the sample should also narrate the story of your population. In practice, we are prone to over-sample some kinds of people and under-sample others, which leads to sampling bias. You might have heard about sampling bias even if we knowingly create certain possible

biases in our model, some of the variables are dropped, and some of the observations are dropped.

There are many ways where sample biases are created. So, we have certain techniques to minimize bias. At this moment I am not exactly heating the solutions for correcting the sampling bias all totally, but some of the approaches we are addressing.

(Refer Slide Time: 05:58)



- ❑ **Over-sampling** in region with small populations ensures that they have a large enough sample to be representative.
- ❑ **Under-sampling** is done in regions with larger population sizes in order for the survey to be cost effective.

So, over-sampling in regions with small populations ensures that they have a large enough sample to be representative.

Whereas in other cases like under-sampling with larger population sizes in order for the survey to be very cost-effective. We take a very specific sampling. So, that hedge actually undermines the sampling requirement and that is why we are saying under-sampling.

(Refer Slide Time: 06:33)

Weighting is a statistical technique for removing bias from a survey sample and ensuring that the results are more representative of the target population.

Weights are used to restore the representativeness of the sample, so the total sample “look like” the actual population.

The diagram shows a golden balance scale. On the left pan, a pie chart labeled 'Actual' is shown. On the right pan, a pie chart labeled 'Survey' is shown. The scale is balanced, indicating that the weighted survey sample matches the actual population.

4

Now, weighting is in fact, a statistical technique for removing bias from a survey sample. Whether it is over-sampled, or it is under-sample in both cases there are possibilities of biases. So, the weighting method helps in correcting those things. The target population weights are used to actually restore the representativeness of the sample, so the total sample looks like the actual population.

(Refer Slide Time: 07:00)

### DHS Definition

**Technical definition:**  
A representative factor applied to each case in tabulations which is in relation with the overall probability of selection and interview for each case in a sample, either due to design or happenstance.

**Practical definition:**  
A number that is multiplied by each case (women, child, household, couple) to “weight up” or “weight down” that observation if under-or over-sampling is applied.

Source: Demographic and Health Survey (DHS)

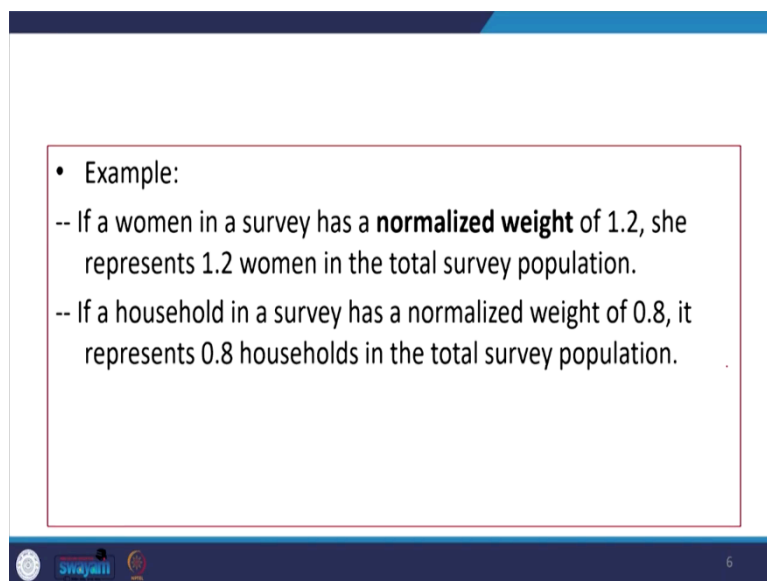
5

Now, I am coming to the exact definition of those technical terms we are using. So, it is taken from the demographic and health survey (DHS).

So, the technical definition of the DHS is that a representative factor is applied to each case in tabulations which are in relation to the overall probability of selection and interview for each case in a sample, either due to design or happenstance.

The practical definition is that a number that is multiplied by each case may be for women, children, households, and couples to weight up or weight down that observation if under or over-sampling is applied. So, weight up or weight down is possible if we apply the exact sampling procedure.

(Refer Slide Time: 08:14)

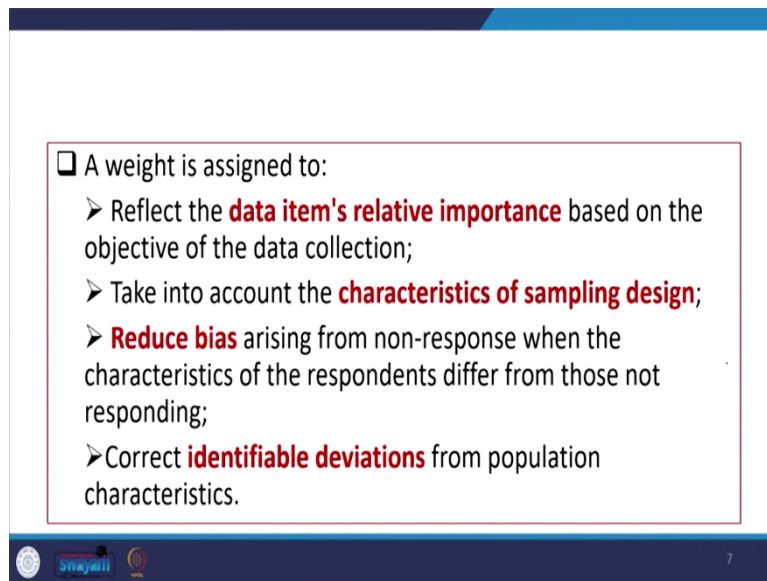


- Example:
  - If a women in a survey has a **normalized weight** of 1.2, she represents 1.2 women in the total survey population.
  - If a household in a survey has a normalized weight of 0.8, it represents 0.8 households in the total survey population.

So, let us understand that further, if a woman in a survey has a normalized weight of 1.2, she represents 1.2 women in the total survey population. So, it is not 1. Usually each person uses 1 without any weight on it. If we carry with the normalized weight, then the coefficient or the weight value matters.

If a household in a survey has a normalized weight of 0.8. It represents 0.8 households in the total survey population. So, accordingly, the weights are interpreted. So, let us understand that weight is assigned to reflect the data items relate to relative importance based on the objective of the data collection. It is of relative importance in the total data collection carried in the total population, considering the characteristics of the sample design.

(Refer Slide Time: 09:18)



□ A weight is assigned to:

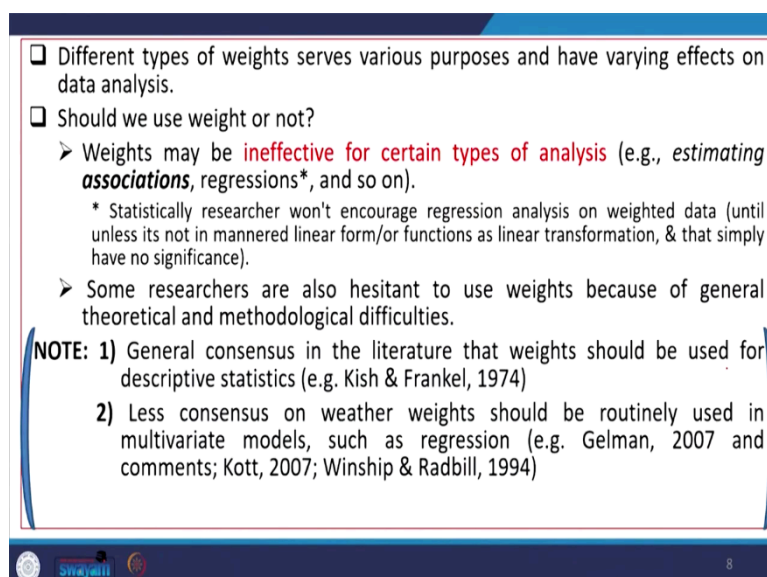
- Reflect the **data item's relative importance** based on the objective of the data collection;
- Take into account the **characteristics of sampling design**;
- **Reduce bias** arising from non-response when the characteristics of the respondents differ from those not responding;
- Correct **identifiable deviations** from population characteristics.

7

So, this also emphasizes or gives proper weight to the sampling design which was initially framed. This also helps in reducing biases arising from non-response when the characteristics of the respondents differ from those not responding.

So, the reduction of biases gives relative importance to some of the sections. Then it helps in proper design as per the expectations and this also helps in correcting identifiable deviations from the population characteristics. If there is some deviation from the population since it is not represented correctly your weight may lead to representing this in a better way.

(Refer Slide Time: 10:17)



□ Different types of weights serves various purposes and have varying effects on data analysis.

□ Should we use weight or not?

- Weights may be **ineffective for certain types of analysis** (e.g., *estimating associations, regressions\**, and so on).
  - \* Statistically researcher won't encourage regression analysis on weighted data (until unless its not in mannered linear form/or functions as linear transformation, & that simply have no significance).
- Some researchers are also hesitant to use weights because of general theoretical and methodological difficulties.

**NOTE:** 1) General consensus in the literature that weights should be used for descriptive statistics (e.g. Kish & Frankel, 1974)

2) Less consensus on whether weights should be routinely used in multivariate models, such as regression (e.g. Gelman, 2007 and comments; Kott, 2007; Winship & Radbill, 1994)

8

So, there are different types of weights that serve various purposes and have varying effects on data analysis. Should we use weight or not? That is an obvious question that arises in any study. Often we are confused about should I go for it or not. Let us try some heat and trial process.

Let us go for the heat and trial process. If it is suitable for my work, then I will go for it else I will not consider it. In fact, that is not the right way researchers usually do in that way, but some we are giving some guidance. There might be many other ways of understanding, but let us go with this approach.

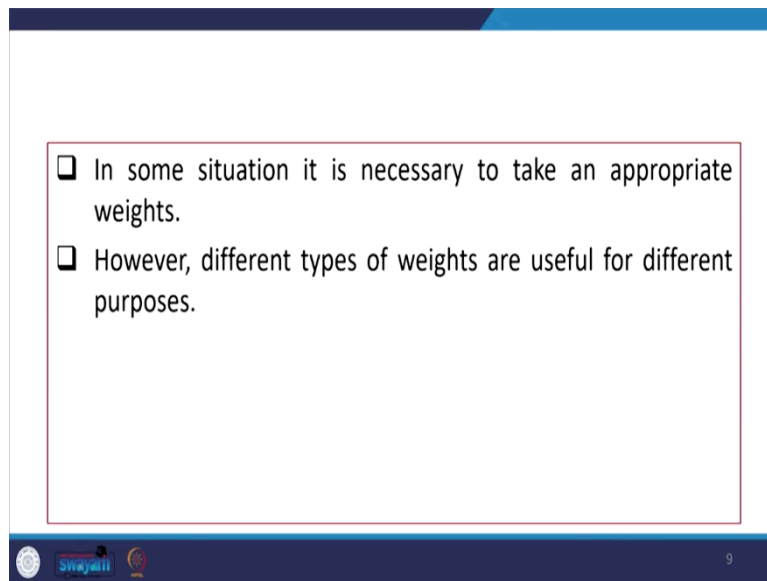
Weights may be ineffective for certain types of analysis such as estimating associations or regressions and so on. On regression, we have certain expert views and literature reviews I am just going to talk about. Let us understand that statistically, the researcher will not encourage regression analysis on weighted data, until and unless it is not in any mannered linear form.

So, the idea is that it is not usually suggested in regression unless it is not in a linear form or functions with a linear transformation and has no significance. So, then the only weight is going to be assigned, otherwise, it is ineffective. Some researchers are also hesitant to use weights because of general theoretical and methodological difficulties. That is in fact true and I was just talking about it.

So, from the 2 references we have highlighted at this moment, there are many other solutions as well as general consensus in the literature that weight should be used for descriptive statistics and less consensus on weather weights. The weather weight should be routinely used in multivariate models such as regression also suggested. So, there is less consensus on this aspect as well.

So, now, let us understand further details of all those things. In some situations, it is necessary to take an appropriate weight.

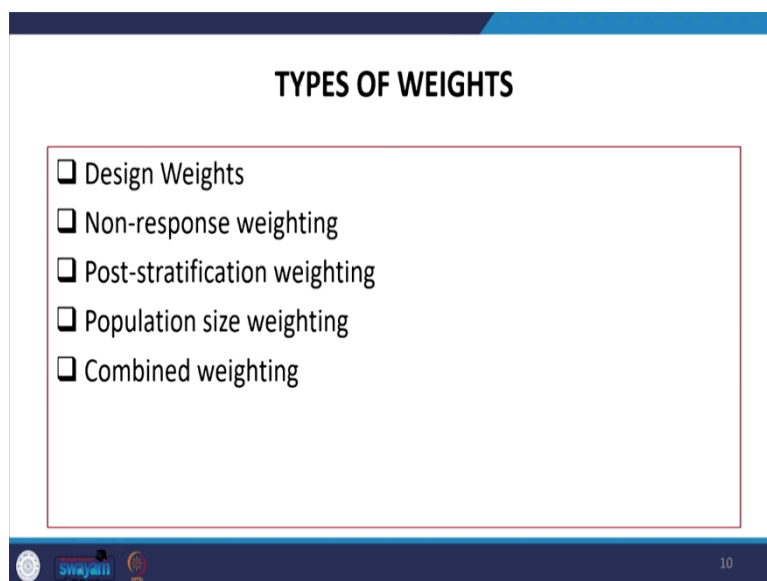
(Refer Slide Time: 12:56)



- In some situation it is necessary to take an appropriate weights.
- However, different types of weights are useful for different purposes.

However different types of weights are useful for different purposes.

(Refer Slide Time: 13:01)



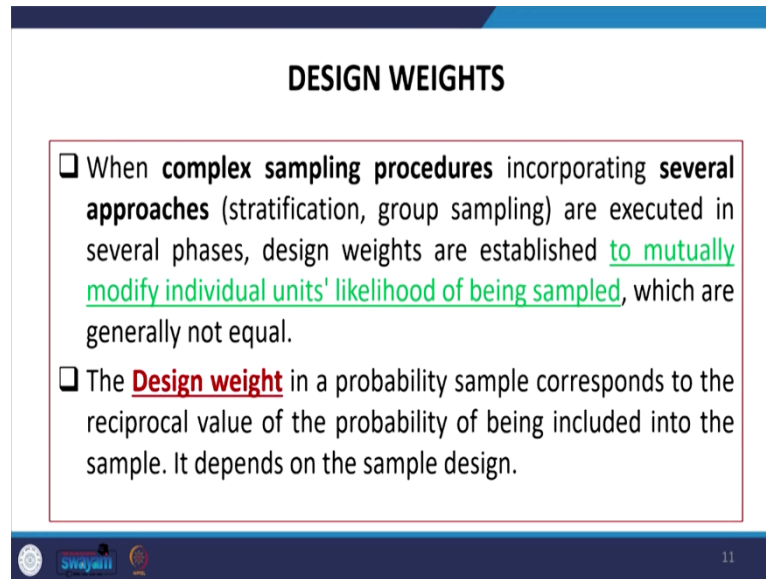
### TYPES OF WEIGHTS

- Design Weights
- Non-response weighting
- Post-stratification weighting
- Population size weighting
- Combined weighting

So, we are going to clarify those details. There are different types of weights. We are indicating here as 5. Design weights, non-response weighting, post-stratification weighting, population size weighting, and combined weighting.



(Refer Slide Time: 13:19)



**DESIGN WEIGHTS**

- ❑ When **complex sampling procedures** incorporating **several approaches** (stratification, group sampling) are executed in several phases, design weights are established **to mutually modify individual units' likelihood of being sampled**, which are generally not equal.
- ❑ The **Design weight** in a probability sample corresponds to the reciprocal value of the probability of being included into the sample. It depends on the sample design.

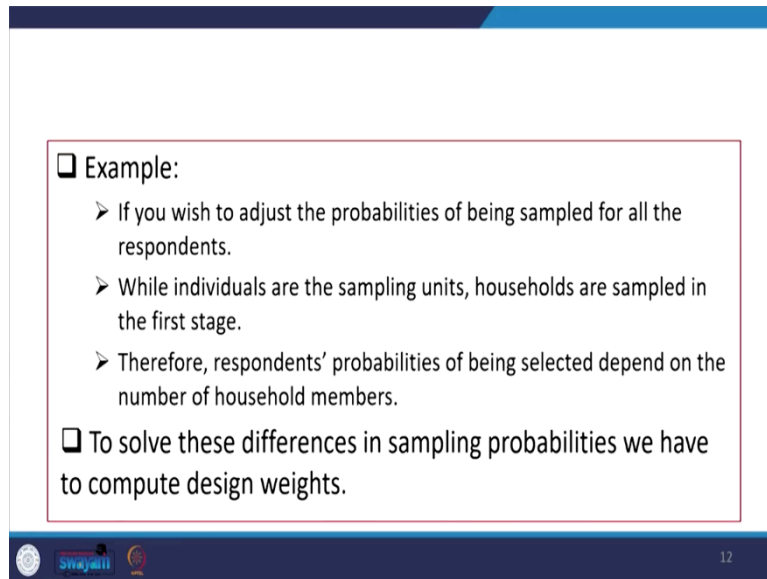
swayam 11

So, we will address them one by one. When complex sampling procedures incorporate several approaches such as stratification, group sampling is executed in several phases, design weights are estimated. If you have a multi-stage sampling format your design weights are most appropriate.

So, design weights are established to mutually modify individual units' likelihood of being sampled, which are generally not equal. So, just try to understand in the simple term that these are applied in multi-stage sampling procedures and complex sampling procedures. They have different likelihoods of being sampled in different layers.

The design weight in a probability sample corresponds to the reciprocal value of the probability of being included in the sample. So, the reciprocal value of their probabilities is going to be used as the sample in that particular category. So, if the probability of carrying that particular variable is  $1/3$ , then the weight that is going to be assigned is 3 by 1.

(Refer Slide Time: 15:02)



□ Example:

- If you wish to adjust the probabilities of being sampled for all the respondents.
- While individuals are the sampling units, households are sampled in the first stage.
- Therefore, respondents' probabilities of being selected depend on the number of household members.

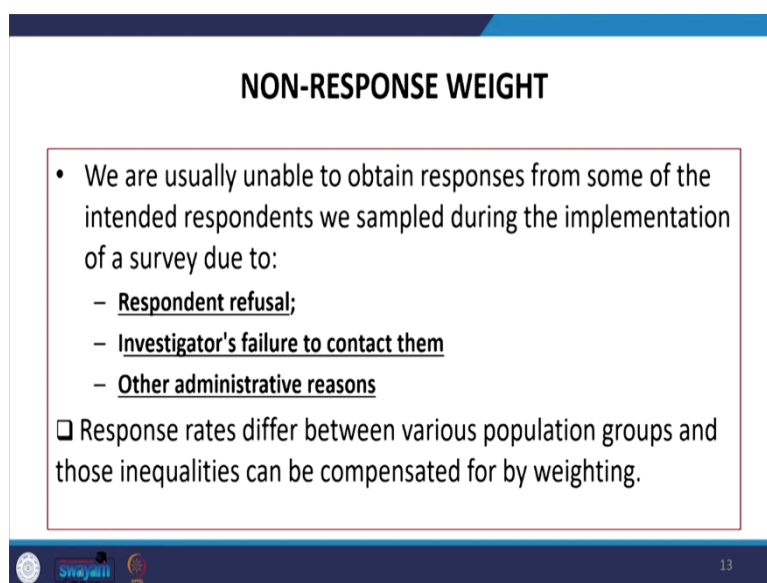
□ To solve these differences in sampling probabilities we have to compute design weights.

swajati 12

So, that is all about design weight and some examples we have also cited here like if you wish to adjust the probabilities of being sampled for all the respondents.

While individuals are the sampling units, households are sampled in the first stage, and therefore, respondents' probabilities of being selected depend on the number of household members. To solve these differences in sampling probabilities we have to compute design weights. So, design weights we will also discuss.

(Refer Slide Time: 15:36)



### NON-RESPONSE WEIGHT

- We are usually unable to obtain responses from some of the intended respondents we sampled during the implementation of a survey due to:
  - Respondent refusal;
  - Investigator's failure to contact them
  - Other administrative reasons

□ Response rates differ between various population groups and those inequalities can be compensated for by weighting.

swajati 13

The next type of weight is non-response weight. We are usually unable to obtain responses from some of the intended respondents we sampled during the implementation of this survey due to maybe respondent refusal. Maybe there is an investigator or enumerator who may not be able to contact the right persons for the response or other administrative reasons. There might be many reasons why responses are not recorded correctly.

So, response rates differ between various population groups and those inequalities can be compensated by weighting. If those are not there, some weighting techniques will actually compensate for these non-responses as well. So, that is why it is going to represent more population as well.

(Refer Slide Time: 16:39)

**POST-STRATIFICATION WEIGHT**

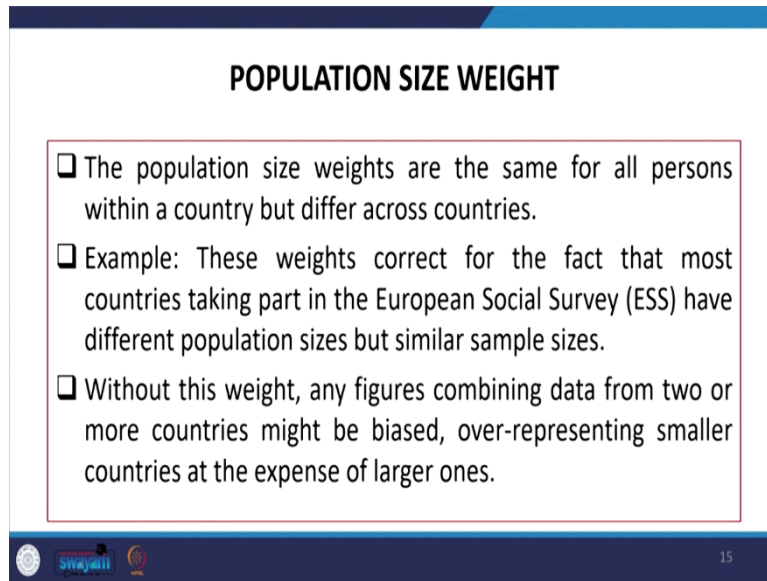
- The distribution of some variables in your sample population, such as sex, age, and education, may differ from the distribution in the actual population.
- For example,
  - Sample → 58% men
  - Population → 48% men
- Post-stratification weighting** is done in order to achieve a distribution equal with that of such known characteristics of the population.

14

Another approach is a post-stratification weight, which is the distribution of some variables in your sample population such as sex, age, and education, which may differ from the distribution in the actual population. For example, your sample is 58 percent men and the population is 48 percent of men, but in the sample, you have got 58 percent of men, and the population is actually 48 percent.

In the post-stratification, weighting is basically done in order to achieve a distribution equal to that of such known characteristics of the population. This proportion which you said might almost resemble your sample estimation. So, in the post-stratification, some of the ratios are going to be different and could also be compensated as well.

(Refer Slide Time: 17:42)



**POPULATION SIZE WEIGHT**

- ❑ The population size weights are the same for all persons within a country but differ across countries.
- ❑ Example: These weights correct for the fact that most countries taking part in the European Social Survey (ESS) have different population sizes but similar sample sizes.
- ❑ Without this weight, any figures combining data from two or more countries might be biased, over-representing smaller countries at the expense of larger ones.

Swajathi 15


The next one is called the population size. What do you mean by population size? It is by the ratio of the population size. The population size weights are the same for all individuals within a country but differ across countries. In a particular case study, it is cited that these weights correct for the fact that most countries taking part in the European Social Survey ESS have different population sizes but similar sample sizes.

Without this weight, any figures combining data from 2 or more countries might be biased or over-representing smaller countries at the expense of the larger countries. So, the smaller countries should get a reciprocal weight because of the fact that the population size differs and there must be respective weights.

(Refer Slide Time: 18:59)

### COMBINED WEIGHT

- The data file may include several different types of weights for different purposes.
- Subsequently, they are combined into a final, combined weight.




In the last one, we are discussing combined weight. The data file may include several different types of weights for different purposes. Subsequently, they are combined into a final, and the combined weight is taken.

(Refer Slide Time: 19:12)

#### WEIGHTING USING DHS EXAMPLE DATA SET

- Sample weights included in the each DHS recode file.
- Example Data set link:  
[https://www.dhsprogram.com/files/Part\\_III\\_Weighting\\_DHS\\_Data\\_Stata.zip](https://www.dhsprogram.com/files/Part_III_Weighting_DHS_Data_Stata.zip)

Unit of Analysis	Weight Variable
Households (HR file)	hv005
Household members (PR file)	hv005
Women or children (IR, KR, BR file)	v005
Domestic violence (IR file)	d005
HIV test results (HA file)	hiv05
Men (MR file)	mv005



So, we will be discussing some of those in our explanation and how different databases are in fact utilizing these weights or did I come up with these specific weights.

As per the DHS data set, we have listed out those weight variables, and sample weights included in each DHS recode file. The DHS file and DHS decode file are different. They have actually processed it and included labeling with their refined form of data and also, they have recorded the variables.

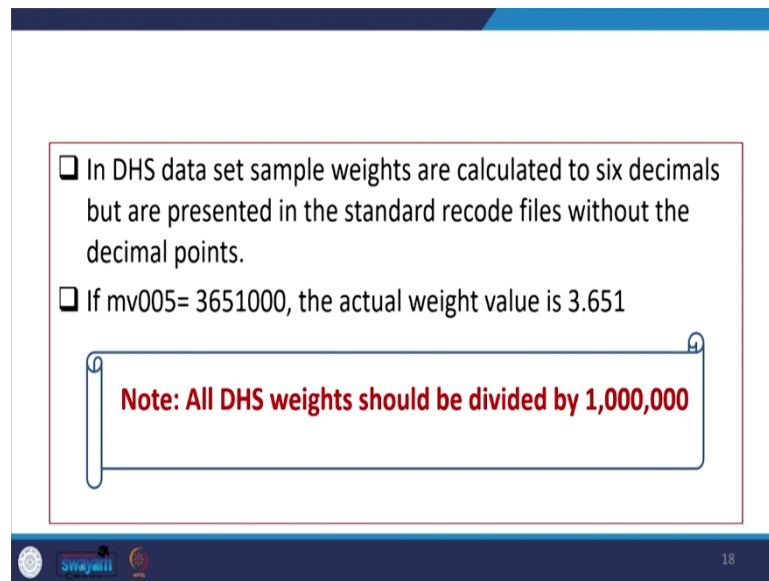
Now, you can also get this information through this link we have given just simply click on this you will be redirected to the respective page for details. Now, the weights in this DHS example data set like the household weight that is written as in the HR file and household recode file is hv005. Household members recode file is hv005.

Similarly, women and child and they carry individual recode. There are different ways of clarifying it. So, the weight variable and its name are listed accordingly. Interestingly if anybody is working on domestic violence, they are supposed to take the domestic violence weight.

For men file recode is different. So, if you are simply taking any individual weight that is v005 and just randomly considering this weight over here interchangeably this will be giving you an error and your data is not going to be represented at all.

So, after explaining all those things we will also help you to guide the right directions for interpreting the weight and then carrying in your own analysis. In the DHS data sample set weights are calculated to 6 decimals but are presented in standard recode files without the decimal points.

(Refer Slide Time: 21:51)



- ❑ In DHS data set sample weights are calculated to six decimals but are presented in the standard recode files without the decimal points.
- ❑ If mv005= 3651000, the actual weight value is 3.651

**Note: All DHS weights should be divided by 1,000,000**

swajati 18

So, you are supposed to take in the decimal point like the mv weight which is the men file. It gives a weight value of 365100. So, you have to divide it by 1000000 and if you divide it that will be your exact weight value. So, that you will be getting the right weight value and rest you can simply use your refined or newly generated weight variable.

We are going to guide you on how you guys can go about understanding the weights and apply them directly to your research. We will surely go through a sample data set and those sample data sets would also be provided to you along with this PPT lecture. So, the steps are mentioned here.

(Refer Slide Time: 23:21)

**STEPS TO WEIGHT (DHS DATA)**

1. Create the weight variable
2. Weight the data or Apply the weight
  - After creating the weight variable you have to include this variable while tabulations.

□ Here we will demonstrate the tabulation of percentage of women living in urban/rural residence

□ Stata Dataset-Example data set, i.e. **ZZIR61FL.dta**  
➤ We will use IR file because our unit of analysis is women

swayamii 19

Please follow it very carefully first we will create the weight variable that is the first step, and we are using the DHS data that is taken from NFHS 4.

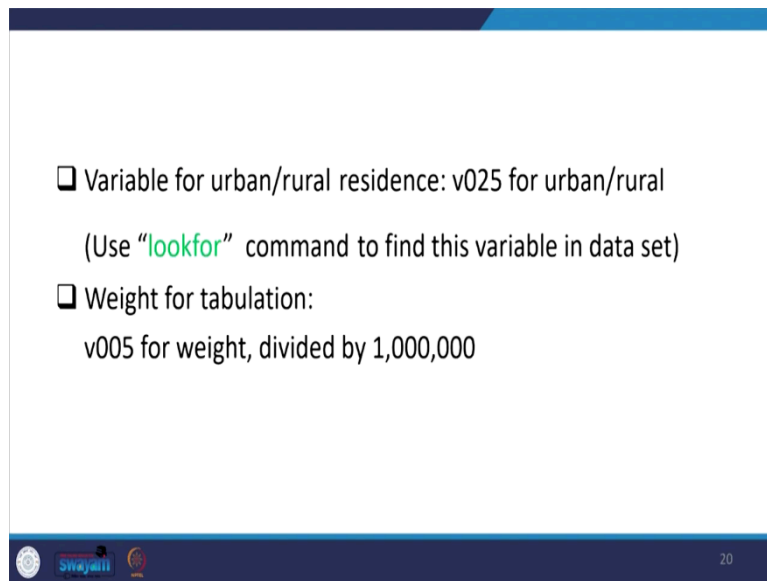
The second step is to weight the data or apply the weight. After creating the weight variable, you have to include this variable while tabulations. Here we will demonstrate the tabulation of percentages of women living in urban or rural resident. So, this is the particular hypothesis we are going to find out.

We will use the individual recode file because our unit of analysis is women. I have already guided you this individual recode file is an IR file.

So, individual recode files should be open not any other file and these are available in the NFHS data. I think we have already guided you about those different data sets on their respective page, where you guys can able to download them, and once you have downloaded them and how you should go for it.



(Refer Slide Time: 25:09)



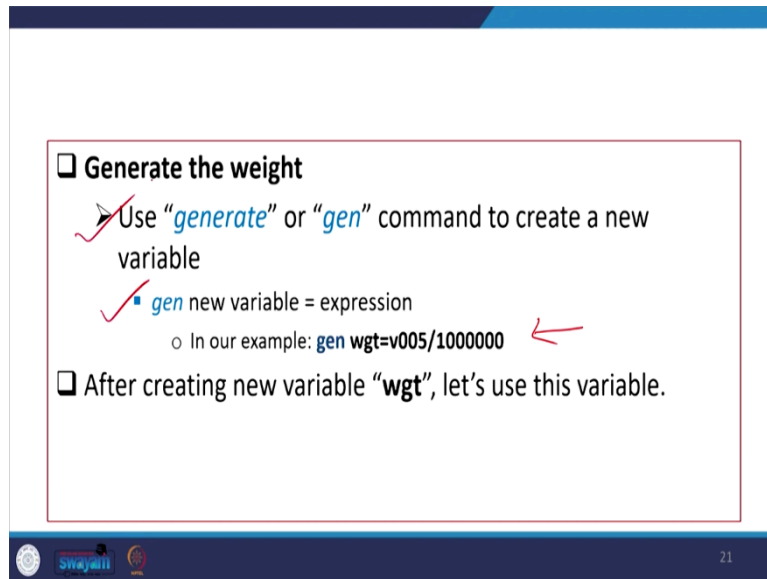
- ❑ Variable for urban/rural residence: v025 for urban/rural  
(Use “lookfor” command to find this variable in data set)
- ❑ Weight for tabulation:  
v005 for weight, divided by 1,000,000

Now onwards, we are giving you all experience of these techniques to go for the weights.

I have a variable for the urban or rural residents in this DHS file. This individual recode file is available with the name v025 for rural-urban with the categorical variable. Now, we will use the command loop if you are not able to tap the exact variable and if you are confused enough, you can simply type it and look for the search on the command page. You will get it.

Then the next step is to go for weight for the tabulation. The weight variable we have already mentioned that it is v005. I think I have already shown you this is v005 weight for that particular variable or for the data set. So, then we have to divide it by 1000,000 that will be generating the right weight for us, alright.

(Refer Slide Time: 26:20)



**Generate the weight**

- Use “*generate*” or “*gen*” command to create a new variable
- *gen* new variable = expression
  - In our example: `gen wgt=v005/1000000` ←

After creating new variable “**wgt**”, let’s use this variable.

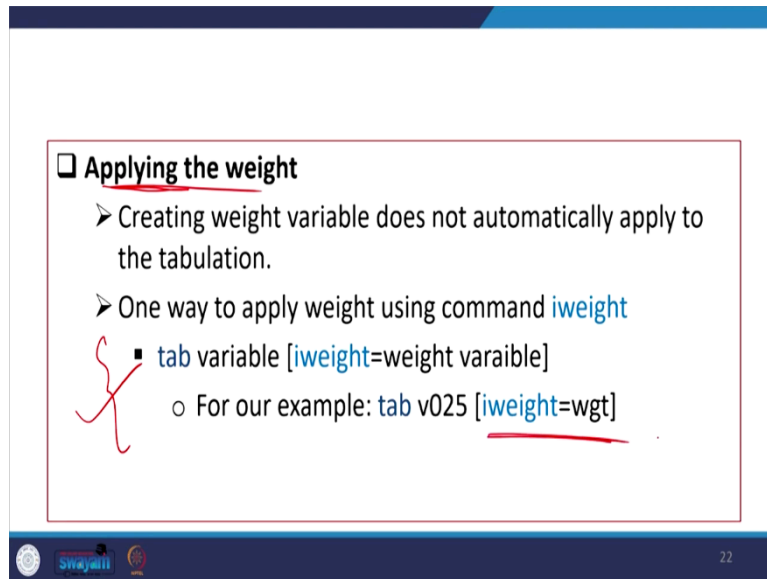
swayam 21

So, to generate the weight we must use the command generate or simply insert a gen command to create a new variable. So, will be using stata I think I suggested to you get ready with stata in the last lecture.

We are soon going to start the operation with the help of stata. One lecture will also guide you about starting with SPSS but not with all applications. Mostly will be operating through stata.

So, the steps you must do it very carefully in the command window. I will show the command window. You should generate the variable as given in the example. Then the variable name we have created as weight is equal to the weight variable divided by 1000000. This is what we have done in this example data set we are going to do it now.

(Refer Slide Time: 27:40)



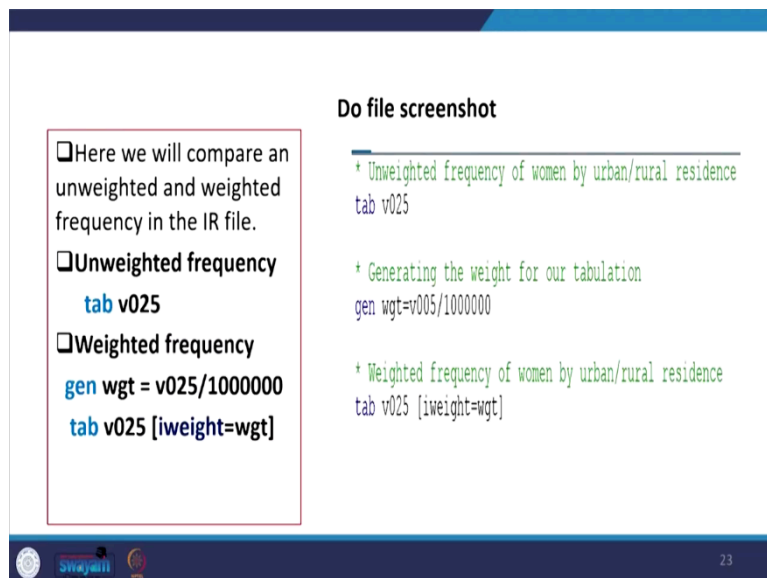
**□ Applying the weight**

- Creating weight variable does not automatically apply to the tabulation.
- One way to apply weight using command `iweight`
  - `tab variable [iweight=weight variable]`
    - For our example: `tab v025 [iweight=wgt]`

22

So, applying the weight we can do that like this.

(Refer Slide Time: 27:46)



**□ Here we will compare an unweighted and weighted frequency in the IR file.**

**□ Unweighted frequency**  
`tab v025`

**□ Weighted frequency**  
`gen wgt = v025/1000000`  
`tab v025 [iweight=wgt]`

**Do file screenshot**

```
* Unweighted frequency of women by urban/rural residence
tab v025

* Generating the weight for our tabulation
gen wgt=v025/1000000

* Weighted frequency of women by urban/rural residence
tab v025 [iweight=wgt]
```

23

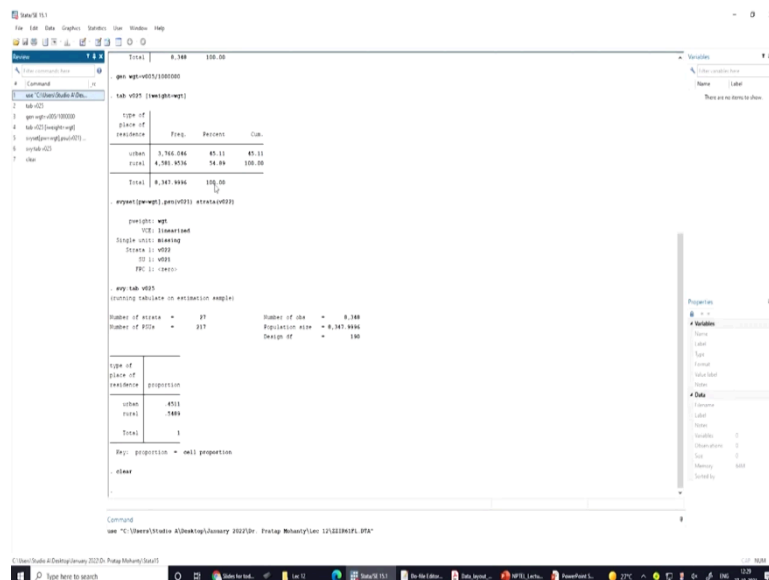
Creating a weight variable does not automatically apply to the tabulation we have to apply it carefully. So, with the command, `iweight` is equal to the weight variable and that variable has to be defined.

Then we can understand the tabulation like how it looks like and what kind of percentage it gives in our example set.

Now, after saying all those things you must be very careful about this command this is what the command I have guided. So, here we have applied the weight exactly as per your data. We have presented the do file screenshot for you, and we are also providing the do file to you.

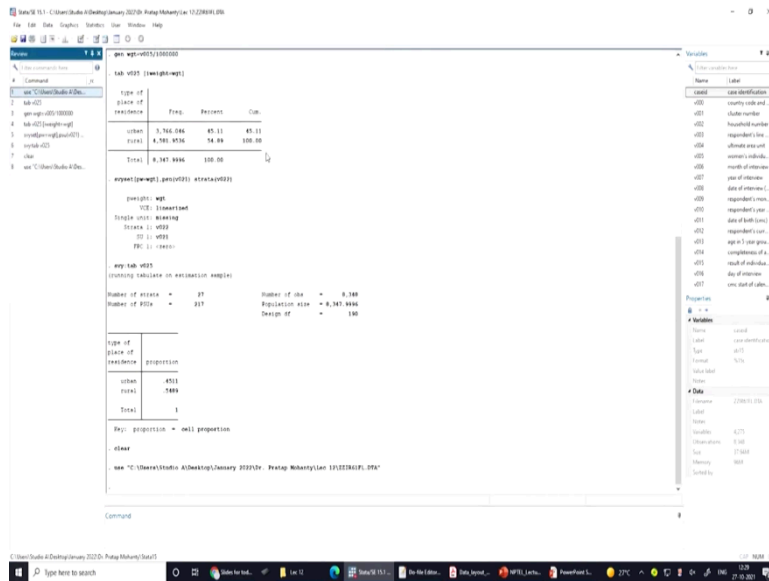
Here we will compare an unweighted and weighted frequency in the individual recode file. The unweighted frequency will map through tab v025 and the weighted frequency with this new variable that is generated and accordingly we will take the tab of this with its weight. So, these are all let us go by this and do some experiments.

(Refer Slide Time: 29:40)



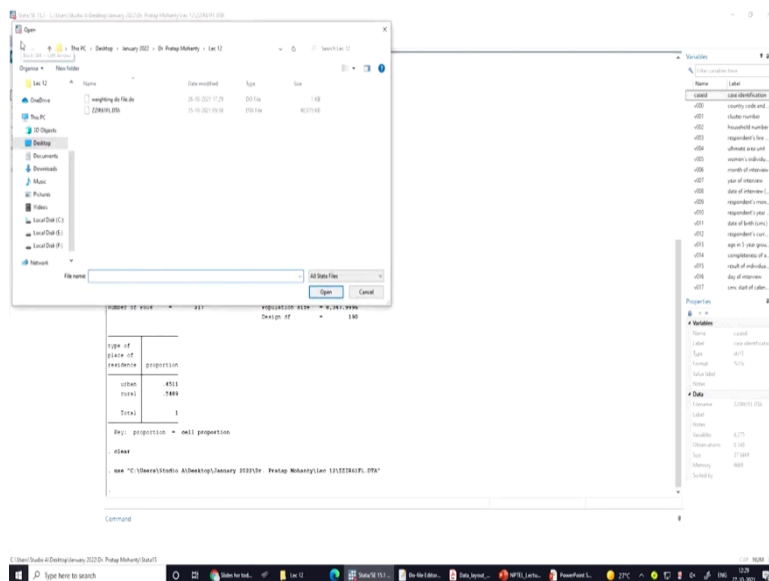
Here is the stata in front of us. So, I am going to just experiment with it and at this moment we do not have data on the screen. So, we are now opening the data.

(Refer Slide Time: 29:59)



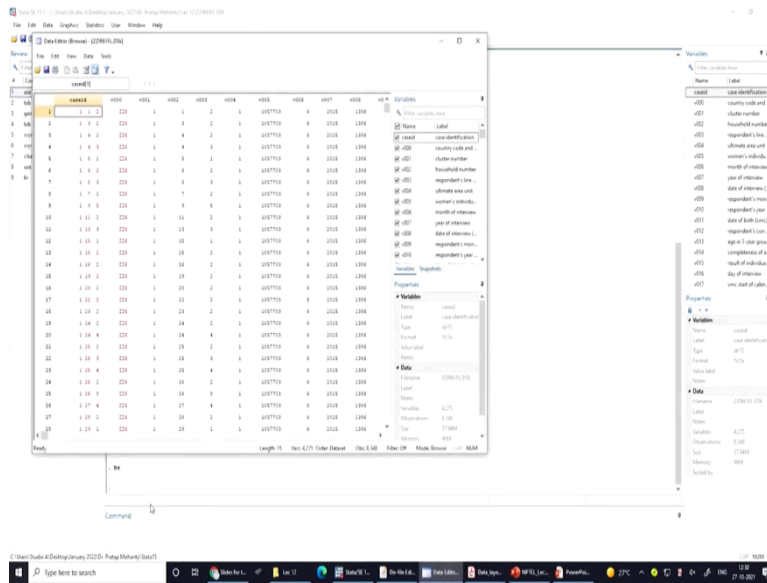
Here is the file I am just opening.

(Refer Slide Time: 30:00)



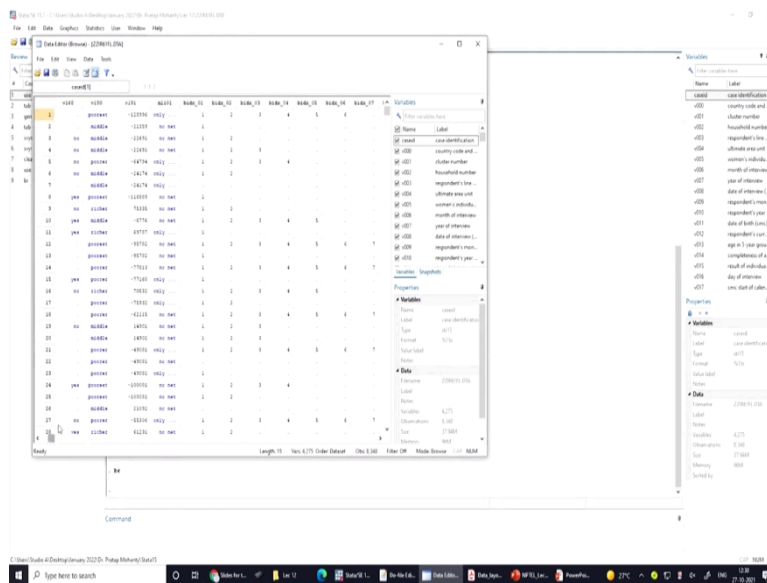
It is the file alright. We have already opened the data on the screen you can just show the data once by typing the br browse command on it.

(Refer Slide Time: 30:17)



So, this is the data sample data we have taken.

(Refer Slide Time: 30:22)

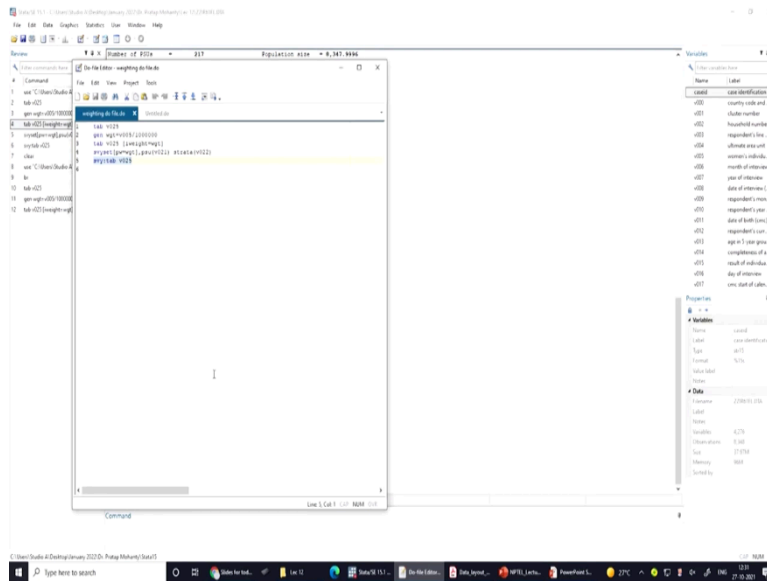


It has 4275 variables and observation is 8348. This is not the exact data of NFHS. We have taken sample data for our analysis to make the file a little shorter.

Now, we are going to show it step by step to go for weight. So, first, we are generating a weight here.



(Refer Slide Time: 31:29)



So, here is the do file and the screenshot of this one has already been shown to you. The tab of the unweighted variable is v025. Now, what weight we have scaled down to its decimal point the weight variable is v005 divided by 1000000.

The third row is the tabulation with the weight variable and then we have tried to compare. So, let us close this and now we can go for comparison, let me just go back to the PPT.

So, this is what is your weight and how you can operate. We have shown you a data file and this data file is also shared with you. The do-file screenshot also we have given, and you may also copy it on your own screen for further experiment.



(Refer Slide Time: 32:36)

### RESULTS IN STATA

#### Un-weighted frequencies

```
. tab v025
```

type of place of residence	Freq.	Percent	Cum.
urban	3,424	41.02	41.02
rural	4,924	58.98	100.00
Total	8,348	100.00	

#### Weighted Frequencies

```
. gen wgt= v005/1000000
. tab v025 [iweight=wgt]
```

type of place of residence	Freq.	Percent	Cum.
urban	3,766.046	45.11	45.11
rural	4,581.9536	54.89	100.00
Total	8,347.9996	100.00	

Now we are comparing the results in stata. Tabulation of the unweighted and weighted frequencies is presented here.

Now, before weight, you can easily see that your urban percentage out of the total is 41 percent as compared to rural. So, after weight, you can see that there is some adjustment or a 4-percentage increase in urban. So, it gives a certain weightage to the population where its size has been reduced due to the sampling procedure.

(Refer Slide Time: 33:25)

## DHS EXAMPLE REPORT

□ Note: Total weighted and unweighted numbers (i.e. at national level) will always be same because the weights have been normalized.

□ **iweight** can be used for frequencies and tables where you don't need significance testing or confidence intervals.

**Table 3.1 Background characteristics of respondents**  
Percent distribution of women and men age 15-49 by selected background characteristics, Model DHS 6 data

Background characteristic	Weighted percent	Women Weighted number	Unweighted number	Weighted percent	Men Weighted number	Unweighted number
<b>Age</b>						
15-19	23.5	1,958	2,041	23.4	771	760
20-24	17.1	1,428	1,371	15.8	523	516
25-29	17.0	1,419	1,357	15.2	502	496
30-34	13.2	1,100	1,110	12.0	397	375
35-39	13.2	1,099	1,108	14.7	485	476
40-44	7.9	660	644	10.1	332	350
45-49	8.2	684	717	8.8	290	311
<b>Religion</b>						
Religion 1	22.9	1,915	1,828	21.6	711	664
Religion 2	76.4	6,379	6,464	78.0	2,573	2,607
Religion 3	0.3	23	27	0.3	10	9
Religion 4	0.0	2	1	0.0	0	1
No religion	0.0	4	4	0.0	1	1
Other	0.0	0	0	0.1	4	2
<b>Ethnic group</b>						
Ethnic group 1	23.1	1,930	1,838	23.7	781	736
Ethnic group 2	33.0	2,588	2,688	29.3	985	986
Ethnic group 3	4.4	367	498	4.8	157	204
Ethnic group 4	35.4	2,956	2,768	36.4	1,202	1,135
Ethnic group 5	5.4	453	500	4.9	163	201
Other	0.4	34	39	0.8	26	16
Missing	0.2	20	17	0.2	6	6
<b>Marital status</b>						
Never married	31.0	2,585	2,488	45.4	1,497	1,420
Married	59.9	5,000	5,130	47.0	1,549	1,620
Living together	3.3	272	250	4.4	146	109
Divorced/separated	3.4	284	271	3.0	98	98
Widowed	2.5	206	209	0.3	9	7
<b>Residence</b>						
Urban	45.1	3,766	3,824	49.4	1,631	1,412
Rural	54.9	4,582	4,524	50.6	1,668	1,872
Total 15-49	100.0	8,348	8,348	100.0	3,299	3,284
50-59	na	na	na	na	322	337
Total 15-59	na	na	na	na	3,621	3,621

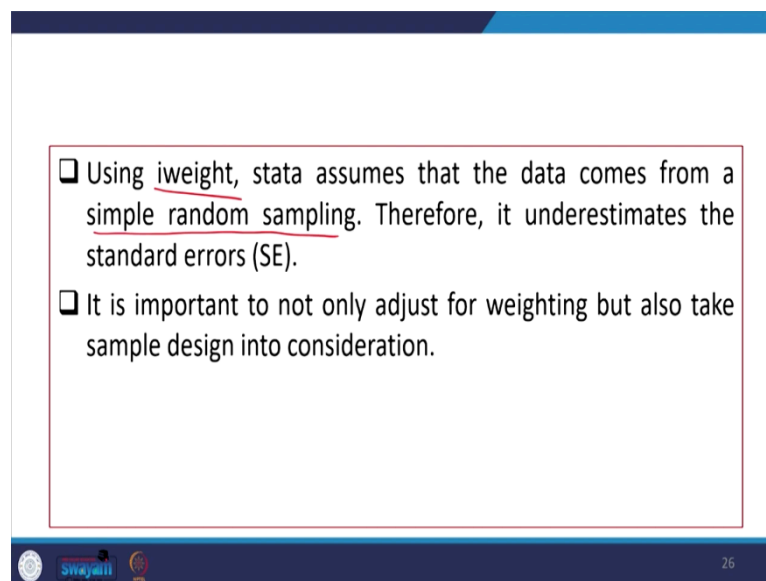
Note: Education categories refer to the highest level of education attended, whether or not that level was completed.  
na = Not applicable.

So, this is one of the interesting aspects we have shown to you, and you can also follow these aspects from the DHS example report. The total weighted and unweighted numbers will always be the same because the weights have been normalized as well.

So, we will also discuss the normalization technique in the next lecture. The `iweight` can be used for frequencies and tables where you do not need significance testing for the confidence interval. Especially without significance testing `iweight` can be easily taken and should be taken in fact, to represent your data carefully.

So, here is the data which we have shown to you for the eligible women from 15 to 49 is provided in the data set. The total women weighted number is 8348 and the unweighted number is all of these. The weighted number after weight it is getting different. So, it is getting a certain adjustment to the data this is what has proved similarly in other categories you can also compare and then find it out.

(Refer Slide Time: 34:44)



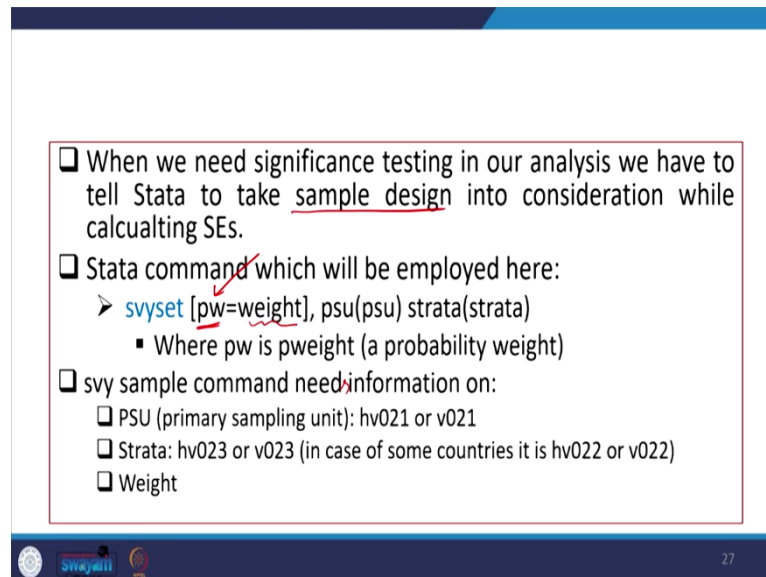
Like using `iweights`, stata assumes that the data comes from a simple random sampling. So, that is one of the interesting aspects when you are using `iweight` with the assumption that is followed by simple random sampling. You should take note of it.

Therefore, it underestimates the standard errors. So, when simple random sampling is taken then there is a high chance of underestimation. If it is simple random sampling, then your

sample size should not be very less. You have followed a sampling procedure and you have taken a sample it has to be lesser. So, therefore, it underestimates the total population size.

It is important to not only adjust for weighting but also take sample design into consideration.

(Refer Slide Time: 35:40)



- ❑ When we need significance testing in our analysis we have to tell Stata to take sample design into consideration while calculating SEs.
- ❑ Stata command which will be employed here:
  - `svyset [pw=weight], psu(psu) strata(strata)`
    - Where pw is pweight (a probability weight)
- ❑ svy sample command need information on:
  - ❑ PSU (primary sampling unit): hv021 or v021
  - ❑ Strata: hv023 or v023 (in case of some countries it is hv022 or v022)
  - ❑ Weight

When we need significance testing in our analysis, we have to tell stata to take the sample design into consideration while calculating standard errors. So, standard errors might be high if you do not consider the sample design weight.

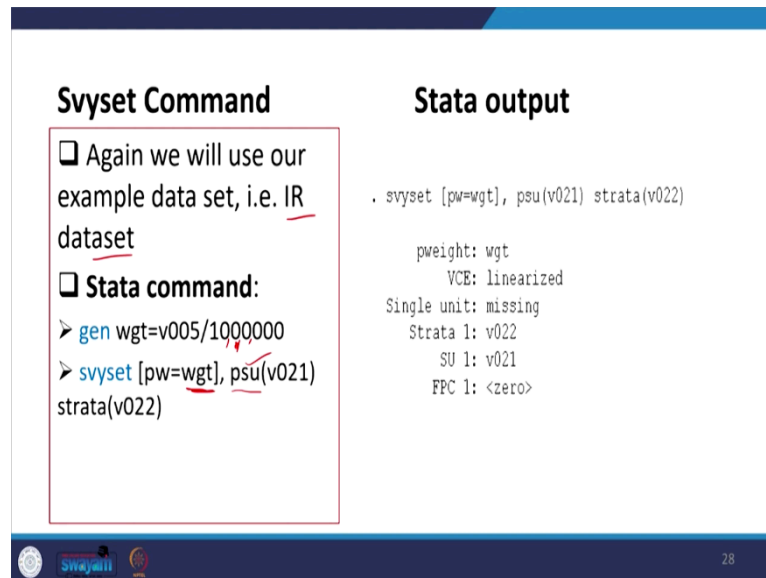
If you simply go by iweight it might not correctly reflect your result it might give you underestimation. Most of the larger databases come with complex sample designs and it must be set data with the sample design command.

To say stata that my data follow a sample design, we have to go for this command. So, the stata command which will be employed here is svyset within bracket pw weight.

So, pw weight is actually important this is your weight variable and pw is in fact your command. Then, followed with primary sampling unit has to be specified and the strata exactly should also be specified. The pw is your p weight, which is also called probability weight, it also gives probability assignment.

The svy sample command needs information such as the primary sampling unit and, in your data, whether it is hv021 or v021. We are going to clarify strata and it is hv022 or v022 as for the DHS data. So, a weight variable is also required.

(Refer Slide Time: 38:21)



**Svyset Command**

- ❑ Again we will use our example data set, i.e. IR dataset
- ❑ **Stata command:**
  - `gen wgt=v005/1000000`
  - `svyset [pw=wgt], psu(v021) strata(v022)`

**Stata output**

```
. svyset [pw=wgt], psu(v021) strata(v022)

pweight: wgt
VCE: linearized
Single unit: missing
Strata 1: v022
SU 1: v021
FPC 1: <zero>
```

28

Again, we will use our sample data set and that is the individual recode data set that we have just shown to you for eligible women. The stata command that we are going to apply is generate weight is equal to v005 dividing it by 1000000. So, the svyset command has to be applied like this p weight and the weight variable we have already generated in our data.

Now, the primary sampling unit is also defined with v021. The primary sampling unit is clearly defined in the report and the strata variable has to be defined for the instructions.

(Refer Slide Time: 39:38)

❑ After running svyset line, now you can run the tabulation

❑ svy: tab v025

```
. svy: tab v025
(running tabulate on estimation sample)

Number of strata = 27      Number of obs = 8,348
Number of PSUs  = 217    Population size = 8,347.9996
                                   Design df      = 190
```

type of place of residence	proportion
urban	.4511
rural	.5489
Total	1

Key: proportion = cell proportion

29

So, here is the starta output. We are once given this command and we get these things on our screen. So, we are also going to experiment just now.

(Refer Slide Time: 39:46)

❑ iweight should be used for simple weighted frequencies. ←

❑ On the other hand, if you need to calculate SEs, p-values, or confidence interval you should use svy command

30

(Refer Slide Time: 39:47)

### STATA COMMAND FOR WEIGHT IN NSS

- ❑ `gen weight=.`
  - This command will generate the empty column with name weight in the dataset.
- ❑ `replace weight= MLT/200 if NSS!=NSC`
  - This command will put the calculated weight by dividing MLT by 200 for those records where NSS and NSC are different.
- ❑ `replace weight=MLT/100 if NSS==NSC`
  - This command will put the calculated weight by dividing MLT by 100 for those records where NSS and NSC are same.

31

So, if you open it, we can do that experiment as well. So, the weight variable we have already defined is not it. So, here is the command.

(Refer Slide Time: 40:09)

The screenshot shows the Stata command window with the following commands and output:

```
use "C:\Users\Student\Documents\Sem5\2020-21\Pratap Mahaling\ Lec 13\1328431.DAT"
clear
key: proportion = well proportion
use "C:\Users\Student\Documents\Sem5\2020-21\Pratap Mahaling\ Lec 13\1328431.DAT"
clear
tab v020
type of |
place of |
residence |
Freq. Percent Cum.
-----+-----
urban    3,434   41.59  41.59
rural    4,766   58.41  100.00
Total    8,200   100.00

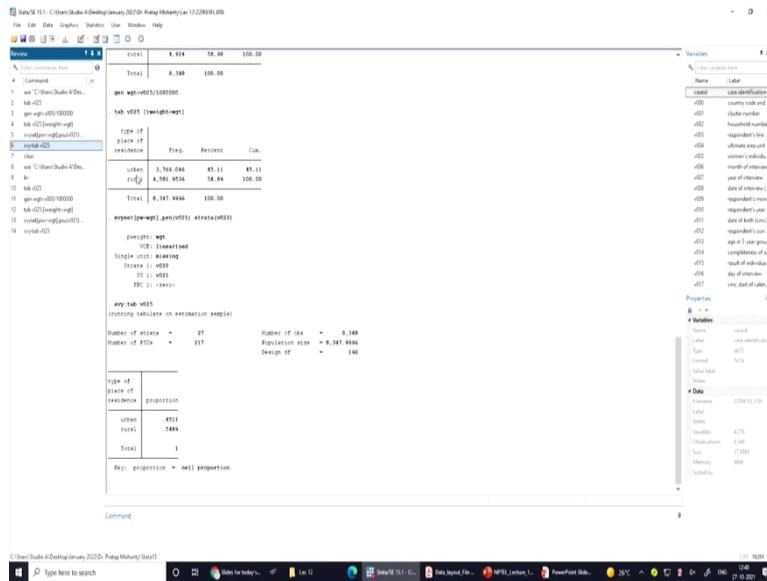
gen wgt=mlt/100000
tab v020 [(wgt=mlt)]
type of |
place of |
residence |
Freq. Percent Cum.
-----+-----
urban    3,766   45.93  45.93
rural    4,434   54.07  100.00
Total    8,200   100.00

svyset [(wgt)]_psu(v020) strata(v020)
dswgt01: wgt
dsc: 10000000
Single v020: MISSING
Strata 1: v020
dsc 1: v020
PFC 1: -resid-
```

The output shows the distribution of the 'wgt' variable across different residence types (urban and rural) and places (type of and place of residence). The total number of observations is 8,200.

The command is `svyset pw weight` and the `psu` is defined and then its strata variable is given. Now Stata has just given you this output now we are simply going to get the result alright in front of you.

(Refer Slide Time: 40:33)



So, now you can find out the proportions of rural and urban. So, this gives in proportion, but you have in your interpretation it has to be in percentages. It should be 45 percent for the urban and for rural it should be 54 percent.

So, now you can understand that your design weight is coming out with a different proportion whereas, in the earlier case it was of a different percentage. So, I am just coming to our PPT then we will clarify further details. We have also given you the command to compare the result after the weight.

Your tabulation should start with svy instead of just tab is not going to give you the result. So, iweight should be used for simple weighted frequencies that are especially for simple random sampling techniques whenever is applied.

So, p weight is applied when you have a design weight. So, especially it helps in giving the right standard error, its p values, confident interval etc. For the last couple of guidance in another 2-3 minutes, I am going to just clarify the rest of the details we will experiment with in other slides as well. I am just referring to the NSS 75<sup>th</sup> and giving you one document in front of your screen here.

(Refer Slide Time: 42:32)

Apply final weight for Sub-sample wise estimates as follows:  
Final Weight =  $MLT/100$

---

For generating Sub-sample-combined estimates based on data of all sub-rounds taken together, both Sub-sample-1 FSU's and Sub-sample-2 FSU's are to be considered.  
Apply final weight for Sub-sample combined estimates as follows:  
Final weight =  $MLT/100$ , if  $NSS=NSC$   
=  $MLT/200$  otherwise.

9. Common Primary Key for identification of a record for any schedule is:  
FSU Serial Number =  $d(5)$  (i.e. offset = 4th byte, length = 5 bytes)

(Refer Slide Time: 42:40)

Government of India  
Data Quality Assurance Division  
National Statistics Office  
164, Gopal Lal Thakur Road, Kolkata-108.  
Phone No. 2577-1128

-----  
*NSS 75<sup>th</sup> Round*  
*Final Multiplier-posted unit-level data*  
*for Schedule- 25.0 of NSS 75<sup>th</sup> round*

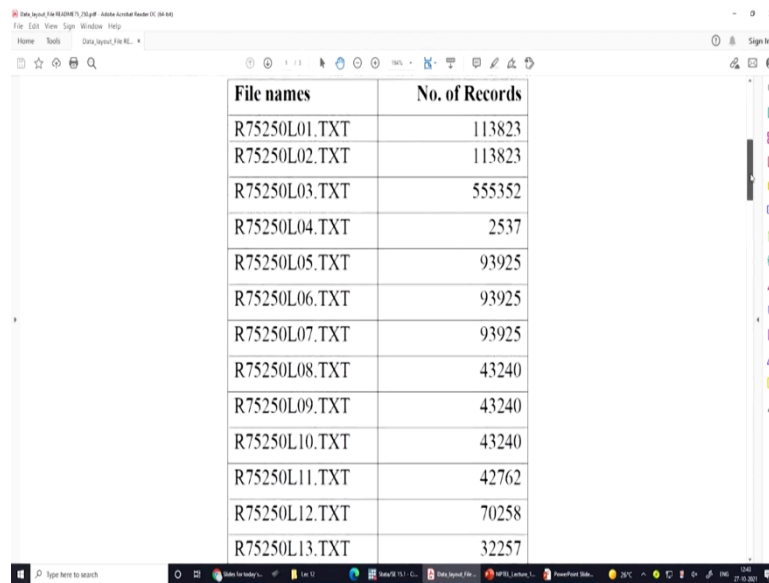
A) Data for Sch. 25.0 (Social Consumption: Health).  
There are 13 data files belonging to 13 different levels as per layout (datalay75\_250.XLS).

File names	No. of Records
...	...

I have already shown the readme file of the 75th round.



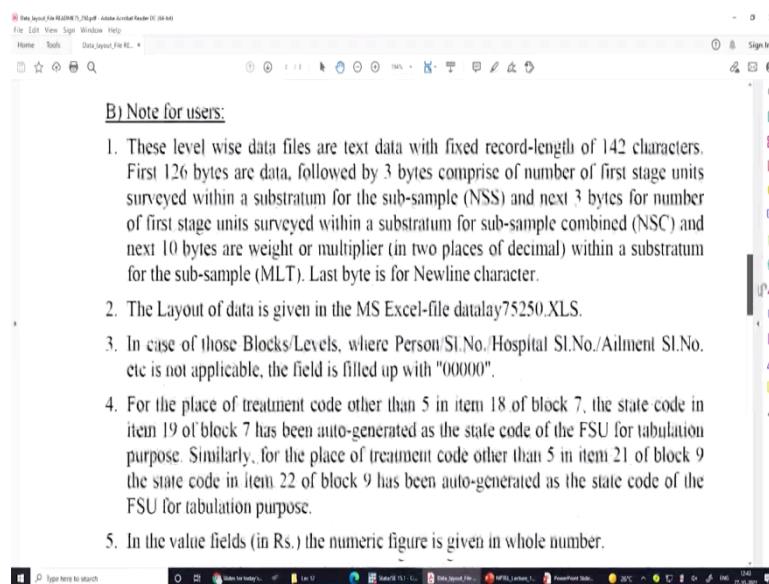
(Refer Slide Time: 42:44)



File names	No. of Records
R75250L01.TXT	113823
R75250L02.TXT	113823
R75250L03.TXT	555352
R75250L04.TXT	2537
R75250L05.TXT	93925
R75250L06.TXT	93925
R75250L07.TXT	93925
R75250L08.TXT	43240
R75250L09.TXT	43240
R75250L10.TXT	43240
R75250L11.TXT	42762
R75250L12.TXT	70258
R75250L13.TXT	32257

This suggests the number of records and especially regarding weight.

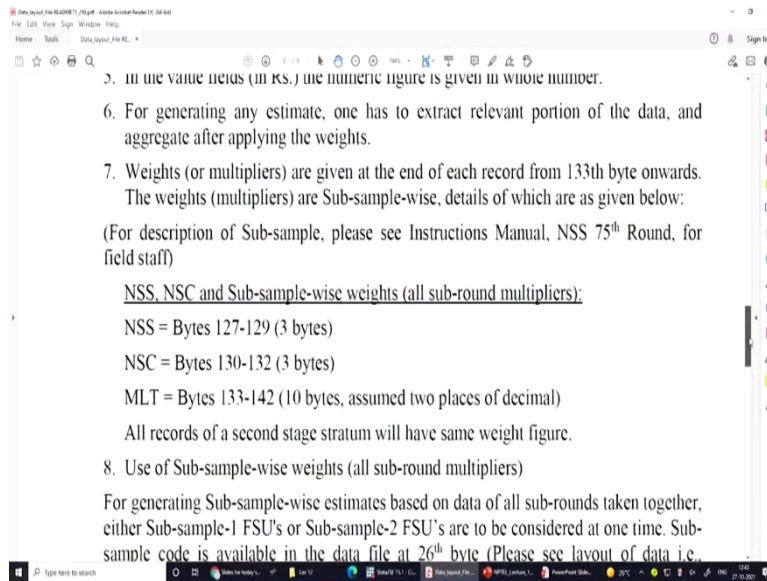
(Refer Slide Time: 42:49)



**B) Note for users:**

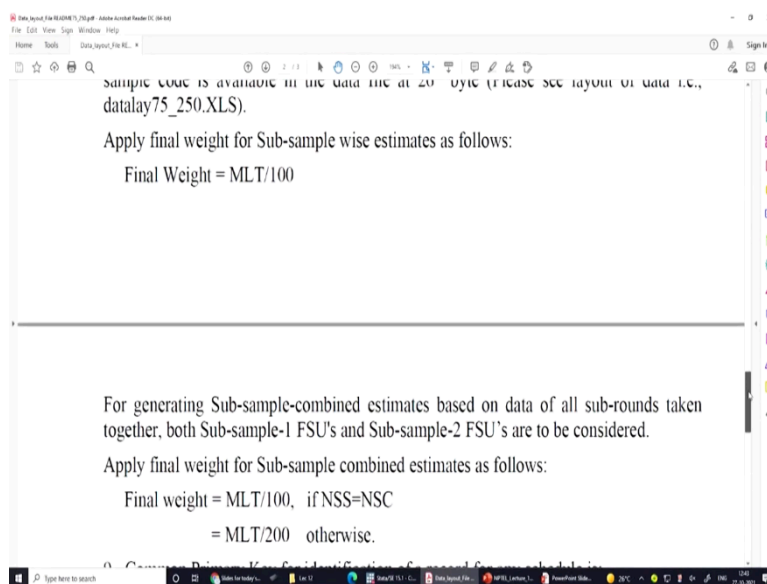
1. These level wise data files are text data with fixed record-length of 142 characters. First 126 bytes are data, followed by 3 bytes comprise of number of first stage units surveyed within a substratum for the sub-sample (NSS) and next 3 bytes for number of first stage units surveyed within a substratum for sub-sample combined (NSC) and next 10 bytes are weight or multiplier (in two places of decimal) within a substratum for the sub-sample (MLT). Last byte is for Newline character.
2. The Layout of data is given in the MS Excel-file datalay75250.XLS.
3. In case of those Blocks/Levels, where Person SI.No./Hospital SI.No./Ailment SI.No. etc is not applicable, the field is filled up with "00000".
4. For the place of treatment code other than 5 in item 18 of block 7, the state code in item 19 of block 7 has been auto-generated as the state code of the FSU for tabulation purpose. Similarly, for the place of treatment code other than 5 in item 21 of block 9 the state code in item 22 of block 9 has been auto-generated as the state code of the FSU for tabulation purpose.
5. In the value fields (in Rs.) the numeric figure is given in whole number.

(Refer Slide Time: 42:47)



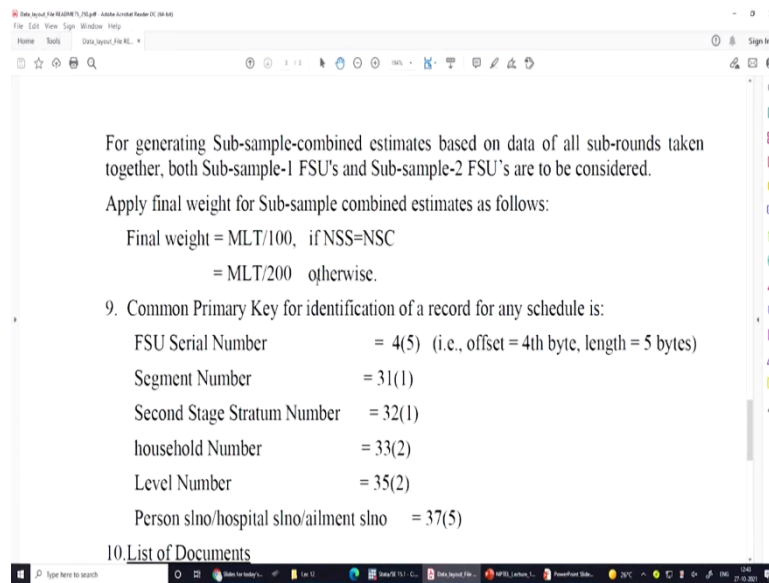
It has guided very clearly which kind of weight should be defined.

(Refer Slide Time: 42:53)



Usually, we go by the multiplier as the variable for the weight, but the final weight should not be just a multiplier it has to be with your certain steps.

(Refer Slide Time: 43:05)



For generating Sub-sample-combined estimates based on data of all sub-rounds taken together, both Sub-sample-1 FSU's and Sub-sample-2 FSU's are to be considered.

Apply final weight for Sub-sample combined estimates as follows:

$$\text{Final weight} = \text{MLT}/100, \text{ if } \text{NSS}=\text{NSC}$$
$$= \text{MLT}/200 \text{ otherwise.}$$

9. Common Primary Key for identification of a record for any schedule is:

FSU Serial Number	= 4(5)	(i.e., offset = 4th byte, length = 5 bytes)
Segment Number	= 31(1)	
Second Stage Stratum Number	= 32(1)	
household Number	= 33(2)	
Level Number	= 35(2)	
Person slno/hospital slno/ailment slno	= 37(5)	

10. List of Documents

The final weight is an MLT by 100 if your sub-sample is equal to the sample combined, otherwise, divide MLT by 200.

So, we are just doing it for your understanding. So, these things will also discuss in other slides as well. I have simply said you have to generate a weight equal to a dot, that dot must be replaced with this command. So, likewise, we generate a new variable.

So, similarly, we are generating a new variable called weight with a dot, and that dot is filled with this information. So, the MLT variable is already available in NSS. I am not experimenting with this at this moment, but I will just guide you to MLT by 200 if NSC is not equal to NSS.

So, stata gives this information an exclamatory mark equal to NSC. If they are equal, then you have to put double equal to and it will generate the right weight.

(Refer Slide Time: 44:38)

**REVIEW OF COMMAND**

- ❑ **Generating weight variable**
  - **gen** new variable = expression
  
- ❑ **Applying weight**
  - **tab** variable [iweight=weight variable]
  - **svyset**[pw=weight], psu(psu) strata(strata)
  - **svy: tab** variable

The review of commands we have already discussed today. We have discussed generating weight; its expression has to be given. In general studies, we have to go by the tabulating variable which is the iweight variable. In the case of design weight, we are supposed to give with the svy set command, then we can tabulate that variable and we will get the right result for our study.

I think these are all guidance so far and we will be repeatedly using them in our coming lectures it will consistently help you to understand the right analysis of the data. I think these are fine at this moment we have probably exceeded a bit of our time. So, we will commence in the next class with this I think I am closing this session here.

Thank you.