

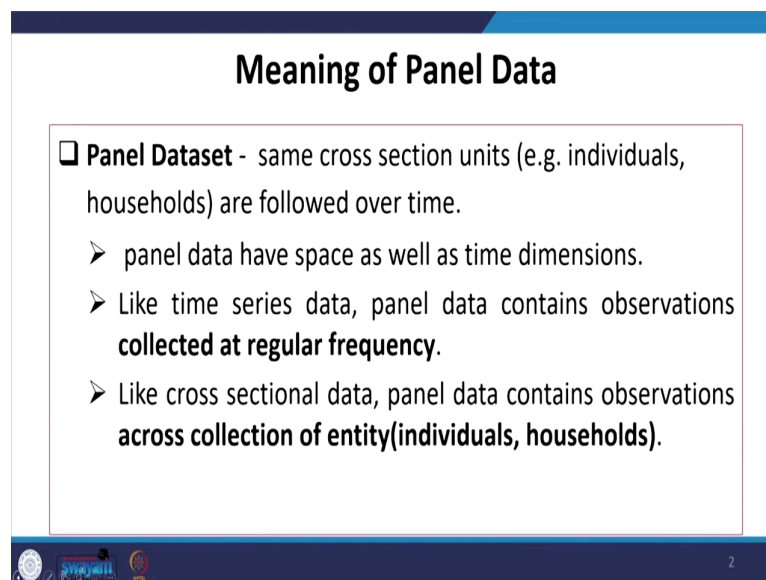
**Exploring Survey Data on Health Care**  
**Prof. Pratap C. Mohanty**  
**Department of Humanities and Social Sciences**  
**Indian Institute of Technology, Roorkee**

**Lecture - 31**  
**Balanced and Unbalanced Panel Data in Healthcare**

Welcome friends, to this NPTEL MOOC program on exploring healthcare survey data. We are in the 7th week, trying to explore the healthcare data and their panel information; I am Pratap Mohanty, at present teaching in IIT Roorkee. This particular lecture is targeted to deal with the panel data of balanced or unbalanced kind; how balanced panel and unbalanced panel is understood in healthcare data.

We will give you the practical handouts to understand the panel data in healthcare. Now, I am clarifying the meaning of panel data; panel data is basically a mixture of cross section and time series. As we have already explained cross section then as compared to the time series, this is basically repeated cross section data and it is also called same cross section units followed over time. Panel data set could be at the individual level and also could be at the household level.

(Refer Slide Time: 01:45)



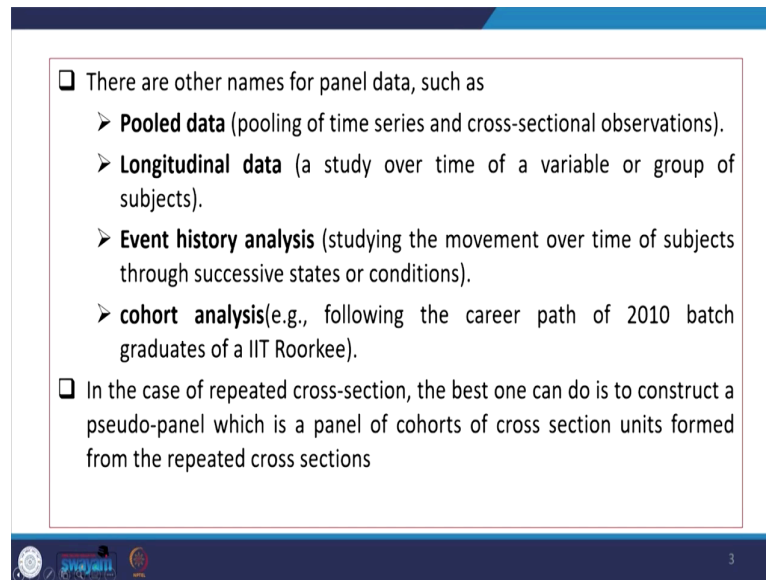
**Meaning of Panel Data**

- ❑ **Panel Dataset** - same cross section units (e.g. individuals, households) are followed over time.
  - panel data have space as well as time dimensions.
  - Like time series data, panel data contains observations **collected at regular frequency**.
  - Like cross sectional data, panel data contains observations **across collection of entity(individuals, households)**.

2

Panel data have space as well as time dimensions. Like time series data, panel data also contains observations collected at regular frequency. Like cross section data, panel contents observation across collections of entity such as individuals and households.

(Refer Slide Time: 02:06)



- ❑ There are other names for panel data, such as
  - **Pooled data** (pooling of time series and cross-sectional observations).
  - **Longitudinal data** (a study over time of a variable or group of subjects).
  - **Event history analysis** (studying the movement over time of subjects through successive states or conditions).
  - **cohort analysis**(e.g., following the career path of 2010 batch graduates of a IIT Roorkee).
- ❑ In the case of repeated cross-section, the best one can do is to construct a pseudo-panel which is a panel of cohorts of cross section units formed from the repeated cross sections

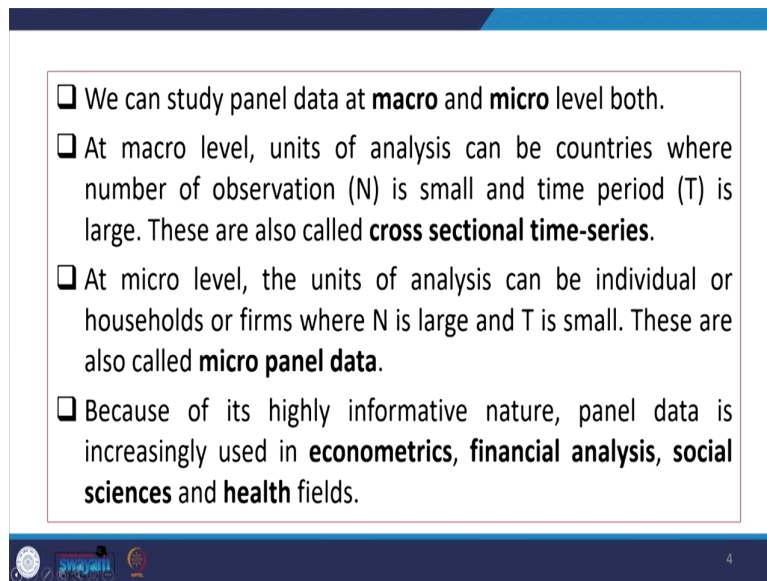
There are other names of panel data, such as pooled data; like pooling of time series and cross-sectional observations, where the same unit of observation may not be repeated over time, let us say simply pooled. Whereas, in case a longitudinal data, a study over time of a variable or a group of subjects. Another one is called even history analysis; we are studying the movement over time of subjects through successive states or conditions.

Another aspect or type of panel data is called cohort analysis; Suppose in this case, we follow the career path of 2010 graduates of IIT Roorkee for example. So, that batch- how it is reflected in different successive periods; when we are catching our analysis for a particular batch, particular cohort, that type of analysis is called cohort analysis.

In the case of repeated cross section, the best one can do is to construct a pseudo panel, which is a panel of cohorts of cross section units formed from the repeated cross sections.

We will have also a dedicate session called pseudo panel; we will be discussing a bit on how to form it, we will also give some example dataset for your better understanding.

(Refer Slide Time: 03:40)



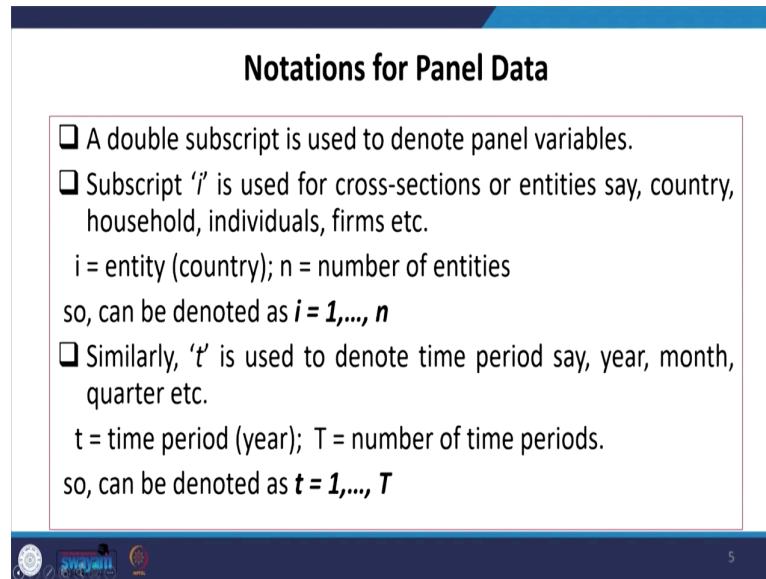
- ❑ We can study panel data at **macro** and **micro** level both.
- ❑ At macro level, units of analysis can be countries where number of observation (N) is small and time period (T) is large. These are also called **cross sectional time-series**.
- ❑ At micro level, the units of analysis can be individual or households or firms where N is large and T is small. These are also called **micro panel data**.
- ❑ Because of its highly informative nature, panel data is increasingly used in **econometrics, financial analysis, social sciences** and **health** fields.

We can study panel data at macro as well as micro level. At macro level, units of analysis can be countries, states etc. The number of observations is usually less and time period is large in case of macro panel. Over the time period, number of observations are available; but the time period horizon is quite large in macro context. These are also called cross sectional time series data.

At micro level, the units of analysis could be of individuals or households or firms, where N is large and T is small, T refers to time. These are called micro panel data.

So, now you can differentiate between macro panel and micro panel; because of this property, it is highly informative in nature, moreover panel data is increasingly used in econometrics, financial analysis, social sciences and health fields these days.

(Refer Slide Time: 04:48)



**Notations for Panel Data**

- ❑ A double subscript is used to denote panel variables.
- ❑ Subscript ' $i$ ' is used for cross-sections or entities say, country, household, individuals, firms etc.  
 $i$  = entity (country);  $n$  = number of entities  
so, can be denoted as  $i = 1, \dots, n$
- ❑ Similarly, ' $t$ ' is used to denote time period say, year, month, quarter etc.  
 $t$  = time period (year);  $T$  = number of time periods.  
so, can be denoted as  $t = 1, \dots, T$

5

There are different notations used for panel data, which is distinctive than that of the earlier conventional regression techniques or cross-sectional data.

In case of panel, a double subscript is used to denote panel variables. So, it is not just one subscript, it is 2 subscripts; if a model has used those, those are actually denoting panel variables. Subscript like  $i$  is used for cross sections or entities; let it be country or household, individuals etc., they are called cross-sectional subscript.  $i$  is your entity and it matters with how many entities are there, that is called  $n$ . So, it can be denoted as  $i$ , where  $i$  stands for 1 to  $n$ . So, maybe  $i$  varies from 1 to  $n$ , and that variation we will get it in panel. Similarly, another subscript that is used is called  $t$ ;  $t$  stands for time, time period over whether the data is available over month or over year or over quarters, accordingly the  $t$  dimension is also given along with  $i$  dimension. So,  $t$  may also vary from 1 to capital  $T$ , i.e., the  $T^{\text{th}}$  period we have mentioned.

(Refer Slide Time: 06:15)

□ So, this double subscript distinguishes cross sections and time series.

□ Panel data with  $k$  regressors can be represented as:  
 $(X_{1it}, X_{2it}, \dots, X_{kit}, Y_{it})$

Where,  
 $i = 1, 2, \dots, n$   
 $t = 1, 2, \dots, t$  }  $nt$

So, this double subscript distinguishes panel from cross sections and time series. In cross section we have only  $i$ ; whereas in time series we have only  $t$ , but in panel, we have both  $i$  and  $t$ . Therefore, the dimension of panel in total is of  $i$  into  $t$  or since  $i$  varies from  $i$  to  $n$  across times; so it should be  $n$  times capital  $T$ .

So,  $nT$  dimensions are actually included in panel data set. Panel data set with  $k$  regressors can be represented like the following, as we have presented here. So, like in cross sections, we used to have  $X_1, X_2, X_3, \dots, X_k$  and here we have  $X_{1it}$  to upto  $X_{kit}$  and at the end we have  $Y_{it}$ ,  $Y$  variable denoted as dependent variable. Here we are mentioning that it is not just  $X_{1i}$ , it is  $X_{1it}$ ;  $t$  component is mentioned, this is your  $t$  in addition to that.

Similarly, in the  $Y$  variable we have also added with the subscript called  $t$ ; says we have already pointed out that  $i$  varies from 1 to  $n$ , then  $t$  varies from 1 to  $t$ . So, total dimensions number of dimension will be  $n$  times  $t$ .

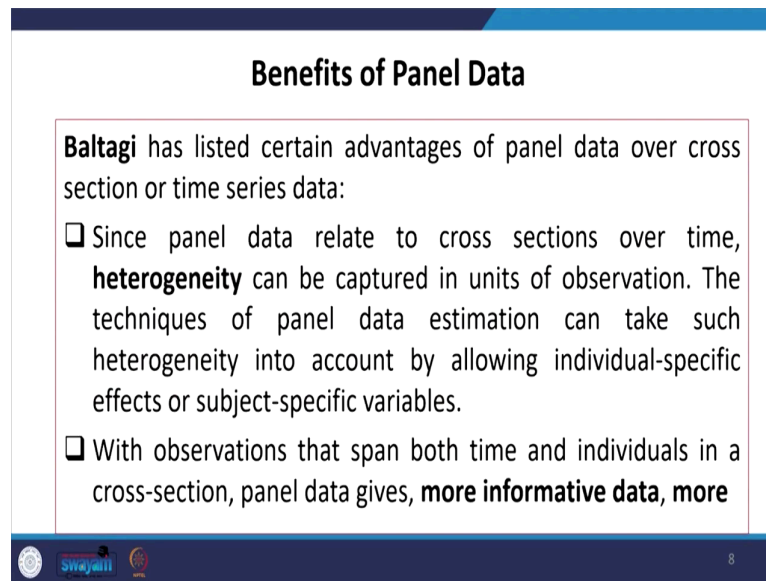
(Refer Slide Time: 07:40)

Entity	Time	Notation
1	1	$Y_{11}$
1	2	$Y_{12}$
1	T	$Y_{1T}$
...	...	...
N	1	$Y_{N1}$
N	2	$Y_{N2}$
N	T	$Y_{NT}$

Now, just look at an example on the screen for your clarity; here entity is given; time dimensions are given, the notations that is used is given.

Entity 1, it is still repeated here, then it is N entities. And then notations are given as if it is time period 2. It is available for the same entity and the same entities is also available till the T time period. Then the notations are r is equal to Y 1 1, since 1 is here and 1 is there (in first row); then this is 1 and 2 is here in second row, then 1 and T in third row. Similarly, if N is the entity, then N is represented in T time period; so, similarly N 1, N 2 and till N t is mentioned.

(Refer Slide Time: 08:40)



**Benefits of Panel Data**

**Baltagi** has listed certain advantages of panel data over cross section or time series data:

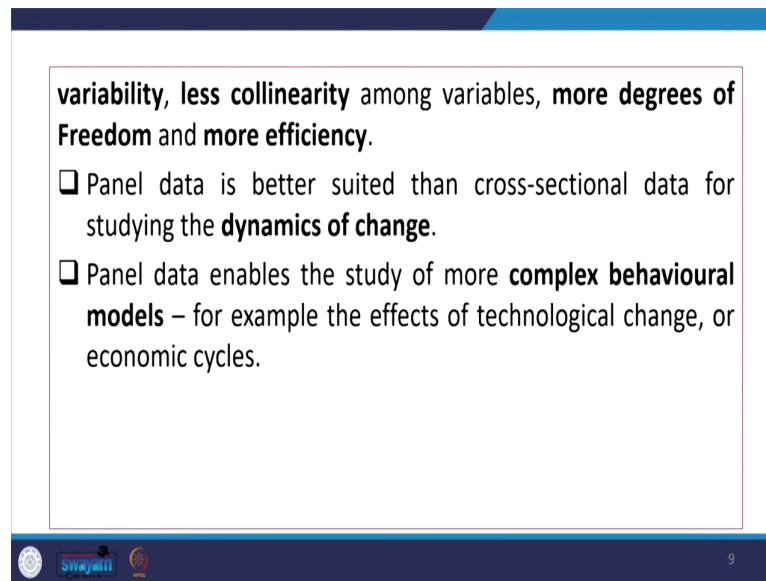
- ❑ Since panel data relate to cross sections over time, **heterogeneity** can be captured in units of observation. The techniques of panel data estimation can take such heterogeneity into account by allowing individual-specific effects or subject-specific variables.
- ❑ With observations that span both time and individuals in a cross-section, panel data gives, **more informative data, more**

swajali 8

Now, we are discussing about benefits of panel data, you will get better idea from the Baltagi's book. This has listed certain advantages of panel data over cross sections or time series data. The first important advantage is that, panel data relates to cross sections over time.

The heterogeneity can be also captured in units of observations, that is most important. The techniques of panel data estimation can take such heterogeneity into account by allowing individual specific effects or subject specific variables. Then the second advantage is that, with observations that span over time and individuals in a cross section, the panel data keeps more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency.

(Refer Slide Time: 09:31)



**variability, less collinearity** among variables, **more degrees of Freedom** and **more efficiency**.

- ❑ Panel data is better suited than cross-sectional data for studying the **dynamics of change**.
- ❑ Panel data enables the study of more **complex behavioural models** – for example the effects of technological change, or economic cycles.

Swajati 9

So, one explanation behind all those reasons is that, we are including more dimensions, more frequencies; since your  $t$  times component is multiplied with the  $n$  number of observations.

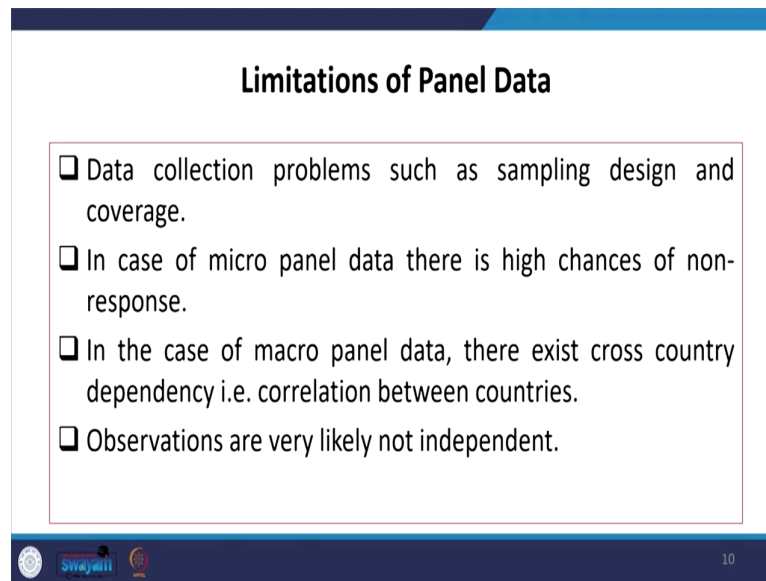
Or to the  $i$ th, “ $i$ ” which varies from 1 to  $n$ th observations i.e., again multiplied with  $t$ , so dimensions actually increase; therefore we will get better results with efficiency. Panel data is better suited than cross section data for studying the dynamics of change. Panel data enables the study of more complex behavioural models; for an example- the effects of technological change or economic cycle. Any sort of effects even in evaluation can also be dealt better in panel data.

Since it also reduces the individual heterogeneity as these are normalized over period of time and is usually present in cross sectional data.

Now, we are also identifying certain limitations of panel data; i.e., these data create some problems, such as framing your sample, then getting its coverage.



(Refer Slide Time: 11:00)



**Limitations of Panel Data**

- ❑ Data collection problems such as sampling design and coverage.
- ❑ In case of micro panel data there is high chances of non-response.
- ❑ In the case of macro panel data, there exist cross country dependency i.e. correlation between countries.
- ❑ Observations are very likely not independent.

swayam 10

In case of micro panel data, there is high chances of non-response in the next round; since you are covering more observations, non-responses would be there. We will also show one example data set from India's human development survey in comparison to the previous one and they have given a panel content.

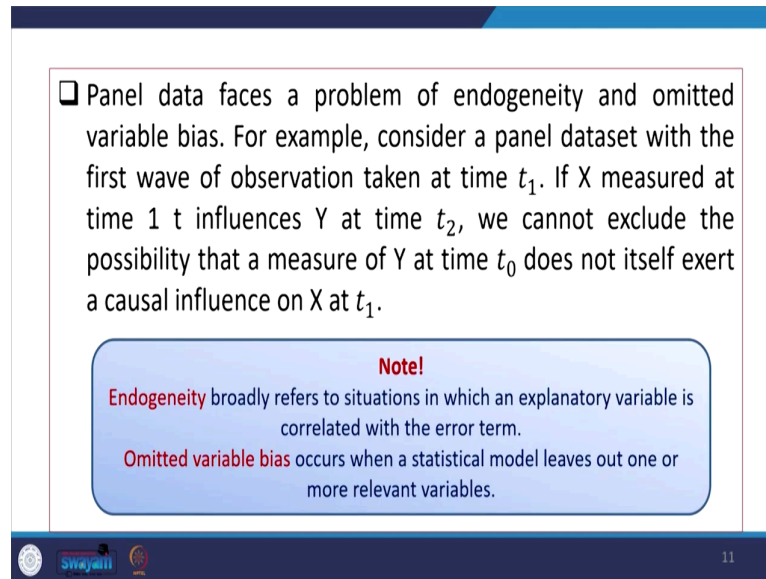
We will discuss that how many observations are actually missing from the previous to the new. And accordingly, the final panel is made and whether that is going to be balanced or unbalanced, we will discuss.

In the case of macro panel data, there exist cross country dependency such as correlation between countries. Since, cross country dependencies are always the case in international trade context; so, correlation between this country might exist, then that may create some problem in the panel explanation.

Observations are very likely not independent so that is another issue; though our assumption is of independent-ness. But actually, they are not independent, so over time they are not necessarily going to be independent.

Now, we are clarifying certain conceptual things like endogeneity issues and omitted variable bias as the limitations of the panel data. Panel data faces a problem of endogeneity and OVB that is omitted variable bias.

(Refer Slide Time: 12:33)



□ Panel data faces a problem of endogeneity and omitted variable bias. For example, consider a panel dataset with the first wave of observation taken at time  $t_1$ . If X measured at time 1 t influences Y at time  $t_2$ , we cannot exclude the possibility that a measure of Y at time  $t_0$  does not itself exert a causal influence on X at  $t_1$ .

**Note!**  
**Endogeneity** broadly refers to situations in which an explanatory variable is correlated with the error term.  
**Omitted variable bias** occurs when a statistical model leaves out one or more relevant variables.

swayam 11

Let us consider a panel dataset with the first wave of observations taken at time period  $t_1$ . If X measured at time 1t influences Y at time  $t_2$ ; we cannot exclude the possibility that a measure of Y at  $t_0$  does not itself exert a casual influence on X at  $t_1$ . So,  $t_0$  and  $t_1$  whatever is the response is over 2 time period, we cannot just conclude that there is no casual influence.

So, here the problem of a and endogeneity arises; because of the fact that the explanatory term might be correlated with the error term or the explanatory variable might be also explaining the model because of some correlation exist.

So, endogeneity bias or endogeneity broadly refers to a situation in which an explanatory variable is correlated with the error term. Like we said Y is a function of X and X is a function of Y. So, in that case your error term is in fact correlated with your explanatory variable i.e., X. And this happens because of the repeated information over time and there might be correlation in the error term.

Another problem that we have in panel data is we are supposed to leave certain important variables or observations, as we cannot take or are unable to take all variables or observations, may be due to unavailability; so, in that way there might be a huge possibility of omitting certain important variables. So, that is in sort called- omitted variable bias. So, there are some techniques to control this biasness.

(Refer Slide Time: 14:35)

**Balanced and Unbalanced Panel Data**

- ❑ **Balanced Panel :**
  - All the subjects have the same number of observations. In other words, it has same number of time observation (T) on each of the  $N$  individuals.
  - A balanced panel has **no missing observations.**
  - The total number of observation is equal to **NT.**
- ❑ **Unbalanced Panel :**
  - Different number of time observations ( $T_i$ ) on each entity or individuals or subjects.
  - An unbalanced panel has **missing observations.**
  - The total number of observations is **less than NT.**

swayamii 12

Now, we are going to clarify you about the balanced versus unbalanced panel data. I am just now going to give you the conceptual background about it or you can say conceptual information about it; then at the end we will be experimenting with the data and I will show you that how we can clarify our concepts practically.

First of all, what do you mean by balance panel? Balance panel are the panel where all the subjects have the same number of observations, all the subjects or the entity have the same number of observations that is entered. In other words, it has same number of time observations i.e., T should be same for each of the  $N$  individuals. So,  $N$  and  $T$  is expected to be perfect and there are no missing observations. So, in that case total number of observations is equal to  $N T$ ;  $N$  times  $T$ ,  $T$  since no missing observations are available or no missing observations are reported.

Then second one is called unbalanced panel; here obviously another side is true, there must be missing observations and that too all  $N$  observations by the time are not available. So, that means different number of time observations on each entity or individual or subject are not available. And unbalanced panel has missing observations for sure; the total number of observations is of course less than that of “ $N$  times  $T$ ” i.e., the total dimensions of the panel.

(Refer Slide Time: 16:17)

N	T	Health expenditure	Nature of treatment
1	2019	5000	Modern
1	2020	6000	Modern
1	2021	7000	Modern
2	2019	2500	Traditional
2	2020	2300	Traditional
2	2021	2600	Traditional

**Balanced Panel**

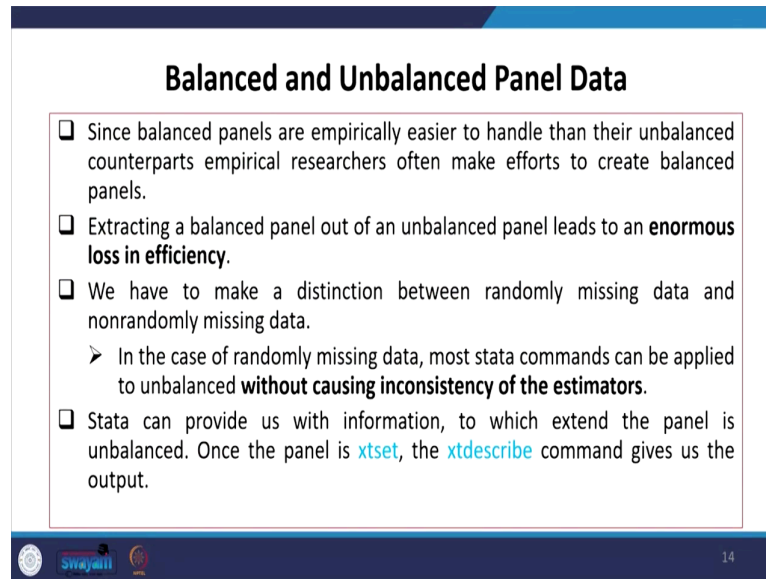
N	T	Health expenditure	Nature of treatment
1	2019	5000	Modern
1	2020	6000	Modern
2	2019	2500	Traditional
2	2020	2300	Traditional
2	2021	2600	Traditional

**Unbalanced Panel**

Here is an example on the screen for you to clarify the balanced panel and unbalanced panel. Here your N, T, healthcare expenditure and nature of treatment is given. The first N observation i.e., entity 1 is repeated over different time periods like 2019, 20 and then 21. This is “N times T” dimension we are discussing here since the second observations is also repeated with the same time period. And all the entries are available, that is why we are saying this is a balanced panel.

Second table is called unbalanced panel because you can just see that here entity 1 is available in 2019 and 20. So, 2021 is missing here, whereas 2<sup>nd</sup> entity of N is available in all the 3 periods. So, 2021 for the first observation is in fact missing and providing missing information; therefore, this is called an unbalanced panel.

(Refer Slide Time: 17:24)



**Balanced and Unbalanced Panel Data**

- ❑ Since balanced panels are empirically easier to handle than their unbalanced counterparts empirical researchers often make efforts to create balanced panels.
- ❑ Extracting a balanced panel out of an unbalanced panel leads to an **enormous loss in efficiency**.
- ❑ We have to make a distinction between randomly missing data and nonrandomly missing data.
  - In the case of randomly missing data, most stata commands can be applied to unbalanced **without causing inconsistency of the estimators**.
- ❑ Stata can provide us with information, to which extend the panel is unbalanced. Once the panel is `xtset`, the `xtdescribe` command gives us the output.

swayamii 14

So, further information about balanced and unbalanced panel are mentioned here that- since, balanced panels are empirically easier to handle than their unbalanced counterparts, the empirical researchers often make efforts to create balanced panels. But, extracting a balanced panel out of an unbalance panel leads to an enormous loss in efficiency.


We have to make a distinction between randomly missing data and non-randomly missing data. So, some data points might have been randomly missed and some might be non-randomly or are actually deliberately missed, so these non-randomly missed data points are actually a case of concern. If it is randomly missed, still the data set can be considered as balanced panel. In the case of randomly missing data, most data commands can be applied to unbalanced without causing inconsistency of the estimator.

So, once Stata finds that missing entries are there, Stata simply reads it as unbalanced panel. But it is not creating inconsistencies in the result or the estimator is not inconsistent. Stata can provide us with information the extent to which the panel is unbalanced. Once the panel is given the command i.e., `xtset` and the `xtdescribe` command; then we can understand whether the data is unbalanced or not and to what extent it is unbalanced.

(Refer Slide Time: 19:05)

### Practical Example


- ❑ WHO data set
  - Observing Health Dataset in Stata, Whether it is Balanced or Unbalanced.
- ❑ Check data is balanced or unbalanced :
  - `xtset,`
  - `xtdescribe`
- ❑ After the `xtreg-` command run `tab ID if e(sample)`. That will give you a list of the observations that are *not* dropped.
  - `xtreg HEXP HC3 DALE`
  - `tab ID if e(sample)`.

 15

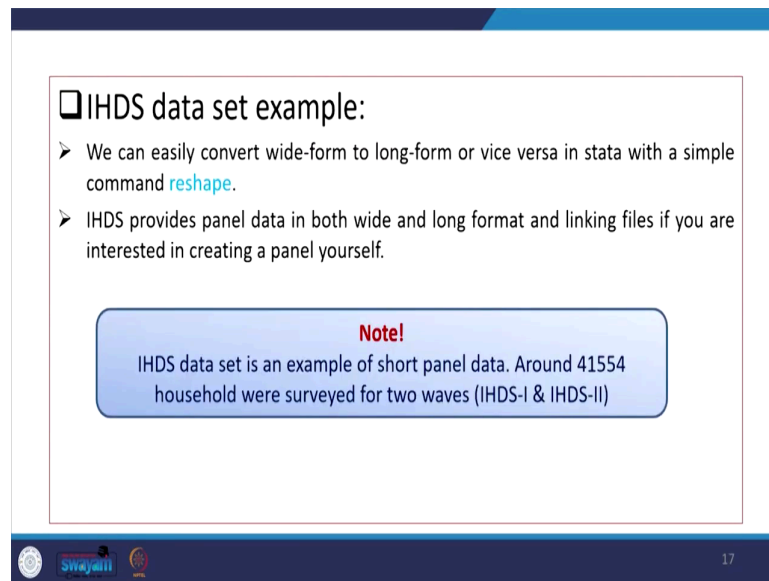
(Refer Slide Time: 19:08)

### Types of Panel Data based on time dimension

- ❑ On the basis of time dimension panel data is classified as **short panel** or **long panel**.
- ❑ Data on many individual units (N) and few time periods (t) are called short panel or micro panel. (**N>T**)
- ❑ In long panel data number of individual units is less than number of time periods. It is also known as macro panels. (**T>N**)
- ❑ Short panels are more common than long panels.
- ❑ The estimation techniques depends on whether we have short panel or long panel.

 16

(Refer Slide Time: 19:09)



□ IHDS data set example:

- We can easily convert wide-form to long-form or vice versa in stata with a simple command `reshape`.
- IHDS provides panel data in both wide and long format and linking files if you are interested in creating a panel yourself.

**Note!**  
IHDS data set is an example of short panel data. Around 41554 household were surveyed for two waves (IHDS-I & IHDS-II)

swayam 17

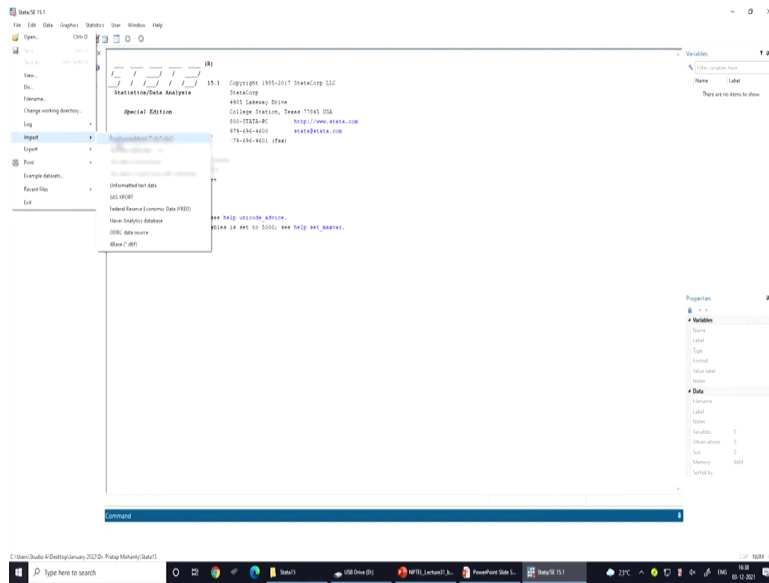
Now, we will have a practical data set; the practical data set we have taken is from WHO sample data set; this information will also be provided to you as an additional document for your own practice. And we are going to explain all those things, I am just referring what are the data sets we are considering here.

We are observing health dataset in data whether it is balance or unbalance. We need to check this through these 2 commands (`xtset` and `xtdescribe`) to identify whether the dataset is balanced or unbalanced.

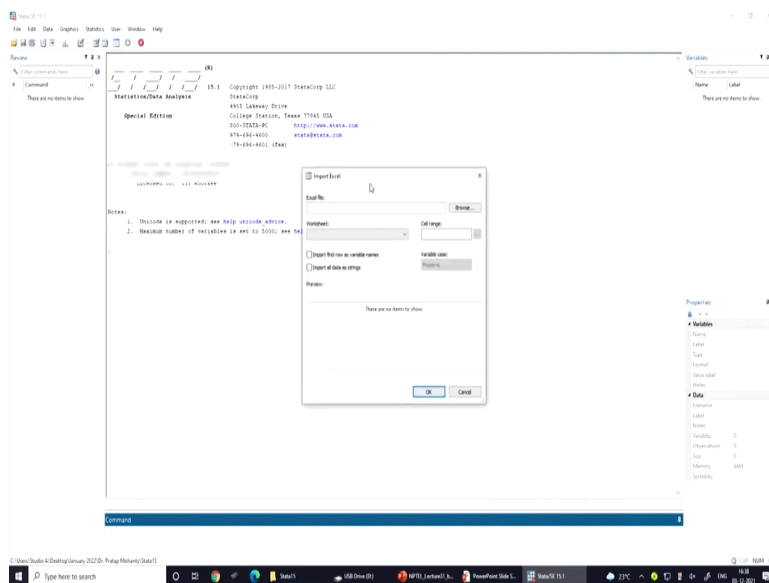
After that in the next lecture, we can think of running regression command through the `xtreg` command. And we can accordingly find out whether that is more suitable or not, that will give you a list of observations that are not dropped.

So, I will try to give the detailed explanation of `xtset` command in the next class; but at this moment I am going to talk about some practical examples, so let us go for it, then we will also look to other aspects.

(Refer Slide Time: 20:21)

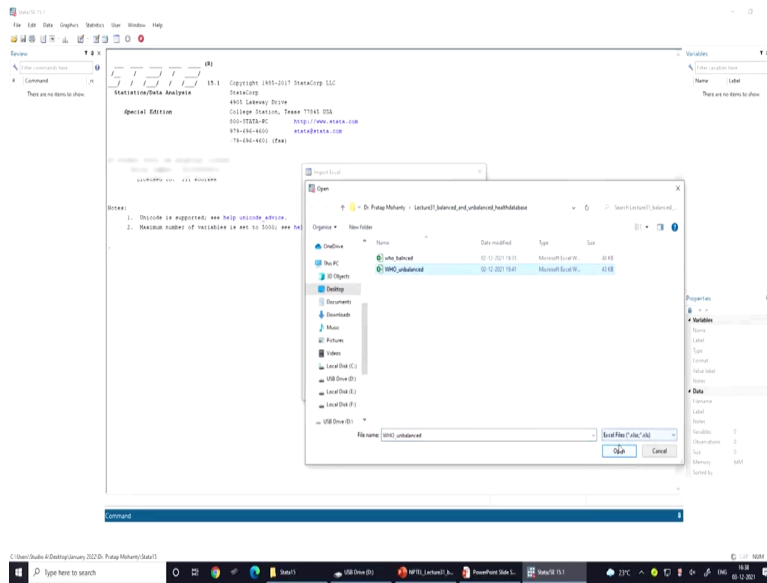


(Refer Slide Time: 20:33)



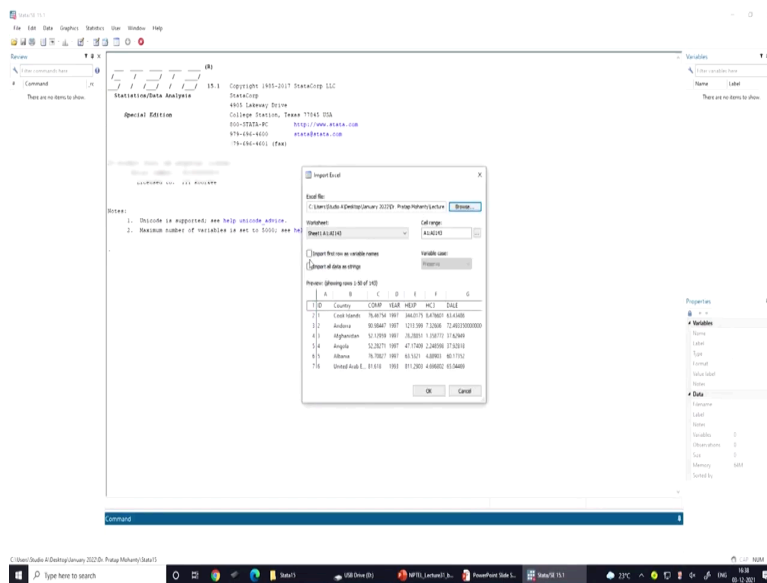


(Refer Slide Time: 20:35)



Now, here is our data, I am just opening it for your reference. Now, I am going to the file, then importing the data; this database we have taken is from the WHO and this is in excel base format.

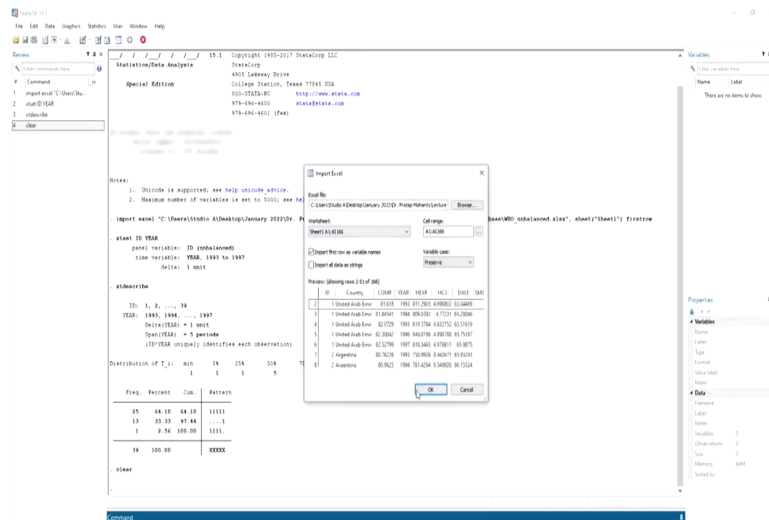
(Refer Slide Time: 20:39)





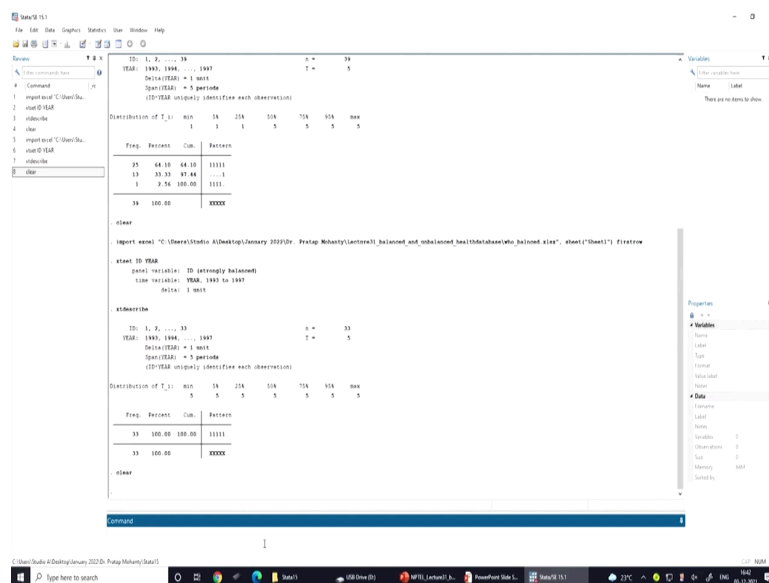
In only one case, one observation where first 4 time period are available, but the last time period it is in fact missing. Therefore, the data has earmarked as unbalanced. So, next we are going to open the data set to highlight or to understand the balanced panel.

(Refer Slide Time: 23:11)



So, so here I am going to open a kind of panel data, which is a balance one; we have again taken from the same source i.e., WHO. So, WHO balanced panel data it is on healthcare issues. We will also share this with you, we have identified here the variables names as in the first row.

(Refer Slide Time: 23:31)



Now, with the same commands, like `xtset`, we will confirm whether this data is balanced or unbalanced with the help of ID variable and the time variable that are defined here.

Now, you can see results of the `xtset` command, `xt` is the standard command in panel; though `xtset` is conforming us that your data from third 1993 to 1997 is a strongly balance panel. You can also check it using the description as well, through the `xtdescribe` command.

These do files will also be shared with you. Now, you can see all those 33 observations are actually repeated in all 5 time periods. So, 1 is mentioned 5 times. So, all 5 time periods, all 33 observations available. So, I think this is not problematic and I am sure that you can easily understand it in any database, whether it is a panel or not a panel data.

So, this is how I have explained. Let us move to the understanding of all the datasets; I will also clarify some other concept related to panel, like short panel and long panel. On the basis of time dimension, panel data is classified as short panel or long panel. Data on many individual units and few time periods like where  $N$  is higher than that of the  $T$ . In that case it is called short panel or that is called micro panel. Just the reverse of this is observed in the long panel or in the macro panel. So, long panel are also called macro panels. The long is the word refers to mostly emphasizing the time; short panels are more common than long panels, the estimation techniques depend on whether we have short panel or long panel.

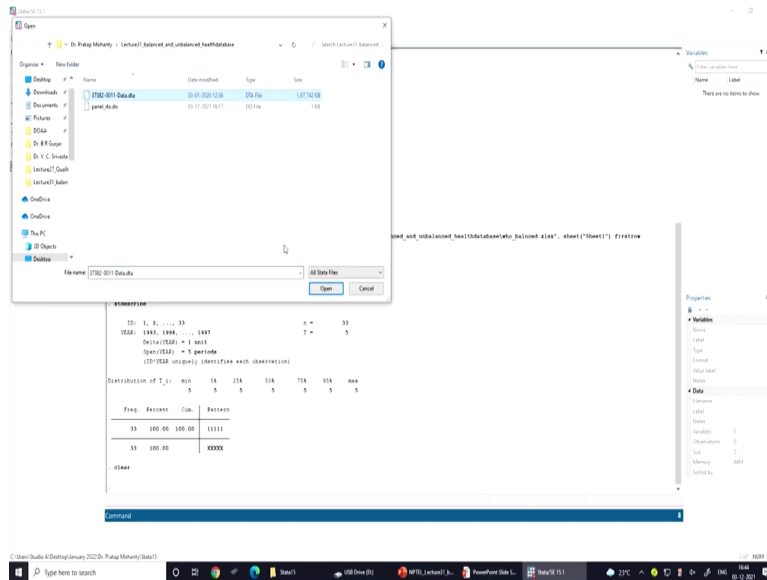
We are now exploring in front of you the data of IHDS i.e., India's human development survey. This keeps panel observation and this is in fact a micro panel data, where  $N$  is higher than that of  $T$ ,  $T$  here is only for 2 periods. So, we are going to open it and here number of observations are of 41554.

Households were surveyed for 2 periods in IHDS I and IHDS II. We have already discussed something about the IHDS. You can also search in any website about India's human development survey and you will get to know more about it. So, IHDS dataset is an example of short panel data.

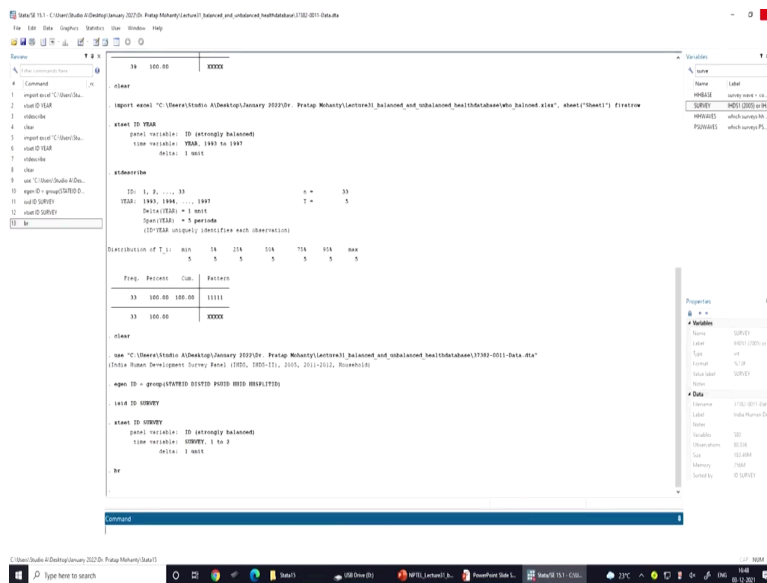
We can easily convert wide form to long form or vice versa in Stata with a simple command called "reshape". So, like a wide panel, where number of cross sections are much higher and our time dimensions are much higher; if it is so, then we can actually flip it through reshape command.

IHDS provides panel data in both wide and long format and linking files. So, if you are interested in creating a panel yourself, this is going to help you better. Now, I am going to show it here about this IHDS data for your better understanding.

(Refer Slide Time: 27:11)



(Refer Slide Time: 27:16)



So, I am going to clear it i.e., clear the earlier data. Now, we are going to open the IHDS data in front of you. First of all, we need to understand which are the common ID or the ID variables and those would be uniquely identified in the panel data.

So, we are generating an ID variable. So, “egen” is the command that is used to generate ID variable as per the instruction given in IHDS, you can search in Google in their webpage. You are making an ID variable through this command and there are 5 important variables i.e., starting from state ID, district ID, psu ID, HH ID, and household split ID.

Now, you can see the state ID is mandatory, then districts, then psu, then primary sampling unit; then you come to the household level, then how household is getting split over time period, who are the final households. So, your split ID is also very important.

So, now we have defined an ID variable. So, ID variable with the name ID; now we can check whether these are actually uniquely identified or not. So, “isid” is the command that is used for ID, ID variable we have defined. Now, at the end it is there. And the time dimension has to be given as well to survey a period as per the data.

So, I know that these 2 are in fact uniquely identified, since it does not throw any error. Now, we can check whether this is a balanced or an unbalanced data. The xtset command is the one through which we can check it. So, it suggests that it is a strongly balanced data.

Strongly balanced data as the time period is from 1 to 2 as I already mentioned. We can also check it from the description command i.e., “xtdescribe” in case if you want and in the way that we have already guided. Also, we can check it through the browsing window; once open the browse window and through the display we are able to get it.

(Refer Slide Time: 29:48)

Variable	Storage	Display	Label	Position
stateid	int(8)	%10.0g	State ID	1
distid	int(8)	%10.0g	District ID	2
psuid	int(8)	%10.0g	Primary Sampling Unit ID	3
hhid	int(8)	%10.0g	Household ID	4
splitid	int(8)	%10.0g	Household Split ID	5

