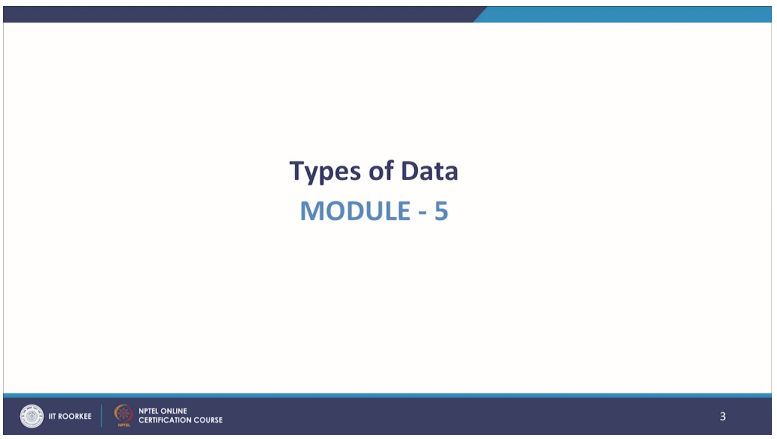
Econometric Modelling
Professor. Sujata Kar
Department of Management Studies
Indian Institute of Technology, Roorkee
Lecture No. 05
Types of Data

Hello, and welcome again to the course on Econometric Modelling. This is module 5.

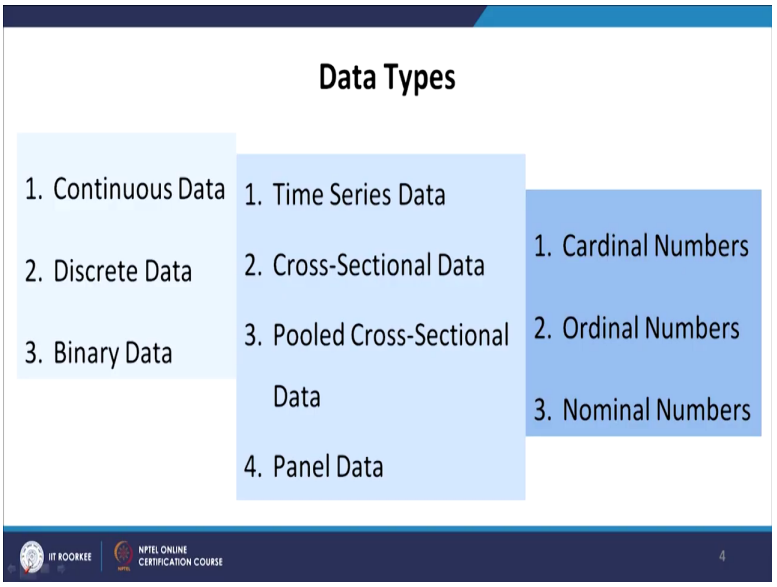
(Refer Slide Time: 0:30)

Above slide shows the organization of the course, it has been divided into 8 parts and this module completes the first part, that is introduction to econometrics. So far, we have in the first module have defined what is econometrics, what are the types, applications, then we discussed formulation of econometric modeling, the steps involved in it, then in the last two modules, we have reviewed the basic concepts. And finally, in the first part itself, the final module is the types of data.

(Refer Slide Time: 1:32)



So, module 5 deals with types of data. Data types are categorized into three alternative categories, and they are not at all mutually exclusive categories, they are actually different types of categorization. So, the first type is continuous data, discrete data, and binary data. We have already discussed continuous and discrete variables.

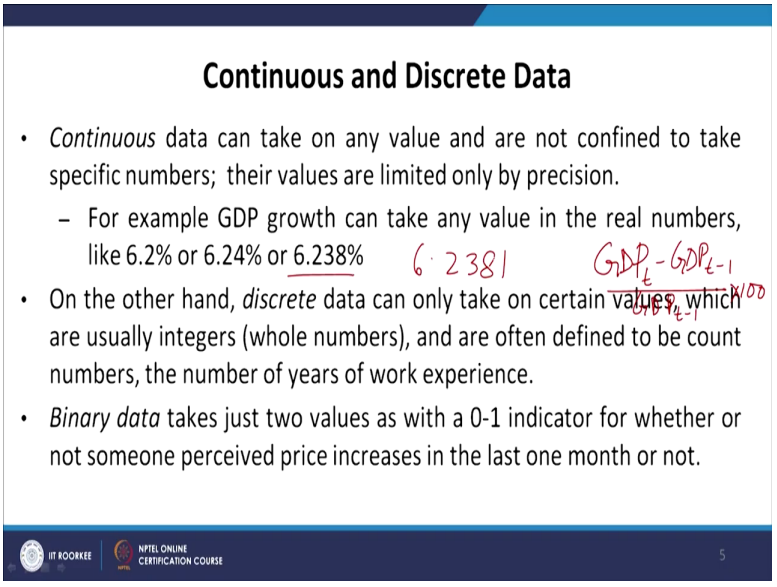


When we collect observations on a particular variable or on a set of variables, then that becomes data. So, if we are working with continuous variables, then, of course, the data collected will be continuous data. If you are working with discrete variables, then the data collected will be discrete data. And the third one is binary data. Now, what are these data, how we define them, that anyway I am going to talk about in detail in the next slides, but here we just briefly introduce what are the types or categories of data we are going to deal with.

So, the first categorization is continuous data discrete data, and binary data. The second categorization is time-series data cross-sectional data, pooled cross-sectional data, and panel data. This is sort of the most important categorization which is most relevant for this course, according to this categorization methods also keep on changing. That is why we call it the most important categorization.

The third categorization is cardinal numbers, ordinal numbers, nominal numbers. We first look in detail at continuous and discrete data.

(Refer Slide Time: 3:36)



Continuous data can take on any value and are not confined to take specific numbers, their values are limited only by precision. For example, GDP growth can take any value in their real numbers like 6.2 or 6.24 or 6.238. So, if we keep on rounding the numbers at the decimal places, then we are handling the same number, but expanding in decimal places are rounding off at decimal places. So, this is an example of continuous data.

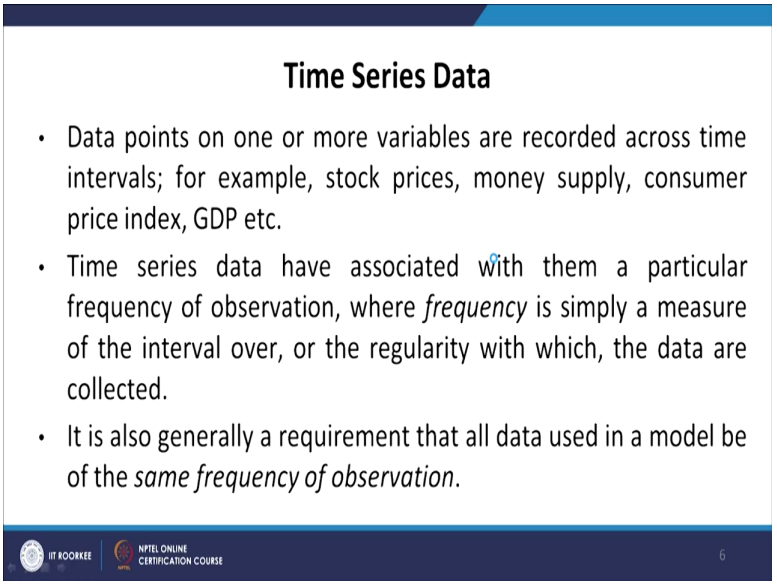
We actually had taken up example of GDP growth while discussing continuous variables as well. If you remember, the growth is calculated, as I mentioned that growth is calculated as GDP at time period t minus GDP at time period t minus 1 divided by GDP at time period t minus 1 multiplied by 100 *(Refer to above slide on Continuous and discrete data)*. And I mentioned that, since this is a calculated number, it is a ratio so that in percentage terms, it can take a value like this we can further expand it to 6.2381 and so on.

Next is the discrete data which can take only certain values which are usually integers or whole numbers and are often defined to be count numbers, the numbers of years of experience, for example. And finally, we have binary data that takes just two values as with a 0 1 indicator for whether or not someone perceived price increases in the last one month or not.

So, whenever we record data in terms of 0 and 1, then that becomes binary data. If the example that is being given here is that whether someone perceived price increases or not, the response would be recorded as 0 and 1. Suppose, no, I did not perceive, so 0, yes, I perceived, 1 and so on, for a large number of individuals.

So, these are the types of continuous-discrete and binary data. Now, we deal with the next category at length as mentioned previously. So, under the next category, we have time-series data, cross-sectional data, pooled cross-sectional, and panel data.

(Refer Slide Time: 6:26)



So, the first type is time-series data. Data points that are recorded on one or more variables across time intervals; for example, stock prices, money supply, consumer price index, GDP, etc. are time-series data. Time series data have associated with them a particular frequency or of observation, where the frequency is simply a measure of the interval over or the regularity with which the data are collected. It is also generally a requirement that all data used in a model be of the same frequency of observation.

Frequency is defined as the regularity or the period over which the data has been collected. So, when we work with annual GDP data, then the GDP figures are observed annually. When we work with quarterly data, say we are working with an index of industrial production, then the observations are quarterly, for each quarter I have an IIP figure or value.

We can also work with weekly data, for example, price index data are available on weekly basis. So, for each and every week of a month or for all the months in a year, we have observations. So, here the frequency is weekly frequency, we can also work with daily data, for example, stock prices, or stock indexes, we can work with data, which are more frequent than even daytime. So, we can have our list price data, so that is also one kind of frequency.

Now, as I probably mentioned in the very first module, that depending on frequency, which is a characteristic of data, the methodology used can also be different. Now, what is stated here is that in a particular model, all the observations should have the same frequency. So, it is not possible that when I have GDP as my one variable, which is measured or observed annually, at the same time, I am having inflation observed monthly or weekly, I must have all the variables having the same frequency. So, all of them should be either annual or monthly or quarterly and so on.

(Refer Slide Time: 9:01)

Private final consumption expenditure (PFCE) & GDP at market prices in India, base 2011-12

Observation	Year	PFCE (₹ billion)	GDP (₹ billion)
1	1990-91	3805.90	5684.24
2	1991-92	4371.31	6534.27
3	1992-93	4928.86	7510.42
4	1993-94	5646.92	8643.08
5	1994-95	6562.23	10138.62
.	.	.	.
29	2018-19	79080.57	139814.26
30	2019-20	83259.07	145659.51

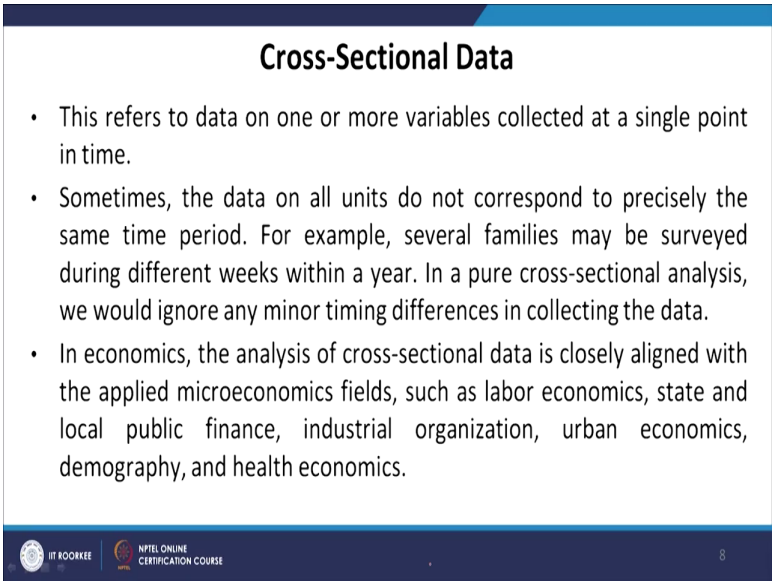
IIT ROORKEE | NPTEL ONLINE CERTIFICATION COURSE

We will take an example; the example is of private final consumption expenditure PFCE and gross domestic product at market prices in India with the base period 2011-12 *(refer to slide*

above: Time ;9:01). And this is an example of time-series data. These examples we had earlier also taken up in or in the context of observing covariance and correlations.

Now, here I am showing you the data, the data period under consideration is from 1991 to 2019-20. We have the PFCE figures in billions, we have the GDP figures in billions and the observations are from 1 to 30. Both the series are observed annually. So, the frequency is annual frequency. So, this is an example of time-series data where the observations are plotted or recorded against the time period, and the time period is the annual year.

(Refer Slide Time: 10:10)



Then, we will talk about cross-sectional data, this reference to data on one or more variables collected at a single point in time. So, just opposite to that of time series data. Sometimes the data on all units do not correspond to precisely the same time period. For example, several families may be surveyed during different weeks within a year. In a pure cross-sectional analysis, we would ignore any minor timing differences in collecting the data.

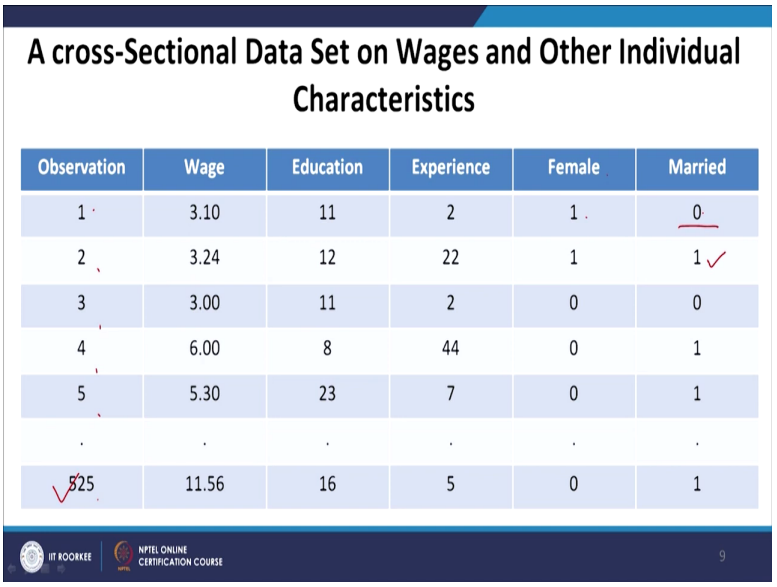
So, when we refer to a single point in time or a snapshot of the data at any point in time, that does not have any specific reference to a particular time, it can be an entire year or it can span more than a year. So, if I take up a project of collecting data, on certain variables from a large number of individuals across a large number of countries, and then it is possible that I have taken a sufficiently long time to collect the data, because of which it is a primary data.

So, I am hand collecting the data, taking information from each and every individual, it is possible that it expands beyond 1 year. And it is also possible that the data points that I am

considering or the variables that I am considering that has less relevance to the time period. So, as time progresses, the data that I have considered does not lose its value much. So, if so, is the case, then we will still consider it to be cross-sectional data. Because I will consider it to be collected at a single point in time, I am continuously collecting data, it is possible that the time period that I have taken is pretty long.

In economics, the analysis of cross-sectional data is closely aligned with the applied microeconomic fields such as labor economics, state and local public finance, industrial organization, urban economics, demography, and health economics, just to name a few, that these are the fields where we use cross-sectional data. Below (*refer to slide time:12:37*) is the example of cross-sectional data. So, this is a cross-sectional data set on wages and other individual characteristics.

(Refer Slide Time: 12:37)



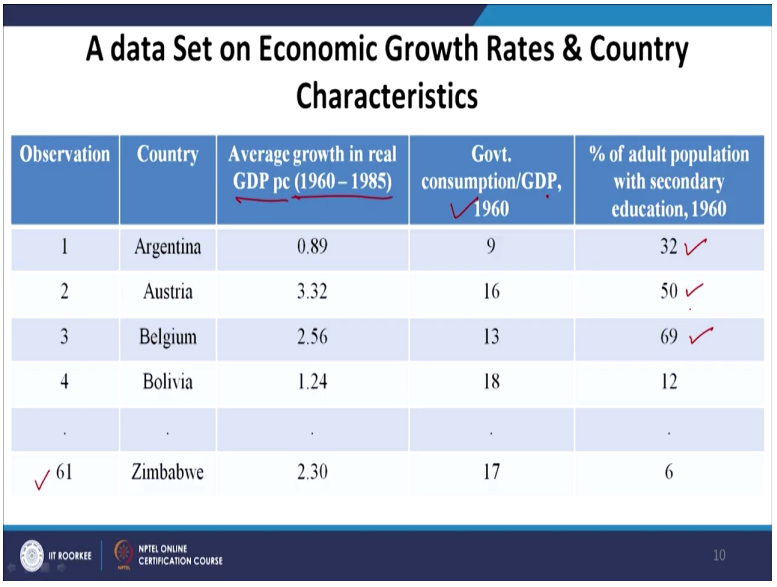
Suppose the number of individuals considered is 525. So, I have recorded data for 525 individuals, individuals 1, 2, 3, 4, 5, and so on, their wages are recorded, which is measured in terms of, these are taken up from different studies. So that is why these wages might be in terms of dollars per hour. That is why they are 3.10, 3.25, and so on.

And then we have also measured their education levels. Education levels are measured in terms of the number of years spent on education, then experience, it could be the number of years, so 2 years, 22 years, 2 years, 44 years and so on, and then we have two more columns;

one is female. If someone is a female then is assigned a value 1, otherwise 0. Then for the married column, if someone is married then assigned a value 1, otherwise 0.

So, these are observations that do not have any reference to any time period. So, that is why we call it cross-sectional data, whatever time has been taken to collect the data on this. 525 individuals that we consider to be collected in a single point of time and present it as cross-sectional data. This is cross-section, because it is collected over a section of individuals. So, there are 1 to 525 individuals on whom the data has been collected.

(Refer Slide Time: 14:22)



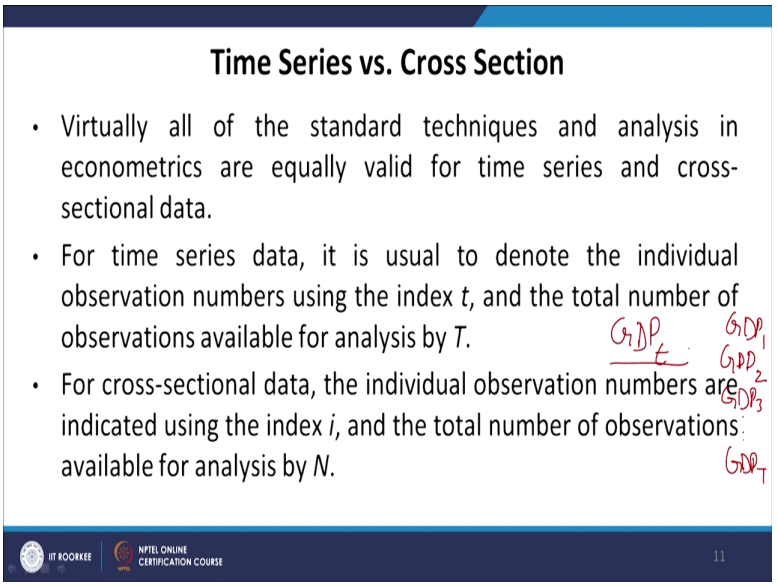
A data set on economic growth rates and country characteristics can also be an example of cross-sectional data. So, here we have different countries. For example, we have considered 61 countries, country 1, Argentina, 2, Austria, Belgium, Bolivia. So, you can see that these countries are arranged in terms of their alphabetical orders. Then, what we have considered is the average growth in real GDP per capita for the period 1960 to 1985.

So, you can see that there has been a reference to the data related to GDP per capita, but he does not talk about when this data has been collected, it is quite possible that I have collected that data in 1992, by that time, the 1960 to 1985 data for all these countries that is 1960 to 1985, average growth in real GDP per capita is available for all these countries. So, again, these are, this is an example of cross-section data.

Similarly, we also have the percent of the adult population with secondary education prevailing in 1960, so percent of the adult population with secondary education in 1960, in

Argentina, it was 32 percent, in Austria, it was 50 percent, Belgium 69 percent and so, on. So, this makes another case of cross-sectional data where we have collected data on certain variables specific to certain years across a large number of countries specifically 61 countries. Then, we compare, time-series and cross-section.

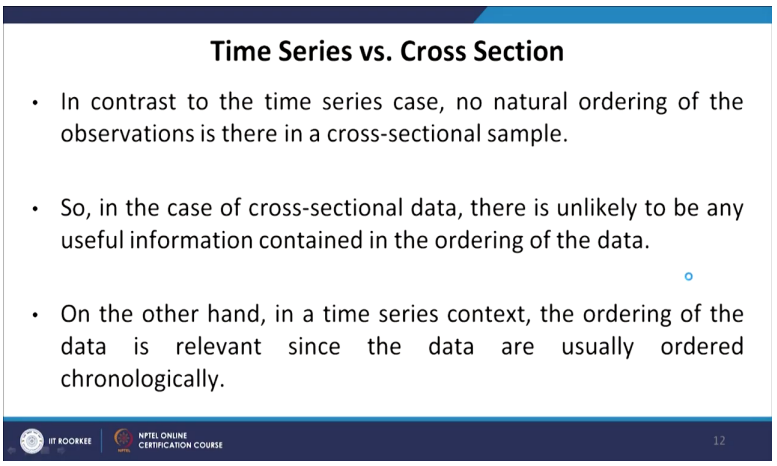
(Refer Slide Time: 16:36)



Virtually all of the standard techniques and analysis in econometrics are equally valid for time series and cross-sectional data. For time-series data, it is usual to denote the individual observation numbers using the index t, and the total number of observations available for analysis by capital T. So, when we calculate GDP growth rates, then we mention it as GDP small t.

So, this t can be anything like a particular year, it can be 1990, 1992, 2010, 2015, 2016 anything, if I generally, come up with a set of observations, then I will write it rather as GDP one the first period, GDP to the second period, GDP three the third period and finally, I come at GDP capital T. So, if it is a 30 years data, then capital T, here is thirtieth observation. So, the individual is generally referred to as small t the total numbers are generally referred to as capital T. For cross-sectional data, the individual observation numbers are indicated using the index i and the total number of observations available for analysis by capital N.

(Refer Slide Time: 18:04)



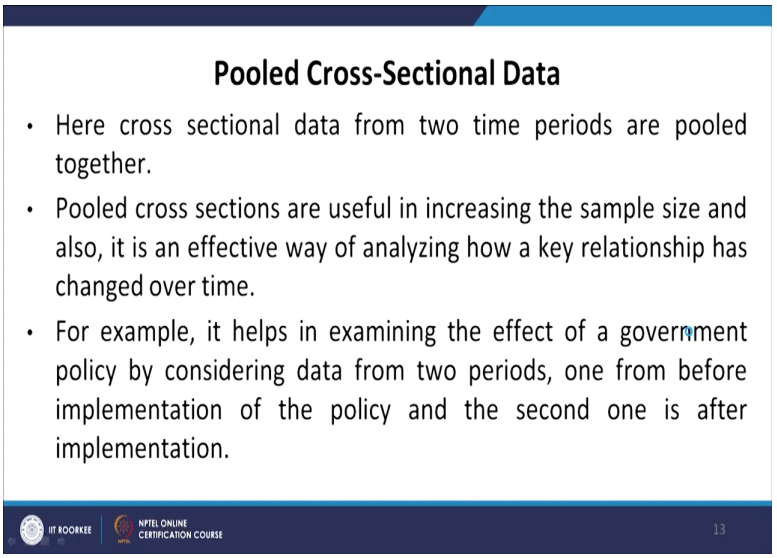
In contrast to the time series case, no natural ordering of the observations is there in a cross-sectional sample. So, in time-series data, ordering is very important, which is not the case with cross-sectional data. For example, when we considered the cross-country data that is data on 61 countries on average GDP figures or GDP growth rates, then what we did is that we just simply ordered them alphabetically.

But, if I put it in reverse alphabetical order, that is starting from Z, then I have Argentina or countries beginning with A that would not make any difference, that is the point or even if I do not order them in terms of any alphabetical order, rather if I ordered the countries in terms of their GDP growth rate, I have the country with the highest GDP growth rate at the top and the lowest GDP growth rate at the bottom, then also the analysis would not make any difference. So, orderings in the case of cross-sectional data are actually not important.

But as you can see that time-series data is plotted against time. So, their time plays a very important role. And we cannot have the haphazard ordering of data, that if I am having 30 years of data, either I must, record them from 1990 to 2020. Or I can, do it the other way around, that is 2020 to 1990. But it is not possible that sometimes I have 2020 followed by 2010, then comes 2011, then comes 1990, and so on. This haphazard thing will not give us good results or appropriate results. So, ordering is very important in the case of time series data.

So, in a time series context, the ordering of the data is relevant since the data are usually ordered chronologically, and the chronology matters, that helps us in understanding how things have improved over a period of time, or deteriorated over a period of time or generally how things have changed over a period of time.

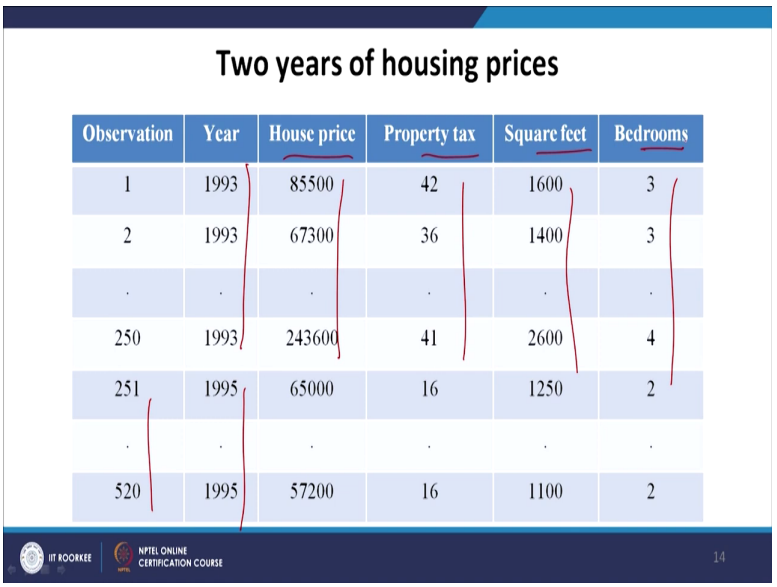
(Refer Slide Time: 20:32)



Now, we talk about pooled cross-sectional data. Here, cross-sectional data from two time periods are polled together. Polled cross-sections are useful in increasing the sample size and also it is an effective way of analyzing how a key relationship has changed over time. For example, it helps in examining the effect of a government policy by considering data from two periods, one from before the implementation of the policy and the second one is after implementation of the policy.

For example, if we want to examine whether midday meal, in the Indian context, has improved school participation, or has decreased the school dropout rates in the rural areas in India, or even in urban areas where midday meals are provided by the schools, then we can consider two periods, the first period being where the midday meals were not introduced or not provided to the children and the second period belongs to the period where midday meals have been introduced. We are considering two periods and then for each period, we are collecting data from individuals or from schools. And then we are collecting them by combining them. So, this is called pooled cross-section.

(Refer Slide Time: 21:58)



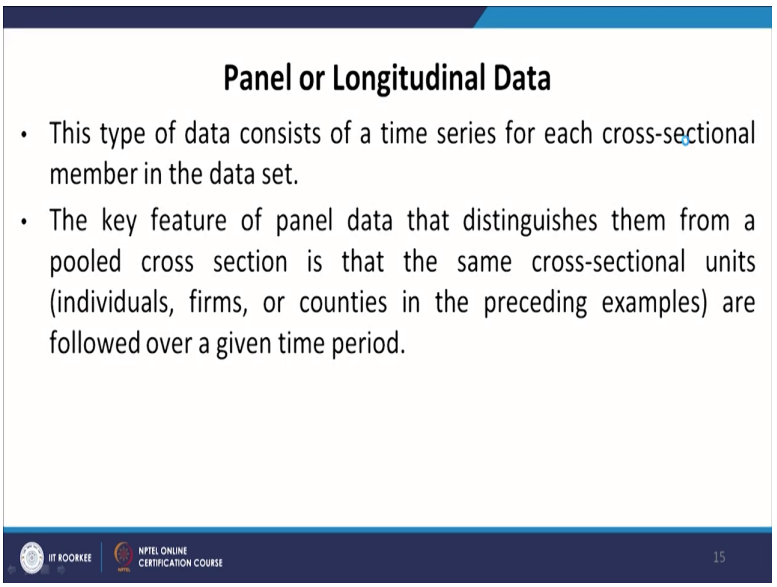
We have an example of two years of housing prices. So, the first year refers to 1993, and the second year reference 1995. Now, in between, there has been a change in the property tax rates. And we are going to see that whether house prices have been impacted by the introduction or reduction of the property tax. So, in the first period in 1993, 250 observations were considered. So, we have data from 1 to 250.

In the second period in 1995, we have 270 observations so that we total have 520 observations. So, it is not necessary that we have the same number of observations in both periods. And for each period or in each period, we have considered or collected data on house prices, property tax, square feet, bedrooms of the houses, and similar information for the other period as well.

So, you can just see that information on these variables i.e. house price property tax square feet, and bedrooms are collected in two different periods from two different sets of individuals or entities. In the first period, we had 250 observations in the second period we have 270 observations. So, this is an example of the pooled cross-section.

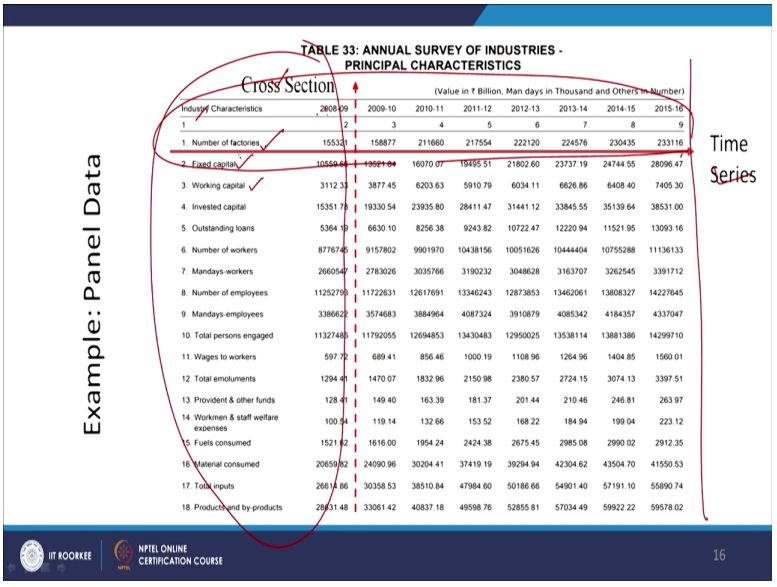
Now finally, we talked about the fourth type of data under the second type of categorization that is a panel or longitudinal data.

(Refer Slide Time: 23:40)



So, this type of data consists of a time series for each cross-sectional member in the data set. The key feature of panel data that distinguishes them from a pooled cross-section is that the same cross-sectional units that are individuals, firms, or countries in the preceding examples are followed over a given time period. As with a pure cross-section, the ordering in the cross-section of a panel data set does not matter. For this, I will also show you how panel data is organized.

(Refer Slide Time: 24:17)



Industrial Characteristics	2008-09	2009-10	2010-11	2011-12	2012-13	2013-14	2014-15	2015-16
1	2	3	4	5	6	7	8	9
1. Number of factories	155321	158877	211660	217554	222120	224576	230435	233116
2. Fixed capital	10559.60	13521.04	16070.07	19495.51	21802.60	23737.19	24744.55	28096.47
3. Working capital	3112.33	3877.45	6203.63	5910.79	6034.11	6626.86	6408.40	7405.30
4. Invested capital	15351.79	19330.54	23935.80	28411.47	31441.12	33845.55	35139.64	38531.00
5. Outstanding loans	5364.19	6630.10	8256.38	9243.82	10722.47	12220.94	11521.95	13093.16
6. Number of workers	8776745	9157802	9901970	10438156	10051626	10444404	10755288	11136133
7. Mandays-workers	2660547	2783026	3035766	3190232	3048628	3163707	3262545	3391712
8. Number of employees	11252793	11722631	12617691	13346243	12873853	13462061	13808327	14227645
9. Mandays-employees	3386622	3574683	3884964	4087324	3910879	4085342	4184357	4337047
10. Total persons engaged	11327483	11792055	12694853	13430483	12950025	13538114	13881386	14299710
11. Wages to workers	597.72	689.41	856.46	1000.19	1108.96	1264.96	1404.85	1560.01
12. Total emoluments	1294.41	1470.07	1832.96	2150.98	2380.57	2724.15	3074.13	3397.51
13. Provident & other funds	128.41	149.40	163.39	181.37	201.44	210.46	246.81	263.97
14. Workmen & staff welfare expenses	100.54	119.14	132.66	153.52	168.22	184.94	199.04	223.12
15. Fuels consumed	1521.62	1616.00	1954.24	2424.38	2675.45	2985.08	2990.02	2912.35
16. Material consumed	20659.82	24090.96	30204.41	37419.19	39294.94	42304.62	43504.70	41550.53
17. Total inputs	26614.86	30358.53	38510.84	47984.60	50186.66	54901.40	57191.10	55890.74
18. Products and by-products	28331.48	33061.42	40837.18	49598.76	52855.81	57034.49	59922.22	59578.02

So, this is an example of panel data, where we have considered annual survey of industries principal characteristics, so these are the industry characteristics which include the number of factories, fixed capital, working capital, and so on. And then we have data on individual

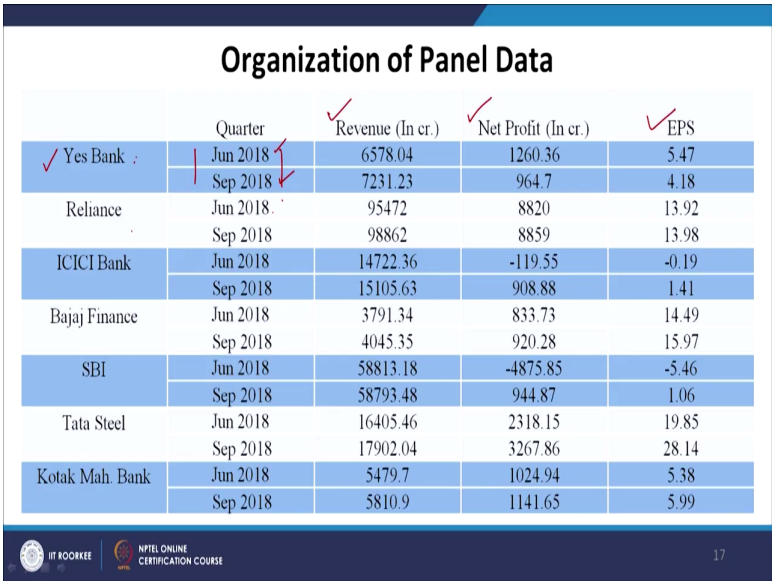
variables as you can see for 7 years i.e. 2008-09 to 2015-16. This makes a cross-sectional data. When I draw this arrow (*refer to above slide*), this implies that this section forms cross-sectional data, because we have a large number of characteristics for only 1 year.

And then we have time-series data, where for one characteristic we have, observations or, on 7 years. So, this is time-series data. But overall, the entire data set is panel data or longitudinal data, because it combines cross-section as well as time series. Now, you can see that the important difference between pooled cross-section and panel is that, here we are collecting data on the same characteristic for each and every year.

So, if I work with individuals, then what would happen is that in the previous example, where we considered housing prices, property taxes, and so on. This would imply that when we are collecting data in 1995, we should consider the same individuals from whom we collected data in 1993. So, similarly, when I give the example of midday meals and their impact on school dropout or school enrollment, I consider two different time periods, then in two different time periods, I consider these same schools in order to make it panel data.

If I consider different sets of schools into different time periods, then it is a pooled cross-section, but if I consider the same schools, then it becomes panel data. So, that is a major difference between pooled cross-section and panel data. Though many times or most often you would say that panel and pooled cross-section or pooled data, these two are used synonymously or interchangeably.

(Refer Slide Time: 26:47)

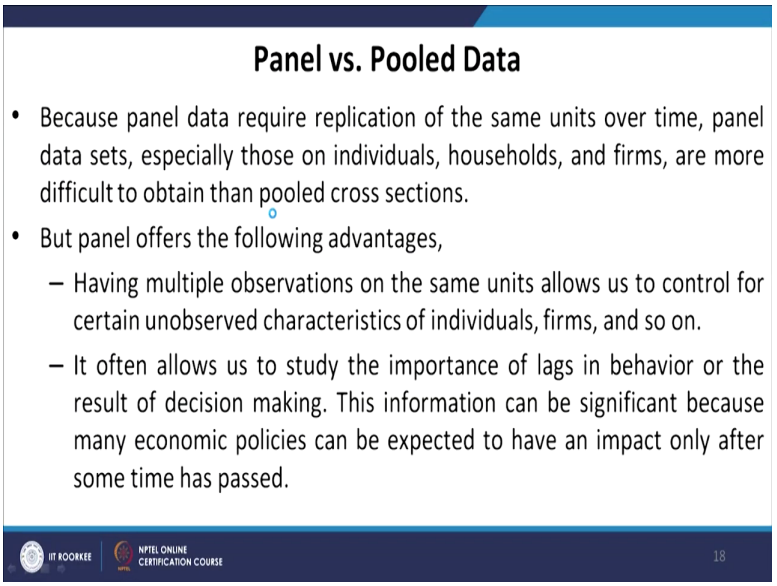


How do you organize panel data? So, you can see that for each and every entity, we have, information for each and every time series *(refer to slide time: 26:47 above*). So, these are the entities like; Yes Bank, Reliance ICICI Bank, etc. And then, for two different quarters, I have considered their revenue, net profit, and EPS. So, what is happening here is that, when I am organizing the data, I am first considering the entities and then I am considering their observations for each and every time period.

So, these time periods would actually or must follow a chronology. So, we have June 2018, followed by September 2018 and that is the case for each and every entity. But it is not necessary that I must have Yes Bank at the top, I can have Reliance at the top of ICICI Bank at the top. So, these banks can be ordered in any manner, there is no specific order in which I have mentioned them, but the time series component should be ordered chronologically.

So, this is our organization, panel data is organized where we consider each and every individual or entity, and then information pertaining to its time series values are first mentioned, then we go to the second entity, and so on.

(Refer Slide Time: 28:15)



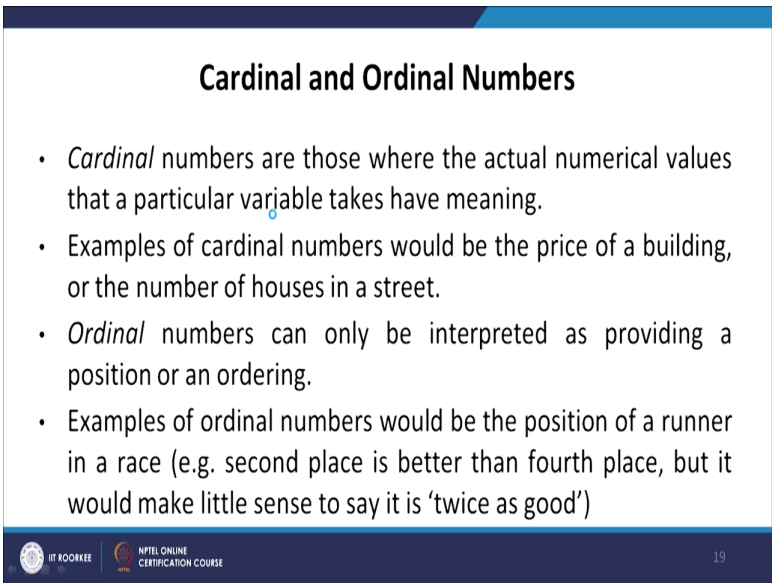
Now, we finally compare panel and pooled data, because panel data require the application of the same units over time panel data sets, especially those on individuals households, and firms are more difficult to obtain than pooled cross-sections. But panel offers the following advantages, first of all, having multiple observations on the same units allows us to control for certain unobserved characteristics of individual firms and so on.

So, when we talk about controlling for certain observations, then we essentially talk about incorporating or rather controlling the impact of those characteristics on the dependent variable or on the variable of our focus. So, since those characteristics remain the same across two sets of observations, it becomes easier for us to control those observations. And the second advantage is that it often allows us to study the importance of lags in behavior or the result of decision making, this information can be significant because many economic policies can be expected to have an impact only after some time has passed.

So, for example, when we talk about the introduction of midday meals in schools, then if I consider any schools before and after, then they would certainly reflect some amount of changes. But if I consider exactly the same schools in both periods, then the changes are more prominent, the changes are more evident and more obvious. Because the same schools have been taken and we can understand whether things have changed after the introduction of the policy or not.

And it takes certain time for individuals to react to a policy, for example, the government may introduce taxes on products that are harmful, health-wise. So, it can impose taxes on fatty foods, it can impose taxes on butter. So, if the tax rates are increased on butter, then all the commodities that use butter as an intermediate product will increase their prices. So, consumers would experience an increase in the prices of many commodities, which use butter as an intermediate product with a time lag. And they would also take some time to react to that price increase, so that they would settle for something cheaper, at least some section of the consumers would settle for something cheaper. So, there is a certain time lag, that lag, or the time that individuals or consumers take in order to react to a policy change.

(Refer Slide Time: 31:38)



And then, we will talk about the last set of data types, that is the cardinal, ordinal, and nominal numbers. So, first of all, let us talk about cardinal and ordinal numbers. Cardinal numbers are those where the actual numerical values that a particular variable takes have meaning. For example, cardinal numbers would be the price of a building or the number of houses in a street.

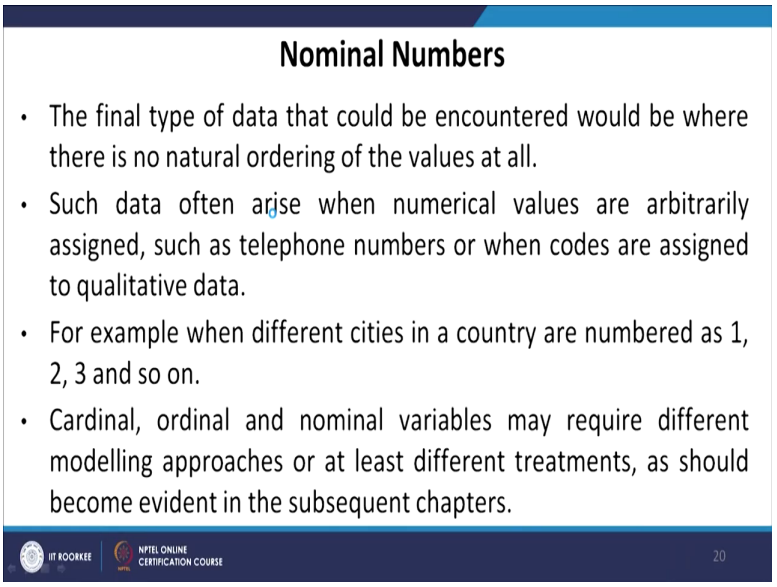
So here, this is something very interesting that the numbers have some meaning, It will become evident, what it means when we talk about ordinal numbers. Ordinal numbers can only be interpreted as providing a position or an ordering. For example, ordinal numbers would be the position of a runner in a race, for example, the second position is better than the fourth place, but it would make little sense to say it is twice as good. So, it just talks about the position.

For example, if I am asked to express my preference for three types of fruits, let us say mango, banana, and grapes. I can always say that I prefer mangoes, then followed by grapes, and then by bananas. So, the rankings are mango 1, grapes 2, bananas 3, but that does not mean that does not talk about how much more I prefer mangoes over grapes or bananas. It is quite possible that the preferences are not very strong.

So, mangoes are preferred slightly over grapes, and slightly over bananas. It is also possible that the preferences are very strong. So, I love mangoes as compared to that I hardly like grapes and bananas. But ordering 1, 2, 3 itself does not talk about how strong the likings or the preferences are.

Whereas, when I talked about cardinal numbers, the numbers have meaning it implies that for example, it is mentioned here that the number of houses in a street is a cardinal number. So, if a street has 50 houses and another street has 100 houses, then we can always say that the second street has double the number of houses compared to the first street. So, here the numbers have a very specific meanings which is not the case with ordinal numbers.

(Refer Slide Time: 34:06)

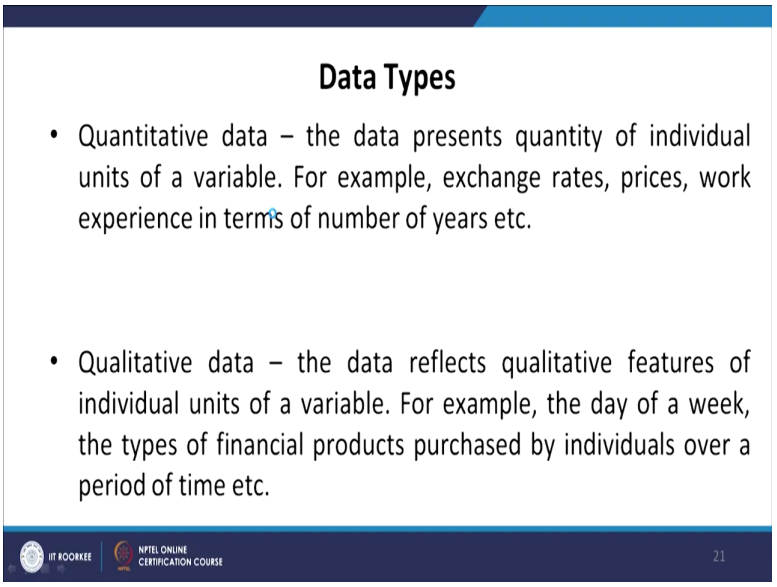


And finally, we have nominal numbers. The final type of data that could be encountered would be where there is no natural ordering of the values at all. Such data often arise when numerical values are arbitrarily assigned, such as telephone numbers, or when codes are assigned to qualitative data. For example, when different cities in our country are numbered as 1, 2, 3 and so on.

We can have other examples also like we can collect data on the profession of individuals and then the professions can be categorized as numbers for analysis. So, those who are in white-collar jobs, are given number 1, blue-collar jobs are given 2 and then manual laborers are given 3, agricultural farmers are given 4, and so on. So, these are some categories where the numbers do not reflect anything specific, not even any order. So, they are nominal numbers,

Cardinal, ordinal and nominal variables may require different modeling approaches or at least different treatments, as should become evident in the subsequent chapters or modules when we discuss models specific to this kind of data variable or number.

(Refer Slide Time: 35:23)

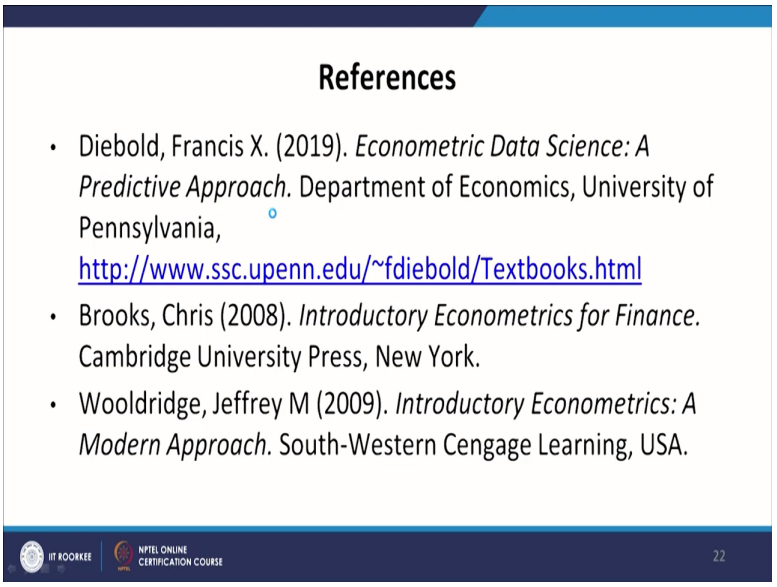


And finally, we have another classification that is prevalent or possible or relevant for all other types of data that have been discussed just now, and that is qualitative and quantitative data. So, quantitative data is the data that presents a quantity of individual units of a variable. For example, exchange rates, prices, work experience in terms of the number of years, etc. While qualitative data reflects, qualitative features of individual units of a variable.

For example, the day of a week, the types of financial products purchased by individuals over a period of time, etc. So, on the day of a week, for example, I want to find out that what are the days on which maximum trading takes place in stock exchanges. We can consider trading volume maybe. And then against that, we record them for five days a week. And when we record the days, then Monday, Tuesday, Wednesday, Thursday, Friday, cannot get into the analysis where you would be using numbers against them.

So, converting them into, say, nominal numbers and then Monday 1, Tuesday 2, and so on, we have five days. So that is an example of qualitative data, as opposed to quantitative data, which are close to cardinal numbers where the numbers themselves give some information in terms of, very specific values, and it has, the numbers themselves has meaning to associated with it. So, that is all about the data types.

(Refer Slide Time: 37:06)



These are the references. Thank you.