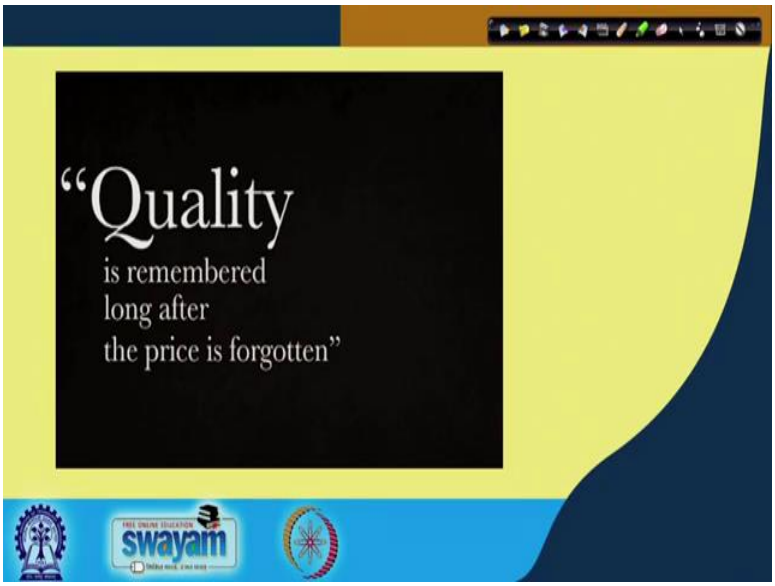
Six Sigma
Prof. Jithesh J Thakkar
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture – 18
Data Collection and Summarization (Part 1)

Hello friends once again I welcome you to the journey of Six Sigma and now we are moving ahead with a new topic in our typical DMAIC define, measure, analyze, improve and control cycle. So, before we actually say talk about today's topic that is data collection and summarization, this topic is divided into two part; part one and part two and lecture 18 is a basically part one on data collection and summarization. So, before we move on this, let us try to appreciate one very good quality quote, quality is remembered long after the price is forgotten.

(Refer Slide Time: 01:07)

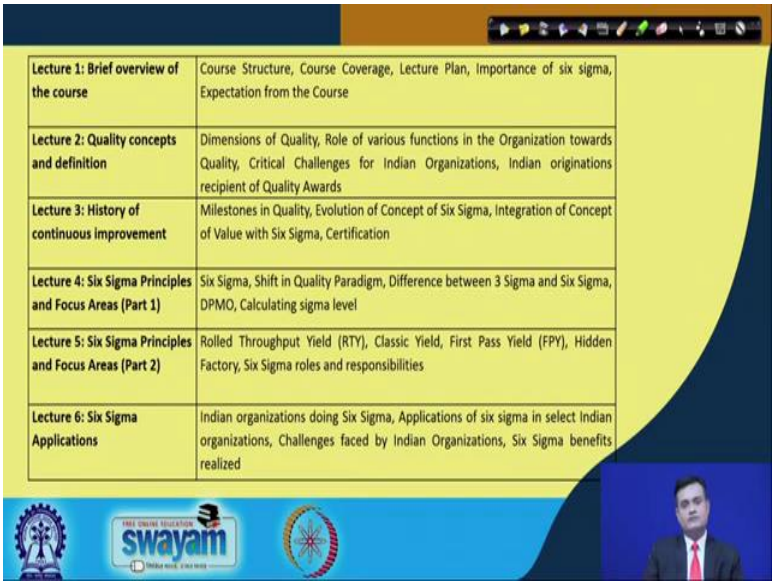


So, this is a very good quote from the view point that quality is an experience.

So, if you visit a restaurant, if you visit hotel or if you go to bank then many a times fine they have charged for certain services we may forget, but the quality is always remembered and in the form of experience it remains in our mind. So, then this leads to word to word, mouth to word publicity and you become the agent of the organization and you market this services or the products.

So, before we move on the discussion on lecture 18 that is data summarization let us try to have a recap and let us appreciate where we are in our six sigma journey. So, if we quickly see then first we started with quality fundamentals and key concepts. Week 1 and week 2 where devoted on the discussion of various topics related to quality, importance, basics and fundamentals.

(Refer Slide Time: 02:40)



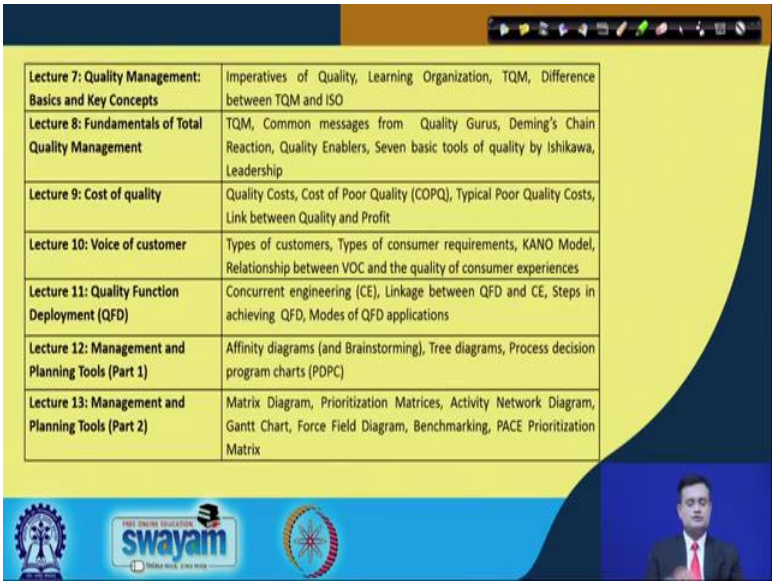
Week 1, we have talked about brief overview of the course. Why this course is important and what you can expect from this course, quality concepts and definitions. So, why quality is important and in the competitive world, how organization can appreciate the importance of the quality and what are the Indian organization they are really striving for quality excellence.

Then we have seen the detailed history of continuous improvement starting from inspection and then moving on to statistical quality control, quality assurance and certification and then say lean six sigma and then lean six sigma. We have also seen some of the critical concepts as a part of lecture 4, six sigma principles and focus areas part 1 and six sigma in focus areas part 2 as lecture five and we talked about what is the difference between three sigma and six sigma.

Then what is DPMO, how to calculate how to calculate the sigma level and judge the present level of quality and there are many other concepts like rolled throughput yield, FPY- First Past Yield and hidden factory and other concepts. We have also seen in

lecture six some six sigma applications where we have just studied couple of Indian organizations which has implemented the six sigma and realized the various benefits.

(Refer Slide Time: 04:17)



In week 2, if you see in lecture 7 we talked about quality management basics and key concepts. So, you cannot really implement any continuous improvement program like six sigma unless you have a proper organizational culture and a system where such kind of initiatives can really flourish. So, we have seen and the quality management principles and TQM and then in detail in lecture 8 we have seen the fundamentals of TQM, the views of various quality gurus their theories like Crosby, Juran, then Ishikawa, Feigenbaum and many others.

Lecture 9 was devoted on cost of quality and we appreciated that there is a cost of prevention, there is a cost of appraisal, internal failure cost, external failure cost and unless you measure the quality in terms of cost it is not possible to convince the management for any quality initiatives like six sigma. Then we had seen voice of customer this is the most important thing because you may be applying six sigma and you may think about the process improvement, but if your process is not customer centric if your product is not customer centric then whatever effort you will put that will simply go waste.

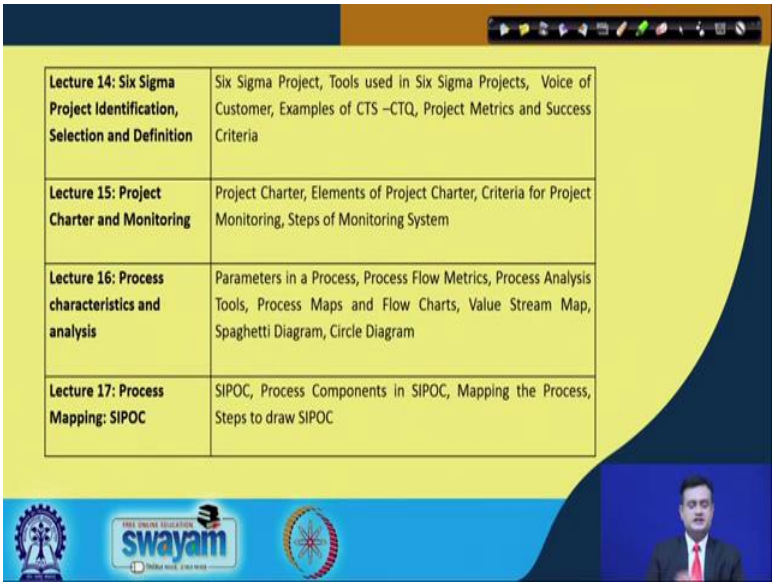
So, it is important to understand customer and capture the voice and we talked about it. Subsequently, we had seen as a part of lecture 11 quality function deployment. So, it is a

cross functional technique concurrent engineering and we try to see that quality can be built right at the design stage. So, we have seen house of quality and many other examples.

Lecture 12 management and planning tools. So, we have seen affinity diagram, tree diagram, process decision, programming charts and subsequently in lecture 13 management and planning tools part 2 we have seen few more tools and approaches to analyze the situation and identify the root cause for which really the six sigma effort is required. So, these are as included matrix diagram, prioritization matrices, activity network diagram, Gantt chart, Force field diagram, benchmarking, PACE prioritization matrix and so on.

Now, in week 3 we have advanced in our journey and we basically dealt with the first phase of my DMAIC cycle and that is the define phase.

(Refer Slide Time: 07:10)

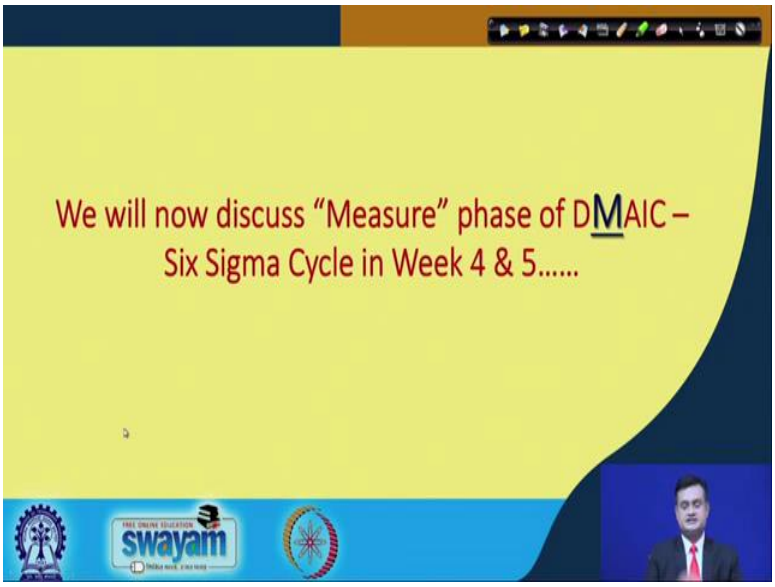


So, we have seen that what is the importance of six sigma project identification, selection and definition and what are the various CTX that could be critical to quality, critical to serviceability, critical to delivery and many others. Then the lecture fifteen we talk about project charter and monitoring, setting the objective, scope responsibility matrix, expected deliverables and other things.

Process characteristics and analysis we conducted in lecture 16 and we had seen couple of approaches like flow diagram, process flow matrix, process analysis tools, flow chart, values stream map, spaghetti diagram and circle diagram. Lecture 17 we devoted on SIPOC because I do not want to confine only to my organization, it is also important to see the implications of various phases in the value chain supply chain on each other and if we can try the various matrices synchronize them then really a value can be generated for the customer.

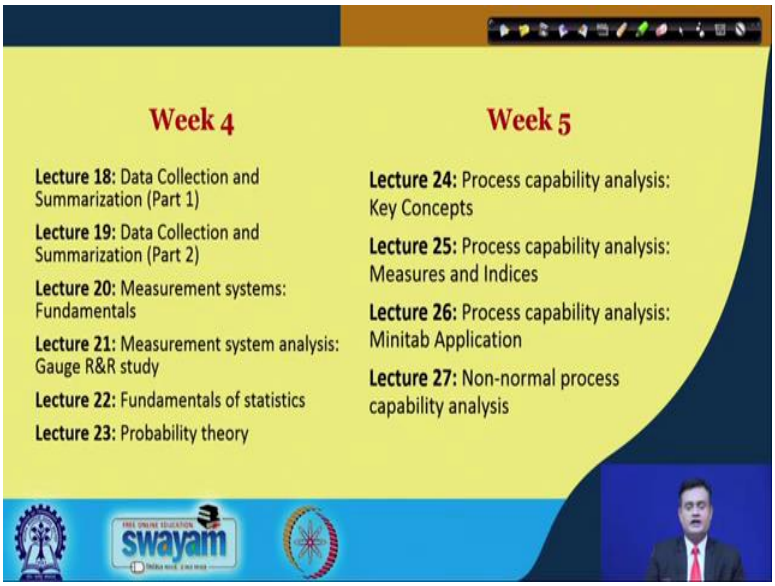
So, SIPOC is an approach which basically help you to understand, define, quantify the entire value chain, supply chain from supplier your internal, then process, then your operations and customers. So, this particular approach really helps to analyze the complete value chain.

(Refer Slide Time: 08:51)



So, we have seen this SIPOC in lecture 17. Now let us move ahead in our journey of DMAIC and we will now discuss the measure phase of DMAIC that is the second phase and this phase we will cover in week 4 and week 5.

(Refer Slide Time: 09:14)



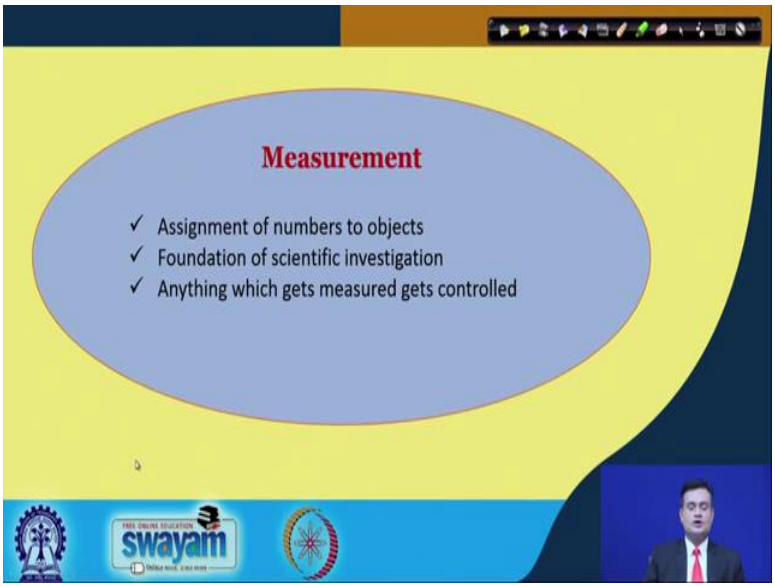
So, our plan for week 4 and week 5 is like this. Week 4, we will have lecture 18 on data collection and summarization basically this lecture part 1, lecture 19 part 2 of the same, lecture 20 measurement system fundamentals, 21 measurement system analysis gauge R & R study, 22 fundamentals of statistics and 23 probability theory.

Week 5 we will deal with some of the important topics like process capability analysis key concepts, process capability analysis measures and indices, process capability analysis some Minitab application and lecture 27 non normal process capability analysis. So, we are also trying to include in measure analyze phase wherever possible Minitab application so that you can conduct the analysis easily for more number of variables or factors and specially as it gives you very good display of the graphs and charts and tables you only need to interpret them appropriately and industry people can easily make use of this researchers, people engaged in the project can easily make use of this software and conduct the analysis.

(Refer Slide Time: 10:40)

So, now let us talk about this lecture which is lecture 18 on data summarization. So, we will focus on variables and measurement scales, data collection methods and population and sampling.

(Refer Slide Time: 10:58)

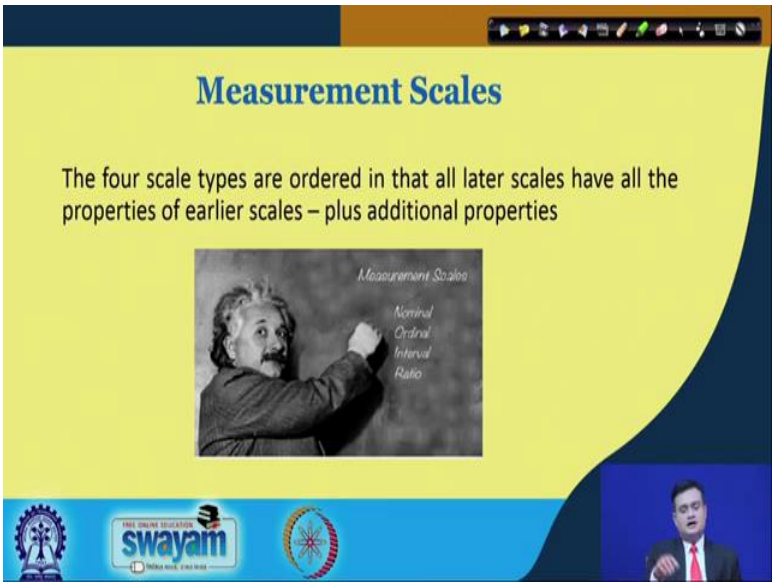


So, measurement the word is very well known to all of us, but what actually it means. So, measurement means assignment of numbers to the objects. So, there is a well saying that you cannot control if you cannot measure.

So, anything if you have to control monitor you have to measure and for that say we have to assign the numbers. Sometimes the phenomena may be subjective, but we try to capture it on a scale of 1 to 5 or 1 to 7 and then try to quantify the various issues so that, measurement and subsequent analysis becomes possible.

Foundation of scientific investigation as I said unless you have some quantified values you cannot further process it and you cannot apply the various statistical tools for gaining the deeper inferential insights. Anything which gets measure gets controlled as I said and it is important that we can have a proper measurement system by which the phenomena can be appropriately captured.

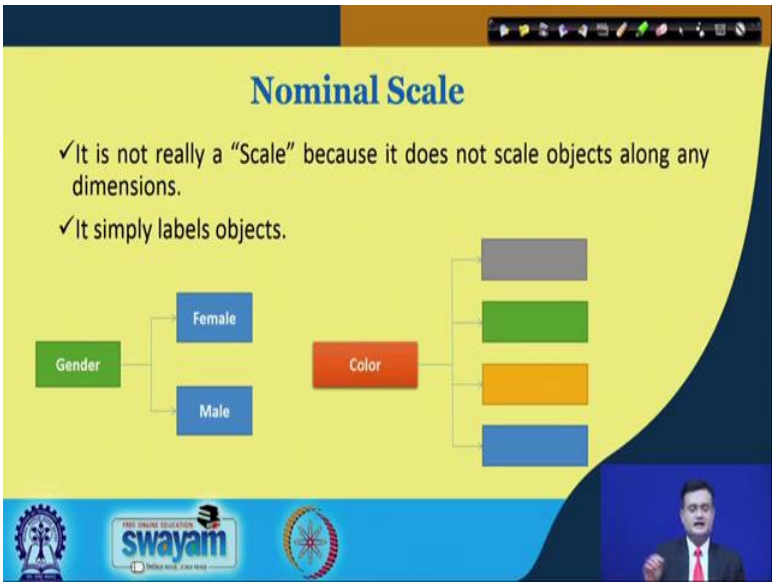
(Refer Slide Time: 12:17)



So, there are various ways or rather scales to measure the phenomena capture the phenomena and there are basically four scales and this types are ordered in that all later scales have all the properties of the earlier scales. So, if you move from nominal the four scales: the nominal, ordinal, interval and ratio then your nominal covers some portion when you go to ordinal it covers all the features of nominal in addition to that it covers something else.

So, that way the various scales are developed and depending upon the need of measurement you make use of the various scales.

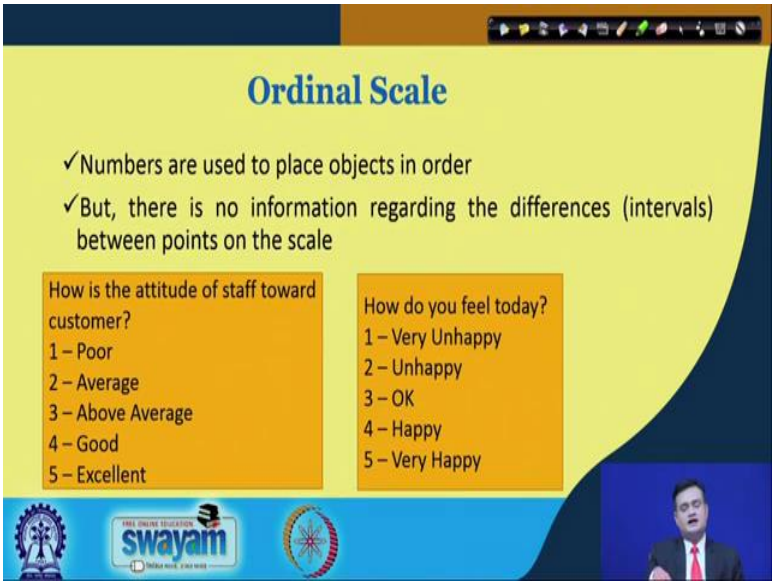
(Refer Slide Time: 13:08)



So, we will see in detail. So, let us see nominal scale it is not really a scale because it does not scale objects along any dimensions please remember. It is simply a leveling kind of approach it levels the objects.

So, just see that, if I take the example of gender then female and male. If I take the example of color fine I may have grey color, green color, orange color, blue color. So, here I am just trying to categorize, I am just trying to label the various categories based on say some classification and if I do this it is called nominal scale.

(Refer Slide Time: 13:52)

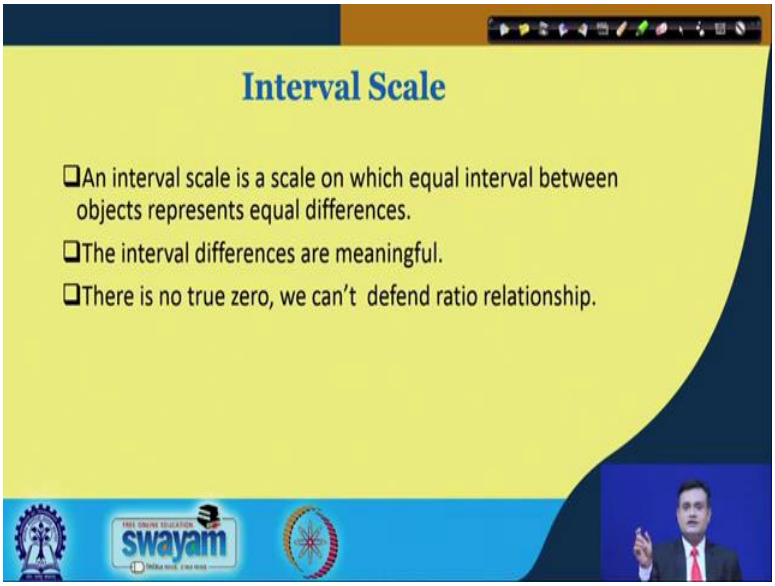


The another scale which is little advanced compared to my nominal scale that is the ordinal scale. Here actually I am using the numbers to place objects in the order. So, whatever phenomena I am studying I want to capture the order. For example, if you see the first box how is the attitude of staff towards customers.

So, I have to have way means by which I can capture and here it is one means poor, two means average, three means above average, four means good and five means excellent. So, ordinal scale basically places the numbers and then you are trying to measure a particular object or the phenomena. Then, the another box you can see how do you feel today. So, fine I am giving you choices one very unhappy, unhappy, ok number three four happy and five very happy.

So, likewise you can put the numbers and this number will help you to quantify that ok as on today what is the rating for a particular situation or the object.

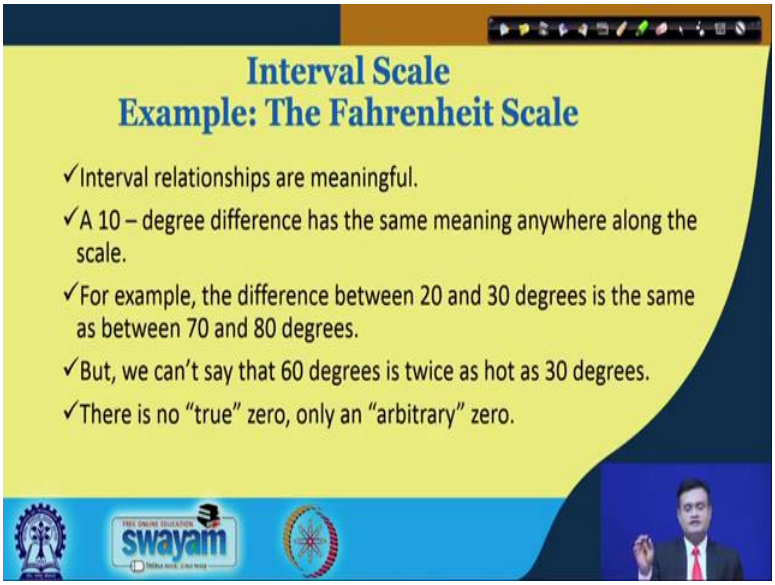
(Refer Slide Time: 15:17)



Then you have an interval scale which is widely used and you have seen. So, interval scale typically it is a scale on which equal interval between objects represent equal differences. The interval differences are meaninful there is some scientific basis and there is no true zero please remember, there is no true zero we cannot defend ratio relationship ok.

So, if I put b by a and then if I put c by b we cannot really defend the ratio relationship here there is no true zero, but through interval it defines the relationship.

(Refer Slide Time: 16:00)

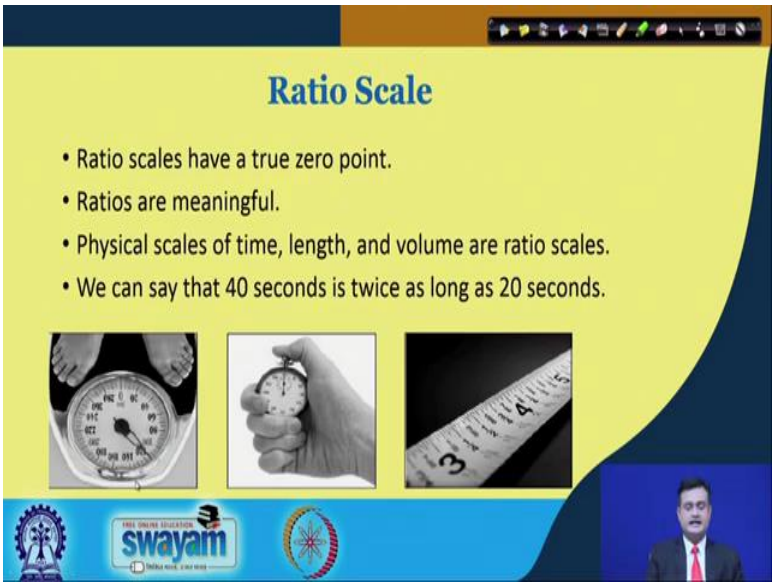


So, if we just see some examples to make it very clear then interval relationships are meaningful on scientific basis a 10 degree difference has the same meaning anywhere along the scale. So, you are using let us say thermometer and you have a scale in Fahrenheit or may be centigrade and then 10 degree difference now for example, difference between 20 and 30 degree is the same as between 70 and 80 degree.

So, there is an increase in temperature of 10 degree, but we cannot say please remember there is a caution we cannot say that 60 degree is twice as hot as 30 degree. There is no ratio relationship. Interval is fixed fine 40 to 50, 10 degree increase, 70 to 80, 10 degree increase this is understood, but the moment I say that 60 degree is twice as hot as 30 degree, this is not correct.

So, ratio relationship is not valid and the reason mainly is there is no true zero it is only an arbitrary zero. So, this is the property of interval scale.

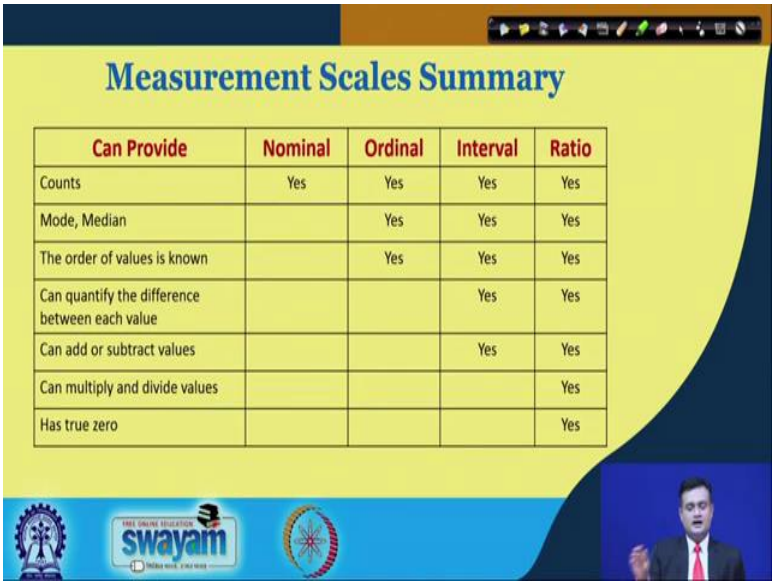
(Refer Slide Time: 17:21)



Now, the last one if you see it is a ratio of scale. So, this is the scale which has a true zero. You use this scale widely your stop watch, your scale measurement at everywhere you are using basically the ratio scale it has a true zero, zero means it is a zero distance it is a zero weight whatever you say zero minute and this is a true zero.

So, ratios are meaningful and physical scale of time, length, volume are ratio scales and we can say that 40 second is twice as long as if you compare it with the 20 second. This is not possible as I mentioned in the interval scale.

(Refer Slide Time: 18:10)



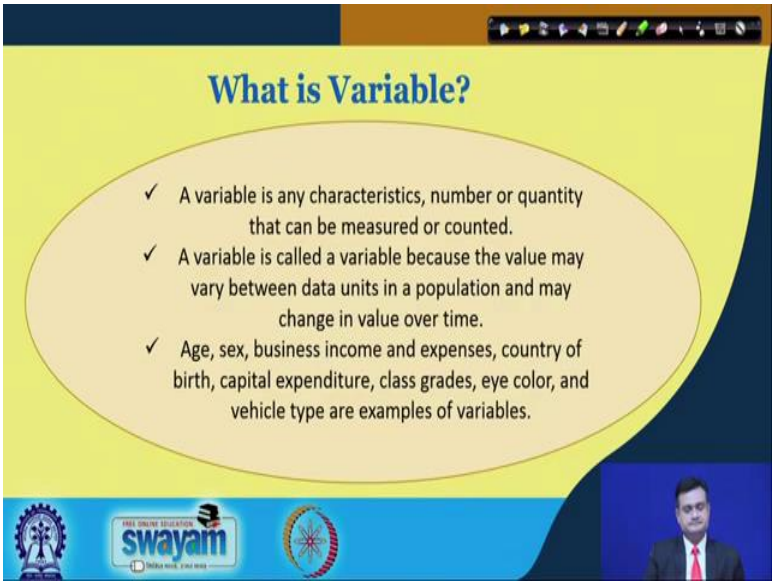
Can Provide	Nominal	Ordinal	Interval	Ratio
Counts	Yes	Yes	Yes	Yes
Mode, Median		Yes	Yes	Yes
The order of values is known		Yes	Yes	Yes
Can quantify the difference between each value			Yes	Yes
Can add or subtract values			Yes	Yes
Can multiply and divide values				Yes
Has true zero				Yes

Now, if we just see the summary and this will really help you and as I mentioned initially that as you advance in the scale from nominal to ordinal, ordinal to interval, interval to ratio each scale is able to cover the features of the previous scale in addition to some more features.

So, if we see counts nominal yes, if we see mod median then this part is true for ordinal interval and ratio count is yes for all the case, if you see the order of value is known again this is true for ordinal interval ratio not for nominal then can quantify the difference between each value, as we have seen not possible with the ordinal, but interval and ratio it is possible can add and subtract value, yes for interval and ratio can multiply and divide value, interval ratio logic is not valid. So, here this is only possible in case of ratio scale.

So, has true zero; obviously, it is a ratio scale. So, here please do not misunderstand that because ratio scale covers everything that is the best scale again I would say that it all depends upon the nature of phenomena the object you want to measure and the particular phenomena which is under investigation.

(Refer Slide Time: 19:41)



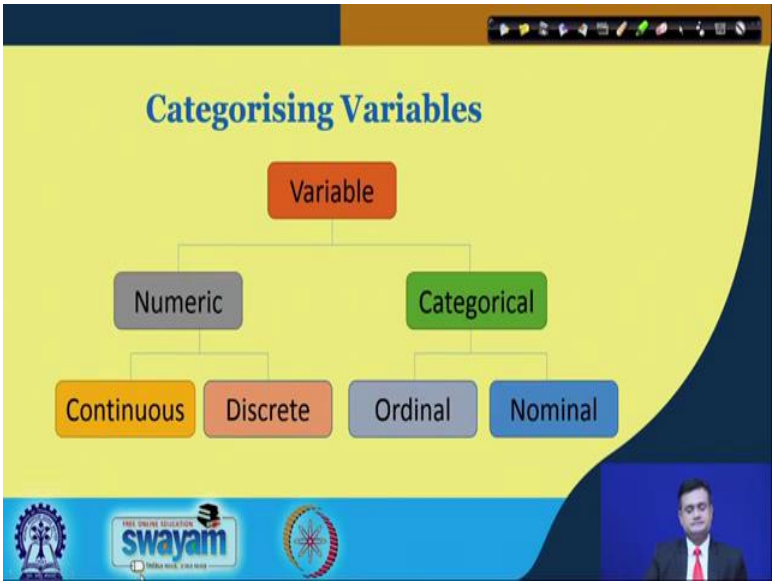
So, I think it would be very preliminary to say, but let us try to appreciate what is variable because we are now discussing the measure phase and subsequently we will see the analyze if we are not through and thorough in our measure phase then analyze phase

will not carry much value that will follow just a simple principle GIGO garbage in garbage out.

So, measurement phase is extremely important once you have defined the project and then once you have the measurement, good measurement, quality measurement you can really opt for the analyzed phase. So, a variable is any characteristic number or quantity that can be measured or counted and variable is called a variable because the value may vary between data units in a population and may change over a period of time.

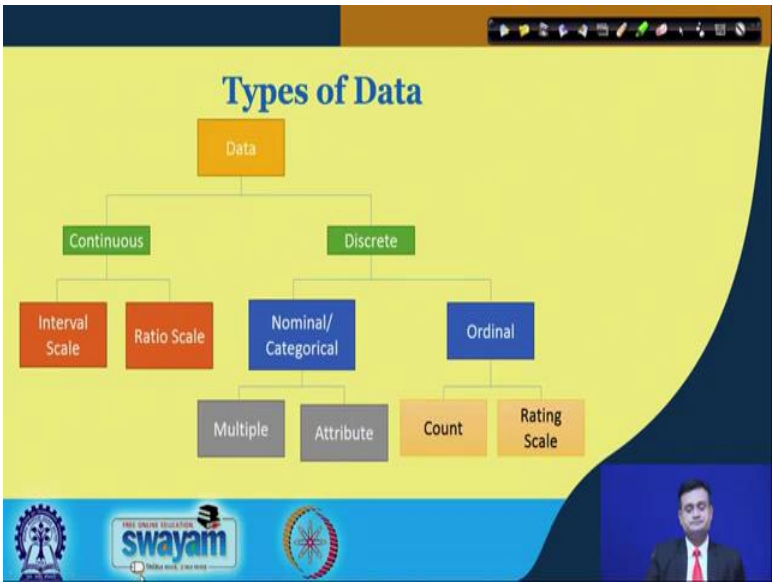
Some examples; age, sex, business income expenses, country of birth, capital expenditure, class grades, eye color, vehicle type, many examples you can think of basically are of variables.

(Refer Slide Time: 20:59)



Broadly there are two categories of variables one is numeric other is categorical and when you say numeric again it could be of continuous and it could be of discrete. Discrete means a point value and continuous means there could be a decimal value and there is an interval within that you can measure the different values. Categorical there is ordinal and nominal. So, we have already seen the scale also.

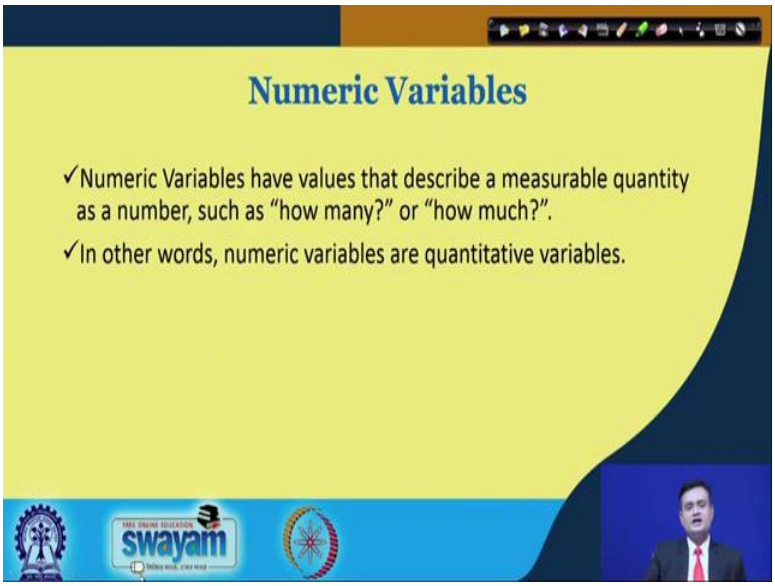
(Refer Slide Time: 21:28)



So, if you have to really analyze the data then, your data could be seen into two categories continuous and discrete as I mentioned and if you are dealing with the continuous data then, the appropriate scale would be interval and ratio scale. If you are dealing with the discrete data what is the gender male or female. Then for you it is the nominal or categorical data or ordinal data.

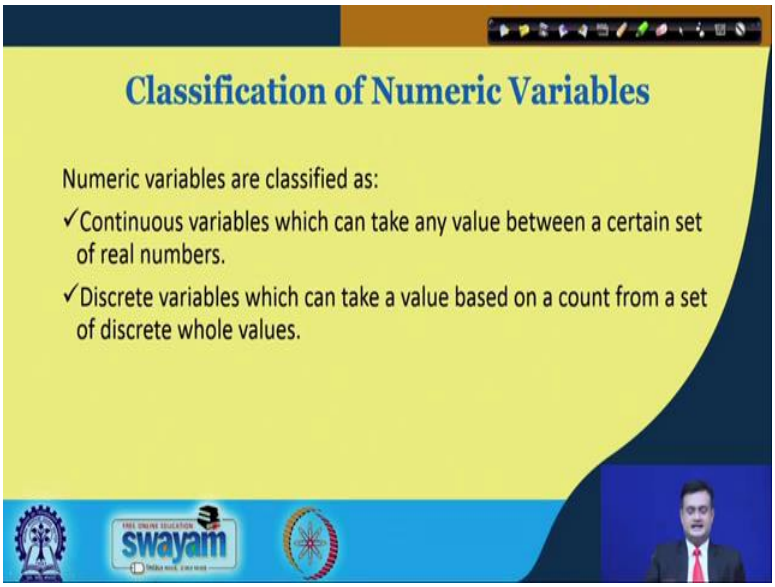
You may have nominal or categorical, multiple or attribute, ordinal count or rating scale we are not going into this detail, but just I am providing the classification that these are the various possibilities.

(Refer Slide Time: 22:15)



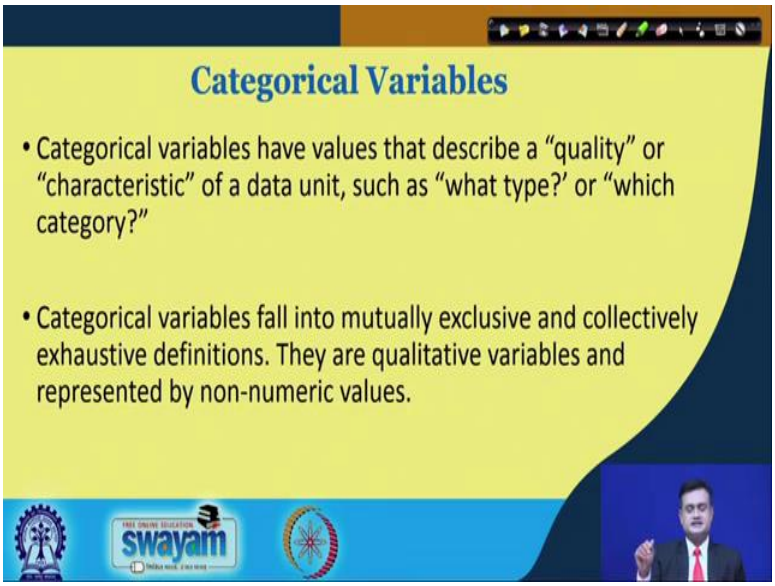
Numeric values basically that describes a measurable quantity as a number such as how many or how much or in other word we can say that numerical variables are quantitative variables you quantify it and you can do the further analysis.

(Refer Slide Time: 22:38)



So, you have the broader classification as I mentioned you have continuous variable and this variable can take any value between a certain set of real number and discrete variable which can take a value based on count from a set of discrete whole values. So, you have two broad categories one is continuous other is discrete.

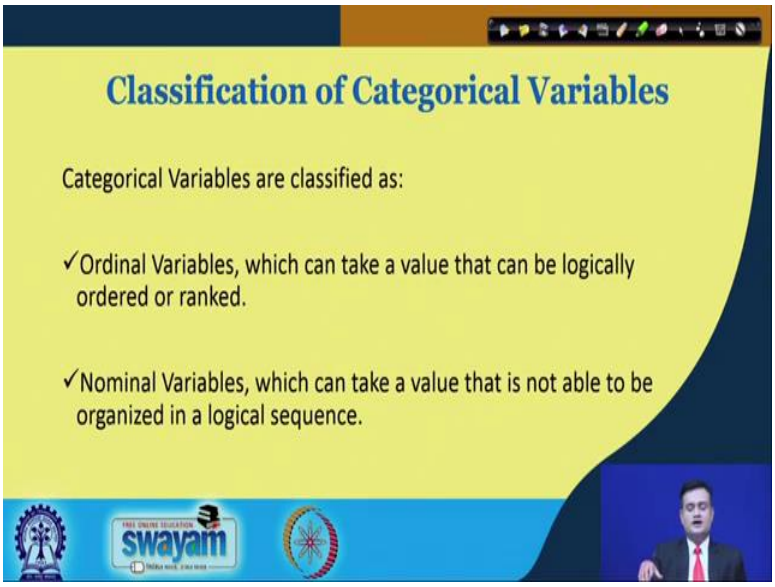
(Refer Slide Time: 23:04)



You have categorical variables. So, basically have values that describe a quality or characteristic of a data unit such type what type it belongs to what category or which category. So, when you are dealing with such kind of phenomena, basically you are trying to capture the categorical variable.

Categorical variable they fall into mutually exclusive and collectively exhaustive definition and they are qualitative variables and represented by non-numeric values.

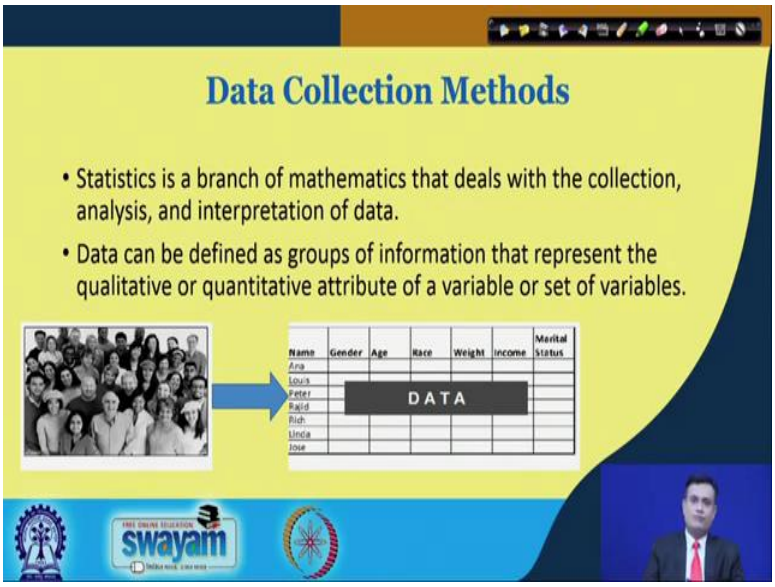
(Refer Slide Time: 23:41)



We can classify the categorical variables also and these are ordinal variables, nominal variables. Ordinal variables take value that can be logically ordered or ranked we have seen and nominal variables take a value that is not able to be logically organized in a sequence, but it just helps you to classify.

There are various data collection methods and let us try to appreciate the importance of couple of them.

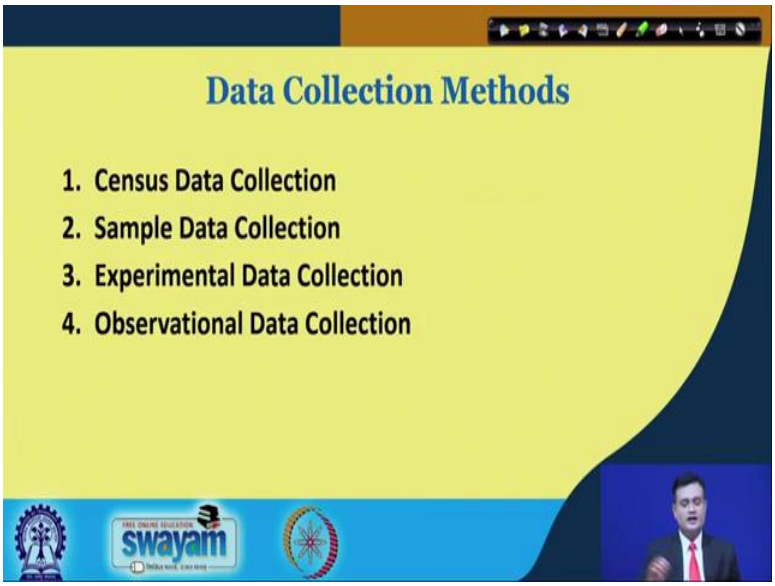
(Refer Slide Time: 24:15)



So, what is the importance of data collection method? See statistics we use extensively to analyze the phenomena and to draw conclusions which are more concrete based on inferential analysis. So, statistics is a branch of mathematics that deals with the collection analysis and interpretation of data and as I mentioned that analysis is only possible if you have appropriate good quality measurement of the object.

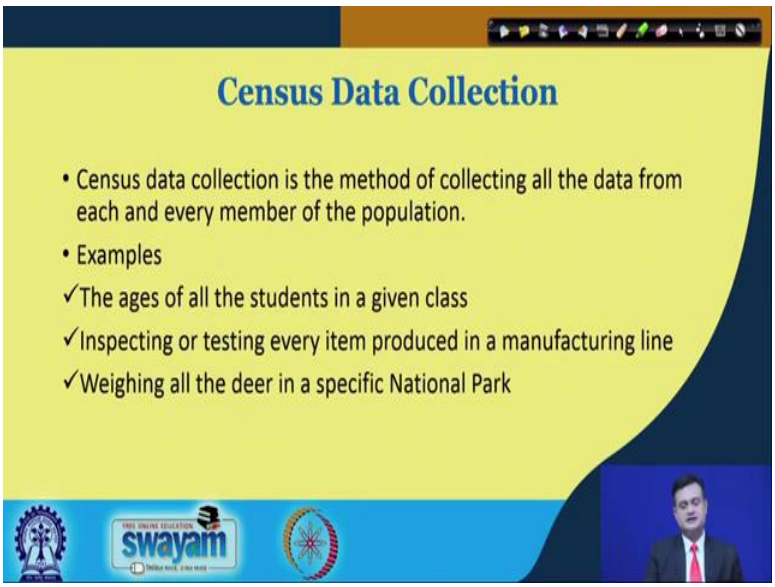
So, data can be defined as groups of information that represent the qualitative or quantitative attribute of a variable or may be set of variables.

(Refer Slide Time: 25:02)



We have different data collection methods; census data collection, sample data collection, experimental data collection and observational data collection.

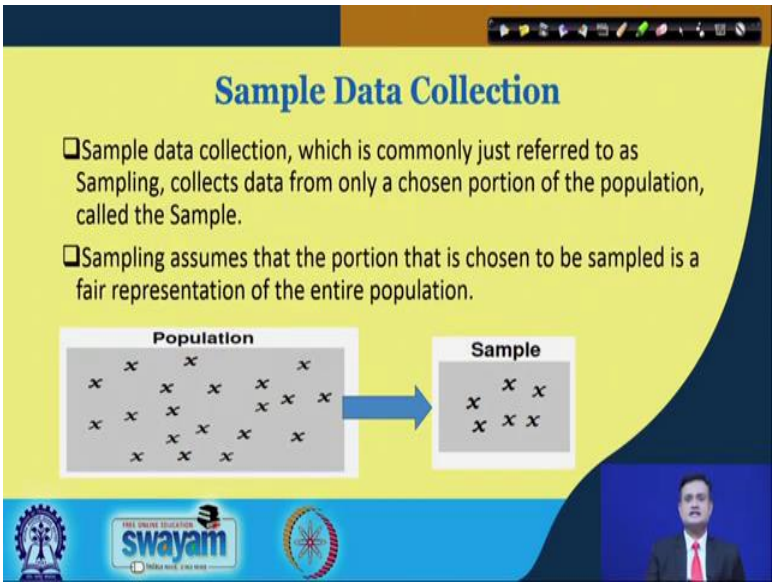
(Refer Slide Time: 25:13)



So, when you go for census data collection you are trying to collect all the data from each and every member of the population.

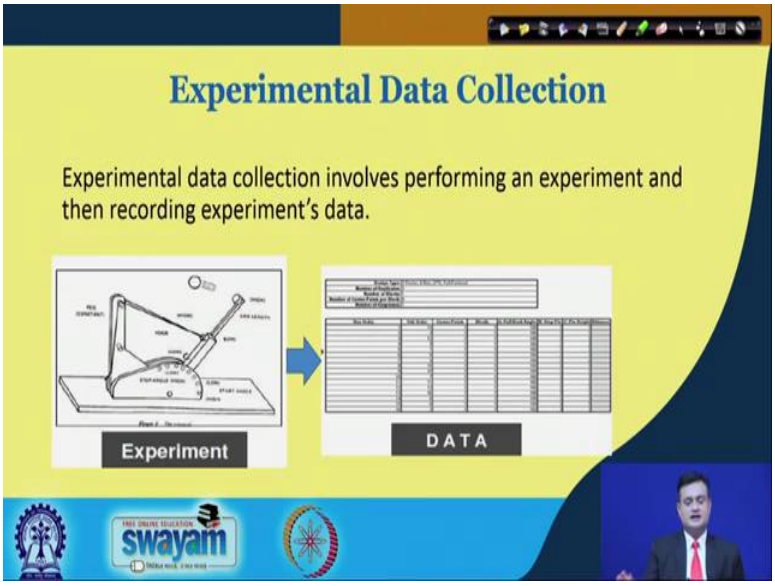
So, we may say couple of example that age of all the students in a given class, inspecting or testing every item produced in a manufacturing line and may be weighing all the deer in specific national park. So, this is your census related data.

(Refer Slide Time: 25:40)



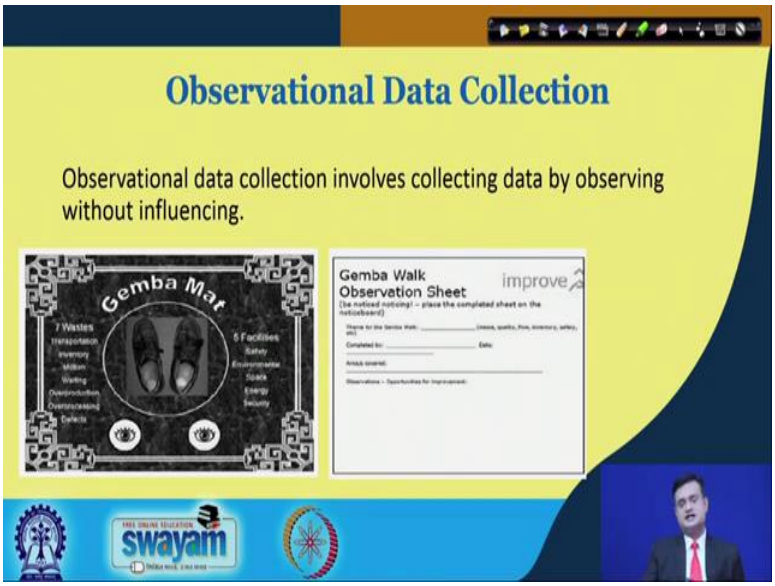
Second one which is very important that is sample data collection. So, it is not possible for me to collect all the data of the population, but I can draw a sample how to draw a sample we will see and based on my sample which is a representative of my population I can do the analysis.

(Refer Slide Time: 26:06)



There is an experimental data widely used in all engineering say branches. So, you conduct the experimentation, tabulate your data in a specific format and based on this you try to derive certain conclusions.

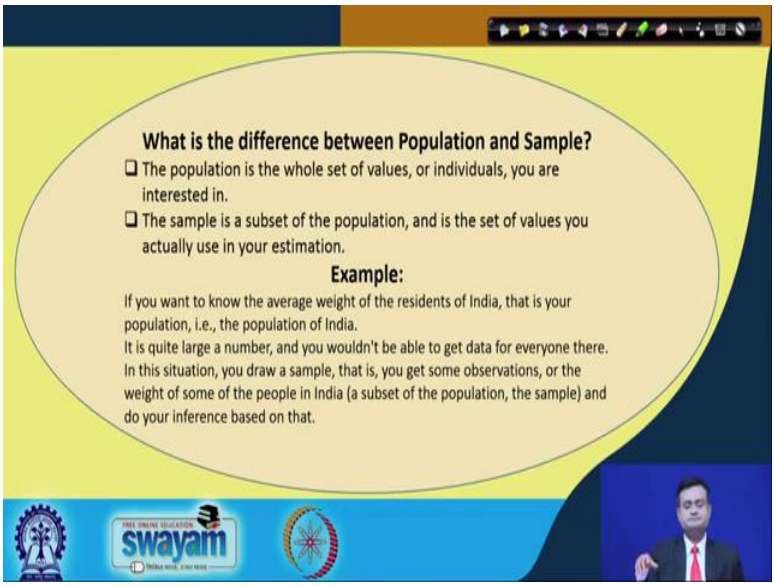
(Refer Slide Time: 26:24)



You have observational data for example, in lean we take about Gemba walk. So, here I will just walk through the company or plant and try to take the observation. For example, you are visiting a super market and just you are trying to observe how the things are arranged how the products are displayed or how the sales people they are behaving and then you are trying to collect some data on some scale or in a verbatism then this is your observational data collection.

Now, let us see sampling as I mentioned, statistics basically do the inferential analysis based on the sample because I cannot deal with the population, it is extremely cumbersome and costly and hence I try to take the sample and based on the analysis of the sample I try to make the inferences.

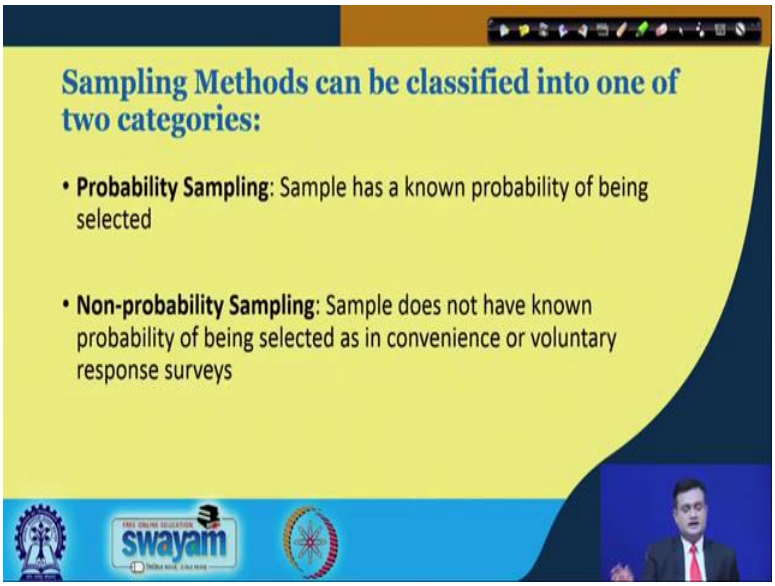
(Refer Slide Time: 27:23)



So, the difference between population and sample is very simple population is whole or individuals you are interested in. So, suppose you want to study let us say the reasons behind HIV let us say fine you can say approach each and every person in the country and then try to analyze, but that is not possible. So, you take the representative sample and based on that you are trying to draw certain conclusions.

So, sample basically is a subset and the word which is very important it represents the population how it represents we will see. So, examples could be if you want to know the average weight of the resident of India that is the population then it is quite large number and it is not possible for you. So, you may take some sample you may divide entire country into some zones and then you can take the sample or you can take the sample randomly for a particular region.

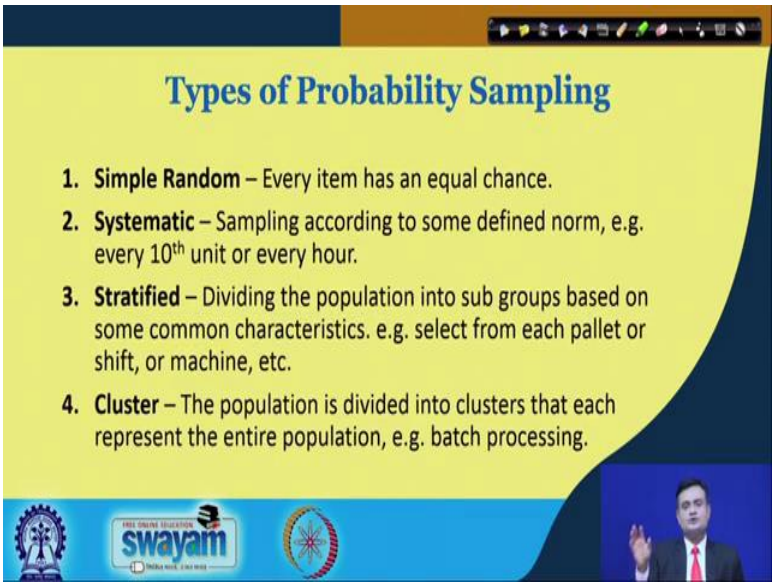
(Refer Slide Time: 28:30)



Sampling method broadly you can classify into probabilistic sampling and non probabilistic sampling. In probability sampling say sample has a known probability of being selected from the population. In case of non probability sampling, sample does not have known probability or being selected in convenience or voluntary response survey.

So, I am I am just trying to do the survey based on my convenience or based on my comfort and then I would only end up with those units those sample which are easily approachable to me and then it is not the non probability sampling in majority of this statistical analysis we mainly use probability sampling. So, let us try to see what are the widely used probabilistic sampling approaches.

(Refer Slide Time: 29:20).

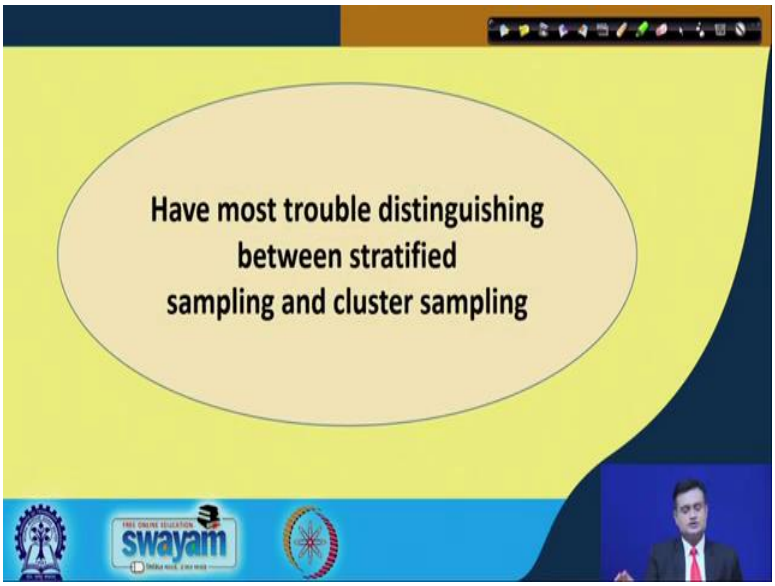


So, number one is simple random every item has an equal chance it is the widely most widely used and I just try to draw the sample suppose you have produced let us say one batch of production during the first shift now, there are 1000 bearings I will just try to draw the sample of desired quantity and then analyze the various features measure the various features may be outer diameter, inter diameter, inner diameter, thickness, surface finish and then I will try to conduct the quality analysis.

Systematic here I will say sampling is done according to some defined norm. Example every tenth unit or every hour. So, instead of taking a sample from the batch what I will do I will inspect the production every 3 hour and then I will try to derive certain conclusions about the quality of the production. Stratified sampling here I divide the population into subgroups based on some common characteristic and select from each pallet or shift or machine etc. Cluster sampling here the population you divide into clusters that each represent entire population.

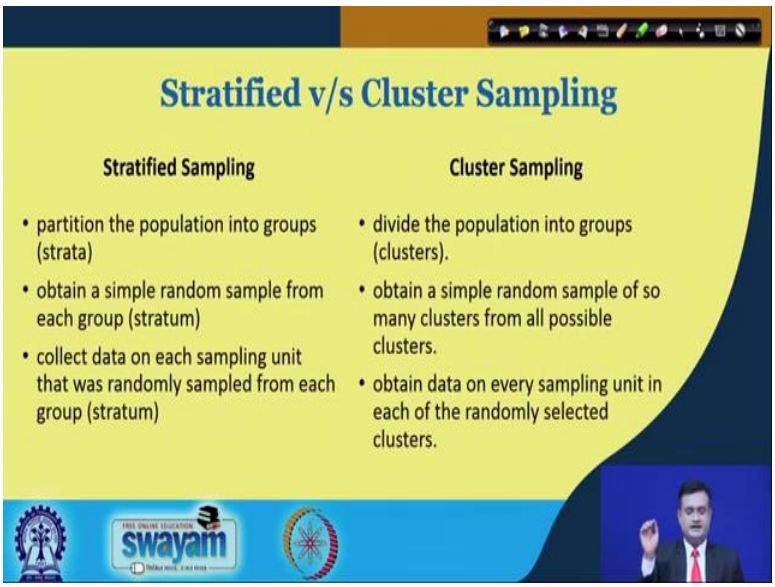
So, suppose if you look at India, then you have eastern zone, western zone, north zone and suppose you want to select the best players in for the cricket team then fine you may try to divide it into zone and then you conduct the investigation. There are certain issues or difficulties in differentiating stratified and cluster sampling.

(Refer Slide Time: 31:19)



So, there is some trouble students they have. So, I will try to help you with example and say differentiation that how these two techniques are really different.

(Refer Slide Time: 31:31)

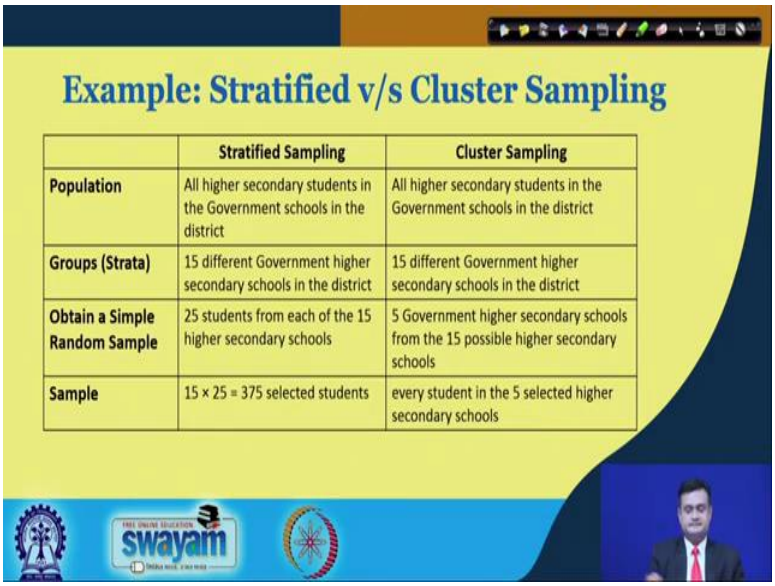


So, when we say stratified versus cluster sampling in stratified sampling, you partition the population into groups. In cluster sampling you divide the population into clusters that is the groups. Here, obtain a sample in case of stratified sampling from each group remember that is a stratum. In case of cluster sampling obtain a simple random sample of

so many cluster. So, here it is about taking a sample from so many cluster from all possible cluster.

In stratified sampling collect data on each sample unit that was randomly sampled from each group. So, you have selected obtained random sample from each group or stratum and then from that you are trying to collect a random sample. In case of cluster, obtain data on every sampling unit in each of the randomly selected cluster. So, please understand the difference in cluster sampling, I am selecting the cluster randomly and once the cluster is selected I obtain the data, collect the data on every sampling unit every member in that particular cluster which is randomly selected.

(Refer Slide Time: 32:57)



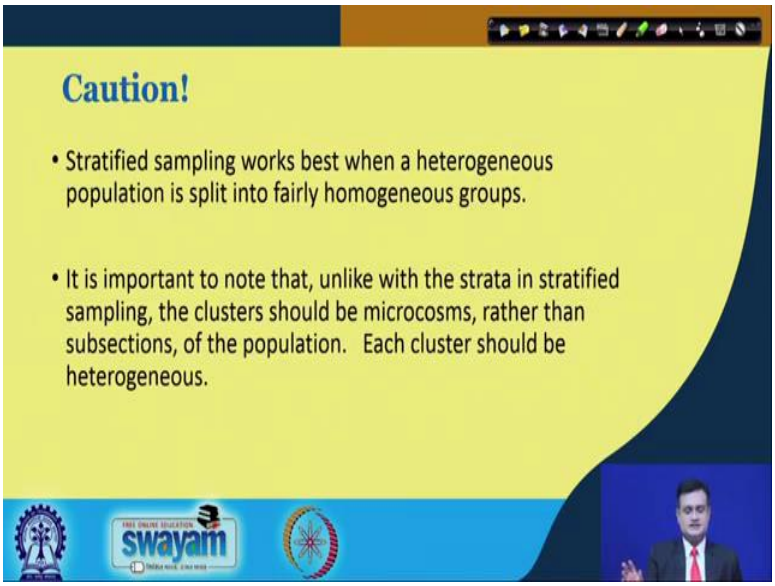
I think this example will help you to make your idea better clear. Let us say stratified sampling and cluster sampling and here I want to select some of the Government higher secondary schools from a particular district. So, let us say and; obviously, from that I want to select some students may be in order to nominate them for some competition or whatever.

Now, population in case of stratified and cluster both is same that is all higher secondary students in the Government school in the district. So, it is same for both group or strata here again it is same I am saying 15 different government higher secondary school in the district for cluster sampling again it is same. Now the difference comes say obtain a simple random sample. In case stratified sampling, 25 students from each of the 15

higher secondary schools. So, each school government higher secondary school I will take 25 students randomly. So, the total sample will be 15 into 25 that is 375. If you see the cluster sampling about obtain obtaining a simple random sample what I will do I will select clusters randomly.

So, I have 15 different government higher secondary schools which are the cluster or groups I will select only 5 let us say randomly out of this 15 and then from these 5 every student in this 5 will be considered as a unit sampling unit for the measurement. So, that is the difference that once the cluster is selected randomly each unit will be measured in case of cluster sampling.

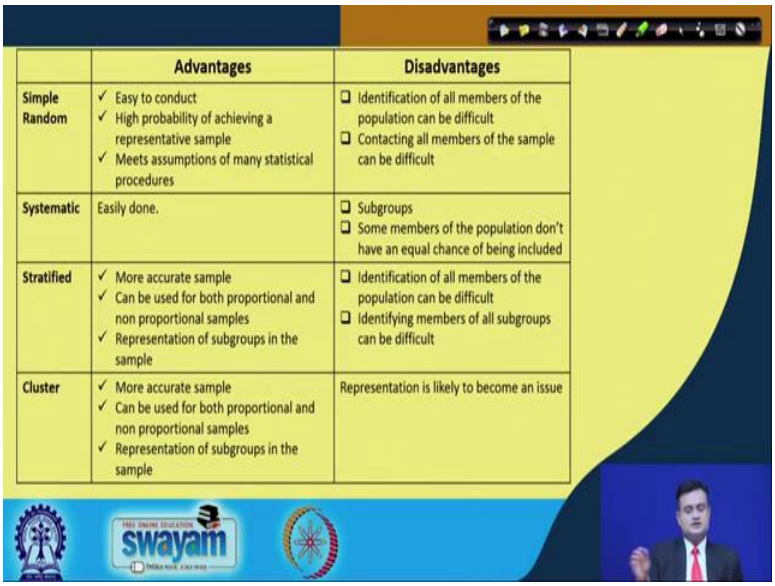
(Refer Slide Time: 34:55)



There is a caution stratified sampling works best when a heterogeneous population is split into fairly homogenous groups and it is important to note for cluster sampling that unlike the strata in stratified sampling the cluster should be microcosms.

So, basically some kind of say units, some kind of clusters rather than sub sections of the population representing some phenomena. So, that is the part and each cluster should be heterogeneous within it.

(Refer Slide Time: 35:35)

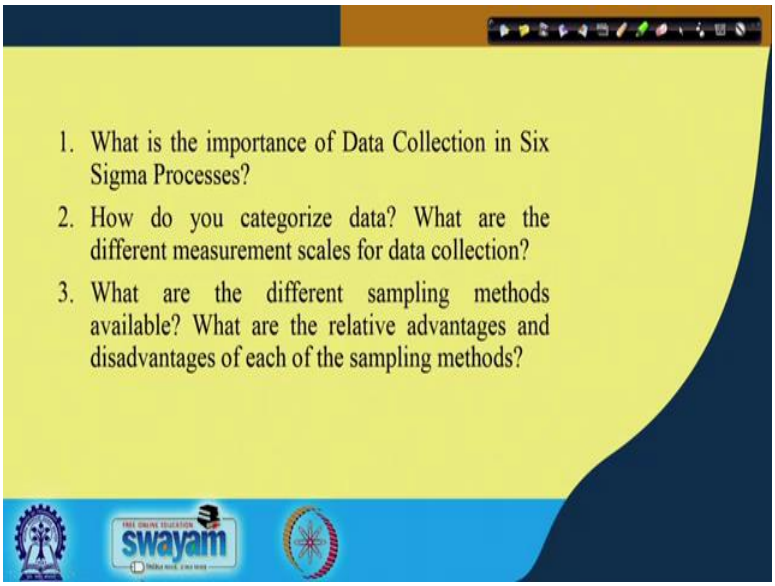


So, just we can finish with some advantages and disadvantages of the techniques sampling methods. So, simple random easy to conduct, high probability of achieving representative sample and meets assumption of many statistical procedure. So, this is widely used. Disadvantage; identification of all members of the population can be difficult and contacting all members of the sample can be difficult.

Systematic fine it is easy to conduct I know every third hour every fourth hour every ten minute, but there could be a difficulty in getting the subgroups or some members of the population do not have an equal opportunity or equal chance of being included. Stratified and cluster we have seen in detail. So, stratified is more accurate sample same applies to cluster and similar way can be used for both proportional and non proportional both stratified and cluster an representation of subgroup in the sample.

If you see the disadvantage identification of all the members of the population can be difficult for stratified and in case of cluster representation is likely to become an issue.

(Refer Slide Time: 37:01)

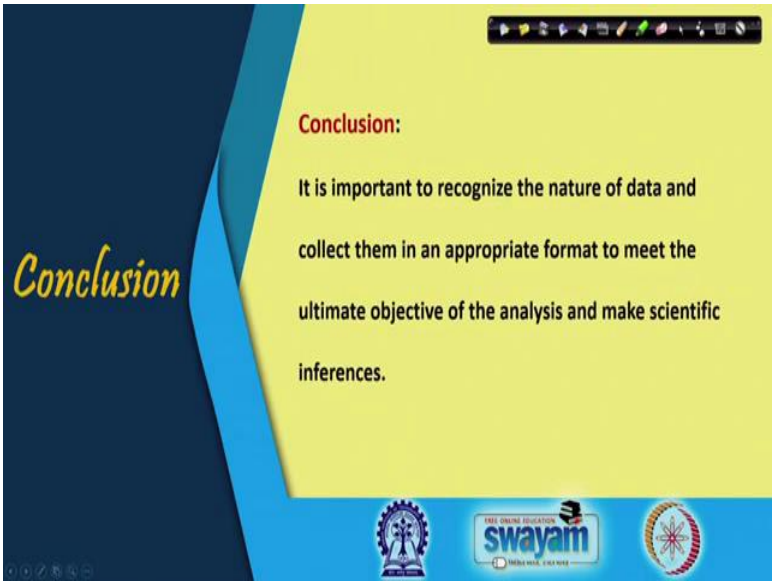


Now, let us finish this lecture with couple of think it question, this should help you to revise the content delivered in lecture 18 and couple of questions I am just floating for your introspection. What is the importance of data collection in six sigma processes. How do you categorize data, what are the different measurement scales for data collection and finally, what are the different sampling methods and what are the relative advantages and disadvantages of each of the sampling methods.

(Refer Slide Time: 37:34).

So, you can take help of these references to go into detail of this topic and see some more examples.

(Refer Slide Time: 37:40)



My conclusion about data collection and summarization is that it is important to recognize the nature of the data and collect them in an appropriate format to meet the ultimate objective of the analysis and scientific investigation.

So, thank you very much for your patience and interest in learning the concept on data collection and summarization; we have part 2 on the same topic as lecture 19 we will discuss next time be with me.