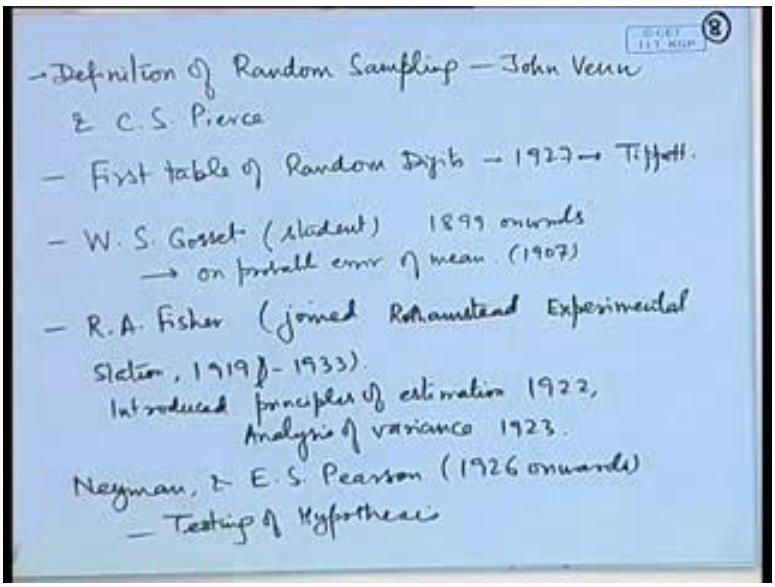
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 50
Descriptive Statistics – II

(Refer Slide Time: 00:21)



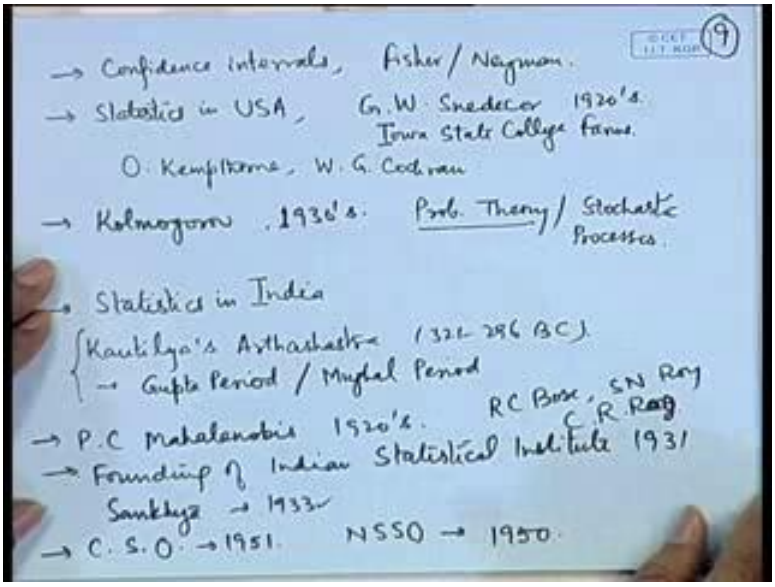
The definition of random sampling is credited to John Venn and C.S Pierce and in 1927 the first table of random digits was printed by Tippett. Around the same time another statistician by name W.S Gosset he was working on various statistical methods. In fact, yesterday I introduce the student t distribution. So, actually it is attributed to W.S Gosset, who wrote a paper on probable error of mean in 1907 and here he introduced the student's t distribution he wrote under the nickname student.

One of the most profound influences on a statistical theory and its development is by R.A Fisher; he is known as father of modern statistics. He is joined Rothamsted experimental station in the year 1919 and we work for several years in this experimental station and their they conducted experiments in agriculture and through experiments is started developing the theory of basic theory of a statistics or foundations of a statistical inference and his papers became very famous in the year 1922 he gave almost all the elementary concepts of estimation that we study today for example, the concept of

maximum likelihood estimation, the concepts of unbiased estimation, in 1923 he introduced the topic analysis of variance which led to the designs of experiments.

In the mean time polish mathematician Neyman and E.S Pearson who was son of the earlier mentioned Carl Pearson, we started developing the query of testing of hypothesis from 1926 onwards, so they published a series of papers where they gave the so called name and Pearson lemma and its applications for testing various kinds of hypothesis around the same time R.A Fisher has given theory of likelihood (Refer Time: 02:46).

(Refer Slide Time: 02:49)



Fisher and Neyman are credited with giving confidence intervals around the same time Neyman gave the development of a stratified random sampling etcetera. Around 1930s the subject a statistics is started to develop in USA also. In fact, G.W Snedecor is credited with introducing the subject in United States of America.

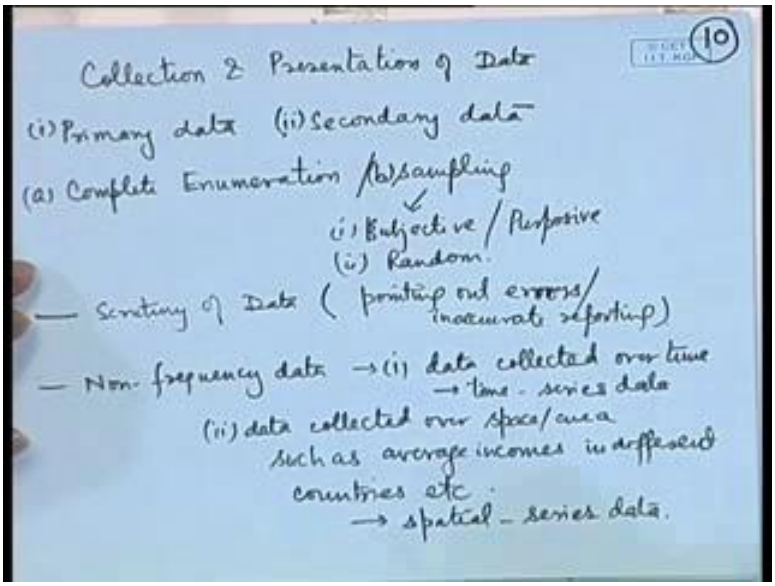
So, in 1920 he was working in Iowa state college and there on the agricultural farms he started using various statistic techniques. Slowly other statisticians like Oscar Kempthorne and W.G Corcoran also joined in and slowly the subject of a statistics is started in USA in 1939 there Heinemann Abraham ward etcetera move to USA and the subject got a big boost USA; meanwhile in the (Refer Time: 03:54) USA Sahara or Russia a n Kolmogorov and many of his associates they develop the rigorous theory of probability and stochastic processes.

In fact, the modern probability theory as we studied today is based on the foundations of Kolmogorov as (Refer Time: 04:12) definition. The statistics in India also did not lag behind. In fact, with the starting of statistics in Europe the subject is started to develop in India also apart from the ancient references as such as Kautilya's Arthashastra which was written in 321 BC onwards, now where he has given detailed description of how to collect data on our tensors of on various aspects of the a state, we see the records of various kinds of tensors and other data in Gupta period as well as the Mughal period and these are recorded. The modern statistical methodology in India got a initiative by P.C Mahalanobis. So, in 1920s he was a student at in England, where he came under influence of Carl Pearson and he saw the biometrical journal.

After coming back to India he started to regularly use the statistical tools for various kind of studies, he was first one to understand the importance of the subject statistics and therefore, with his efforts the irritated Indian statistical institute in 1931; they prominent journal 1933 and he collected a group of very talented people around him who later on contributed greatly to various aspects of a statistical methodology. Peoples such as R C Bose who contributed made great contributions to the subject of designs of experiments SN Roy C.R Rao who contributed greatly to multivariate analysis designs of experiments linear models etcetera and therefore, Indians are not lagging behind in the development of statistics.

With the efforts of Mahalanobis, central statistical organization was established in 1951 by government of India to keep, to collect the data on various aspects of the state. In 1950 national sample survey organization was established to collect data on various socio economic aspects of the life.

(Refer Slide Time: 06:55)



After the historical development now we look at that what is the data, what kind of data we have and what we what is to be done with the data.

So, first aspect is the collection of data. Now collection of data refers to two types of things: we have the so called primary data and the secondary data; for example I want to study the growth of a sector of economic say services sector in past 10 years, how the development or you can say what are the revenues through the export of the softwares; now for this I myself do not have to go and ask the companies say various companies which are situated in say Bangalore or in Pune or in Hyderabad which the so called software cities where these companies are their they produce and then they have a record of how much earning they have to expose.

However one himself does not have himself or herself does not have to go and collect this data, this data may be available by the economics ministry of the ministry of economic of ways or a export wing of the ministry of finance and we can simply collect the take the data from their for our inferences, this is called secondary data; that means, the data has already been collected by a government agency or by private agency, but it is already there and the person who is going to use it simply borrows that data for doing a statistical analysis.

However there may be various other kind of aspects where one needs to collect the data himself or herself for example, if today I want to know the consumption pattern for a

particular product, so that I want to know that whether a similar product will be successful in another area if we introduce, then maybe I will myself make a survey and collect the data on that, this is called primary data primary data collection.

So, in general we deal with two types of data sets: one is primary data, one is secondary data and depending upon what kind of your objective, what kind of data is already available we go for either of them. Now when we collect the data on certain aspects then there are 2 types of collections, one is complete enumeration; that means, we take the data on each unit of the population or we adopt sampling, so what is the difference between these two? For example when I say we want to know the consumption patterns of certain product say suppose I say- I want to know about consumption of a certain south Indian dish which is been taken by a people. Now here it is impossible to know the consumption pattern for each individual of a locality, because we may not able to go to each person and ask him rather what we can do is, we can look at the restaurants in the locality and we can go there and ask for the sales of that particular item.

Now, this does not represent full data rather it is a subset of the population so we call it sampling method. Whereas for certain other things complete enumeration is done for example, when both are taken for certain preference by the government for certain elections then it a complete enumeration, the government carries out sense of the every 10 years, which gives the data on each individual of the entire countries population and various aspects of those people for example, what are their income, what are their ages and various other kind of information.

So, what happens that in depending upon the kind of a study that we are having, we will go for either complete enumeration or for sampling; by sampling we mean that we take only a subset of the population. Now in most of the practical a study it is not possible to look at the complete enumeration, because complete enumeration may be to time consuming too lengthy and it may not fulfill the objective because of the inaccuracy coming into because of the employing area of a large number of personal and also the time taken for the complete enumeration. So, in most of the cases it is a very practical thing to take a sample or you can say a representative sample of the population and draw the conclusion based on that sample.

Now, the sampling methodologies are also of two types: one is a subjective or purposive or judgmental sampling for example, I can take a set of people and get their responsive on a certain opinion, we may take their opinion for a certain thing and to suit our needs for example, if we want to we are introduce a new legislation and now I want to know that how many people are happy with it. So, if I have the target group and such that they are happy with this one, we will take only their opinion.

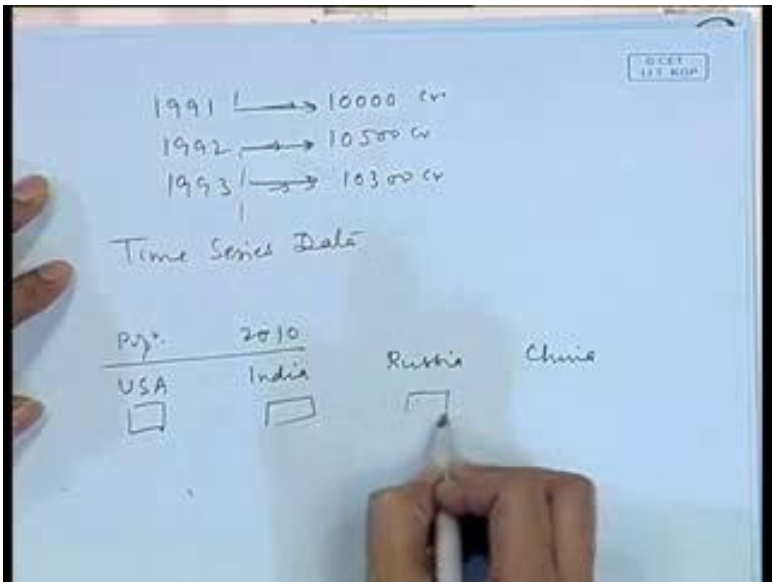
So, this type of sampling method is called subjective or purposive or judgmental and it is not having any statistical base, it is used for certain purpose by various organizations, people, politicians or any other type of people who want to get certain response based on their preferences; and second aspect is random sampling, where we take a sample which represents the whole population; that means, each unit of the population has a certain probability of getting selected in the sample. So, in a statistics we are concerned with random sampling.

So, now if we have a data whether it is through complete enumeration or through sampling, we meet to analyze it. Now before analysis we have to firstly make a presentation of the data to make it suitable for a statistical analysis. The first job of a statistician when we gets the data is to is scrutinize the data, that is for possible inaccuracy which are too glaring for example, if it is heights of say school children and the heights are measured in say centimeter, so may be all the heights are in the vicinity of say 50 centimeter; say a school going child in a primary school and there is an entry which shows say 50 or say I am sorry we are talking about inches. So, 50 inches and suppose there is an entry 150 inch, now what does one 50 inch represent? It means a person which is of more than 12 feet height.

So, obviously, a school children school child cannot have this height. So, it will show an any courage in accuracy for example, if we have kept the record of certain dates and so we are recording the dates in terms of say day month and the year, and there we suddenly show for a day the number is written as 45; obviously, it cannot be correct. So, first of all a statisticians job is to see the data carefully and point out actual or you can say glaring in accuracy, which may be made due to wrongly typing or inputting the data or various kind of things may be there and come to a and analyze that the why this error is here.

So, if it is a genuine error it should be removed or it may still be accurate for example, in a set of students whose heights are in the vicinity of say 70 inches and there is a person with height 80 inch then maybe there is a genuine person who is very tall now there are various types of data.

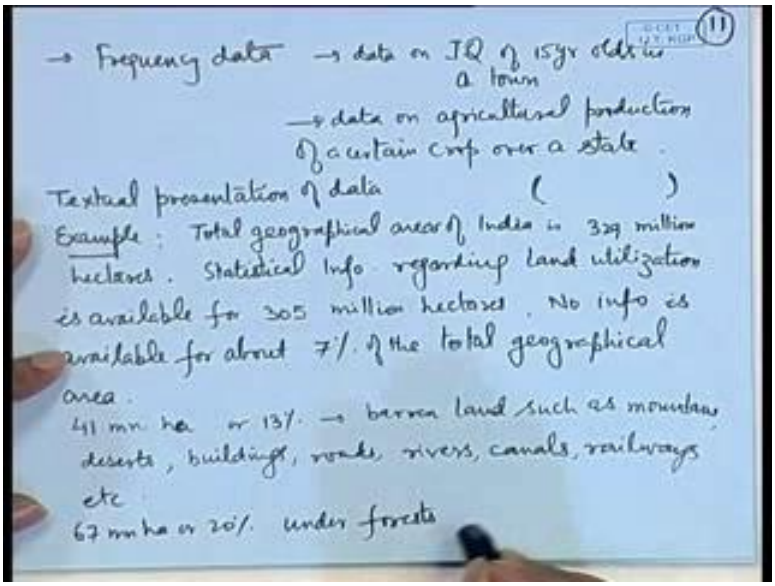
(Refer Slide Time: 16:23)



So, one is the data which is collected over time or collected over a space for example, we keep that in a year say 1991 the exports where say of the tune 10000 crores. In the year 1993 the exports where say 10500 crores, in the 1993 the exports where say 10300 crores; that means, we are collecting the data or recording the data based on years or time this is known as time series data.

Another data of the similar nature is collected by area space for example, what is the population in the year say 2010 in countries for example, what is the population in USA, what is the population in India, what is the population in Russia, what is the population in china etcetera. So, this data which is collected according to the space if it is known as a spatial data; so these are known as non frequency data, because it is collected according to time or space or area etcetera.

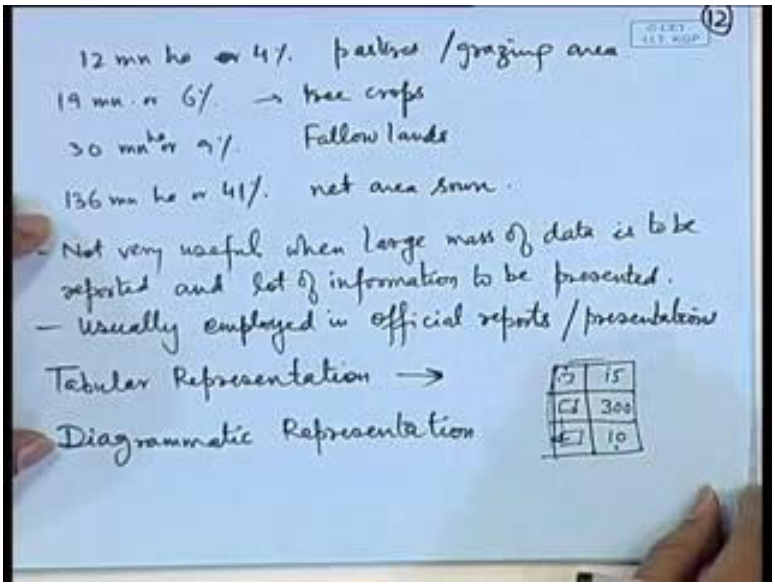
(Refer Slide Time: 17:37)



The other kind of data is known as frequency data for example, if we are recording the IQs of 15 year olds of a town, the data on the agriculture production of a certain crop over a state, these kinds of data are known as frequency data; so in this particular discussion we are concerned about the frequency data. Now how to represent the data? The simplest or you can say the most common way of the presentation of data is a textual representation of the data for example, read a report in a government of India does that or a government of India economic report or a report on say agriculture area.

So, the report read something like this, the total geographical area of India is 329 million hectares, the statistical information regarding land utilization is available for 305 million hectares, no information is available for about 7 percent of the total geographical area; that means, above 24 million hectares. 41 million hectares or 13 percent of the geographical area is barren land; that means, agricultural production is not there in this land and this land corresponds to something like mountains, desserts, rivers, canals, buildings, roads, railways etcetera; that means, effectively it cannot be used for any type of cultivation.

(Refer Slide Time: 19:22)



67 million hectares or 20 percent geographical area is under forests. 12 million or four percent area is pastures or grazing area for cattle's, 19 million or 6 percent of the area consist of 3 crop; that means, the trees are there they may give crops or may not give the crops, but they are under the trees. 30 million hectares or 9 percent is follow lands and 136 million hectare or 41 percent is the net area which is shown under certain crops.

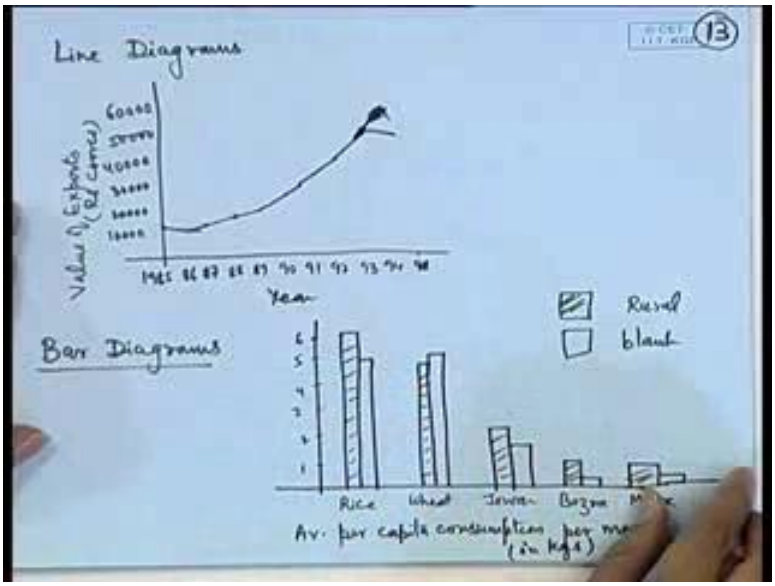
So, basically we wanted to show how much is the area which is actually under the crops and that is represented here that is 41 percent of the area.

Now, this type of information or you can say data, one cannot use this kind of textual presentation if we have large masses of the data when there are large number of variables and when there is a lot of information to be presented, then this kind of descriptive presentation is not very convenient. Of course, this is usually presented in the official reports or the presentations for the management when the management tells with the company or to the share holders what are things this, simply give a textual report. The disadvantages apart from that one cannot draw too much information is all another drawback is also that one may give a selected or selective information to the persons; from the actual data there may be lot of information, but the management may present only what is favorable and they may hide certain fact which may not be very nice to know or nice to tell.

So, the most popular form of presentation of data is the tabular representation. So, in a table so for example, we may say that the persons with see there may be certain forms which have agricultural production of a certain capacity. So, suppose the number is 15, certain other agricultural form they have production between certain things to certain things such number is say 300, and certain higher income or you can say high productive fields suppose there are 10 in number. So, a routine kind of table may look like this which tells something about the values and the frequencies corresponding to that, we will see about these things more later on.

One of the most popular forms of representation of the data is diagrammatic representation, where we make use of certain diagram to effectively tell the situation about that particular data. This type of presentation is extremely useful when you are making presentation in front of an audience so visual presentation they can immediately make out the change or you can say various kind of information that can be drawn from the data.

(Refer Slide Time: 22:34)



For example we are looking at the volume of exports year wise from 1985 onwards up to year 1994. The volume of the exports in crores and we are presenting using a line diagram. So, in 1985 the export was to the tune of say 12000 crores, in the year 86 it became a little less say 11000 crores, in the 87 it became again to the tune of say 12500

crores, in the year say 90 it became to the tune of say 30000 crores, in the year 93 say it became to the tune of 50000 crores, in the year 94 it became likely less than that.

Now, this gives a extremely (Refer Time: 23:31) position of the volume of the exports year wise and you are can clearly see a pattern that from the year 1986 to 93 there is clearly an increasing trend. One may even say that what kind of trend it is it is slightly parabolic or you can say quadratic increase in the volume of the exports and from 93 to 94 there is again a drop and one the government may try to analysis that why there is a drop. So, this line diagrams for the time data or for the spatial data give a very nice way of presenting and one can evenly compare the things. A similar kind of representation is by bar diagrams, now this bar diagrams; through this example I will show, we want to know average per capita consumption per month by different populations. So, here we have two populations that is ruler or urban population.

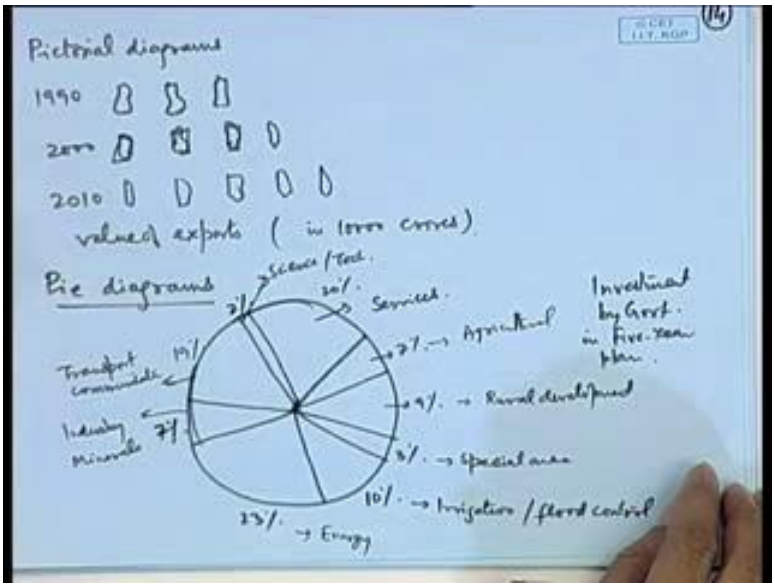
So, what is the per capita consumption you want to have comparative studies for various serials say Rice, Wheat, Jowar, Bajra, maize? So, if you see per capita consumption per month of rice for ruler people it is 6.5 kg and correspondingly if you see urban then the consumption is around 5 kg. So, this tells clearly that the rice consumption by ruler people is per capita is slightly more than the urban people. Similarly if you see wheat consumption you find that the urban per capita consumption is slightly higher than the ruler per capita consumption; where as for other crops such as Jowar, Bajra or maize the rural consumption per kg per head is much more than the urban consumption.

Now, people may draw interesting conclusions from there and then they may try to verify it to other sources. So, one may say that per capita consumption for each of the serials expect say wheat is higher for the urban people low for the ruler people; now what does it mean? Does it mean that the ruler people have more income are they simply wheat more. So, one may like to investigate these phenomena; similarly for wheat you have more consumption for the urban people so that means, either urban people have more access to the wheat, are they prominently heat wheat product that is why like bread which is normally available in the market.

So, they are consumption of the bread or wheat goes up. So, one may like to investigate this further. So, this bar diagram gives a very clear cut comparative study and also this tells the relative consumption of various serials for example, rice consumption is

maximum compared to wheat, then Jowar, Bajra and maize and this also may tell that about the a relative importance of the crops for growing and therefore, the government can also decide the policies, the farmers can decide their policies about these things.

(Refer Slide Time: 27:16)



Some of the various extremely simple types of representations are through pictorial diagram which may not be very useful, but sometime there may be helpful for example, in the year 1990 volume of exports for nearly 30000 crores, in the year 20 into the 2000 it was nearly 40000 crores, in the 2010 it was nearly 5000 crores. So, here each unique here represents a volume of 10000 crores, but this type of representations has clear cut limitations because you cannot use that for various purposes mostly if you have to use fractional data and also if the number of units cannot be represented in integer type of things .

Another very useful pictorial representation is pie diagram. So, if a particular population is divided into various categories, then how much is the share of the each category we can represent nicely in a pie chart or a circle, and here we make the sectors of the circle so this particular diagram represents the investment proposal by the government in the 5 year plan to various sectors of the economy.

So, here for example, the government will spend 20 percent on services, it will spend say 19 percent on transport and communication say infrastructure. It will spend 7 percent on agricultural or related fields, it will spend 9 percent on rural development, the

government will spend some 3 percent of its amount on a spatial area developments, 10 percent on irrigation of flood control exercise, some 23 percent on energy and some 2 percent on science and technology etcetera, some 7 percent investment will be in the area of industry, minerals etcetera.

So, this kind of pie chart gives a very clear representation comparative expenditure by the government on various things, and it is extremely useful. In the next lecture we will be concentrating on how to actually now represent frequency distribution, how to classify the frequency distribution and what basic measures of central tendency are the dispersion can be evaluated from there. So, in the next lecture we will be taking up these.

Thank you.