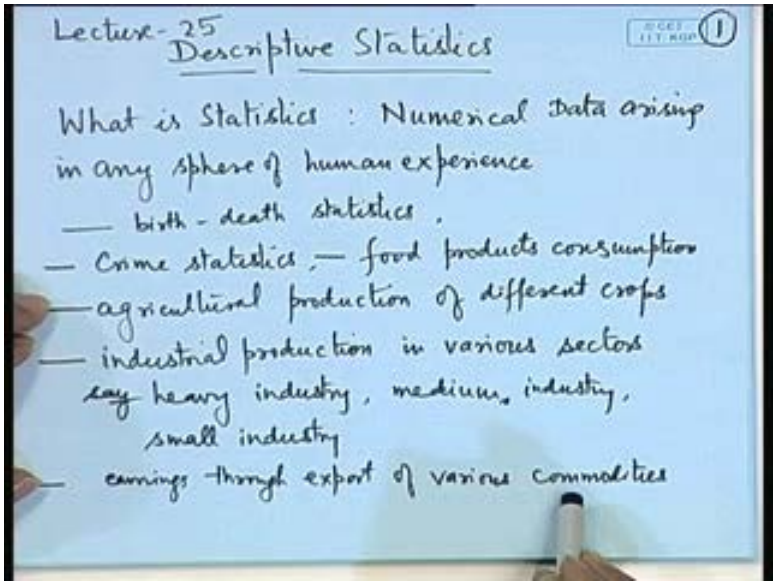
Probability and Statistics
Prof. Dr. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Module No. #01
Lecture No. #25
Descriptive Statistics – I

In the course of Probability and Statistics, till now we have concentrated on the topic probability, random variables and their distributions. We studied various kinds of discrete and continuous distributions, we discussed joint distributions and we discussed sampling distributions, which arise in the study of certain population and sample.

So, now it brings us to the second aspect of this course that is, Statistics. So, we firstly introduce what is the term, statistics referring to and its historical development. In this particular section, we will tell the various types of statistics that are used and also the representation of the data through statistics.

(Refer Slide Time: 01:22)



So, let me start with the term, "what is statistics". So, in plural sense, when statistics word is used it refers to numerical data, which arises in any sphere of human experience. So, in our day to day life from the following examples I will try to show that everywhere we are making use of statistics. For example, the records of birth and death are kept in every municipality or in village or in town, office. So, these data represents birth and

death statistics. Every state or a town or a police station keeps the record of crime statistics. That is, the number of crimes committed under various categories during a given period. A certain market may keep record of the consumption of food products of various types. For example, what is the market for say south Indian dishes, what is the market for consumption of say, Chinese dishes? So, the data on that comes under food products consumption.

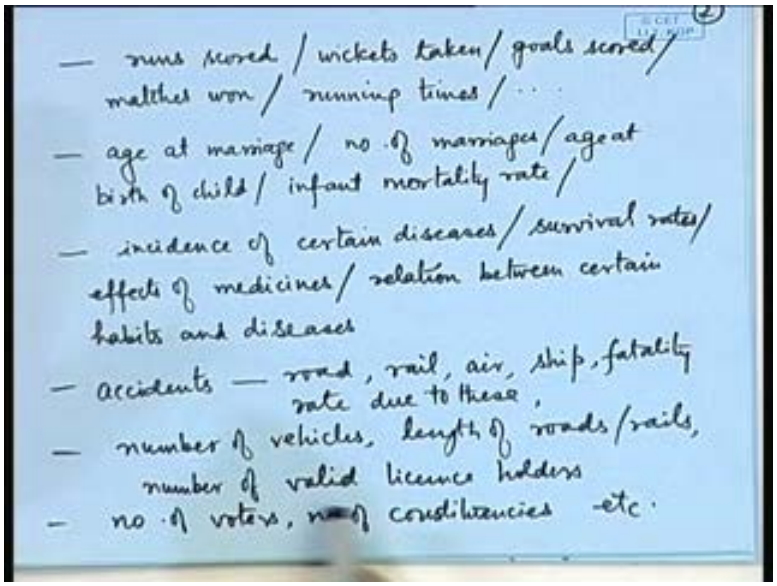
Government regularly keeps the record of agricultural production of different crops. So, how much production will be there under say wheat, how much production will be there for rice, how much will be the production for the pulses. And, this type of information is extremely important for making the policies by the Government. How much should be released in the market, what should be the pricing, whether there should be any import of these products from other countries to meet the short fall or if there is a surplus, then should the government be exporting it, should be encouraging the exports. So, the agricultural production data is one of the most important data that is kept by the Government.

Industrial production in various sectors for example, how much capacity of the heavy industry sector is there and how much it is actually producing and then how much it is contributing towards the growth of economy; in the sense that, whether we are able to give enough things for our local consumption as well as for international exports. For example, metal industries, steel industries. Similarly, what is the production for the medium sectors or the small sector industries are. Complete industrial production data is of extreme importance to the Government for formulating its various economic policies and also to tell that how much Government should invest in various kinds of industries, what kind of facilities it should give to those industries.

So, from this point of view, the industrial production data or the statistics is very important. And, it is regularly collected by the Government. For example, another data which is regularly collected is earnings through export of various commodities. For example, the Government gives certain incentives for export of an export item of a particular type and then it wants to see the effect on that, what is the increase in the export of those categories. For example, it gives certain relief to those manufacturers who are producing a certain technical item, then for example, softwares and then whether it leads to the increase in the export of that particular item, now these has some

furnishing consequences. For example, if you see that there is an increase in the export; that means, there must have been a better opportunity for the employment in that sector. People must have earned more and so on, so forth.

(Refer Slide Time: 05:39)



So, this type of data is of importance. Statistics are kept in area such as sports, where we keep record of the runs scored by players, wickets taken in a cricket match. In a football match, how many goals are scored or in a hockey match, how many goals are scored, how many matches are won by different teams. In say athlete athletics, what are the running times of the top athletes for certain events say hundred meter race or two hundred meter race, etcetera.

In Social Studies, we may keep record of the age at marriages of the women, age at marriages of the men and on the average, how many marriages the persons are doing in a lifetime, what is the age of the parents at the birth of the child, what is the infant mortality rate, various of various statistics of this kind are helpful for formulating various cultural policies and also social policies by the Government for health and family welfare and for various other kind of things.

In medical, we keep track of the statistics on the incidence of certain diseases. For example, what is the incidence of tuberculosis, what is the incidence of malaria? If it is more, then Government will try to formulate a policy so that, it can reduce the incidence. What steps should be taken, what are the pockets, where it is more. So, they can study

the socio-economic profile of that and try to create certain reforms there; so that, the incidence of those diseases can be reduced. If there is a new epidemic, then how to control that. So, the data on the incidence of diseases is regularly monitored by the Government agencies by the medical council.

What are the survival rates for certain diseases, so whether if the survival rate is less, whether new drugs should be introduced so that the survival rate can be increased. What are the effects of the medicines, for example, if any medicine cures a disease, but it creates certain side effects so that the person dies of something else; so, we need to have a data on the effects of the medicines, relation between certain habits and diseases. For example, is there a positive correlation between the habit of smoking and say lung cancer or tuberculosis, therefore those causes should not be encouraged, which give rise to certain diseases.

For example, if a locality is unhygienic, then it may give rise to cholera disease, etcetera. So, the data on the sickness or the diseases or the deaths due to certain diseases is regularly supplied by the hospitals, by the medical agencies. And, it helps Government and other medical research organizations to formulate various kinds of policies for curving the incidence of those diseases or curing those diseases or controlling the diseases.

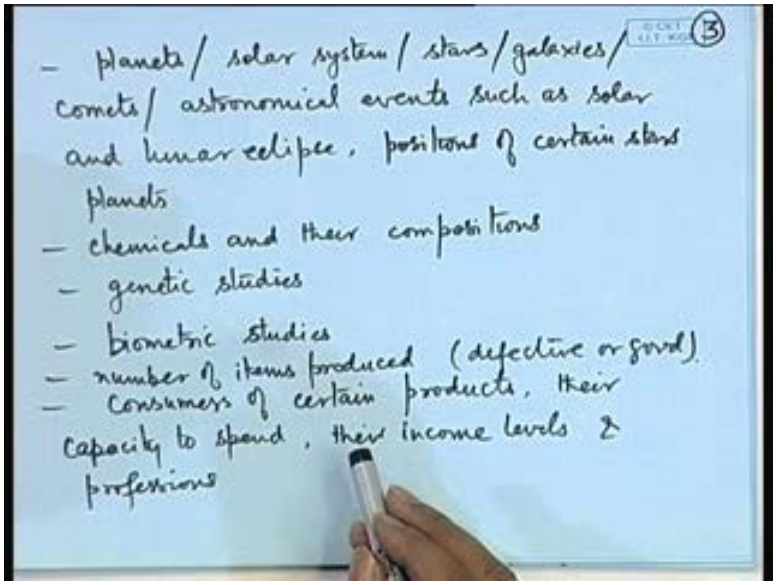
The number of accidents due to various reasons such as road accidents, rail accidents, air accidents, ship accidents and what is the mortality or fatality rate due to these accidents. For example, if it is observed that there are large numbers of accidents due to a particular aircraft, then the aircraft may be grounded and that particular company will be asked to rectify that kind of, those kinds of defects. So that, such accidents are avoided. For example, if accidents are occurring due to collision of trains, then certain devices should be installed, which we need to detecting that there is another train on the same rail line nearby and therefore, the driver should be alerted. What steps can be taken to reduce the number of road accidents. So, this information on actual accidents and also the fatality rates due to these is collected to make policies regarding these things.

Another kind of information which is extremely useful to the companies, which produce automobiles; so, for example, already how many automobiles are available in the country, then what is the length of the roads how many people are in the age group

which can hold valid license for driving those vehicles. So, these types of information are useful for the automobile companies to make an estimate that, how much production they should do for these things.

Another kind of data that is kept is the number of voters in different constituencies for parliamentary elections or assembly elections and the description of those types of voters. For example, how many males are there, how many females are there, how many are between age group from eighteen to twenty five, how many are between twenty five to thirty five; so that the leaders or political parties can make plans accordingly that, those should be giving benefits to the voters of that type, which are there in their constituencies. The collection of the statistics or the utilization of the statistics is not only done in the Government sector or social sector, it is extremely useful in the Sciences or Engineering or Technological studies also.

(Refer Slide Time: 11:44)



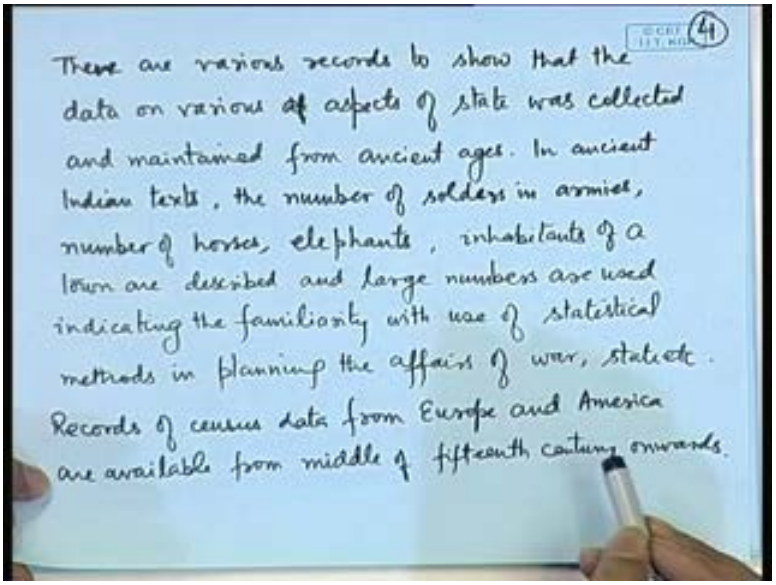
For example, Physics, Chemistry, Biotechnology, in each of the or say Economics, everywhere the use of statistics is there. For example, physicists will be interested in knowing the movement of the planets, the number of planets, what is there in the solar system, what is the number of the stars in a particular galaxy, what is the number of galaxies, comets, astronomical events such as a movement of a certain comet or hitting it to another planet, etcetera. Events such as solar or lunar eclipse positions of certain stars

and planets, etcetera. The data on all these things are extremely useful to the cosmologists.

Physicists also use the data in other discipline such as moment, the number of say electrons or number of neutrons or various protons. And, also they are using in various meteorological studies, their composition. And, the people in chemistry or chemical sciences, they study various kind of chemicals and their compositions. The biologists, they study genetic makeup of certain human beings, the genetic makeup of certain item. There are biometrical studies. In industry, when a manufacturer is being done, then what is the percentage of the defective items produced by that particular manufacturing process? It helps us to rectify or improve the manufacturing process so that the number of defectives can be reduced.

In Economics, the consumers of certain products, what is their capacity to spend, what are their income levels, what are their professions, so these things are useful for creating products needed for certain particular type of consumers. So, for example, if a new soap is to be sold, then they find out that what should be the price of the soap, so that it is affordable to most of the people in that particular income group. If there are people with high income in particular place, then they can introduce a soap which is of high values. And therefore, they will like to make it more fashionable or more attractive, so that those groups of people are attracted to that.

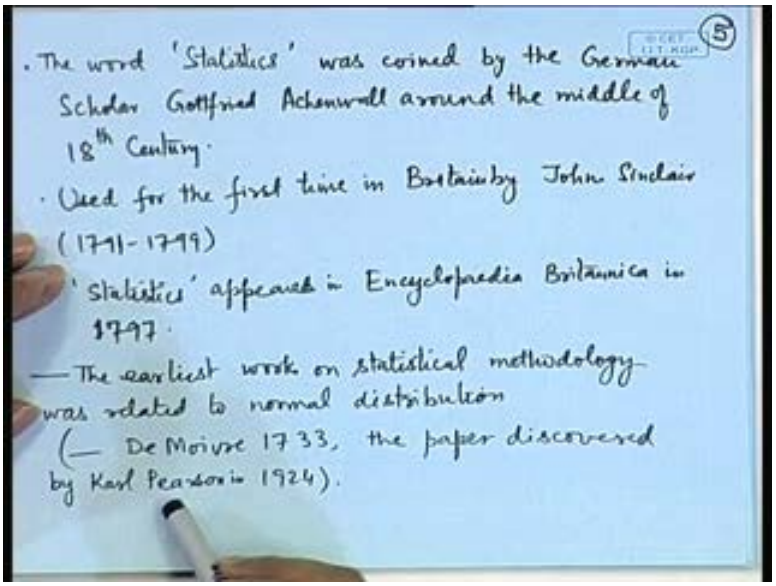
(Refer Slide Time: 15:37)



So, we can see that almost every area of human activity involves collection of a statistics. And, it is actually done. It is used by agencies, by individuals, by organizations, by people, which are relevant to those statistics. And, we can actually say that the term statistics or the usage of statistics is as old as the civilization itself. In fact, ancient text shows that. For example, in a town or in a capital of a certain state, how many inhabitants are there, how many houses were there or what is the size of the army, how many infantry people are there, how many horses or elephants were there in a particular army or in the army of the enemy. This type of data we read in the ancient text either in Indian or European or American text. Such data was available. It is recorded.

Therefore, the term statistical information of the data has been in use from ancient ages. And, it also shows that people were familiar with the statistical methods for planning the affairs of the war, state, etcetera. Records of census data from Europe, America, etcetera are available from middle of fifteenth century onwards. There is data from England or from Scotland on certain professions that these many people work in this profession, etcetera. This kind of breakup, detail breakup of the data is available from European countries or even from USA from fifteenth, sixteenth and seventeenth centuries, etcetera. That means, the people have been aware of the importance of the data as well as statistical methods to how to use that data, how to interpret that data from quite longtime. It is not a new subject.

(Refer Slide Time: 16:36)

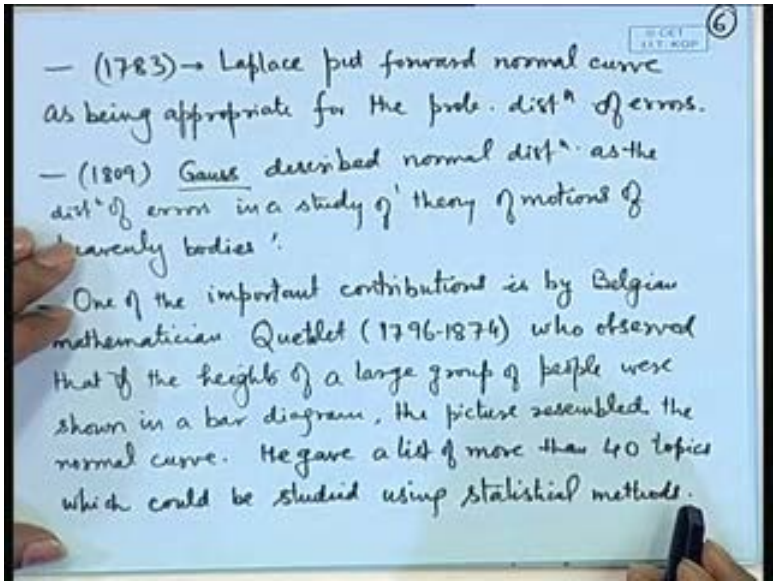


However, the term ‘statistics’ as we use it for the subject Statistics as in a singular, it was coined by the German scholar Gottfried Aschenwall around the middle of eighteenth century. And, first time it was used in Britain by John Sinclair, who wrote a series of volumes in from 19 from 1791 to 1799, which gave the communication between the ministers from Scotland on various aspects of the data.

In the year 1797, the term statistics was used in the Encyclopedia Britannica. Now, related to the statistical methodology. So, now you can understand that there are two types of meanings of the word statistics. One is the data statistics; means the data, the numerical values of the observations, which are taken on various aspects of human activity. As we have given examples of almost every area of human activity. The second is the term statistics, which is referring to the subject of statistics. And, this means the statistical methods which interpret a certain data, which analyze certain data and give inferences on based on that data. That is called the subjective statistics. So, now we will refer it to by saying statistical methods or statistical techniques or statistical inference.

The modern statistical methodology, the earliest references are to the normal distribution. Most probably De Moivre in 1733, he wrote a paper which appeared in the... journal. And therefore, it was not known. It was later discovered in 1924 by British statistician Karl Pearson. In this one, most probably for the first time, he gave what is known as normal distribution. And, he gave it as a… or arising in a large number of Bernoullian trials. He obtained it as a limit of certain sums of random variables. It is known as one of the first central limit theorems.

(Refer Slide Time: 19:06)



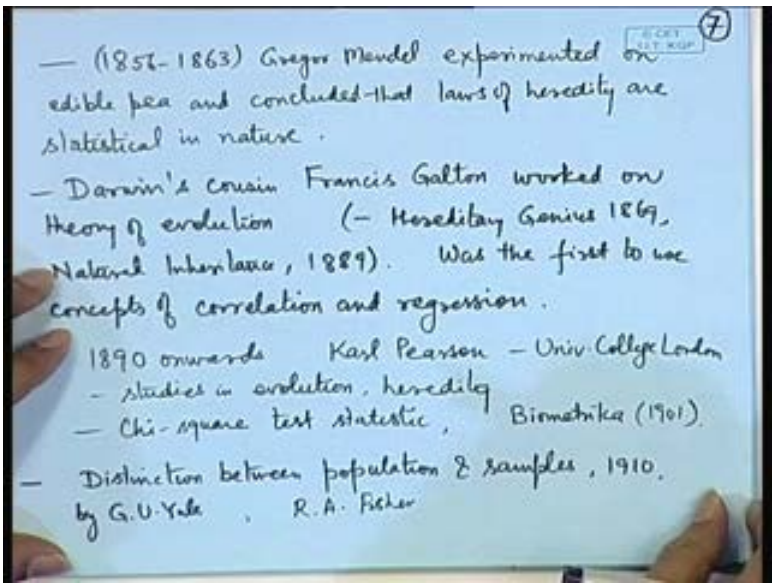
The great French mathematician Laplace in his manuscript in 1783, he said that normal curve can be used for describing the probability distribution of errors. So, when we are doing certain measurements and those measurements are not accurate because of the measuring instrument or by the person who is measuring that. So, there may be errors and you take several observations on that. So, when you take several observations on that, you have several values corresponding to those errors. So, if we plot the frequency curve of that, it looks like a normal distribution curve. And, this theory was propounded by French mathematician Laplace.

Later on the great mathematician Carl Friedrich Gauss, he considered the study of planetary motions that is, the theory of motions of heavenly bodies. And, he also came up to the same conclusion that the distribution of the errors for those astronomical measurements can be nicely described by the normal distribution. So, his data is came in 1809. But, he credited Laplace by telling that this idea was from Laplace. But, in fact the normal distribution has a name Gaussian distribution because of the gauss Gaussian study or the study by gauss on this topic. However, the most appropriate usage of a statistical methodology was probably done first by Belgian mathematician Quetelet, who lived from 1796-1874.

So, he studied apart from this planetary motions etcetera, he studied the usual statistics such as heights of a large number of people. So, he said that, if we represent this

distribution of heights using a bar diagram, then the picture resembles that of a normal distribution. He prepared a list of various topics more than forty topics and he submitted to it to British statistical association. And, he said that all these topics can be studied using statistical methods. So, as early as earlier nineteenth century, people were aware that there are various phenomena where statistical methods can be used. That means the phenomena are not deterministic. This idea was there much of the year.

(Refer Slide Time: 22:07)



George Mendel the Gregor Mendel, he experimented, he did various genetic experiments and mostly on peas, etcetera. And, he concluded that the laws of heredity are statistical in nature. Influenced by the Darwin's theory of evolution, lot of people started studying the laws of heredity. And, one of the famous works on that is, by Francis Galton, who happened to be a cousin of Darwin ever done the theory of evolution. And, his data is Hereditary Genius came in 1869 and Natural Inheritance in 1889. And, he observed that many of these hereditary laws are actually statistical in nature. And therefore, he introduced the concepts of correlation and regression in his studies.

In particular, he studied the heights of the children in comparison with the heights of their parents. One of his famous studies is that the children of taller parents are taller or tall, but less tall, less taller than their fathers. And, again the children of shorter parents are short, but taller than their parents. So, he termed this as regression towards normality. That means you are converging towards. From the height, you are coming down little bit

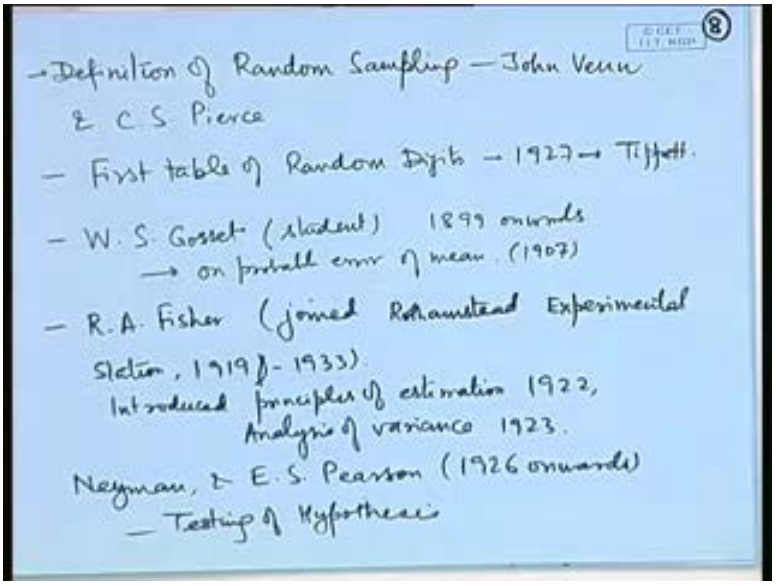
and from lower side, you are going little bit up. That means there is something called normal heights. And, he called it regression towards normality. So, probably these are the first references to the terminology of correlation and regression, which is used in modern statistics.

From 1890 onwards, two centers of statistics got developed. One was in University College of London and another was the Rothampsted Experimental Station in England. So, in the University College of London, the mathematician or the biologist or the economist Karl Pearson, he started studying or you can say introducing the concept of statistics in a very systematic way. He was doing studies in evolution, heredity and also various other aspects. In 1901, he started the famous journal called Biometrika, which is the one of the top journals for the study of statistics.

He was the first one to talk about the chi-square test statistic. That means not only that, you say that this data fits a particular distribution; say, a given data we say that it is fitting nicely by the normal distribution not by a gamma distribution. But, how to do a statistical test for that it is not just by observation, but by developing a proper procedure for that. So, he was a first one, who introduced this chi-square test statistic or you can say chi-square test for goodness of fit.

Till this time, there was still not very clear-cut distinction between population and a random sample. That means, we are talking about full population or a subset of the population. There was not much distinction, but it was the 1910 book by G. U. Yule, who gave a clear-cut distinction between population and samples. Around the same time, R.A. Fisher also gave a clear-cut distinction between population and the samples.

(Refer Slide Time: 26:08)



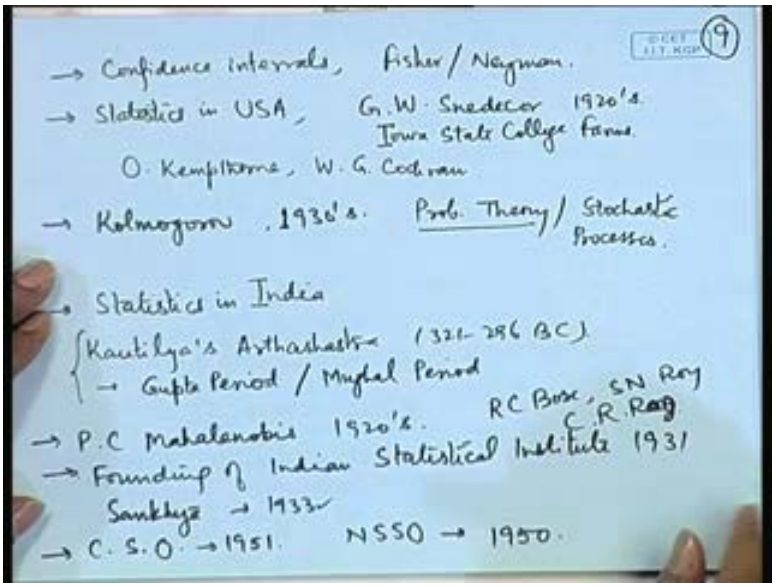
The definition of random sampling is credited to John Venn and C. S. Pierce. And, in 1927, the First Table of Random Digits was printed by Tippet. Around the same time, another statistician by name W. S. Gosset, he was working on various statistical methods. In fact, yesterday I introduced the student's 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 statistical theory and its development is by R. A. Fisher. He is known as father of modern statistics. He joined Rothampsted Experimental Station in the year 1919 and he worked for several years in this Experimental Station. And there, they conducted experiments in agriculture. And, through his experiments, he started developing the theory of basic theory of statistics called, "Foundations of statistical inference". And, his papers became very famous in the year 1922. He gave almost all the elementary concepts of estimation that we studied today; for example, the concept of maximum likelihood estimation and the concept of unbiased estimation.

In 1923, he introduced the topic analysis of variance, which led to the designs of experiments. In the meantime, a polish mathematician George Neyman and E. S. Pearson, who was son of the early mentioned Karl Pearson, they started developing the theory of testing of hypothesis from 1926 onwards. So, they published a series of papers where they gave the so-called Neyman-Pearson lemma and its applications for testing

various kinds of hypothesis. Around the same time, R. A. Fisher has given his theory of likelihood ratio tests.

(Refer Slide Time: 28:35)



Fisher and Neyman are credited with giving confidence intervals. Around the same time, Neyman gave the development of stratified random sampling, etcetera. Around 1930s, the subject statistics is started to develop in U S A 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 statistical techniques. Slowly, other statisticians like Oscar Kempthorne and W. G. Cochran also joined in. And, slowly the subject of statistics started in USA. In 1939, George Neyman and Abraham wall, etcetera moved to USA and the subject got a big boost in USA.

Meanwhile in the… USSR or Russia and Kolmogorov and many of his associates, they developed the rigorous theory of probability and stochastic processes. In fact, the modern probability theory as we study today is based on the foundations of Kolmogorov's schematic definition.

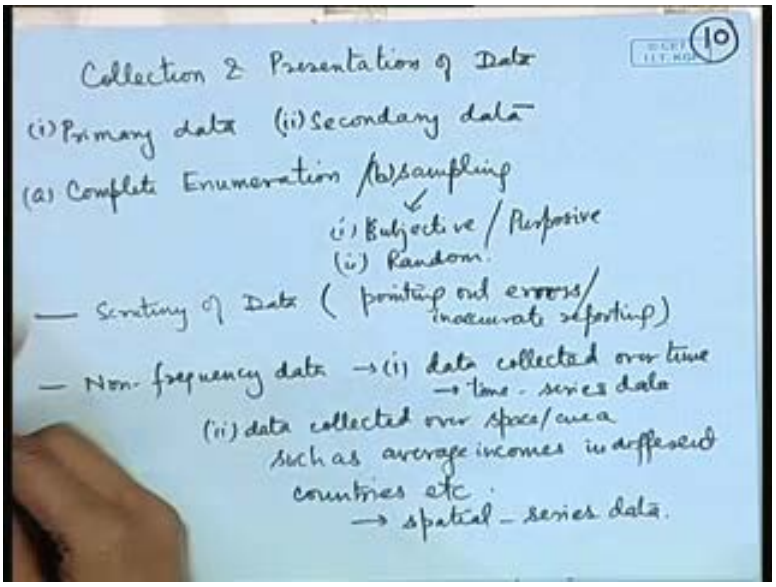
The statistics in India also did not lag behind. In fact, with the starting of statistics in Europe, the subject started to develop in India also. Apart from the ancient references, such as Kautilya's Arthasastra which was written in 321 B C onwards; where he has given detailed description of how to collect data on or census of on various aspects of the

state. We see the records of various kinds of census and other data in Gupta period as well as the Mughal period. And, these are recorded. The modern statistical methodology in India got an initiative by P.C. Mahalanobis. So, in 1920s, he was a student at in England, where he came under influence of Karl Pearson and he saw the Biometrika journal.

After coming back to India, he started to regularly use the statistical tools for various kinds of studies. He was first one to understand the importance of the subject statistics. And, therefore with his efforts, he initiated Indian Statistical Institute in 1931, the prominent journal 1933 and he collected a group of very talented people around him, who later on contributed greatly to various aspects of statistical methodology. People such as R. C. Bose, who contributed, made great contributions to the subject of designs of experiment. S. N. 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 the 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: 32:40)



After this 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 economy, say services sector. In past ten years, how the development or you can say what the revenues through the export of the software are.

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 is the so-called software cities; where these companies are there, they produce and then they have a record of how much earnings 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 or the Ministry of Economic Affairs or an export doing of the Ministry of Finance. And, we can simply collect the take the data from there 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 statistical analysis.

However, there may be various other kinds 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 introduced, then maybe I will myself make a survey and collect the data on that. This is called primary data or 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 is your objective, what kind of data is already available, we go for either of same.

Now, when we collect the data on certain aspect, then there are two 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 will be which is being taken by people. Now, here it is impossible to know the consumption pattern for each individual of a locality because we may not be able to go to each person and ask

them; 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 votes are taken for certain preferences by the Government for certain elections, then it is a complete enumeration. The Government carries out census for every ten years, which gives the data on each individual of the entire country's population and various aspects of those people. For example, what are their incomes and what are their ages and various other kind of information.

So, what happens that in, depending upon the kinds of 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 studies, it is not possible to look at the complete enumeration because complete enumeration may be too time consuming, too lengthy and it may not fulfill the objective; because of the inaccuracy coming into because of the inclined of a large number of personnel 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 representatives, sample of the population and draw the confusion 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 responses on certain 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 introduced 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.

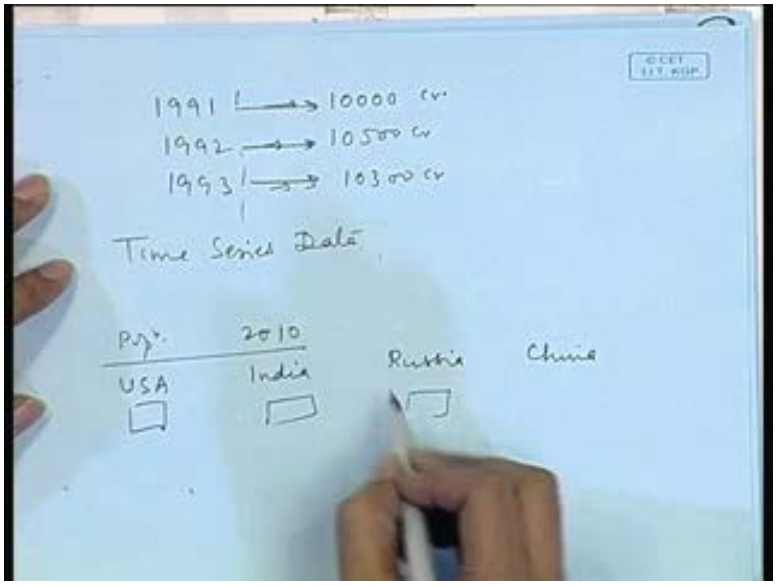
And, second aspect is random sampling; where we take a sample, which represents the whole population. That means each unit of the population has a certain probability of getting selected in the sample. So, in statistics, we are concerned with random sampling.

So, now if we have a data, whether it is through complete enumeration or through sampling, we need to analyze it. Now before analysis, we have to firstly make a presentation of the data to make it suitable for statistical analysis. The first job of a statistician when he gets the data, it is to scrutinize the data that is for possible inaccuracy, which are too glaring.

For example, if it is the heights of, say school children and the heights are measured in say centimeters, so may be all the heights are in the vicinity of say, 50 centimeters. 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 of 150 inches, now what does 150 inches represent? It means a person who is of more than 12 feet height. So, obviously, a school child, school child cannot have this height. So, it will show an encourage 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 a year and there we suddenly show for a day, the number is written as 45. Obviously, cannot be correct.

So, first of all a statistician's job is to see the data carefully and point out actual or you can say glaring inaccuracies, which are 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 that, why this error is there. 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 inches, then maybe there is a genuine person, who is very tall. Now, there are various types of data. So, one is the data which is collected over time or collected over space.

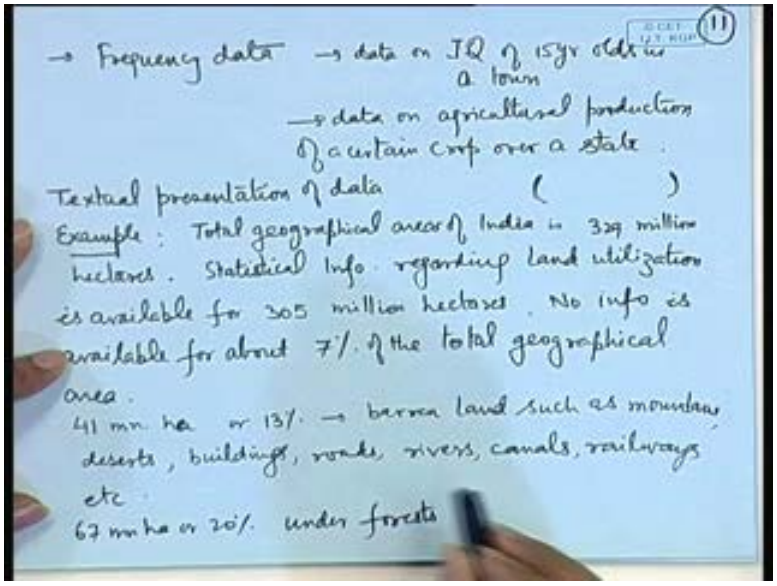
(Refer Slide Time: 42:10)



For example, we see that in a year, say 1991 the exports were say, of the tune 10,000 crores; in the year 1992 the exports were say 10,500 crores; in the year 1993 the exports were say 10,300 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 or 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, it is known as spatial data. So, these are known as non frequency data because it is collected according to time or space or area, etcetera.

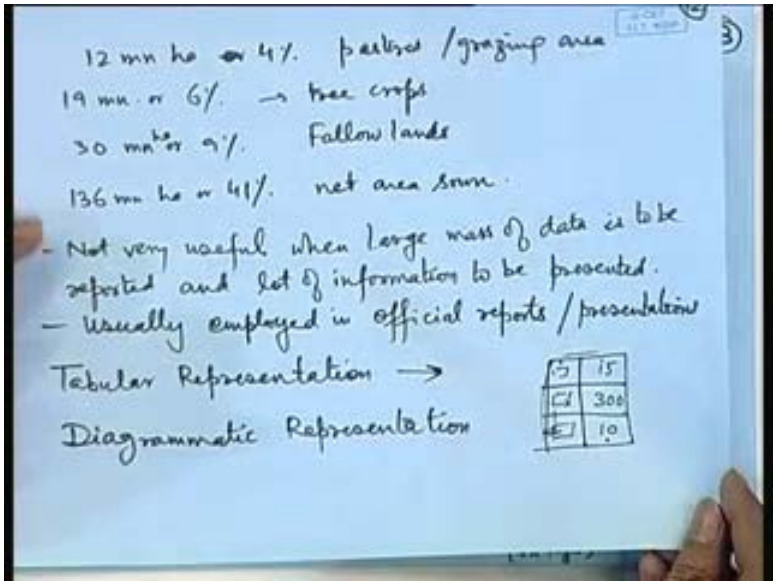
(Refer Slide Time: 43:25)



The other kind of data is known as frequency data. For example, if you are recording the I Q s of fifteen year olds of a town, the data on the agricultural 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 representation of the data is a textual representation of the data. For example, read a report in a Government of India Gazette or a Government of India Economic report or a report on, say agricultural 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. And, no information is available for about 7 percentage of the total geographical area; that means about 24 million hectares. 41 million hectare or 13 percentage 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, deserts, rivers, canals, buildings, roads, railways, and etcetera. That means, effectively it cannot be used for any type of cultivation. 67 million hectare or 20 percentage geographical area is under forests.

(Refer Slide Time: 45:09)



12 million or 4 percentage area is pastures or gazing area for cattles, 19 million or 6 percentage of the area consist of tree or crops. That means the trees are there. They may give crops or may not give crops, but they are under the trees. 30 million hectare or 9 percentage is fallow lands and 136 million hectare or 41 percentage are the net area, which is sown 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 percentage of the area.

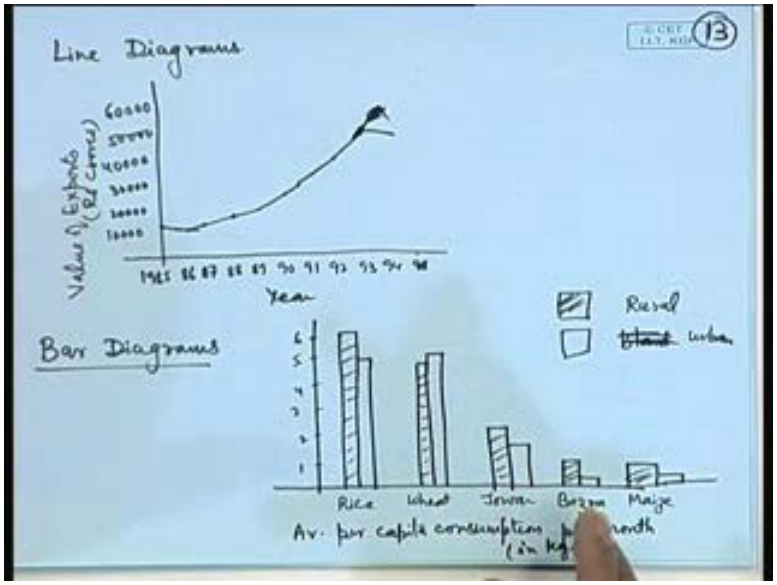
Now, this type of information or you can say data, one cannot use this kind of textual representation if we have large masses of the data, when there are large number of variables and when there is a lot of information to be presented. Then, this kind of descriptive presentation is not very convenient. Of course, this is usually presented in the official reports or the presentations for the management, when the management tells to the company or to the shareholders what are the things, they 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 the data is the Tabular representation. So, in a table, so, for example, we may say that the persons with.. see

there may be certain farms, which have agricultural production of a certain capacity. So, suppose the number is 15. Certain other agricultural farms, they have production between certain thing to certain thing. Such number is, say 300. And, certain higher income or you can say high producing 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. And, through 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: 48:20)



For example, we are looking at the volume of exports year-wise from 1985 onwards up to year 1994. The volume of the exports are in the, in crores. And, we are representing using a Line diagram. So, in 1985 the export was to the tune of say 12000 crores; in the year 1986 it became a little less, say 11000 crores; in the year 1807 it became again to the tune of several 1500 crores; in the year, say 90 it became to the tune of say, 3000 crores; in the year 93, say it became to the tune of 50,000 crores; in the year 94 it became slightly less than that.

Now, this gives an extremely clear position of the volume of the exports year-wise. And, you all can clearly see a pattern that from the year 1986-93, 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 93-94 there is again a drop and when the Government may try to analyze 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 representing and one can easily compare the things.

A similar kind of representation is by Bar diagrams. Now, this bar diagrams. So, through this example, I will show. you want to know average per capita consumption per month by different population. So, here we have two populations that is, rural or urban population. So, what is the per capita consumption? You want to have a comparative study for various cereals, say rice, wheat, jowar, bajra or maize. So, if we see per capita consumption per month of rice, for rural people it is 6.5 kilograms. And, correspondingly if we see urban, then the consumption is around 5 kilograms. So, this tells clearly that the rice consumption by rural people is per capita is slightly more than the urban people.

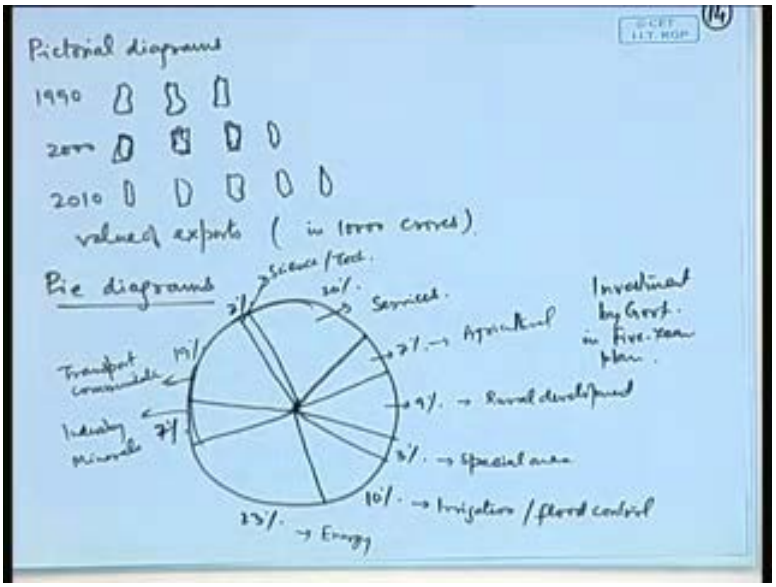
Similarly, if you see wheat consumption, you find that the urban per capita consumption is slightly higher than the rural per capita consumption. Whereas, for other crops such as jowar, bajra, or maize; the rural consumption per kilogram, per head is much more than the urban consumption.

Now, people may draw interesting conclusions from there. And, then they may try to verify through other sources. So, one may say that per capita consumption for each of the cereals except, say wheat is higher for the urban people and for the rural people. Now, what does it mean? Does it mean that, the rural people have more income or they simply eat more. So, one may like to investigate this phenomena. Similarly, for wheat you have more consumption for the urban people. So, that means, either urban people have more access to the wheat or they prominently eat a wheat product. That is why like bread, which is normally available in the market, so their 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 cereals. For example, rice consumption is maximum compared to wheat than jowar, bajra and maize. And, this also may tell that about the

relative importance of the crops for growing. And therefore, the Government can also decide the policies; the farmers can decide their policy about these things.

(Refer Slide Time: 53:04)



Some of the various, extremely simple type of representations are through pictorial diagrams, which may not be very useful. But, sometimes they may be helpful, for example, in the year 1990 volume of exports was nearly 30,000 crores; in the year 1920 into year 2,000 it was nearly 40,000 crores; in the year 2010 it was nearly 50,000 crores. So, here each unit here represents a volume of 10,000 cores. But, this type of representation has clear-cut limitations because you cannot use it for various purposes; mostly, if you have to use the 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 five year plan to various sectors of the economy.

So, here for example, the Government will spend 20 percentage on services; it will spend, say 19 percentage on transport and communication, say infrastructure; it will spend 7 percentage on agricultural or related fields; it will spend 9 percentage on rural development; the Government will spend some 3 percentage of its amount on special

area developments; 10 percentage on irrigation or flood control exercise; some 23 percentage on energy and some 2 percentage on Science and Technology etcetera; some 7 percentage 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 represent frequency distribution, how to classify the frequency distribution and what basic measures of central tendency or the discussion can be evaluated from there. So, in the next lecture we will be taking up these things. Thank You.