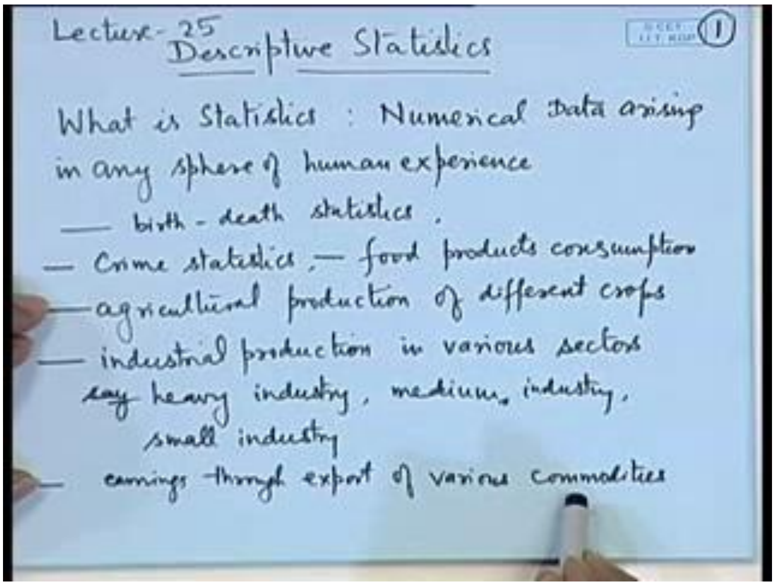
Probability and Statistics

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

Lecture - 49
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 kind of discrete and continuous distribution, we discussed joint distributions and we discuss 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 a statistics that are used and also the representation of the data through statistics.

(Refer Slide Time: 01:24)



Lecture-25
Descriptive Statistics

What is Statistics : Numerical Data arising in any sphere of human experience

— birth - death statistics .
— Crime statistics , — food products consumption
— agricultural production of different crops
— industrial production in various sectors say heavy industry , medium industry , small industry
— earnings through export of various commodities

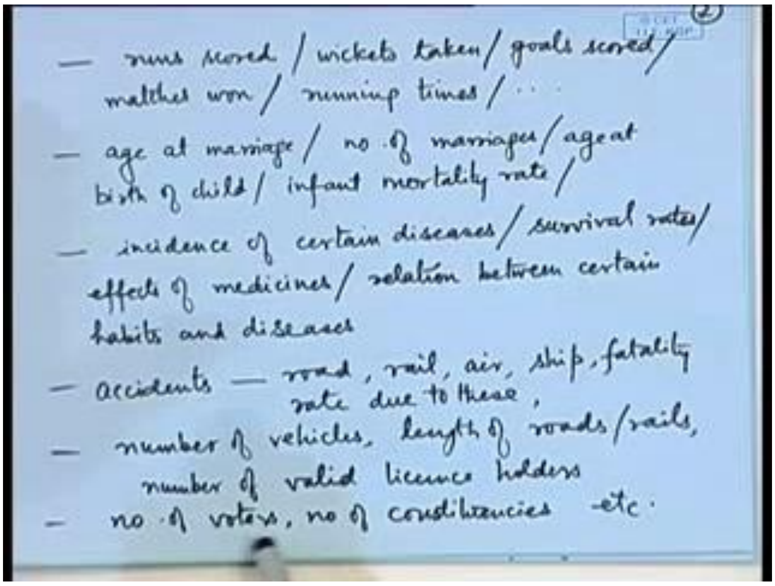
Let me start with the term, what is a statistics? So, in plural sense when statistics word is used it refers to numerical data which arises in any sphere of human experience. So, in our day to day life from the following examples, I will try to show that everywhere we are making use of the statistics. For example, the records of birth and death are kept in every minutes by this or in village or in town of this. So, this data the presents 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 Chinese dishes? So, the data on that comes under food product 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 product from the other countries to need 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 important 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 industry, steel industry. Similarly what is the production for the medium sectors or the small sector industries complete industrial production data is a 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. From this point of view, the industrial production data are the statistic 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 incentive for export of on export item of that particular type and then it want to see the effect on the 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, software and then whether it leads to the increasing the export of that particular item. Now these are some (Refer Time: 05:25) thing consequences; for example, if you see that there is an increasing the exports; 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. So this type of data is of importance.

(Refer Slide Time: 05:41)



- runs scored / wickets taken / goals scored / matches won / running times / · · ·
- age at marriage / no. of marriages / age at birth of child / infant mortality rate /
- incidence of certain diseases / survival rates / effects of medicines / relation between certain habits and diseases
- accidents — road, rail, air, ship, fatality rate due to these,
- number of vehicles, length of roads / rails, number of valid licence holders
- no. of voters, no of constituencies etc.

The 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 or in hockey match, how many goals are scored, how matches are won by different teams. In say athletics, what are the running times of that of athletes for certain events, say 100 meter race or 200 meter race, etcetera. In socially studies, we may keep record of the age at marriages of the women, age at marriage of the men and average how many marriages percents are doing in a life time, 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 or formulating various cultural policies and also social policies by the government for health and family welfare and for you have a kind of things.

In medical we keep track of the statistics on the incidents of certain diseases. For example, what is the incident of severe scoliosis? What is the incident of malaria? If it is more, then government will try to formulate a policy so that it can reduce the incidents 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 incidents of those diseases can be reduced if there is a new epidemic then how to control that. So, the data under incidents of diseases is regularly monitored by the government agencies by the medical council.

What are the survival rates for certain diseases? Whether if the survival rate is less that is new drugs should be introduced so that the survival rate can be increased, what are the effects of the medicines? For examples, if any medicine cures a disease, but it creates certain side effects so that the person dies after something else. So, we 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 a smoking and say lung cancer or TB.

Therefore, those causes should not be encouraged which give raise to certain diseases, for example, if locality unhygienic then it may give raise to cholera disease etcetera. 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 curbing the incidents 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 and ship accidents and what is the mortality of fatality rate due to these accidents. For example, if it is observed that there are large numbers of accidents due to a particular air craft then the air craft may be grounded and that particular will be ask you rectify that kind of those kind of defects so that such accidents are avoided.

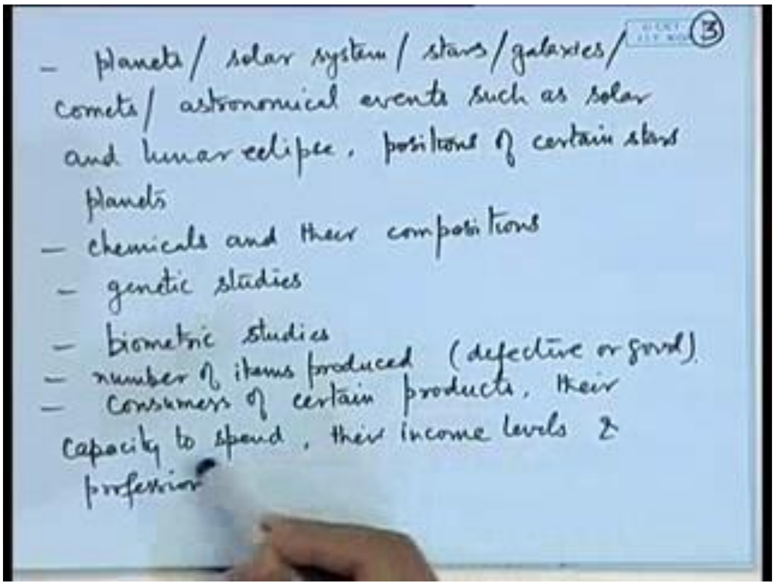
For example, if accidents are occurring due to collision of trains then certain devices should be installed which will need to rejecting that there is another train on the same rail line nearby and therefore, the driver should be alerted. What steps can be taken to reduce the number of road accidents? So, this information on actual accidents and also the fatality rate 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; 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, this type of information evaluate is 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 at between age group from 18 to 25? How many of a between say 25 to 35 so that the leaders are 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 technology collected is also.

(Refer Slide Time: 11:46)



③

- planets / solar system / stars / galaxies / comets / astronomical events such as solar and lunar eclipse. positions of certain stars planets
- chemicals and their compositions
- genetic studies
- biometric studies
- number of items produced (defective or good)
- Consumers of certain products, their capacity to spend, their income levels & profession

For example, physics, chemistry, biotechnology, in each of they are say economics everywhere the use of a statistics is there. For example, the physicist will be interested in knowing the movement of the planets, the number of planets, what are there in the solar system? What is the number of stars in a particular galaxy? What is the number of galaxies comets astronomical events such as a movement of a certain comet are hitting it into another planet, etcetera. We want such as solar or lunar eclipse position of certain stars and planets etcetera the data on all these things are extremely useful to the cosmologists.

This is also use the data in other discipline such as movement in the number of say electrons or number of neutrons are various protons and also they are using in various mettler metallurgical studies the composition; the people in chemistry or chemical sciences they study various kind of chemicals and their compositions; the biologist base study genetic may cup of certain human beings genetic may cup of certain item, there are biometric studies.

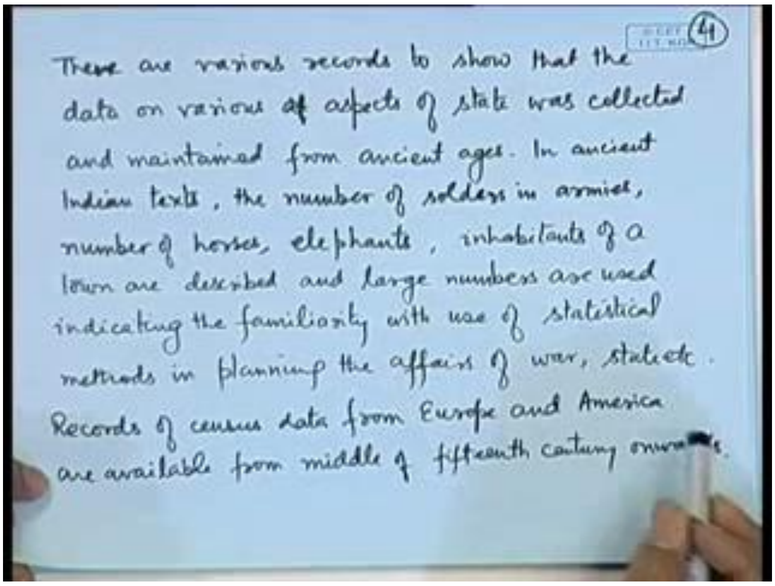
In industry when an manufacturer is doing done then what is the percentage of the defective items produced by that particular manufacturing process. It helps us to rectify or improve the manufacturing process so that the number of defective can be reduced. In economics the consumer of certain products, what is their capacity is 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.

For example, if a new soap is to be sold then they find out that what should be the price of soap so that it is affordable to most of the people in that particular income group. If there are people with high income in a particular place then they can introduce a soap which is of high values, and therefore they will like you make it more fashionable or more attractive so that those group of people are attracted to that.

So, we can see that almost the every area of human activity involves collection of a statistics and it is actually done, it is used by agencies, by individual, by organization, by people which are relevant to those statistics. And we can actually say that the term statistics or the usage of a statistics is as old as a civilization itself. In fact, ancient text so that for example, in a town or in a capital, a certain a state, how many inhabitants are there? How many houses were there or what is the size of the army? How many in country 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 will read in the ancient test either Indian or European or American test such data was available it is recorded.

Therefore the term is statistical or the data has been used from ancient test.

(Refer Slide Time: 15:41)

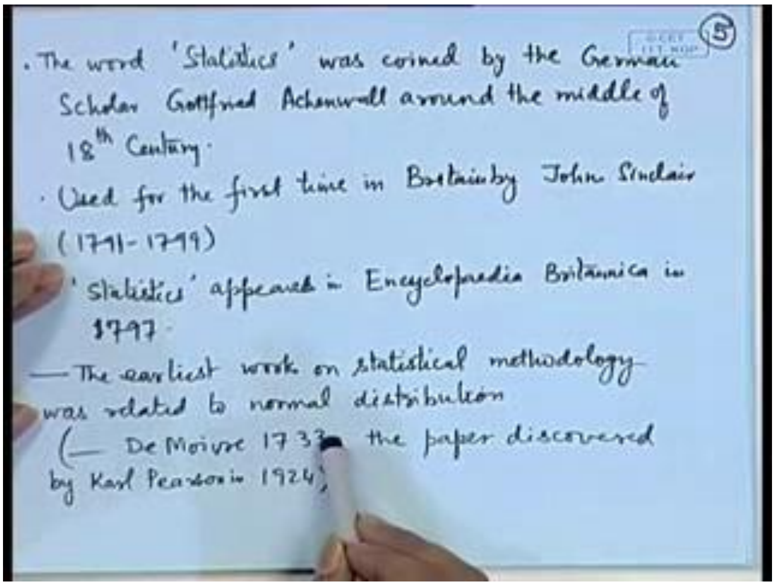


41

There are various records to show that the data on various ~~of~~ aspects of state was collected and maintained from ancient ages. In ancient Indian texts, the number of soldiers in armies, number of horses, elephants, inhabitants of a town are described and large numbers are used indicating the familiarity with use of statistical methods in planning the affairs of war, state etc. Records of census data from Europe and America are available from middle of fifteenth century onwards.

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 15th century onwards. And there is a data from England or from a Scotland and certain professions that these many people work in this profession etcetera, this kind of break up detailed breakup of the data is available from European countries are even from USA, from 15, 16 and 17th centuries etcetera. That means, the people have been aware of the importance of the data as well as a statistical methods to how to use that data, how to interpret that data from quite long time; it is not a new subject.

(Refer Slide Time: 16:38)



5

- The word 'Statistics' was coined by the German Scholar Gottfried Achenwall around the middle of 18th Century.
- Used for the first time in Britain by John Sinclair (1791-1799)
- 'Statistics' appeared in Encyclopaedia Britannica in 1797.

— The earliest work on statistical methodology was related to normal distribution

(— De Moivre 1733 the paper discovered by Karl Pearson in 1924)

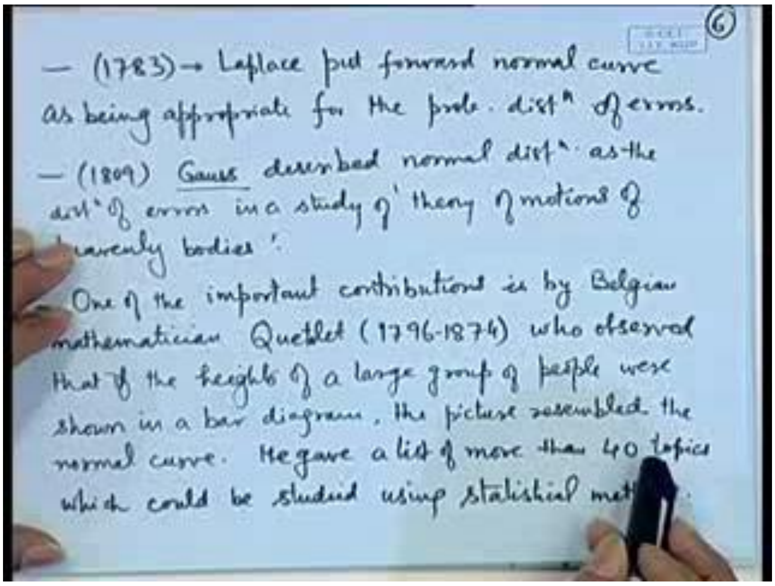
However, the term statistic as we use it for the subject statistics, he was in a singular, it was coined by the German scholar Gottfried Achenwall around the middle of 18th century. And first time it was used in Britain by John Sinclair, who wrote a series of volumes in from 1791 to 1799 which gave the communication between the ministers from a 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; now we can understand that there are 2 types of meanings of the word statistics; one is the data the 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 a 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 subject statistics.

Now we will refer it to be saying statistical methods or a statistical techniques or statistical inference. The modern statistical methodology the earliestreferences are to the normal distribution most probably De Moivre in 1733, he wrote a paper which appeared in a Oxford journal and therefore it was not known, it was later discovered in 1924 by British statistician Karl Pearson. In this most probably first for the first time, he gave what is known as normal distribution and he gave it as a or a arising in a large number of

Bernoullian trials, we obtained there is a limit of certain some soft random variables which is known as the one of the first central limit theorems.

(Refer Slide Time: 19:11)



6

— (1783) → Laplace put forward normal curve as being appropriate for the prob. dist^{n} of errors.

— (1809) Gauss described normal dist^{n} as the dist^{n} of errors in a study of 'theory of motions of heavenly bodies'.

– One of the important contributions is by Belgian mathematician Quetlet (1796-1874) who observed that if the heights of a large group of people were shown in a bar diagram, the picture resembled the normal curve. He gave a list of more than 40 topics which could be studied using statistical met

The great French mathematician Laplace in its manuscript in 1783, he said that normal curve can be used for describing the probability distribution of errors.

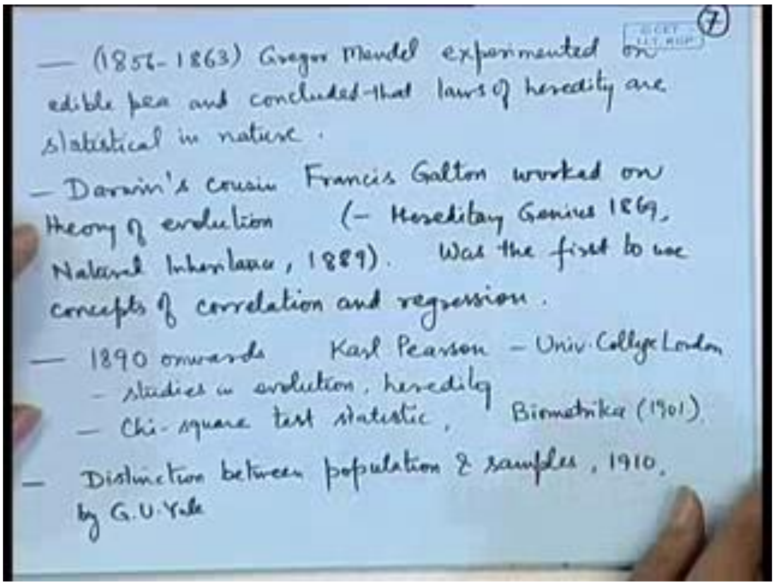
So, when we are doing certain measurements and those measurements are not accurate because of the measuring the instrument or by the person who is measuring that. So, there may be error. And you take several observations from that. So, when you several observations from that you have several value corresponding to those errors. So, if you 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 consider by study of planetary motion that is the theory of motions of heavenly bodies and he also came up with the same conclusion that the distinction of the errors for those astronomical measurements can be nicely described by the normal distribution. So, his (Refer Time: 20:29) came in 1809, but he credited Laplace by telling that the idea was from Laplace. But in fact, the normal distribution as a name Gaussian distribution because of the Gauss's marginal study or the study by the Gauss on this topic.

However, the most appropriate usage of a statistical methodology was probably done by first by Belgium mathematician Quetlet, who lived from 1796 to 1874. So, he studied apart from this planetary motion, etcetera, he studied the usual statistics such as heights of a large number of a people. So, he said that if we present this distribution of height using a bar diagram then the picture resembles that of a normal distribution. He prepared a list of various topics more than 40 topics and he submitted to British statistical association and he said that all these topics can be studied using statistical methods.

As early as earlier 19th century, people were aware that there are various phenomena where a statistical method can be used; that means, the phenomena are not deterministic this idea was there much earlier.

(Refer Slide Time: 22:13)



7

— (1856-1863) Gregor Mendel experimented on edible pea and concluded that laws of heredity are statistical in nature.

— Darwin's cousin Francis Galton worked on theory of evolution (— Hereditary Genius 1869, Natural Inheritance, 1889). Was the first to use concepts of correlation and regression.

— 1890 onwards Karl Pearson — Univ. College London
  - studies in evolution, heredity
  - Chi-square test statistic, Biometrika (1901)

— Distinction between population & samples, 1910, by G.U. Yule

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 is started to 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, he was on the theory of evolution and his (Refer Time: 22:51) 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 establish.

In particular he studied the heights of the children in comparison with the heights of their parents. One of this famous study is that the children of taller parents are taller are tall, but less tall less; taller than their father. And again the children of shorter parents are short, but taller than their parents. So, he term 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 height and we 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 1819 onwards two centers of a statistics got developed: one was in university college of London and another was the Rothamsted experimental station in England.In the University college of London, the mathematician or the biologist or the economist Karl Pearson, he started studying or we can say introducing the concept of a statistics in a very systematic way. He was going a studies in evolution heredity and also various other objects. In 1901 he started the famous journal called Biometrica it is the one of the top journals for the study of statistics. He was the first one to talk about the chi square test statistics; that means, not only that you say that this data of test the particular distribution say a given data we said a that it is fitting nicely by the normal distribution or by a gamma distribution, but how to do a statistical test for that. It is not just by a observation, but by developing a proper procedure for that.

Here the first one who introduced a chi square test statistics or you can say chi square test for goodness of this. Till this time there was a still not very clear cut distinction between a population and a random sample; that means, we are talking about full population or a subset of the population there was not much a distinction, but it was the 1910 book by G U Yule, who gave a clear cut distinction between population and sample; around the same time R A Fisher also gave a clear cut distinction between population and the samples.

Thank you.