

Essentials of Data Science with R Software -1
Professor Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur
Lecture No. 01
Data Science – Why, What and how?

Hello friends welcome to the course Essentials of Data Science with R Software 1. This is about probability theory and statistical inference. So, this is our first lecture and nowadays this word this Data Science this has become a very popular word. And if you try to see this word was not there a couple of years back but it has originated only in the last couple of years and nowadays this is one of the most popular topic to learn to get a job, etc. So, it has gained a lot of importance.

So, my first question to you is that what is this Data Science? And what do you really understand? Well, Data Science word was not there when I was a student, so how it developed how it originated and how it is related to the basic fundamental topic of statistics that is what I thought and I would like to explain you in this lecture.

So, in this lecture we are going to talk about the topic of Data Science why, what and how. So, let us begin our lecture. So, you see that the topic of my this lecture here is Data Science, why, what and how right and as I told you earlier now I change the color of my pen so you do not get confused if you see these types of action during your lecture.

(Refer Slide Time: 01:55)

What is Data Science?

What is Science?

- Special approach to find the answer of a query.
- We want to know the reason - How, Why, Where, Who...?

What is Data?

- Source of reliable information.

What is Data Science?

- A scientific approach of retrieving the information from data.

Now, the first question comes- what is really a science? What do you really mean by science? Now, in case if you ask me that what is science I will say simply suppose there are 100 students in my class and I am teaching them statistics and suppose my basic objective is that I would like to give them grades as A, B, C, D, E, F and so on. So, now I have two approaches, I can say, those students who are coming to my class with this sky blue shirt, they will be getting a grade A, those who are coming with white shirt, they will be coming they will be getting a grade, B those who are coming with yellow shirt ,they will be getting a grade C and so on.

Do you think that you will agree to these things? Surely not. What you will say this is not the scientific way. Now, then my question is this once you know what is the way which is not scientific then you also know what is the scientific approach. For example, if I give you an option that you come to my class I will take your exams, I will give you some questions based on that what I am going to teach you and then you are going to solve them and I will try to give you some marks.

And then I will take say a couple of examinations of 15 minute duration, half an hour duration, 1 hour duration, 2 hour duration and so on and then at the end I will try to add all the marks and then I will try to see that whoso however is getting say more than 80 percent marks that student will be given a grade A, any student who is getting a marks between say 70 percent and 80 percent that student will be getting a grade B. Any student who is getting a marks between say 60 and say 70 percent, that student will be getting a grade C and so on.

Do you think that if I give you this option, will you agree with me? I am sure your answer will be yes. Now, if I say is this a scientific approach. Once again I am confident that you will say yes this is the scientific approach. Now how to define the scientific approach I do not know but at least I am 100 percent confident that you know that how to discriminate between the scientific approach and non-scientific approach for finding out the correct approach and correct reason and correct outcome.

So, now in case if I say what is the scientific approach? In very simple words, I can say this is an is, this is a special approach to find the answer of a query. Now, the question is this what type of query? Why do we want to answer the queries? You see that is the advantage of being the human being. As a human being we are always interested in finding out the reason, why, how, where, etc. And that is what which makes us very different and valuable than many others in this nature.

So, one of the basic fundamental question comes that how to find the answers of these queries in a scientific way. And statistics is one of the subject, this is one of those science, which helps you in getting the answers in the logical, correct and scientific way. And that is why when people realized about statistics, they started calling it Data Science. Well, when we are talking of Data Science, our basic objective is to get an answer to our query.

Now, what are the different subjects what are the different types of tools which are going to help us in getting the correct answer in drawing the correct inference on the basis of given sample of data they all together create the subject of Data Science. For example, if I say suppose I get a data, now I need to know what statistical tool I have to implement on it. For example, if I want to determine the marks of the students then I know from the mathematical point of view I need to do addition, I need to have to find out percentage, etc. So, mathematics is helping us in giving us an idea that what type of mathematical operations are needed.

Similarly, when you come to statistics, then there are different types of tools which can be used there can be arithmetic mean, geometric mean, harmonic mean, variance, standard deviation, etc. That you know. But now, which of the tool is going to give you the correct outcome? When you are trying to determine the grades of the students on the basis of the marks obtained in the examination so here statistics helps you now.

When you are trying to deal with very large size of data means nowadays data size is becoming large. You can imagine that those shopping websites they are being accessed by how many customers every day, there are some applications like Google, Google Maps, etc. If you try to see that how many people around this world are using those websites those applications per day this data is very huge, enormous and it is very difficult for us to handle that data with our hands.

Means if you have 5 values, possibly you can find out the arithmetic mean very easily but if you have 5 million, 5 billion, 5 trillion observations it will take a very long time to find out the arithmetic mean of those observations. Remember, I am not questioning the capability of human being that is enormous, but there is something called convenience. So, this computer is going to help us but now once again the question comes how the computer is going to help us, means how I have to write down the program what I have to inform the computer so that the computer does the same thing what I want? So, here comes the intervention and help from computer science.

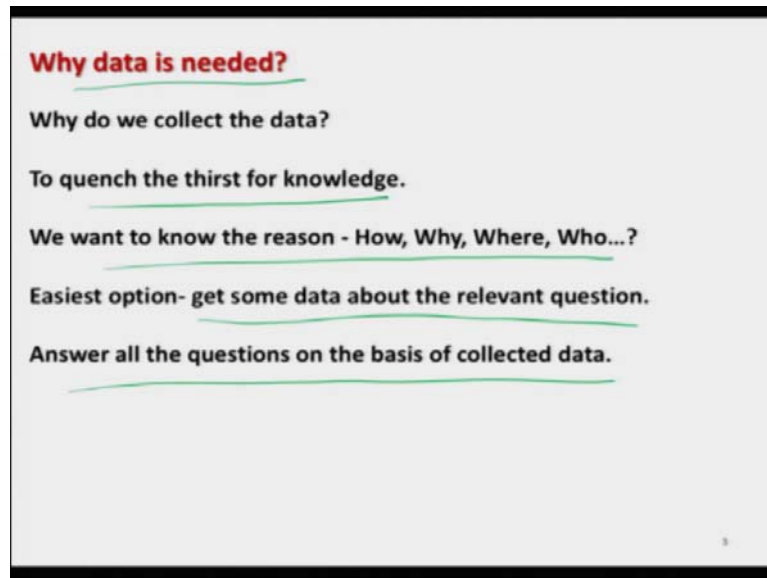
And similarly, there can be some other type of subjects, for example, if you are trying to simulate an equipment from the medical science, possibly the medical science will also come into picture. They will try to give a different types of information different pieces of information that how the instrument is going to work and then we will try to program it, we will try to simulate it, etc., etc., etc. So, now you can understand why this is called a science because each of this subject is going to give you some information and approach, each of the subject to help u,s guide us, in knowing that what should we do so that we get the correct outcome.

All this statistics, mathematics, computer science, etc., etc. they are all the constituents of this Data Science, at least in my opinion. Because if you try to see when I was a student at that time there was no word like Data Sciences and I was trained only in mathematical statistics. So, now over a period of time, this, how this statistics became Data Sciences or how stats started contributing in the Data Science. That is an issue that I am going to handle here and this I am going to make you understand and from this point of view only I am going to handle this course.

So, if you try to see as I said now what is the science? Science is only a special approach to find the answer of a query and we want to know the reason how why where who did it. Now, the next question is, the word here is Data Science so now we have understood what is here science. But what is here, data. What is here, data. Data is source of very reliable information and what is Data Science a scientific approach of retrieving the information from the data.

For example, if you ask me suppose if there are two student and I have to decide, which student is good and which student is bad? One option is that I will try to look into the marks obtained in the examination and then I will try to say that okay whosoever student has got the higher marks that student is a better student. What is that? The marks are nothing but the data. And based on that I am trying to use here a scientific approach to determine that which of the student is better that is, I am simply trying to compare the marks, so this is a part of Data Science.

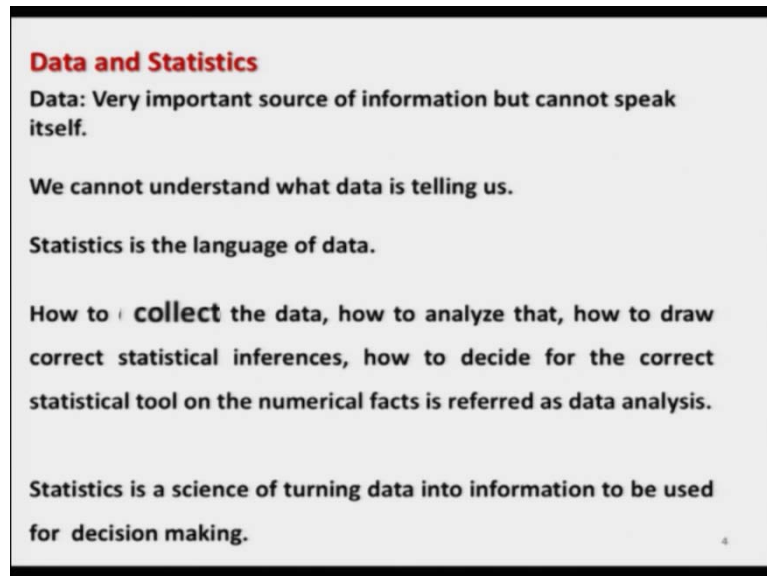
(Refer Slide Time: 12:14)



Now the question comes here, why data is needed? Why cannot we take the same inference or same conclusions without looking into the data? So, this data is collected to quench the thirst for knowledge. And we want to know as we said that we want to know the reason that how, why, where who did it. And one of the easiest approach easiest option to get the answer of such question it is to get some data about the relevant question and then we try to analyse it and then try to answer all the questions on the basis of the collected data.

For example, if I want to know whether a student is good in, say this language or mathematics or say biology then what are we going to do? We are simply going to collect the data on the marks obtained in various examinations and then I will try to say that in whatsoever subject the student is getting better marks the student is better in that subject. So, now I am trying to answer the question that whether the student is good or bad in any subject based on the collected marks.

(Refer Slide Time: 13:25)



Data and Statistics

Data: Very important source of information but cannot speak itself.

We cannot understand what data is telling us.

Statistics is the language of data.

How to collect the data, how to analyze that, how to draw correct statistical inferences, how to decide for the correct statistical tool on the numerical facts is referred as data analysis.

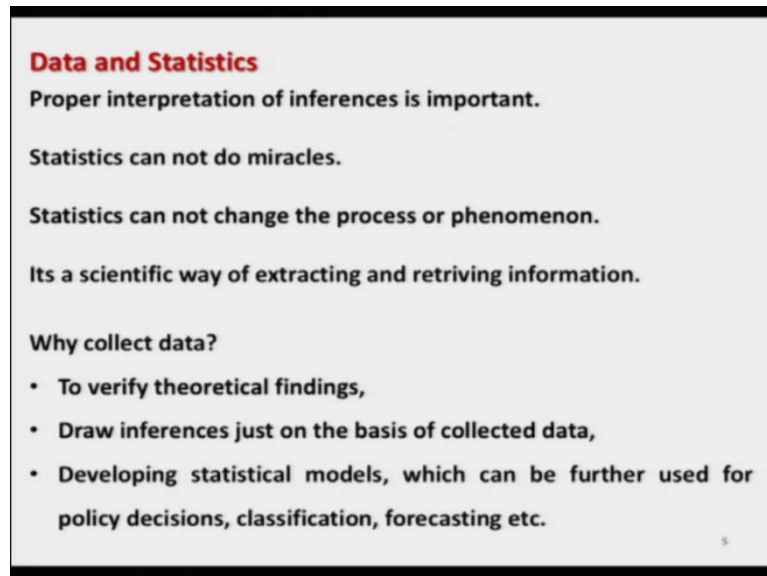
Statistics is a science of turning data into information to be used for decision making.

So, now the question comes, what is the relationship between statistics and data? So, data is a very important source of information but the problem is this the data cannot speak itself. But data has enormous information if there is somebody who has a disability to listen or to speak, the data is just like that but it does not mean that that person is less than anybody else. That person may be far more superior than many other human beings that is my promise to you. The only thing is this we have to learn, we have to understand the language that how to communicate.

So similar is the story with data that we cannot understand what data is trying to inform us and statistics is the language of data which communicates or establish a communication between the human being and the data.

Now, the next question is that how are we going to collect the data? How to analyse the data? and how to draw correct statistical inferences out of this? And how to decide for the correct statistical tool that has to be implemented on the basis of numerical facts? This is all the part of data analysis. Now, what is the statistics doing here? So, I can say now that here that statistics is a science of turning data into information which is to be used for decision making.

(Refer Slide Time: 15:09)



Data and Statistics

Proper interpretation of inferences is important.

Statistics can not do miracles.

Statistics can not change the process or phenomenon.

Its a scientific way of extracting and retriving information.

Why collect data?

- To verify theoretical findings,
- Draw inferences just on the basis of collected data,
- Developing statistical models, which can be further used for policy decisions, classification, forecasting etc.

5

As a human being we are always interested in making a decision. But, when we are trying to make that decision, it is very important that we have to understand, what data is trying to inform us. And based on that proper interpretation of the inferences is very important. For example, if I say somebody has got say 70 marks in mathematics and say 30 marks in say this biology. Does this always mean that the person is better in mathematics?

Do not you think that it may happen at the question paper in biologic biological sciences or in biology is very difficult because that is students is unable to score say higher marks like 70-80. But, the question paper in say this mathematics subject is very simple or easy where the student can get more marks. So, do you think that just by looking at the marks, you can always conclude the student is good and bad in that subject or not. This is how you have to think and this is what I mean when I am trying to write down here proper interpretation.

And for that statistics comes to our rescue and it helps us a lot but remember one thing that statistics cannot do miracles. And as a statistician, we are not allowed to change the process by which we are trying to collect the data and also you have to understand and always keep in mind that statistics cannot change the process or the phenomena. That can only observe the phenomena and give us the information in the form of data that is collected from that process or that phenomena.

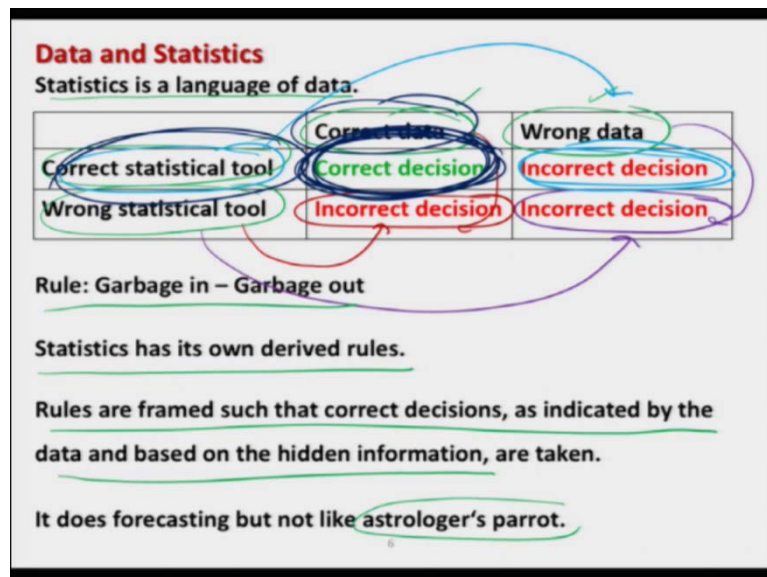
So, statistics is a scientific way of extracting and retrieving the information which is hidden inside the data. So, now the question is why should I collect the data? So, there can be many reasons for this. The first can be to verify the theoretical findings. For example, if I say that in

case if the salary or the income of a person increases the quality of life becomes better. Well, that is my statement and how will you support it or how will you verify it whether my statement is right or wrong.

So, for that we can collect the data on the salary, we can observe the living standards whether they are increasing or decreasing and based on that we can conclude whether by increasing the income or the salary, whether the things are going to improve in terms of standards of life or the living conditions. And then after that we are going to draw the statistical inferences just on the basis of the collected data. That for example, in case if my data collected on the standard of leaving and say this income, they are indicating that as the income is increasing, the standard of living is also increasing. So, that is our statistical inference.

Now, how to draw this is statistical inference, what type of statistical tools are needed, what type of models are needed and how do you classify that whether the standard is really increasing or decreasing, etc., etc. Those things are obtained on the basis of collected data. So, developing statistical models which can be further used for policy decision, classification, forecasting, etc. that is the ultimate objective of statistics or equivalently the Data Science.

(Refer Slide Time: 18:44)



So now, we have understood that this statistic is essentially a language of data. Now whenever you are trying to answer a question or a query, now we have understood that we need to collect the data. Now, there are two options the data can be correct or the data can be wrong. What do you mean by this? For example if I want to measure the average income and we collect the data on the height of the human being who is earning the income.

So, the heights of the human being is going to give us the information about their income, do you think that is it the correct data? Well in case if you want to know the average height of the human being then possibly the collecting the data on the heights of this of the human being will make sense. But, if you want to know the average income then obviously the collection of data on the income is going to make a sense. So, from that point of view, the data can always be classified as correct data or a wrong data.

And now, whenever you are trying to collect the data, after that the next step is to use a statistical tool. So, this statistical tool can also be correct or the statistical tool can also be wrong. So now, based on the two classes of correct and wrong data and two classes of correct and wrong statistical tool application, we have 4 possible options.

That first we try to use here a wrong statistical tool on the wrong data. So definitely this is going to give us then incorrect decision. Now, we try to use the wrong statistical tool on the correct data so that is again going to give us the an incorrect decision. Now, I try to use the correct statistical tool but on a wrong data, so this is again going to give us the incorrect decision. Now, finally I try to choose here the correct statistical tool and correct data and this is going to give us the correct decision, right. That you can see.

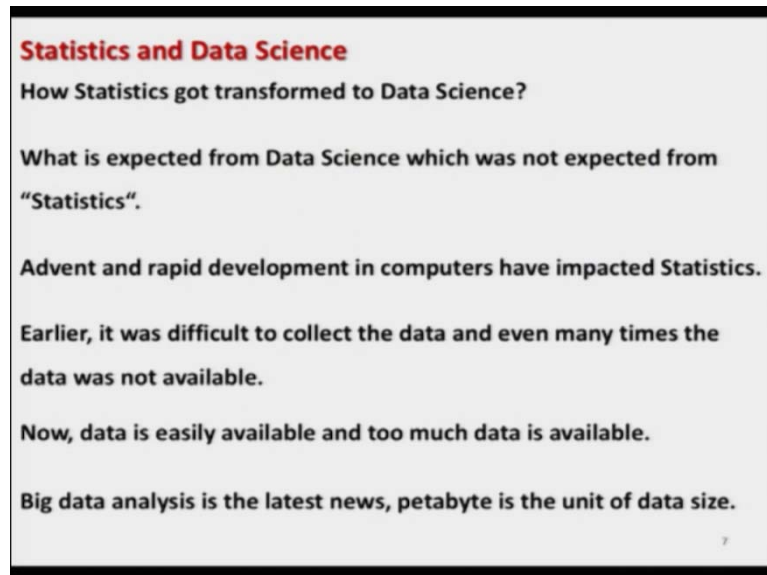
So now, you can see here this that this Data Sciences or statistics they are going to be useful in the sense that they are going to give you the correct statistical outcome only when you try to use the correct statistical tool over the correct statistical data and that is the objective of this course. And I want to make you learn that how to decide that which of the data is the correct data and which of the data is wrong data and which of the statistical tool is correct and which of the statistical tool is wrong.

With right and wrong, I means the application the statistical tool is not wrong but the application of statistical tool is wrong. So, the simple rule here is Garbage in - Garbage out. Means if you give a wrong data and if you use a wrong statistical tool means it is going to give us the wrong outcome and the only outcome is that, that you have to use the correct statistical tool on the correctly observed data.

And statistics has its own rules which are found or derived when we are trying to develop the rule and those rules are the need from the statistical point of view, from the mathematical point of view or for the scientific development of the tool and these rules are framed such that the correct decision as indicated by the data and based on the hidden information are taken.

So, right, based on these tools one can all one can always do a for example, forecasting. But, when it comes to forecasting means you must have seen some time there is a parrot which is helping an astrologer to tell the future. That parrot will take out a small chit and the astrologer will try to read from the chit and it will try to forecast the future. But do you think that that is the scientific way of forecasting, but definitely if you try to get the data try to develop a model and can try to do a forecasting that will be a scientific approach.

(Refer Slide Time: 23:03)



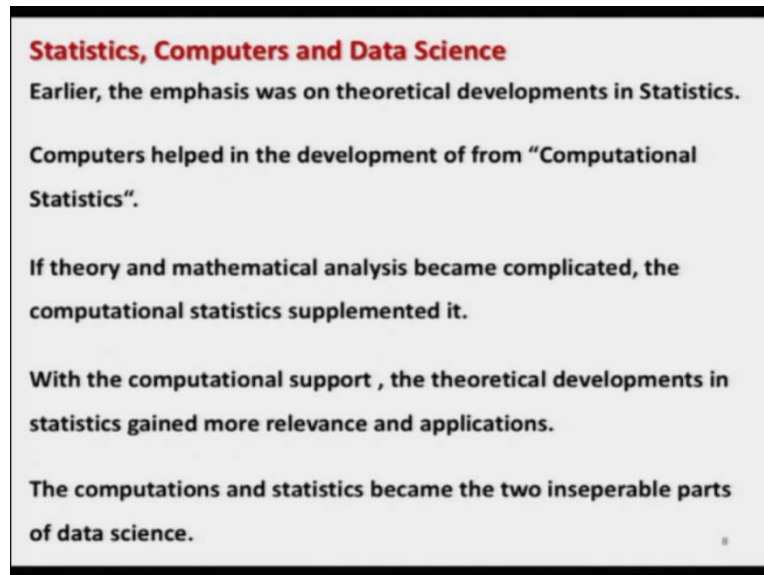
Statistics and Data Science
How Statistics got transformed to Data Science?
What is expected from Data Science which was not expected from "Statistics".
Advent and rapid development in computers have impacted Statistics.
Earlier, it was difficult to collect the data and even many times the data was not available.
Now, data is easily available and too much data is available.
Big data analysis is the latest news, petabyte is the unit of data size.

So, this statistics help you in forecasting in a scientific way. Now the question is this how this statistics got transformed into Data Science and what is the, what is expected from the Data Science which was not expected earlier from statistics. So, you know that in the last two decades, there was a rapid development in the computers and computational techniques and that has impacted our lives. So, 25 years back there was no internet in India and after that now you can see that internet has changed our lives.

So, this advent and rapid development in the computers have impacted the statistics also. Means earlier it was very difficult to collect the data and even many times the data was not available and as a student of statistics, many times people try to read it only in the book and we assume that suppose if this type of data is available, then this type of tool is going to be used. But now, now the data is easily available and is actually too much data is available you can imagine the data what a google server or amazon servers are storing every day.

Now, there is no more a question of small data but we have huge amount of data and means earlier that started with megabites, gigabyte and now we are talking of not only the terabytes but petabytes is the unit of the data in the context of the big data analysis, right.

(Refer Slide Time: 24:42)

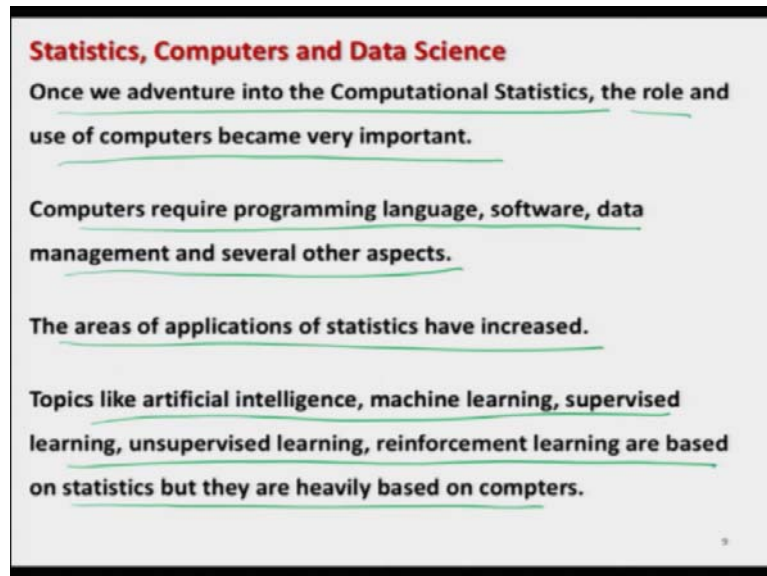


So, earlier there was an emphasis only on the theoretical development in statistics and you will see still that is there because theoretical development still is the foundation for the Data Sciences and big a big data analysis. But, many times after developing the theory it was not possible to judge whether the tools are really working on a given set of data or not or what is the quality of statistical inferences that we are drawn by the application of those statistical tool. So, then computers helped us and computer helped in the development of so called, an area in statistics which is now termed as computational statistics.

And nowadays, we have an advantage that in case if the theory or the mathematical analysis becomes complicated, the computational statistics help us and it supplements us. Because, you see all the tools which you are going to use in statistics, they are not coming from sky but they are developed on the basis of some mathematics and then we have to use a mathematical tool and many time it becomes too complicated for us to understand or to apply those tools.

So, with this computational support that theoretical developments and statistics gained and more relevance and applications were created. Actually nowadays, the computation and statistic they have become the two inseperable parts of Data Science and they together are trying to give you all the information.

(Refer Slide Time: 26:15)



Statistics, Computers and Data Science

Once we adventure into the Computational Statistics, the role and use of computers became very important.

Computers require programming language, software, data management and several other aspects.

The areas of applications of statistics have increased.

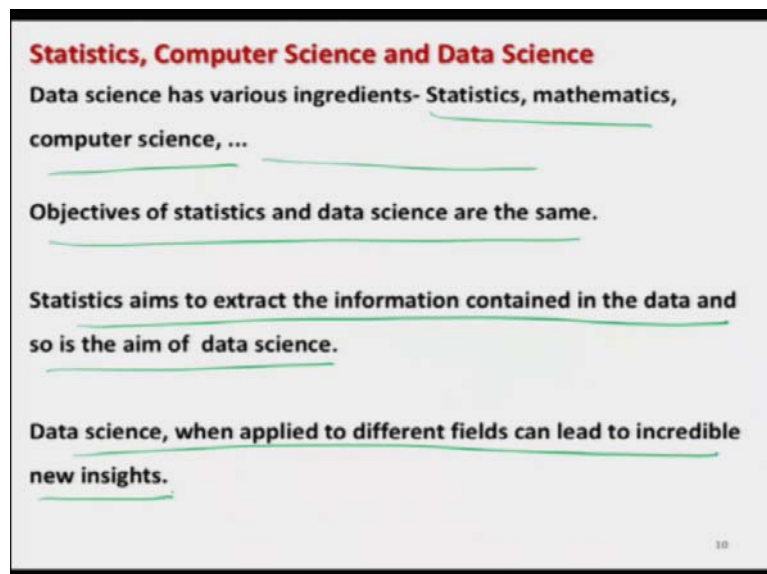
Topics like artificial intelligence, machine learning, supervised learning, unsupervised learning, reinforcement learning are based on statistics but they are heavily based on computers.

9

And once we adventure into the computational statistic, the role and use of computers become became very important. Now, once you try to take the help from computers, the computers do require programming language, software, data management and several other aspects. And because of these things the areas of application of statistics, they have also increased tremendously in the last decade.

Topics like artificial intelligence, machine learning, supervised learning, unsupervised learning, reinforced learning statistics they all are based on statistics they are trying to use the concept which are from statistics but they are heavily based on computers. You cannot do it manually, I am not challenging that you cannot do but it is very difficult.

(Refer Slide Time: 27:11)



Statistics, Computer Science and Data Science

Data science has various ingredients- Statistics, mathematics, computer science, ...

Objectives of statistics and data science are the same.

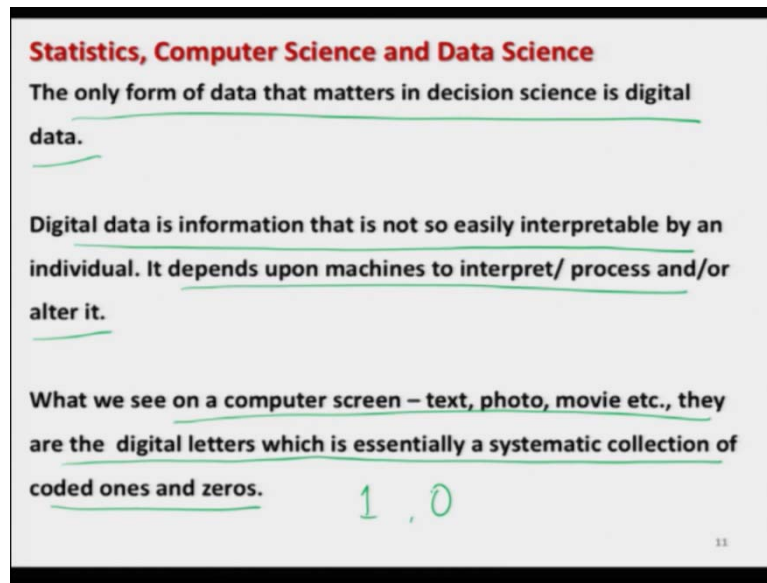
Statistics aims to extract the information contained in the data and so is the aim of data science.

Data science, when applied to different fields can lead to incredible new insights.

10

So, this Data Science has various ingredients, this is statistics, mathematics, computer science and different types of sciences but the objectives of statistics and data sciences they are the same. Statistics aims to extract the information which is contained in the data and that is the same aim of the Data Science tool. And this Data Science where when applied to different fields can lead to incredible new insights, believe me, right.

(Refer Slide Time: 27:44)



And the only form of data that matters in decision sciences is the digital data because what we want that the data should be available in an automated way and possibly means automated way in the sense that whenever we want to use any statistical tool, we are going to use the software and that data should be available in a format which is directly applied to the software.

Now, the digital data is the information that is not so easily interpretable by an individual, it depends on the machine to interpret or process it or alter it, right. So, what we see on a computer screen that can be a text that can be a photograph that can be a movie, that can be a video they are all digital letters which are essentially a systematic collection of the coded ones and zeros. Ones and zero means number 1 and 0.

Those who are from the computer science background they will understand it very easily that whatever we are seeing on the screen, a picture, a jpeg file or a this dot avi file etc., they are they have got different formats and what they are looking on the screen that is only a combination of 0 and 1s in a proper way.

(Refer Slide Time: 29:09)

Expectation from Data Scientist

What is needed to become the data scientist?

First decide what we want to become- A Doctor or a Compounder?

Decide-

Want to only use the tools?

Want to understand the utility of tools?

Or want to develop the tools?

In my opinion- all are needed.

12

So, now what are the expectations from a data scientist? What is really needed to become the data scientist? Before we decide for this thing, we have to first decide that suppose somebody wants to go into the medical science, then first, the person has to decide, what he or she wants to become a doctor or a compounder? Do you know what is the difference between a doctor and a compounder? The compounder will only follow the instructions of a doctor and doctor is the one who is going to decide that which of the medicine is going to be going to work in a particular element or a problem or which of the medicine is going to help the human beings.

The compounder will simply prepare the compound and will pass it on to the patients. So now, you can see I am not saying that both are not important. Both are equally important but they have different roles. The compounder cannot take a call that which of the decision is going to or which of the medicine is actually going to work. So now, between these two tools now you have to decide what you want to do, do you want to use only the tools or you want to understand the only the utility of the tools or you want to develop the tools?

If you ask me, I will say you need to know all the things means you should know how to use the tool, you should know where to use the tool and you must also now know how to develop the tools and that is the same story with the Data Science. That you should know how to collect the data, you should know how to implement the data and you should know how to develop a tool for a given problem to understand the answer of a particular query because the questions in the real life, they are not coming from sky they are generated by the human being and according to their need one needs to develop a proper statistical tool.

(Refer Slide Time: 31:15)

Role of Statistics in Data Science
Statistics is the soul of data science.

• Descriptive statistics ✓	• Nonparametric inference ✓
• Probability theory ✓	• Multivariate analysis ✓
• Statistical inference ✓	• Linear regression analysis ✓
• Decision theory ✓	• Nonlinear regression analysis ✓
• Bayesian inference ✓	• Simulation techniques ✓
• Frequentist inference ✓	• Monte Carlo methods ✓
• Parametric inference ✓	• ✓

13

So, when you come to statistics, statistics is the soul of Data Sciences and there are different types of subjects which are contributing in in the understanding of Data Science that can be descriptive statistics, probability theories, statistical inference, decision theory, Bayesian inference, frequentist inference, parametric inference, non-parametric inference, multivariate analysis, linear regression analysis, non-linear regression analysis, simulation technique, Monte Carlo method ... so many. So, but there is always a point where we start.

(Refer Slide Time: 31:55)

Role of Statistics in Data Science

The theoretical developments are essential which are needed to be exposed to computational procedures.

Computational procedures have their own limitations and so optimization methods are required.

The implementation of statistical, mathematical, optimization methods etc. are to be simultaneously implemented over a data set and for that, data management is required.

All these aspects are logically implemented in a systematic way and correct statistical inferences are drawn.

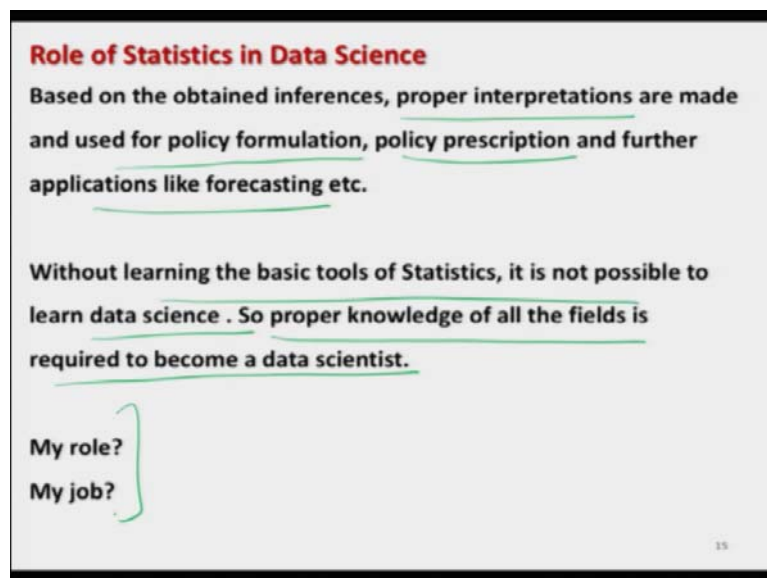
14

So, and this course is going to give you a starting point to learn these topics. The theoretical developments what are needed in the Data Sciences they are essential, why? Because they are the tools which are exposed to the computational procedure means if you do not know

whether you want to find out the arithmetic mean or median, I mean how what are you really going to compute in the computer through the software. Also, the computational procedure have their own limitations. For example, if you want to invert a matrix of 1000 by 1000, you just cannot do it manually or you cannot use by that by the usual procedure, you need some optimization procedures, some numerical techniques.

So, some optimization procedure numerical techniques, they are also important and the implementation of statistical, mathematical, optimization method, etc. and all of them are being exposed simultaneously they are implemented simultaneously over a data set. So now, it means how to manage your data sets so that these tools can be implemented that is another issue so for that data management is required and when all these aspects they are logically implemented in a systematic way, then only the correct statistical inferences are drawn after the use of appropriate statistical tool, right.

(Refer Slide Time: 33:28)



So, once you get such an outcome based on those outcomes you are going to take a correct statistical inference and when you are trying to draw the statistical inference, you have to keep in mind that they have got the proper interpretation. And which are like that that they can be used for policy formulation, policy prescription and they can be used in different types of application like say for example forecasting.

And without learning the tools of the basic tools in statistics it is not possible to learn Data Sciences. So, proper knowledge of all the field is required to become a data scientist. So, now the question here is what is my role and what is my job. Now in case if you try to recall that

many times we here or people try to tell us different types of jokes about statistics. Well, there they are very popular. Now, I ask you a question. Try to think about them and once you learn the basic tools of statistics then you try to think about those jokes once again, all the jokes will finally give you an answer that the person is using a wrong statistical tool at the right place.

Well, somebody wants to know something but for that the correct statistical tool is needed. But if you try to use a wrong statistical tool that is definitely going to give you a wrong interpretation. For example, there is a very popular joke that some person was dissecting the frog. So what that person does, first the one of the leg of the frog was taken away that was chopped, so the frog was still jumping, then that person chopped the second leg of the frog, the frog then the frog was still jumping then the third so-called hand or leg whatever you want to call for the frog that was chopped off but still the frog was jumping but then finally the last hand or the, say leg of the frog was chopped off then finally the frog stopped jumping.

So, the person concluded that if you chop off all the 4 legs of the frog, then the frogs cannot hear. Because, he was asking he or she was asking again jump, jump, jump and the frog was not jumping so that is a wrong statistical interpretation of the data, that is what I always mean that once you get a numerical value you have to interpret it properly. So, my role, my job here is very simple, I want to make you learn the basic topics of statistics which are going to lay the foundation of this Data Science.

Now, as I said there is not one, there is not two, there are various topics which are needed for the learning of Data Science my problem is that I cannot teach you all at the same time means I am also a human being and you are also a human being, so we have to choose a point from where we can start. Whenever you want to climb there is always a first stair, first step once you come to the first step, then you take another step, then you take one more step and after taking couple of steps, you will reach to a certain height.

So, this course is something like the first step. My objective in this course will be that I will be dealing with the topics of probability theory and statistical inferences. I will not be giving you the mathematical details in depth but my objective will be that these are the tools. I will try to give an idea that how they are developed, what are the condition under which they can be used and now when those theoretical concepts, they have to be implemented over a software, then how are you going to implement it and how are you going to take the correct statistical outcome. And how are you going to interpret it correctly that is my only objective.

Well, I am not saying that what I am going to tell you from the computer point of view, from the simulation point of view, from the programming point of view or for the implementation of a theoretical concept in a software, there can be hundreds and thousands of ways but my objective is only to initiate a thought process inside you. I am sure that I cannot make you a data scientist in 30 hours but definitely I can promise you I can start a thinking process for becoming a data scientist, data scientist inside you.

At least you will know that how you have to learn how you have to proceed further. Once you take the first step then second is then third step and finally I am sure that the climbing will continue and then you will reach to much greater heights in your career as a data scientist.

So, from the next lecture, I will try to begin with the topics which are essentially needed for this course and you try to have a quick revision of the concepts of R software. Although I will try to give you the minimal basic concept which are needed to learn this course from the R point of view but it will be nice if you can come with a quick revision of your concepts or the I will say R software. So, I will see you in the next lecture, till then good bye.