

**Essentials of Data Science with R Software - 2**  
**Sampling Theory and Linear Regression Analysis**  
**Prof. Shalabh**  
**Department of Mathematics and Statistics**  
**Indian Institute of Technology, Kanpur**

**Lecture - 01**  
**What is Data Science?**

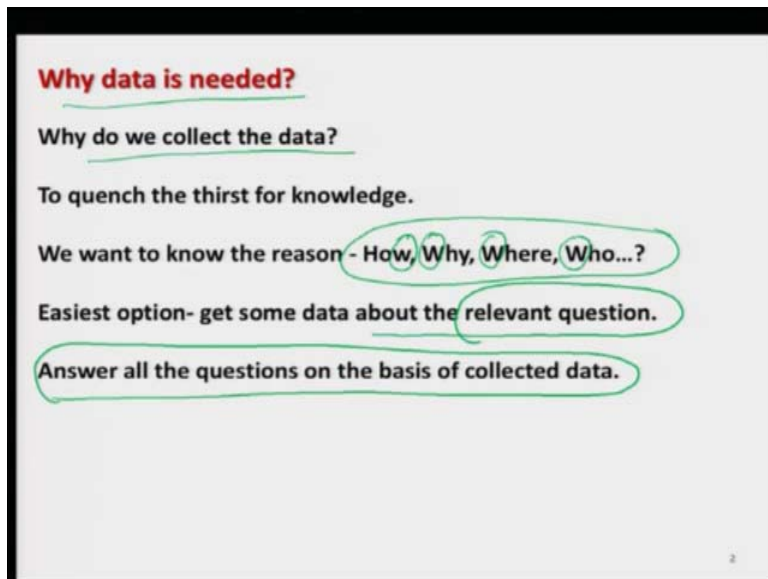
Hello, welcome to the course Essentials of Data Science with R Software- 2, where we are handling the topics of Sampling Theory and Linear regression analysis. And in this lecture we are going to talk about what is Data Science. I think this has become a very popular word in the last couple of years. And now in case if you try to see one of the most popular ambition of a student is to become the data scientist.

Well, what is data science? How do you become a data scientist? These are the questions which different people are trying to answer from different perspective. And similarly, I am one among those. My background is from statistics means I am trained as a mathematical statistician, but now with the time this decision sciences, data sciences these types of different terminologies have come into picture.

So, now I am going to address here a small lecture on what is data science and my objective will be that being trained as a mathematical statistician how do I pursue it, how do I view it, and how will I communicate it to the students. Different people from different background will try to address the topics in data sciences in a different way.

So, similar to those these are the views which I am going to present here they are my personal views means I have thought over this process, I have thought over this topic and then I collected my own ideas which I am going to present in this lecture.

So, let us start to understand what is data science from the perspective of a mathematical statistician? (Refer Slide Time: 02:40)



Now, the first question comes here, why the data is really needed? Why do we collect the data? That is the first question. One of the most simple answer to this question is that we want to quench our thirst for the knowledge.

What type of knowledge? Whenever something happens, as a scientist, as a student of science stream or even as a normal human being you always want to know what is the reason, why it happened, how it happened, where it happened, who did it and so on? So, there are enormous question related to w.

What is this here w? This you can see here this here is w, this here is w, this here is w, this here is w. So, there is a w in how, there is a w in why, there is a w in where, there is a w in who and these are the simple questions which you would like to answer.

Now, the question is how to answer them? First thing is this you will ask your teachers, you will ask your elders and they will try to explain you from their experiences, there may be some processes which are unknown to us which are happening possibly for the first time. For those people will only make a guess, they will not be confident that why the process is happening in that way.

So, under those situations particularly which are happening for the first time or the process which you do not know what is really going to be the outcome, in those cases when we want to answer the question the best thing or the most simple thing is to collect some data, right.

And this data has to be collected on some relevant question. Suppose if I want to know what will be the growth of India or the industrial growth of India in the next 2 years or next 5 years,

then I have to collect the data on those factors which are going to impact the industrial growth of this country, right.

So, what you have to do? That you have to keep in mind that you are trying to collect the data only on the relevant questions and then whatever are my questions which I have which I have framed here under how, why, where, who etc. etc., they have to be answered on the basis of the collected set of data. This is our basic idea, this is our basic objective that we want to achieve on the basis of the data that we collect, right.

(Refer Slide Time: 05:55)

**Data and Statistics**  
Data: Very important source of information but deaf and dumb.  
Data has its own language like sign language.  
Statistics is the language of data.  
Collection, analysis and drawing inferences from numerical facts, referred as data analysis.  
Statistics is a science of turning data into information to be used for decision making.

Now, the next question comes, we always talk of data. We always say collect the data, find the data, get the data, but what is this data? Say, if I try to explain it or I try to understand it in a most simple way, I would say that data is a very important source of information.

Data contains lots of information inside it, but the problem is this- data cannot speak data cannot explain to us just like a human being. In simple words means I can say that suppose data is deaf and dumb. Data cannot speak, data cannot hear, cannot listen to us.

So, now on the one hand, the data is containing lot of information inside it and on other hand, the data cannot speak data cannot listen to us. So, the question is how to communicate with the data, how to get the information out from the data which is contained inside it.

For example, you might have seen on TV, there is a special news for physically challenged people and in that news the speaker, the anchor uses some sign language they will try to

use their fingers or hands to communicate to those people who do not have the ability to hear or speak.

Well, I do not know that language, but those people who know that language they can communicate with those people and they can understand what they are trying to convey. Similar is the case with the data. Data has also got a language. My language is different it might be English, it might be Hindi or it might be any other language.

And I do not know the language which data speaks, it is just like suppose I do not know Bengali language, Tamil language, Telugu language etcetera. But if I learn it, if I understand it then it is very easy for me to communicate means anybody who is speaking in Bengali, Tamil, or Telugu, I can easily communicate and I can easily understand what they are trying to explain.

So, similar is the language of the data. So, in case if we can understand or if we can learn the language of the data then we can easily communicate with it and whatever information which is hidden inside the data that we will be able to understand. So, that is the basic idea.

So, in this slide, this is exactly what I have written that data is a very important source of information, but it is deaf and dumb. But data has got its own language, just like a sign language. And what is the statistics? A statistic is actually the language of data. So, if you understand statistics, you will understand what the data is trying to communicate to us.

What information the data is trying to convey to us and when we try to collect the data, we analyze it and then using some statistical tool we try to draw statistical inferences from those numerical observations, numerical fact, then we refer to as collectively as a data analysis. So, data analysis is not one thing, but data analysis is a collection of several steps which are conducted in a sequential way, in a logical way to find out the answer of our question.

So, briefly I can say that statistics is a science of turning the data into information, which can be further used in decision making .

(Refer Slide Time: 11:02)

**Data and Statistics**

Proper interpretation of inferences which are drawn is important.

Statistics can not do miracles.

Statistics can not change the process or phenomenon.

Its a scientific way of extracting and retring 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.

So, this is how I can briefly explain what is data analysis, what is statistics and what is the relationship between the two.

Before going further let me address one simple issue. Whenever we are trying to get some data; data will have different type of types of information contained inside it and we want to take out that information out, that can be done with the help of statistical tool. And in order to get the right information out of the data we have to use the right tool.

You might have heard that there are different types of stories, different types of comments, different types of jokes about statistics, but let me assure you. Those things are coming from those people who do not know the statistics well . And believe me this is my experience that in case if you are using the statistical tool in the correct way on the correct data, there is a very high probability that you will get a very good outcome, very good result which is reasonably close to the truth.

So, now onwards if you come to know about any such comment, joke possibly first you have to think that this person is making some mistake and possibly he or she is not using the right tool at the right place on the right data and this will give you a new thought process.

So, let me find out this mistake and let me try to correct it and let me try to find out the correct outcome, correct statistical inference out of the same data which is being criticized, right. So,

ok. So, ok, it is very important for us that, when we are trying to use the data, we should use the right data right tool and we should draw the right statistical inferences out of that, right.

So, it is very important that we have a proper interpretation of the inferences which are drawn and remember one thing that statistics cannot do miracles it is not like a magician, but it is based on certain scientific principles and those scientific rules, scientific principles show us the way, they tell us how should we extract the information from the data so, that it is correct.

And one of the biggest challenge before statistics is that statistics cannot change the process or the phenomena. Whatever is happening that will continue to happen if there is a prediction if that prediction will continue to be there if the wind is blowing, the wind will continue to blow.

If there is a change and the atmospheric pressure that will continue, we cannot change it, but we have to observe the phenomena, we have to collect the data based on that and then we have to do our statistical analysis. So, statistics is a scientific way of extracting and retrieving the information. Now, the next question come why should I collect the data? why? Means if by simply looking at a picture if I can draw the inference out of that then why should I collect the data on that picture.

Let me take a very simple example. Suppose in my class if I declare that all those students who will come in the class wearing a blue shirt, they will be getting 100 out of 100 marks in the examination, those people who are wearing white shirt, suppose they suppose we take a call that ok they will get 80 out of 100 marks, those who are wearing a green shirt they will get say 60 out of 100 marks and so on.

Will you agree to this decision? Will you agree that this type of marking scheme should be followed in the examination? Your first answer will be no, but now my argument is this why? What is wrong? Think about it. If you argue I will say well I like the blue color most. So, that is why anybody who is wearing the blue shirt I will give him 100 out of 100 marks, that is my argument.

But, definitely your argument will be that this argument is not acceptable, but then my question will be why? Why this is not acceptable? Now, you have to think and tell me why this cannot be done? Why this is not acceptable? Right. In order to find out the correct answer let me move in the opposite way. Now, I take another criteria, means I say ok I will take a topic say on this statistics. I will try to teach you in the next 20 hours.

And after those 20 hours and during those 20 hours, I will give you certain assignment, certain quizzes, certain exams, half an hour exam, 1 hour exam, 2 hours exam, 3 hours exam. And whatever we are studying in this course I will try to ask some question based on those topics, you will be required to answer those questions.

And, all the questions will have a specific marks depending on the difficulty level, higher the difficulty level higher the marks, lower the difficulty level lower the marks. Now, I will prepare a question paper I will give it to you. You solve it and it will be get checked and then you will get some marks based on those question papers.

Now, at the end I will try to sum up all the marks and any student who is getting say more than 80 percent marks, he will be getting a grade A means anybody who is getting a marks between say this 60 percent and say 80 percent he or she will be getting a grade B means anybody who is getting a mark between say 40 percent and 60 percent he or she will be getting a grade C and so on. Is this acceptable? I am sure that your answer will be yes, this is acceptable.

But, now what is the difference between the two? What is the difference between the two criteria where I am trying to grade the students on the basis of color of the shirt and these marks? In my view, there is only one difference.

In the second case, I have associated a number a data means I have collected the data on the performance of students in their examinations and then I am trying to use some mathematical tool say simply summation or some weighted mean where I give say 10 percent weight is to quiz, 30 percent weight is to exam, to 10 percent weight to projects and so on.

So, that can be a weighted arithmetic mean also. Is this acceptable? You will say yes and this is what we are following. So, now, you have to think once again from the beginning and try to see what is happening between the two criteria.

In the first criteria in the case of deciding the grades on the basis of colors of the shirt, we are not using any mathematics, we are not collecting any quantitative data, we are not essentially collecting the data which is going to reflect the capability of the students that is the most important question.

By collecting the color of the shirt we are collecting some data, no doubt about it, but this data is not relevant to the question which is being asked had this been a fashion show or something like this then this question makes more relevance, but in this case the color of the shirt is not

really going to going to express the capability of the students in the subject statistics or R software.

So, that is why it is very important that we have to collect the data on the relevant question in a correct way. So, why do we collect the data? Sometime we want to verify the theoretical findings and sometime we want to draw statistical inferences just on the basis of collected set of data and based on this we would like to develop a statistical model, which can further be used for policy decision, classification, forecasting etc. and different types of things.

What do you really mean by these three point which I have just said? The first point was to verify the theoretical findings. For example, in economics, people try to propose different types of models, for the country, for the world, for the city and so on. Those models are based on their theoretical findings, their observations on the process and we really want to know whether this model is correct or not.

So, one simple solution is that we try to observe the phenomena and we try to collect the observation on that phenomena and we try to fit those observation to the proposed model. And we try to see whether the model is also generating the similar type of data on similar conditions and if the answer come yes that mean the model is good. And then the model can be used for further processing.

Second was to draw the inferences on the basis of the collected data. For example, I just took an example of evaluation of the students based on their examinations. As a teacher, one of my job is to classify the students into different groups, those who are good, those who are bad, those who are excellent and so on. So, what we try to do? We try to first make a rule that, that we will classify them into suppose 5 grades. Grade A, grade B, grade C, grade D and say grade F.

Then how to do it? Then we decided ok we will try to first teach them, then we will try to conduct some examinations, we will try to obtain their scores in those examinations and then I will try to do some mathematical manipulations, statistical manipulation, statistical calculations on those data on those scores that we have collected which are essentially your data and then based on that we will try to draw inferences out of it.

The rule for inference can be any student who is getting more than 80 percent marks, he will be he or she will be classified into grade A, any student who is getting the marks more than 60 percent, but less than 80 percent he or she can be classified into grade B and so on.



So, these are the statistical inferences and by looking at your grade finally, I can decide whether the student is good or bad. For example, if a student comes to me, I will simply ask what was your grade and suppose the student says A then I would classify that student inside my mind that ok he is a good student, right ok. And then based on that whatever is this thing we try to develop different types of statistical model which are used further in various types of application.

So, now after discussing these aspect now we are at a situation, where you can understand that statistics is not doing anything like a magician. We have to think, we have to formulate certain rules; those rules are to be exposed on the data set and then based on that we have to use our judgment to draw the correct statistical inference. And believing that statistics is a science of data, statistics is a language of data and statistical tools are going to help us in drawing the correct statistical inferences, there are four conditions. I have two types of things, one is a statistical tool and another is data. There are four options, the correct statistical tool is being used on the right data or on the wrong data, or the statistical tool is wrong which is being used on the right data or on the wrong data and based on that we have four possible options, four possible decisions.

So, let us try to see that how those decision will be used and I have compiled those decision in a 2 by 2 matrix here for example, if you see here 2 by 2 table here for example.

(Refer Slide Time: 25:56)

**Data and Statistics**  
Statistics is a language of data.

	Correct data	Wrong data
Correct statistical tool	Correct decision	Incorrect decision
Wrong statistical tool	Incorrect decision	Incorrect decision

Rule: Garbage in – Garbage out

Statistics has its own derived rules.

Rules are framed such that correct decisions, as indicated by the data and based on the hidden information, are taken.

It does forecasting but not like astrologer's parrot.

So, we have here data on one side and statistical tool on this side. So, I have two categories the data is correct or the data is wrong.

Well, what I mean by saying the data is correct or data is wrong that the data has been collected according to the objectives of the study or data is collected in a in some vague haphazard way which is not fulfilling the objective of the study. For example, if I want to know the average height of the students in a class. So, so, in case if I am collecting the data on the heights of the students in the class then it is possibly correct and if I am collecting the data on the age of the students then this is possibly wrong. And similarly when I am talking of the statistical tool, we will always have a choice to employ the statistical tool.

So, now I am taking here two options- the correct statistical tool is used or the wrong statistical tool is used. So, based on these four options- two on data and two on tools- I have four possibilities of answers, for example, now if I try to take one by one. If I try to use the wrong statistical tool on the wrong data, then obviously it will give us incorrect decision.

Similarly, in case if I try to use the wrong statistical tool on the correct data then it is again going to give us a incorrect decision. Similarly if I take wrong data, but if I use the correct statistical tool then again this is going to give us a incorrect decision, but in case if I use the correct statistical tool and if I use the correct data then the decision is going to be correct.

So, now you can see here out of four options there are there is only one option, which is useful that you have to use the correct statistical tool on the correct data. The rule is extremely simple garbage in garbage out, means if you put a wrong data in a wrong statistical tool or vice versa then it is going to give us a wrong outcome, but if you are using the correct data on the correct statistical tool possibly they would this will give us the right outcome.

And then the question comes here that how are we really going to draw the statistical inferences for that we have statistics . And statistics is based on certain rules, those rules are not coming from the sky, the rules are, are found using some mathematical derivations, using mathematics, using the rules of science and so on.

They are not something like happens that you can just make any rule, right. They are based on certain theory, certain concept, using the mathematical tool, right and when we are trying to make those rules, we always keep in mind that the outcome of the rule should be the same which is the correct outcome of the experiment.

For example, if I say if we want to know whether studying at home improves the grades in the examination or not; well, from our experience we know, answer is yes if you study more, there is a very high chance that your grades will get better.

But, now if you try to create a tool, statistical tool and if you claim that ok this is based on certain mathematical tool, mathematical theory what is this telling you that if you study at home possibly your grades will become bad then you have to think, they are not getting say convinced by the facts of the data which data is trying to tell you, right.

So, this is what I mean. So, I can see here that in statistics the rules are framed such that the correct decisions, as indicated by the data are taken. And those information which are hidden inside the data they are extracted they are taken out. So, remember one thing the statistics also does forecasting, but it does not do the forecasting like a astrologers parrot.

What is an astrologer parrot? You have seen that many times there is some person who is sitting with some chit and the person has got a parrot, you go him you give him some money and the parrot will take out one chit and the person will read out your chit and that is taken as this is your fate, this is your luck and so well so, that is a very random phenomena, right.

So, the statistic does not do like this statistics is based on certain rules . So, now, having covered these small topics and after interconnecting them, then the next question come how suddenly statistics became data science. If you ask me as a student, I always studied the statistics as a subject.

And yeah, data science was understood, but possibly it was never mentioned, it was never used, but nowadays, this data sciences or the data science has become a very popular word. So, how this statistics got transformed into data sciences and what is really expected from data sciences, which was not expected from statistics.

(Refer Slide Time: 32:37)

**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 is the topic which now I am going to address here in this slide. So, the question here is how this statistics got transformed to data sciences and what is really expected from data science which is or which was not expected from statistics.

What is statistics did? That people who are not happy or people who are not satisfied, that they started calling the statistics as a data science. One thing you have to you, I am sure that you have observed that this.

The role of computers, the advent of computers has impacted the human lives enormously. And believe me I am the one who has who is possibly growing up with the students, sorry who is growing up with the computers when I did my masters possibly at that time that was the time when these computers were coming to India.

And when I was doing my, this graduation and master possibly I used to see from away that what is the computer, from my eyes and after; that means, I have seen that growth with my own eyes in the last nearly 25 years. So, now what happened that these computers impacted the different parts, different phases, different aspects of the life. So, the same thing happened in the statistics also.

When I was a student, I was usually taught that ok we have to collect the data, but the data collection was a very big thing and in most of the cases we used to get answer from the from our teacher that the data is usually not available, but theoretically you can assume that data is available and then you have to develop the statistical tool.

But, now nowadays, now the time has come where data is not an issue. The data is everywhere, the availability of data is enormous; think about yourself I am sure that you must be having a Smartphone; that Smartphone is connected to internet, there are Google maps.

So, if you if you are moving you are you are being tracked means after a month you will see sometime you will get an email, ok, this is the summary of your last month where you have travelled. So, every moment you are generating a data which is getting stored in some server. For example in Google server or somewhere and that data is being used in different ways.

So, now, a time has come where there is data, data and data. There is so much of data which your hands cannot handle it, your pen and paper cannot work on that data. Size of the paper, the size of the notebook they have become smaller to enter those data. And now once you have so much of data how to handle it?

Definitely statistics cannot be avoided, because statistics is the one who is providing us the rule that what we have to do on the data. So, that we get a correct inference; statistics is the language of the data. So, there is no doubt, there is no question, there is nothing like that the statistics is getting lower priority or statistics is getting under evaluated no statistics is still remain as a statistics, right.

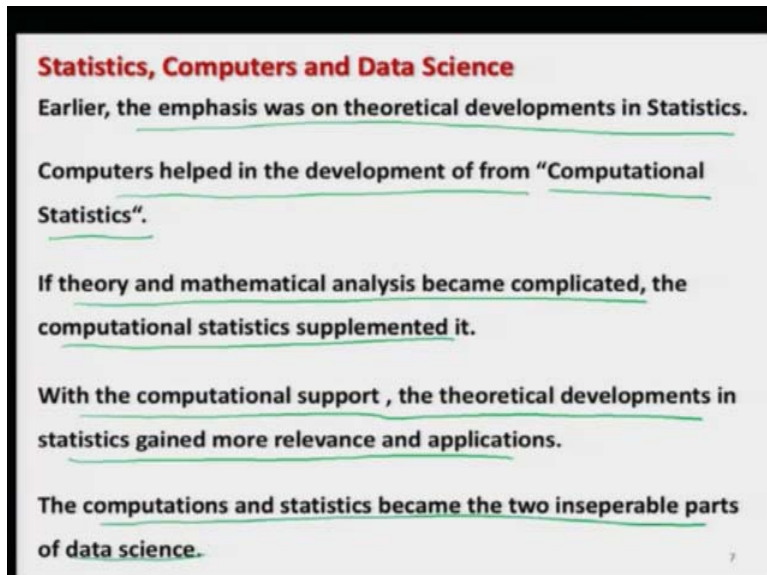
So, now if you try to see how this data science came into picture now, we have computers who can do our job easily which we were not able to do earlier, but now these computers can handle it and when we try to combine the statistics and computers, possibly this gives you the first picture what can be a data science. This is statistics, this is computer, I join it together they become one.

So, they are the two inseparable sides of the data science, this is my perspective. I mean different people have may have different perspectives, I am not contradicting them, but this is what I personally think. So, this the advent and rapid development in computers have impacted statistics. Earlier it was very difficult to collect the data and even and most of the times we always say that data is not available, data was not available, but now, but now the data is available.

And in fact, too much data is available and the data is so much that there is a new area, which is called as big data analysis even that has come into picture. And what is big data? Big data is not byte, not kilobyte, not megabyte, not gigabyte, but possibly the starting unit in the case of big data byte is petabyte.

You can imagine that how this, how the data collection has become so easy and how easily the data is available and what is the possible size of the data, right.

(Refer Slide Time: 38:35)



So, now that the next question comes that what is the role of statistics, computers and data science? How they are interconnected? So, earlier the emphasis was basically on the theoretical developments in statistics then came computer and computer said ok I can help you.

For example, if you are thinking of inverting a 100 by 100 matrix possibly, manually it will take a very long time the computer said, ok do not worry I will help you. And that the role of computer started coming into statistics and a new area what is called as computational statistics, a new aspect of statistics was developed that is called as computational statistics .

Now, when we are trying to do the mathematical analysis, sometime the theory and the mathematical analysis becomes too complicated and we are unable to go further. Under those situation, the computational statistics supplements it and it came as a saviour and it helped us.

And with the help of these type of computational support the theoretical developments in the statistics gained more relevance and applications, right. And this and the computation and statistics became the two inseparable part of the data sciences.

(Refer Slide Time: 40:03)

**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.

If you ask me many times, people do ask me, that what is more important theoretical developments or the applied statistics. I say both are the two sides of a coin, unless and until there is a theoretical development there cannot be an application. And if there is no demand of an application why should there be any theoretical development? What will we do with those theoretical developments that may, that are not usable?

So, this computational power of computers help statistics a lot. Whatever we were thinking and in those cases we are unable to handle them we are unable to solve them. And in those cases where the data size was becoming too large and it was not possible for us even to compute for a smaller or a simple tools, tool then computer came into picture.

For example, if I have to simply find out the arithmetic mean finding out the arithmetic means of millions and billions of data by hand manually it is very difficult, but computers helped us, right.

So, these theoretical developments and applied say statistics developments and these computers they all come together, they all come together and once you have a more, once you have more options in your hand, that you can experiment with new things using the software, using the computer, your thought process also changed.

The way we used to think in statistics from the theory point of view that also got changed and we became more courageous. We will try to think in a very different way which possibly

people might not have think long time back. So, now once we adventure into computational statistics, the role and use of computer became very important.

Now, once it comes to computers; the computers are the slaves of human being and they require proper programming language, they require a software, they need data management and several other aspects. Because the data is becoming too high and every software has a prerequisite that the data has to be inserted in a particular way and using the some appropriate programming language.

So, all these things are needed if you want to use the computers in statistics and because of those these advantages, the areas of application of statistics also increase. And nowadays the topics like artificial intelligence, machine learning, supervised learning, unsupervised learning, reinforcement learning, these all came into picture they have become very popular and they have become an integral part of the statistics.

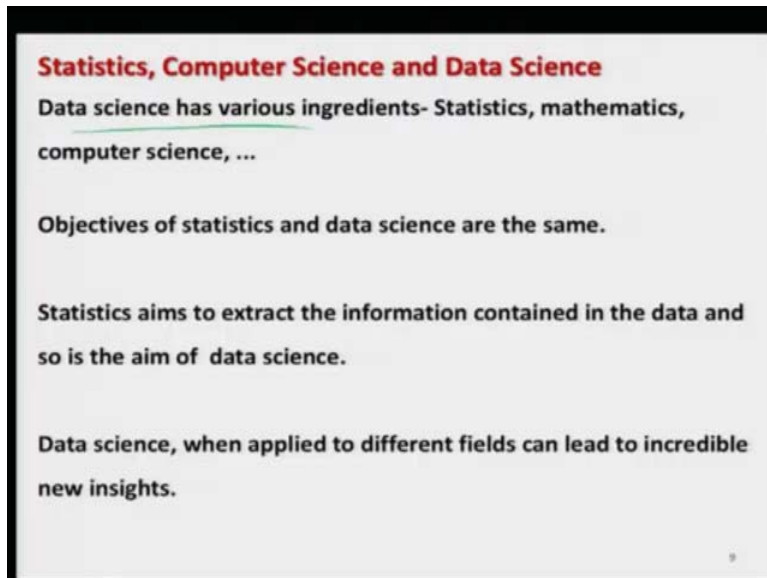
And one important point which you have to always keep in mind that these techniques which are very popular they are based on statistics their concept is coming from statistics, but they are heavily dependent on computers they are essentially the computer intensive techniques where the data will come, the statistics will guide them how to use the tool, which tool have to be used and then the atmosphere of the experiment the and the information coming from the experiment that will help us in getting a, in getting and applying the proper rule to get a proper outcome. So, that is what started happening.

So, now, after understanding this aspect you now you have understand, now you can understand very easily that data sciences is not one thing, but data science is a combination of different aspects this can be mathematics, this can be statistics, this can be computer science and this may also involve the computer architecture.

The hardware, software part of the computer, the database management system means everything together, right, but whatever it is, the objectives of statistics and data sciences are the same. We have to collect the data we will have some data, from this from those data we have to get the right inference, right decision, right picture, right information, whatever is hidden inside the data that we have to take out, right.



(Refer Slide Time: 45:10)



So, this data science has now various ingredients in my opinion that can be statistics, mathematics, computer science etc. And the objectives of statistics and data sciences are the same, right and both of them statistics as well as data science, they aim to extract the information which is contained inside the data and that is the same aim as of the data sciences or data science.

So, and the biggest advantage of data science is that when data sciences is applied to different fields this can lead to incredible new insights, means if you ask me I will be very honest in accepting that using the software, using the computers, means I was able to understand the statistics in a much better way what I had understood as a student of a statistic which was trained only in theory.

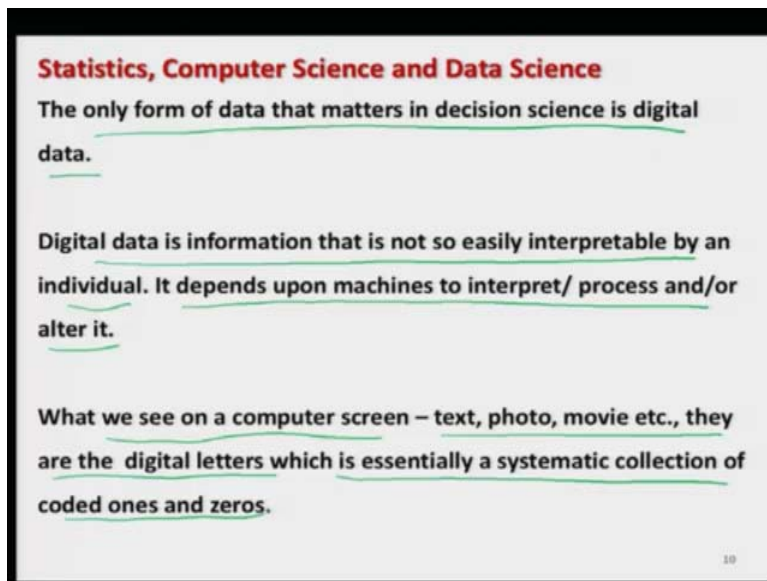
And there is one more change. I remember as a student I used to write my data with my hand, once I write the data by hand then it is transformed it into a or it is transmitted into a calculator or into a computer or something like this. But nowadays, the data, the form of the data is also changing nobody is recording the data manually by writing with the help of a pen, but data is getting transmitted directly and it is in the form of digitals, the data is digital.

In fact, nobody thinks even writing the data, whenever the data we means our brain has been trained to think that data has been collected automatically at and the data is in the form of a digital data, it is not a manual data . Now, once the data is in the form of digits means that is a,

that is a sort of electronic data digital data you have to take the help of computers to read them that data is many times encrypted and we cannot read it.

For example, if you see a photograph, a movie or a picture on the computer screen, this is not really a screen or a picture or a photo, but behind the screen, there is some data which is encrypted in the form of zeros and ones and when those zeros and ones are combined in a particular way they will give you different types of effects, right.

(Refer Slide Time: 48:00)



So, nowadays the only form of data that matters in the decision sciences is the digital data and digital data is information that is not so easily interpretable by an individual it depends on machines to interpret for processing and or changing it altering it. Whatever we see on a computer screen like as text, photo, movie etc., they are also the digital letters which are essentially collected in a systematic way and they are coded as ones and zeros, right.

(Refer Slide Time: 48:33)

**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.

So, now the, now means I have explained you all the smaller parts of the data sciences. Now, if you ask me, what do you need to become the data scientist? So, my question is first you have to decide what you want to be- doctor or a compounder. You better or you very well understand the difference between the two. The doctor decides the medicine and compounder only gives the medicine whatever the doctor has said.

The compounder is usually not allowed to decide the medicine because, he has because he or she has not been trained in the science of medicine. He has been trained only to follow the instruction how the medicine has to be prepared and how it has to be given, the this your lies only with the doctor.

So, that is the same question to you also from my side. First you have to decide what you want to be a doctor or a compounder? Why I am asking this question because now I see on internet on various website they want to make you data scientists in 1 week, 2 week, 1 year, 2 years and so on right.

So, you have to decide what you want to be and how you want to understand this data sciences, what you want to become in your life. So, now I am trying to address the question what is needed to become the data scientist? So, as I said first you decide what you want to be- doctor or the compounder.

And now I can say that you have to decide whether you want only to use the tool, whether you only want to understand the utility of the tools or you want to develop the tools. In my opinion

you need actually all to become a good data scientist. You should know what is the tool, what is the utility of the tool and in and if needed, you can change the rule change modify the rule.

(Refer Slide Time: 50:40)

**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 ✓	• ... .. ✓

12

Now, I try to connect the statistics with your data sciences. When I come to statistics; statistics is the soul of data sciences.

There are many topics in the statistics which are very important, which are taught in the undergraduate, masters program in statistics, we start with the descriptive statistic, probability theory, statistical inference, decision theory, Bayesian inference, frequentist inference, parametric inference, nonparametric inference, multivariate analysis, linear regression analysis, non-linear regression analysis, simulation technique, Monte Carlo method and there is a very long list, means I do not have a space even in the slides to write all those things.

(Refer Slide Time: 51:31)

### 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.

13

And well once you come to the data sciences you need all these things, right. And once you are trying to learn these topics, there are two aspects that you can just quickly learn those topics or you want to learn them in depth.

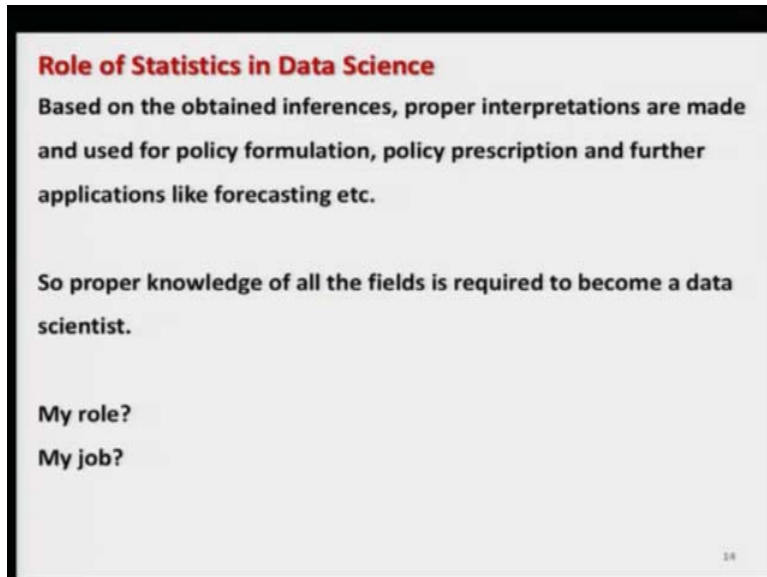
So, addressing that issue I would say that the theoretical developments are essential and why they are essential because the theoretical developments will tell you what is the rule, what is the formula, what is the mathematical form which is to be exposed to the computers for the computation or which is to be implemented on the computational procedures.

Now, once I come to the computational procedures every computational procedure will have some limitation, for example, if you want to invert a matrix every software, every programming language will have a limitation means any software or any matrix cannot for example, invert a matrix of infinite by infinite order, right. So, for that some optimization methods are required.

Now, once you try to develop some statistical tool from theory, the implementation of that theoretical rule will require the knowledge of statistics, mathematics, optimization methods computers etc.

And more importantly all of them have to be implemented simultaneously over a data set and once you said on the data set, then obviously the data management is also required once you are getting the data in of the sizes gigabyte, terabytes, petabytes, you also need to know, should have the knowledge how to store it, how to handle it, how to prepare it for the further means application.

And all these aspects which are logically implemented in a systematic way and then only they will give the correct statistical inferences out of that and you will be able to draw the correct statistical inferences after doing a proper statistical analysis, right. (Refer Slide Time: 53:38)



**Role of Statistics in Data Science**

Based on the obtained inferences, proper interpretations are made and used for policy formulation, policy prescription and further applications like forecasting etc.

So proper knowledge of all the fields is required to become a data scientist.

My role?  
My job?

14

And once you obtain the correct statistical inferences then you have to take the proper interpretations and those proper interpretations are further used for policy formulation, policy prescriptions and further application like as forecasting etc.

So, in my opinion the proper knowledge of all such fields is required to become a successful data scientist, right. So, now, the next question comes what I am going to do here. My objective is very simple, my role is very simple, my job in this course is very simple, up to now there have been development of statistics, there is a new concept of data sciences.

So, unless and until you know the statistics, you cannot extend it to the data science. So, what I will try to do? That I will try to take some important topics of statistics. I will try to show you or explain you the basic concepts, basic fundamental. I will try to show you what is the mathematical background behind those rules, sometime I will be quick, sometime I will try to give you the proof also, sometime I will skip the proof also and I will refer you to some books, some notes.

And I will try to extend it to show you that how the classical concept of statistics has been extended to data science or computers or computational part. So, I would try to or I will try my best to give you a feeling. So, that you get connected from statistics to data science and you

will have a fair idea that, that what type of training do you really require to become a data sciences, data scientist.

So, that is my objective and now I will stop here. I tried my best to use my concept to give you a brief idea that how you can change yourself or how we can transform ourselves from statistics to data sciences. Well, I will say once again, these are my personal views I am not contradicting any other views people have different things, but these are my views on which I am going to run this course.

So next time I will try to see you with some relevant topic of statistics and then we will move further. Till then you also try to think, you also try to look into other type of literature, books, etc. and try to create your own concepts. So, so you stay safe enjoy it and I will see you in the next lecture.

Thank you, goodbye.