Essential of Data Science with R Software-1
Probability and Statistical Inference
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Lecture No. 08
Introduction to Probability

Hello friends welcome to the course essential of data science with R software 1 where we are going to deal with the topics of probability theory and statistical inference. So, you know up to now we have tried to have a quick revision of the topics of the R software that possibility we will be needing, whenever we are trying to do calculations and computations. Now from this lecture we are going to start with the topics in statistics. And the first part which I am going to handle is related to the probability theory.

Now you know what happen. As soon as I say the topic of probability theory, people starts thinking of computing the probability, what they have learnt in their elementary classes. Definition like m upon n or total number of favorable cases divided by total number of cases. And they finally end up thinking of some complicated problems which are very difficult to solve numerically or say analytically. Well, I am not going to do this thing, that is my first assurance to you.

My objective here in this topic is very different. I want to create a foundation. I want to create a statistical thinking which will lead to the creation of data science. Now if you try to see, what are you really trying to do in data science? You have got a data and you are trying to built up different types of models. That is one of the most fancy words that we are trying to create a model for this process. Now if you try to think what do you really mean by a model and how do you characterize it.

Suppose if I ask you very simple question. Do you know what is the model for the weather conditions? Or in case if I ask you the same question in much simpler way. Whenever you leave your home, how do you decide that whether you have to carry an umbrella or not? Now you will think about it. What you really do? You try to look outside and then you try to think if the color of the cloud is black or dark grey. That would possibly indicate yes, it may rain. And then you try to take your umbrella, is not it? Or sometime even our parents our mom always say, okay, even the clouds are dark grey or black please try to carry umbrella it may rain. And always followed this advice. Now please try to think, what are you trying to do.

First word about you use had may, what is this “may”? It may happen or it may not happen. Suppose if the colors of the cloud is just pure white means sky blue, sky. Will you say that there is going to be rain? Now can you really say, no, you are 100 percent confident that it will not rain during the entire day? Very difficult. And even some time we have seen that the clouds are there, sun is there still it rains.

So, that means what you will inform me. You will simple say there are very small chances or less chances that it will rain. So, we should not carry the umbrella. But if the clouds are grey, dark grey, you will say oh the chances are very high that it may rain. So, please try to carry an umbrella. Now what are you trying to do? Do not you think that you are simply trying to create a probability model and based on that you are trying to compute the probability?

You will not believe on me means I know. Well, let try to argue. When you try to see the color of the clouds, depending on whether it is white, blue, grey, dark grey, you try to guess whether there will be rain or not. And you try to quantify it. In case if the clouds are dark grey or towards black, you say the probability of raining today is very high. But in case if the clouds are clear, the sky is blue, then you try to say the probability of rain is very, very low.

And based on this probabilistic statement you try to take an, take a decision whether to carry the umbrella or not. So, now I have given you an argument where your decision are going to be based on the probability. And it is not only the cloud this I can promise you I can assure you if you try to analyze your day to day activities, at every moment you are trying to create up or compute a probability, you are trying to create a probability model and you are trying to compute probability.

Why are you trying to create a probability model? Now you tell me, who told you that when the clouds are dark grey, it may rain. And who told you that if the clouds are white or sky blue, then it may not rain? You observed the phenomena. You observed in your life many many times that whenever the clouds are dark grey then it rains. And whenever the clouds are say clear sky blue, then usually does not rain. So, what you have done.

You have observed this phenomenon and you have collect this some data. But you are so intelligent that without any proper training you statistics you created a probability model inside your mind. And you yourself do not know how you have created it. And as soon as you get the information on the input variable that the color of the sky is or the color of the cloud is white blue or grey, you immediately compute the probability of happening of the event. And if this probability comes out to be pretty high, say more than 80 percent 90 percent, then you say it may rain. But in case if you see the color of the clouds is white or blue, you try to feed this information inside your mind where you already have stored a probability model and mind tells you that the probability of rain is very very low.

So, now if you try to see, who says that you do not know statistics or you do not know data science or you do not know probability. You know better than anybody else. The only thing is this, you do not know how you have created that model. Means if I give you some data, possibly you can create the model inside your brain but you will not be explain me that what is the mathematics behind it or what is the statistics behind it.

And we had discussed in the beginning that unless and until our observations and they were decision are based on scientific approach, people will not believe on us. Means if you tell somebody that the probability of rain today is just 5 percent and one should carry an umbrella. People will not follow your advice. Because they will say that okay means say the frame work by which the person is trying to take a decision is itself wrong.

So, now if you try to see whether you know the probability theory or not, you have to decide. Means I know that you know it. Means I am confident. The only thing which you do not know that what are the ingredients of this model which you have created yourself. And there is another aspect. Well, if you try to see in your day to day life, you are everyday creating such models very quickly and you are implementing it, means if you have say couple of friends and if you decide that we are going to meet at 5 pm at this place, you know that there will be some friends who will always come late. And you will say he will come 30 minutes late.

But there will always be a friend who will always be coming at 5 o’clock or somebody may will come 5 minutes before. So, now depending on which of the friend is coming, you will also try to reach at the that place according to that data. Means if the friend come always late, you will also try to reach your there all little late. But if the friend is very punctual, you will also try to reach there punctual. What is this? You have created the probability model. How? First you collected the data on all your friends and you have created not one not two but one probability model for each of the friend. And as soon as I tell you the name of the friend you immediately compute it inside your mind and take a call yourself that whether you have to reach at 5 o'clock or 5 minutes earlier or 5 minutes later.

So, now can you still argue that you do not know probability theory? When you can create the probability models, how will you convince me that you do not know the probability theory? So that question is out. I hope you will agree with me. Now the question is, here is very simple. Whatever we know god has gifted us brain, intelligence and a built in package of statistics. But here we want to do it on a mathematical way statistical way. So, what are the different components? What are the different ingredients which will help us in constructing this type of probability model?

And when I am talking of here probability theory, it is not only that the probability of an event which you cannot solve. You already are solving so many probabilities every day. So, the only thing is this you have to now understand that how you can convince others that this the probability that you have computed using this rule, this concept, this philology and that is why one has to believe on your decision, number 1.

Number 2; that whenever we are trying to develop any tool there is a thought process. First we have to convince ourselves that whatever we are thinking that is correct or wrong and then that thought process has to be translated into a statistical model or mathematical model. So, how to get it done? And after that once you are trying to use the tools mathematics and the statistics, there are some theoretical tools. How you can justify them on the basis of a real set of data. That is another issue. So, now how to interpret it?

Thirdly, many times we say and we feel that the, that the computation of probability in this event is very difficult. That we are talking actually analytically. That if you ask me to find out the probability that may involve some complicated summation, some complicated integration and I may not be able to solve it explicitly. But with the advent of the software's we can at least approximate that probability by computation very well. So, how to get it done? And whenever you are trying to think about the probability theory from the decision sciences or from the data science particularly, you cannot move even a single step without understanding the probability theory.

For example, in case if you take an example of a shopping website where are millions of customers which are excessing the website every day or possibly every hour, they are trying to go through with different types of things available on the website. And you want to get us to each and every customer very carefully and in the correct way.

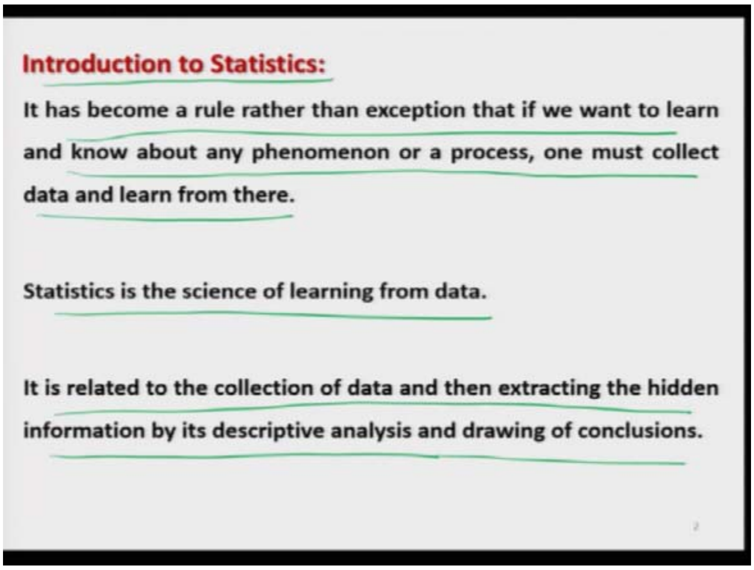
In case if some person is browsing for the shorts and trousers, you would like to send an email or an offer to that customer only for short and trouser. You will not send the offer or a coupon code that is valid for some electronic item. So, how do you do it? You simple try to compute the probability. What is the probability that this customer seems to be interested in clothing's and particularly in shorts and trousers?

And if this probability comes out to be more than a threshold value that you have to decide. Then you will try to send an email or an offer to that customer. Now do you think that anybody will be doing it manually? Or is it like something when you enter into a physical shop a customer care representative or somebody from the shop comes to you to show you the things inside the shop. No, all these things have to automated, these automated things can be done only with the probability theory.

So, this lecture I will simply try to give you the basic background I will try to connect several things together, so that you have a fair idea. Once you are convinced that there is a need to learn the probability theory. I promise you nobody can stop you in learning the probability theory much beyond what I am doing in this course that is my promise to you. So, with this hope let us begin our lecture.

So, this lecture we are going to talk about different small topics and I will try to interconnect them that will only be a theory, theory and theory, no mathematics in this lecture. And from the next lecture we will talk about this mathematics and I will try to connect this mathematics with the real application and with computation. So, let us begin our lecture.

(Refer Slide Time: 17:13)



Introduction to Statistics:

It has become a rule rather than exception that if we want to learn and know about any phenomenon or a process, one must collect data and learn from there.

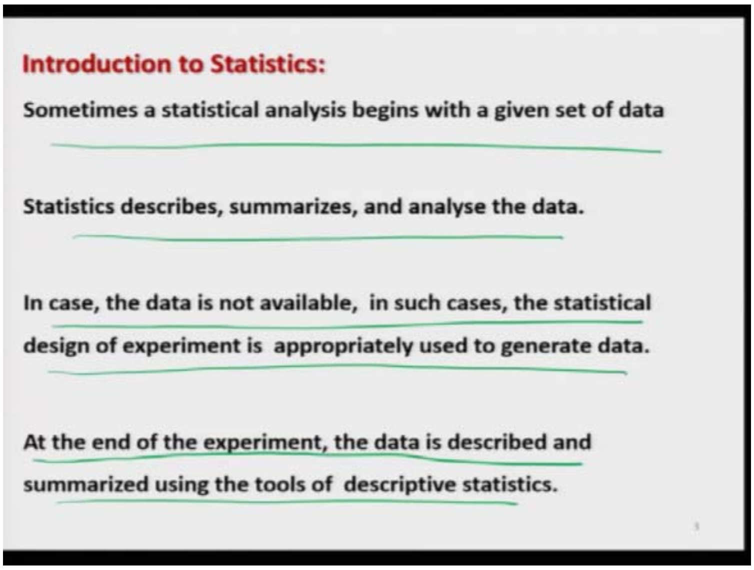
Statistics is the science of learning from data.

It is related to the collection of data and then extracting the hidden information by its descriptive analysis and drawing of conclusions.

So, now the first question comes here what is the statistics and why do we need the statistics. Actually, now if you try to see this question becomes say completely relevant. Because it has become a rule rather than an exception that if we want to learn and know about any phenomenon or a process, we always try to learn it by collecting the data. We collect the data and we try to learn what data is trying to inform us.

And statistics is the science of learning from data. It is related to the collection of data and then extracting the hidden information by its descriptive analysis or different types of several other types of analysis and finally drawing the conclusions. The conclusions are called as statistical inference.

(Refer Slide Time: 18:04)



Introduction to Statistics:

Sometimes a statistical analysis begins with a given set of data

Statistics describes, summarizes, and analyse the data.

In case, the data is not available, in such cases, the statistical design of experiment is appropriately used to generate data.

At the end of the experiment, the data is described and summarized using the tools of descriptive statistics.

Whenever you are trying to use the statistics, the basic food of statistics is the data. So, some times what happen that if somebody has an objective which the person wants to investigate from the statistical point of view, then there are two options. Whether some data related to that objective has already been collected. And we try to begin the statistical, statistical analysis with the given set of data. And then we try to apply different types of statistical tool which try to describes, summarizes, and analyze the data.

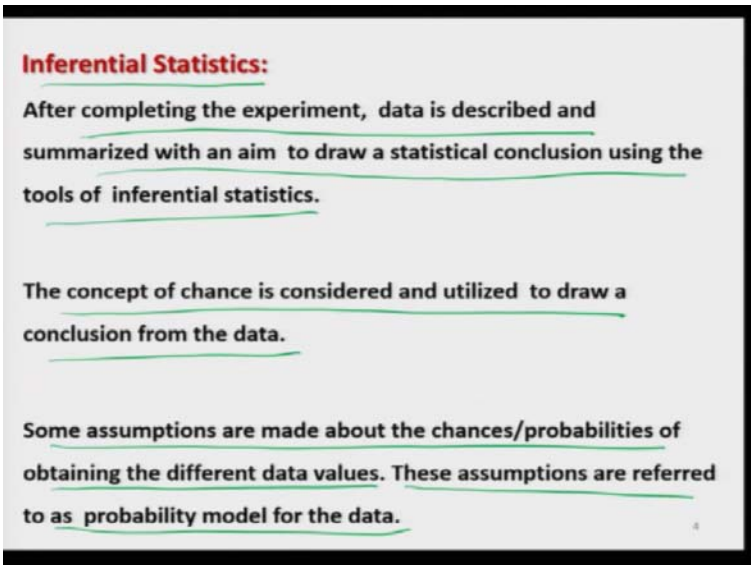
But other possibilities this there is no data available. But there is some statement, some hypothesis, some assumption something what we expect based on that we want to verify it for that the data is not available. In such a case, we try to collect the data by using the statistical design of experiment and we use it appropriately to generate the data.

Now let me try to give you a very simple example of now a days. People are looking forward to developed a vaccine for Covid. Now how to get it done? They will try to give the vaccine which is being developed to certain number of people and then based on their responses they will try to judge whether the vaccine is effective or not. Think about it what people are trying to do, there is no data, but they will try to design it. They will try to give the vaccine to some people who are expected to get Covid. And then they will try to see whether the person really catches the Covid, virus or not.

Do you think that they should give that vaccine to somebody who is always staying at a particular place behind the close room? That may not be able to help them in getting the correct data. So, that is correct data is very important and actually correct means appropriate. So, in statistics, we have a branch what we call as statistical design of experiment. So, in the Covid case also they are trying to conduct the designs of experiment and they try to collect the data. One they have collected the data then they try to analyze it.

So, this is what I meant that when we have the data that is already collect or when we try to collect the data. According our need but statistics is based on both the assumptions and it get us to both the conditions. So, now whenever you either you have the data or you are conducting the experiment to collect the data, at the end of the experiment, you will always have the data. So, this data is described and summarized using the tools of descriptive statistics. This is the starting point.

(Refer Slide Time: 21:27)



Inferential Statistics:

After completing the experiment, data is described and summarized with an aim to draw a statistical conclusion using the tools of inferential statistics.

The concept of chance is considered and utilized to draw a conclusion from the data.

Some assumptions are made about the chances/probabilities of obtaining the different data values. These assumptions are referred to as probability model for the data.

And after that you move forward and there is no end the end comes only when you are satisfied that beyond this you cannot improve it. Now once you have the data you try to use here different types of tools and you try to draw various type of conclusions. These conclusions from the statistical point of view they are called as statistical inference. So, as soon as we start applying different types of tools on the given set of data, we start venturing into the inferential statistics.

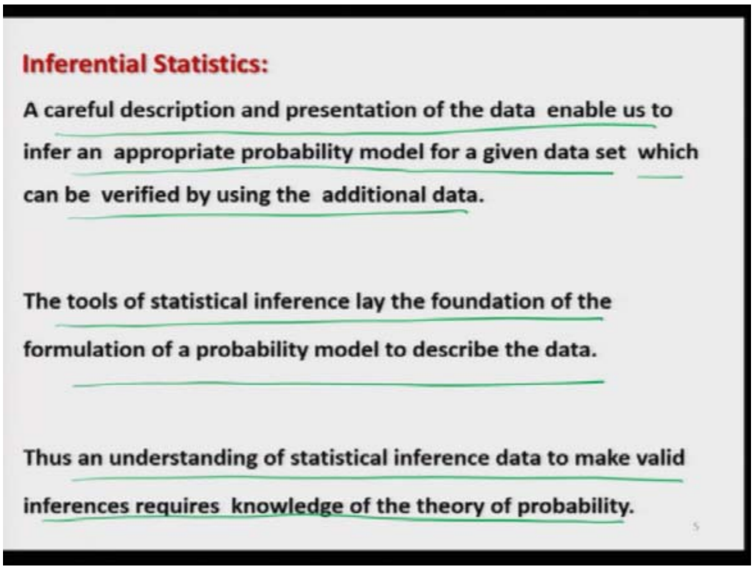
You see when the data comes, the data do not tells us that whether we have to apply the automatic mean or the median or the geometric mean or the standard deviation for example. But this is only you who has to decide that which of the tool when applied on the given set of data is going to give the correct conclusion. So, for that these inferential statistics comes into picture. Where we have to decide that which tool is going to give us the correct statistical inference.

So, after completing the experiment, data is described and summarized with an aim to draw a statistical conclusion using the tools of inferential statistics. Now when you trying to think about the inferential statistics means definitely what data is trying to say that we cannot understand. Because data is simply deaf and dumb in term that cannot speak and we cannot listen to its voice or we do not know its language.

So, whatever we are going to understand, there will always be chance that we may be wrong. So, whenever we are trying to use the tools of inferential statistics, the concept of chance is considered and utilized to draw a conclusion from the data. We always say that well this are the chances or this is the probability that our conclusions are going to be correct. For example, if the clouds are grey, then you say that you are inferring that there can be rain and the probability or the chances of rain are very very high.

Now whenever you are trying to compute such chances or the probabilities some assumptions are made for obtaining the chances and probabilities for different data values. And these assumptions are generally referred to as probability model for the data. And this is what you actually try to construct inside your mind very quickly for which god has possibly trained you.

(Refer Slide Time: 24:02)



Inferential Statistics:

A careful description and presentation of the data enable us to infer an appropriate probability model for a given data set which can be verified by using the additional data.

The tools of statistical inference lay the foundation of the formulation of a probability model to describe the data.

Thus an understanding of statistical inference data to make valid inferences requires knowledge of the theory of probability.

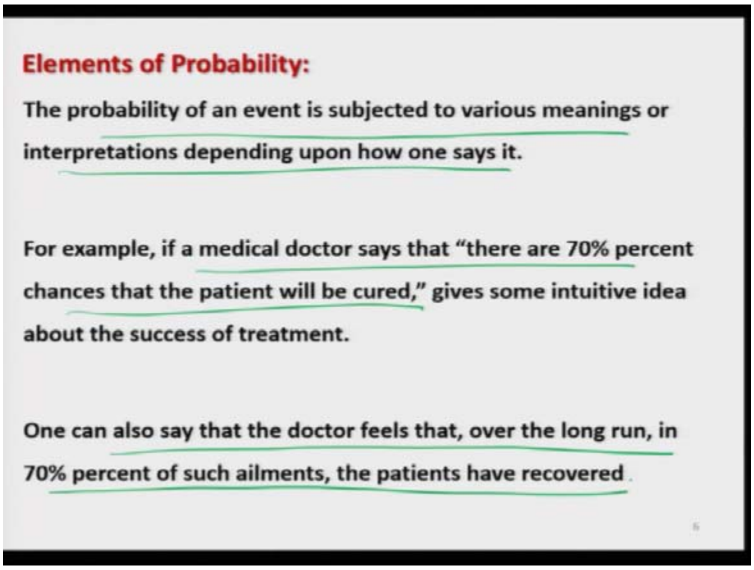
So, a careful description and presentation of the data helps us in inferring an appropriate probability model for a given set of data. And once we have created the data we have to be sure enough that this model is correct. So, what we try to do? First, we try to collect some data then would or we simply try to verify the model on the basis of some additional data so that we are confident enough that our model is working well.

So, the tools of statistical inference lead the foundation of the formulation of a probability model to describe the data. And to extract the hidden information which is hidden inside the data. So, if you want to do it, then obviously and understanding of statistical inference is very important so that you can make valid statistical inference from the given set of data. And that is based on the knowledge of the theory of probability. Because at every step you are trying to judge whether your model is correct or not on the basis of probability. That whether it is going to give you the correct statistical inference or not.

For example, if I ask you that inside your own class what is the average age of your classmates? Means you can say from values say 22 years. Do you think that the age of all the students in your class is exactly 22 year? No. Actually as soon as I ask you this question what do you do. You simply try to think and you simply try to take a sample from that population inside your mind and possibly you try to compute the automatic mean. And then you tell me that this value is coming out to be 22 and you try to see that most of the student, their age are close to 22 years that may be 21 years 7 months, 8 months or 22 years, 2 months, 3 months and so on.

So, now you say that the probability that the average age of the class is 22 is very high. But on the other hand, if I say the average age of the class student is 70 years, you will say the probability is very low. So, you can see here all your conclusion they are based on the probability models and in every case, you are trying to compute the different types of probabilities. And definitely when you want to do it, you need some foundation, you need some theory, you need some concept.

(Refer Slide Time: 26:37)



Elements of Probability:

The probability of an event is subjected to various meanings or interpretations depending upon how one says it.

For example, if a medical doctor says that "there are 70% percent chances that the patient will be cured," gives some intuitive idea about the success of treatment.

One can also say that the doctor feels that, over the long run, in 70% percent of such ailments, the patients have recovered

6

So, now what are the different elements of probability? The probability of an event is subjected to various meanings and interpretations depending on how one says it or how one wants to use it. For example, if the medical doctor says that there are 70 percent chances that the patience will be cured. What does explains you, the patient will be cure. What explains you? This gives the patient and intuitive idea about the success of the treatment. Means if the doctor say, do not worry there are 99 percent chances that the person will cure. The person will go very happily to his home or her home and will sleep peacefully.

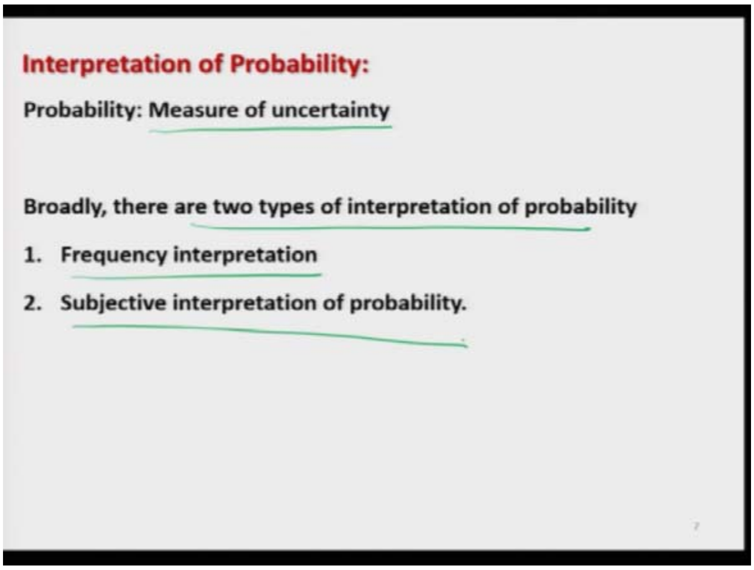
But in case if the doctor says the chances of cure are just 5 percent or 10 percent, possibly the person cannot sleep in the night. Why? Because the person will be thinking that possibly he may not, he or she may not survive after some time. So, you can see that this probability is statement how it affects us. And how we try to judge about the treatment means, if our doctor gives us a patient two option that the person has can cured with treatment number 1 and treatment number 2 and the probability of getting cured with the treatment number 1 is 90 percent. And the probability of getting cured with the treatment number 2s just 30 percent. What do you think? What decision the patient is going to make?

The patient will simple try to choose the treatment number 1. Why? Because the probability of success or the probability of being cured is much higher. Well remember one thing being cured is much much higher. Well, remember one thing 30 percent is still there. Which is a risk. But people are optimistic in their life. And they always try to see that the glass is half full, not half empty. So, now you see here with the probability statement how people try to make different types of conclusion.

Now there is one thing more when the doctor is saying that your 70 percent chances that the patient will be cured. What does this mean? The patient is only one and that patient is going to the doctor. And patient thinks that there are 70 percent chances that the person will be cured. But what doctor is trying to say by the 70 percent is that if there are large number of patients which the doctor has cured, then the total number of patient which got cured is simply close to 70 percent.

I am not saying that or I am not asking what happen to remaining 30 percent. So, from this point of view one can also say that the doctor feels that over the long run in 70 percent of such ailments the patients have recovered.

(Refer Slide Time: 29:55)



Interpretation of Probability:

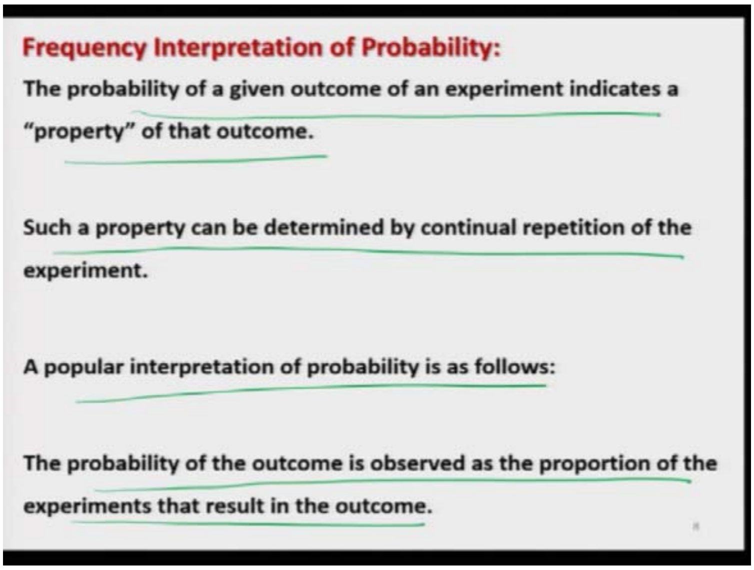
Probability: Measure of uncertainty

Broadly, there are two types of interpretation of probability

1. Frequency interpretation
2. Subjective interpretation of probability.

So, now you can see here that whenever we are trying to measure the probability the probability is somehow trying to measure the uncertainty. And based on that this aspect we have to decide how we are going to interpret the result. Basically, when we talk of the interpretations of the probability, there are two types of interpretations. One is the frequency interpretations and second is the subjective interpretations.

(Refer Slide Time: 30:27)



Frequency Interpretation of Probability:

The probability of a given outcome of an experiment indicates a "property" of that outcome.

Such a property can be determined by continual repetition of the experiment.

A popular interpretation of probability is as follows:

The probability of the outcome is observed as the proportion of the experiments that result in the outcome.

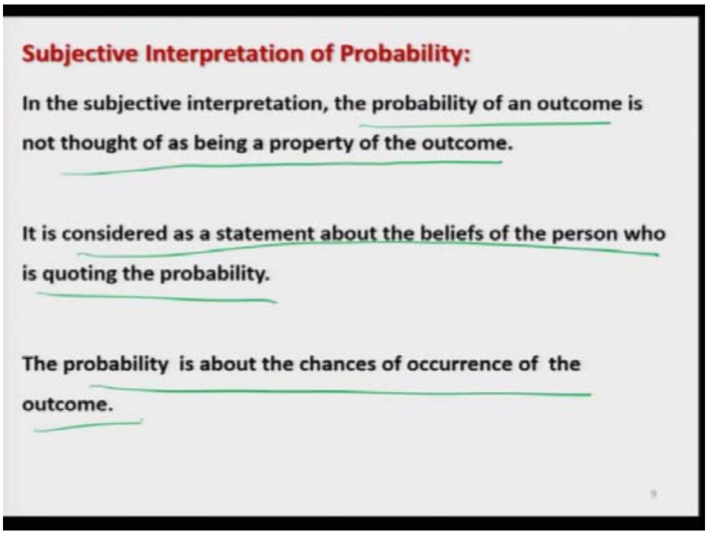
So, now let us try to understand what are these things. So, first we try to take up the topic of frequency interpretations of probability. The probability of a given outcome of an experiment indicates the property of that outcome. Do you agree with me? With this statement. That in case if the patient is going to the doctor this is the property of the doctor that how successful can it be in treating the patient successfully. So, that is the property. And such a property can be determined by the continual repetition of the experiment.

The probability is will not make much sense if that is based on single observation. Means if there is a doctor who has treated two patients, one survived and one passed away, what you will see whether the doctor is good or bad. But in case if the patients are going to the doctor continuously and suppose sufficiently large number of patients have gone to the doctor for the treatment and suppose 95 percent of the patient got cured then we always say that that doctor is very good.

So, a popular interpretation of the probability is as follows. The probability of the outcome is observed as the proportion of the experiment that results in the outcome. This means what, suppose if there are 1000 patients which have gone to the doctor and out suppose 950 patients have got cured. Then what are we trying to do? We are simply trying to compute the proportion. 950 divided by 1000 that means the total number of patient is cured divided by the total number of patients that went to the doctor. Or to whom the doctor treated.

So, this is also probability but if you try to see, we are trying to see the probability as an outcome in the form of a proportion. There the proportion of the experiment that resulted in that particular outcome.

(Refer Slide Time: 32:53)



Subjective Interpretation of Probability:

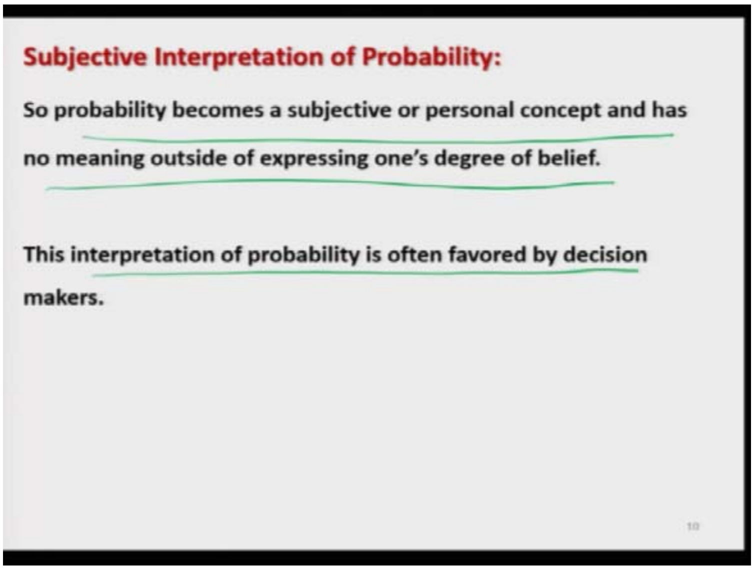
In the subjective interpretation, the probability of an outcome is not thought of as being a property of the outcome.

It is considered as a statement about the beliefs of the person who is quoting the probability.

The probability is about the chances of occurrence of the outcome.

Now in case if you try to consider the subjective interpretations of the probability, then in this case the probability of an outcome is not thought of as being a property of the outcome. But it is considered as a statement about the belief of the person who is quoting the probability. The probability is about the chances of occurrence of the outcome.

(Refer Slide Time: 33:19)



Subjective Interpretation of Probability:

So probability becomes a subjective or personal concept and has no meaning outside of expressing one's degree of belief.

This interpretation of probability is often favored by decision makers.

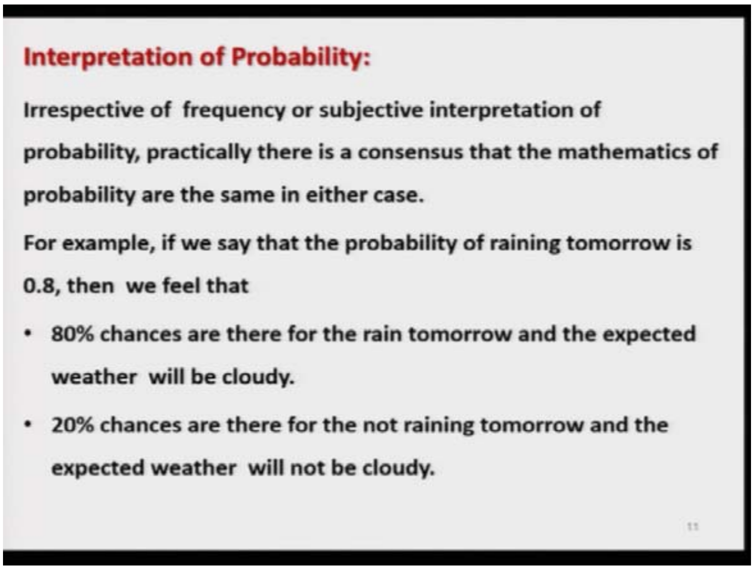
10

What you mean by this? In this case the probability actually depends on the personal belief of the person who is stating it. So, the probability becomes a subjective or personal concept and has no meaning outside of expressing one’s degree of belief. For example, suppose you have got four friends and one of the friend comes always 15 minutes earlier with the given time. Another friend comes exactly at the given time. Another friend comes 15 minute late but always. And there is fourth friend who sometimes comes, sometimes not.

Now if I ask you a simple question about a particular friend who come 15 minutes earlier. That what is your opinion about that friend. You will say yeah, yeah he comes always 15 minutes earlier before time. And if I try to ask you about that friend who sometimes comes, and sometimes not. You will say, I do not believe on that whether the person will come or not. What is that? That is your belief.

And this probability is related only to your belief. Means it might be possible that if you ask some other person about that friend, they may have a different opinion. So, this is what we mean by the subjective interpretations of the probability. But this type of interpretations for probability is often favored by the decision makers. The Prime Minister feels or the Finance Minister feels that if that particular policy is implemented then the country will progress. What is that? That is their belief.

(Refer Slide Time: 35:22)



Interpretation of Probability:

Irrespective of frequency or subjective interpretation of probability, practically there is a consensus that the mathematics of probability are the same in either case.

For example, if we say that the probability of raining tomorrow is 0.8, then we feel that

- 80% chances are there for the rain tomorrow and the expected weather will be cloudy.
- 20% chances are there for the not raining tomorrow and the expected weather will not be cloudy.

11

Now when we come to the on the aspect of interpretations of probability, then irrespective of frequency or subjective interpretation of the probability, practically there is a consensus that the mathematics of probability are the same in either case. Either it is subjective interpretation or the frequency interpretation. We have to use the actually we use the similar type of mathematics to find such probability.

So, you do not have to worry for these things. But you have to understand it is very important in our life to understand, what are we trying to do? For example, if we say the probability of raining tomorrow is 0.8, then we feel that there are 80 percent chances that are there for the rain tomorrow. And the expected weather will be cloudy. And there are 20 percent chances that there will not be rain and expected weather will not be cloudy that is how we try to see.

Now I come to an end to this lecture but if you try to see what we have done, the type of probability theory what was in your mind, I have given you a very different type of interpretations or very different type of process to think about the probability. Well, we are always looking forward for the probability of success. Because as human being we all are optimistic.

But remember one thing that when we are trying to deal in statistics, we always try to develop the tools which are trying to minimize the risk. So, sometime people say that statisticians are pessimistic but that is our choice, either we want to take the risk in our life against the maximum profit or again the maximum risk. Some people say high risk high gain, some people say low risk low gain. Well that is the personal thing, I cannot interfere in this thing that is your choice.

But as far as the statistics or the probability theories concerned, I have tried my best to give you a different type of picture which you have to think because it may happen that when you are trying to compute the probability in very huge bigger data sets. Then your personal belief is going to dominate the value of the probability. Your personal belief that which of the method has to be applied for computing the probability will dominate there.

Sometime you hear all or you listen or read at the newspaper that somebody in the company has taken a wrong decision. Because of this all troubles came. What is that? That is the belief of that person. Which worked and this belief is nothing but your probability. So, you try to think about this lecture, try to have a revision of this lecture couple of times and try to see, how you can interpret this probability for yourself in your case.

And from the next turn I will try to introduce some little bit mathematics into this concept. So, that I can formulize it. Why? Because you will not belief on me unless and until I explain you these things in terms of mathematics. So, you try to have a quick look on the lecture and I will try you see tomorrow once again in this lecture till then good bye.