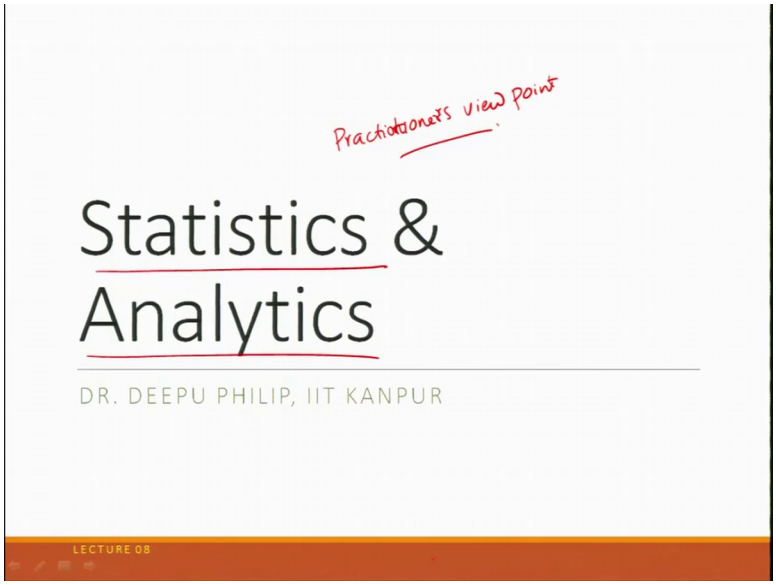
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Department of Industrial and Management Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture – 08
Statistics and Analytics (Private)

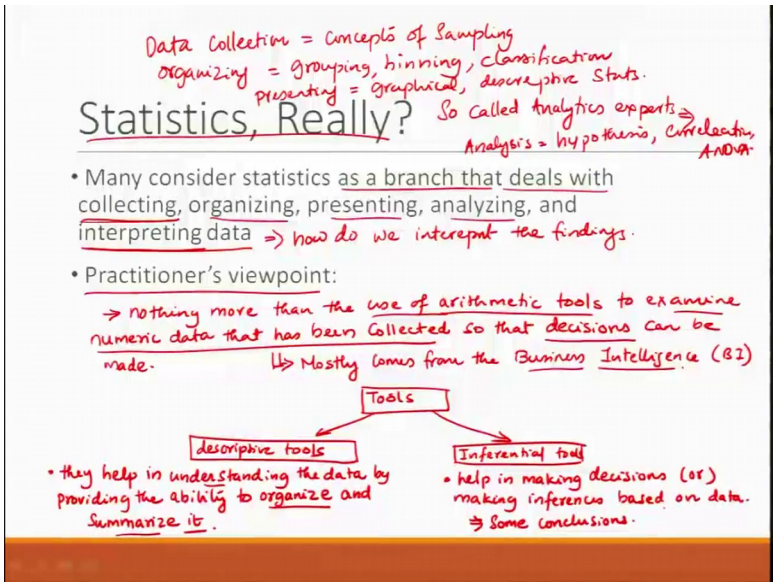
Good evening, welcome to the another lecture of a applied analytics a practitioners approach on descriptive prescriptive and predictive analytics and today, we are getting into the new concepts; new aspects of descriptive stats and other related tools because so far we seen in the philosophy and other aspects of the analytics and now, we are going to see, how we learn some tools and how we apply them to set of data. So, today, we will start looking into the importance of the concepts of statistics and analytics and we will see what statistics for us means.

(Refer Slide Time: 00:55)



So, this we are going to look at from a practitioners view point, ok, sorry I cannot spell properly. So, you guess are to find the right spelling.

(Refer Slide Time: 01:09)



So, let us take a look into statistics, ok, the title itself is kind of funny statistics really is everybody thinks that statistics is something that they know about what it is and many of the people the who are involved in a analytics so, the so called analytics expert; so called analytics experts ok, there definition of statistics is a branch of a branch of science or maths that deals with collecting, organizing, presenting, analyzing and interpreting data.

So, it involves data collection ok. So, that is one of the reasons why sometimes you talk about data collection which sometimes use the concepts of sampling. Then you have organizing and where we talking about grouping, binning, classification, etcetera ok. So that aspects; so you will study some of those tools; then presenting ok, you are talking about graphical presentations, then we talking about descriptive stats, that aspect then analyzing. So, we talked about analysis. So, analyze analysis will be like hypothesis testing correlation stuff like that Anova, they all come in this ok, interpreting is the how do we interpret the results for the findings ok.

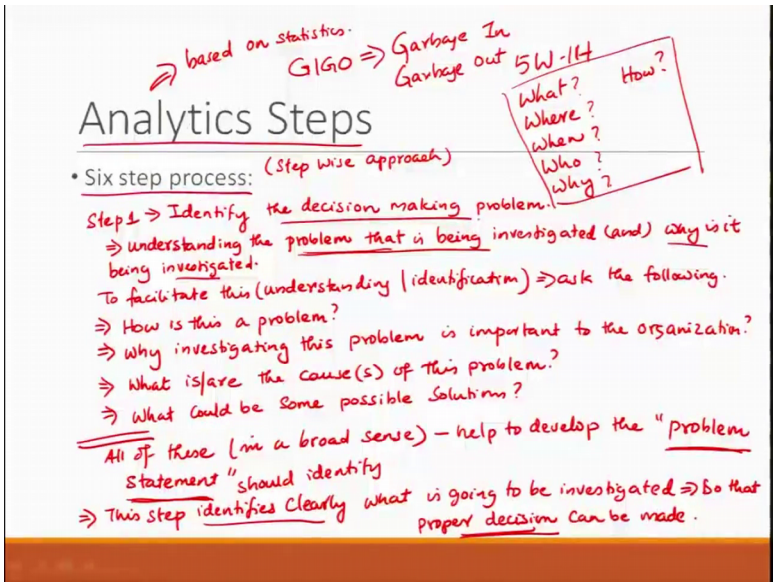
So, this is the classical viewpoint of what is statistics, but for a practitioner. We can think about statistics it has a less thing about us nothing more than the use of arithmetic tools to examine numeric data, data that has been collected. So, that decision can be made. So, the practitioners view point or the practitioners approach focuses on the decisions and it is more about for them or the practitioners, it is more about a set of arithmetic tools or the usage of arithmetic tools to examine the numeric data that has been collected. So,

here you will ask; obviously, then what happened to the other data alphanumeric data that is different part which is not too much of importance to us, but here the importance is mostly on the numeric data because the data collection has been done.

So, this numeric data who collected the numerical data, mostly, it comes from the mostly comes from the business intelligence or what we call as the BI as such it is basically involved collecting organizing and storing the data and these tools these arithmetic tools you can think about these tools to be divided into two forms and what we are going to see is that the tools can be thought about into fashion, the first one is lets what we called as the descriptive tools descriptive tools and the second one we can call it as the inferential tools inferential tools. So, the descriptive tools what does descriptive tools mean these tools they help in understanding the data understanding the data by providing they help in understanding the data by providing the ability to organize the ability to organize and summarize it summarize what summarize the data.

So, the I am here in the descriptive tools is they help us to understand the data by providing this ability to organize and summarize. So, here the aim is to more to organize and summarize data whereas, the inferential tools it is they help in making decisions or they help in making or making inferences based on data. So, here you are thinking about making inferences based on the data that is what is going on in this part ah. So, inferences means some conclusions you are concluding based on what you are thinking about this point now.

(Refer Slide Time: 07:36)



The analytics in a way; so you think about analytics as based on statistics ok. So, since it is based on statistics what we are trying to do here is we can think about it is a 6 step processes because for practitioners, we like to have a stepwise approach ok. So, since it is a 6 step processes. So, we can talk about as the step 1 being identifying or identify let us call it as a verb identify the decision making problem ok.

So, this implies understanding the problem the problem that is being investigated and this is important and why is it being investigated. So, at this juncture what we are supposed to do is we have to identify the decision making problem identified why it is a problem and why is it being investigated ah. So, that this aspect in this and to facilitate this to facilitate this or which means this understanding our identification either one of them we can think about asking few set of questions ask the following. These questions will help in understanding making us understand why this is a problem and why is it to be investigated.

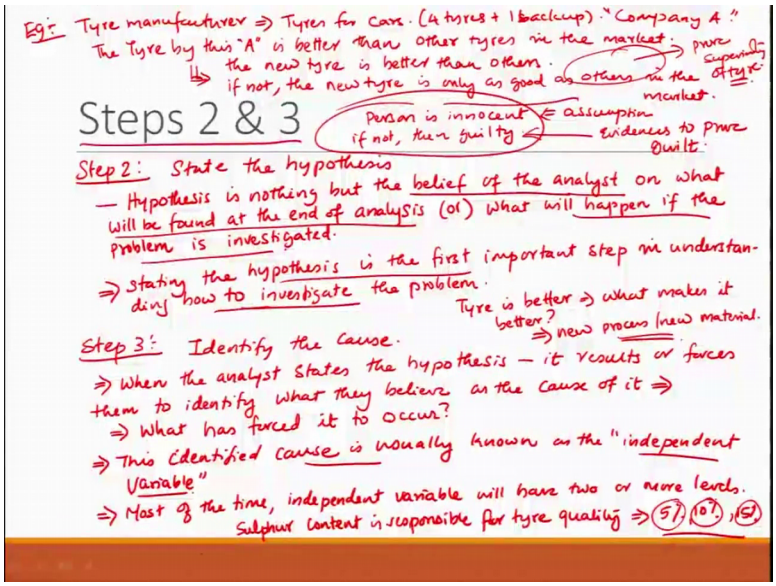
First question is; how is this problem? Second question is; why investigating this problem is important is important not to you important to the organization because we are thinking from a applied sun point then the third question is what is or are the causes for the course the causes of this problem ok, what is the reason behind the problem. Next question you can think about this; what could be could be some possible solutions ok? So, all these questions all of them all of these in a broad sense you are not looking for

exit answers, but you looking for a broader answers help to develop the what we call as the most important thing problem statement ok.

So, all these questions. So, the Japanese philosophy in this case is there talk about 5 w 1 h there is one approach the Japanese do? What, where, when, who, why, these are the 5 Ws, and then the next one is how and Japanese philosophy says that if you follow this approach this 5 w 1 h approach you would be able to reasonably define, what is the problem statement? What we are saying here is similar one, but somewhat broader questions not really in the particular pin pointed questions, but these questions to large extent help us to broadly develop the problem statement and why do we need the problem statement because this step identifies clearly or it should identify clearly or it should identify ok, clearly what is going to be investigated what are we going to study investigated this is to be clearly stated why. So, that proper decision can be made.

So, if the problem statement is not clear then the decision will also not be clear. So, this is where the phrase that the most famous phrase called Gigo comes into picture which means garbage in garbage out. So, if you have a clear problem statement you have been able to identify the problem properly then he because of that proper decision can also be made. So, the first step aim is to identify the decision making problem. So, understand what the problem is being investigated and why is being investigated and then there are set of questions that will help us to develop the problem statement and from there, once you have the problem statement clearly defined, then proper decisions can be made.

(Refer Slide Time: 14:18)



Now, we get to the second steps step two and step three ok.

So, the step 2 is to large extent n is state the hypothesis or what we can call it as; what does it means is hypothesis. Hypothesis is nothing, but the belief of the analyst on what will be found will be found at the end of analysis it is your belief what will be found the end of analysis or what will happen if the problem is investigated. So, the I am here is that you have stating your belief the unless belief on what will be found at the end of analysis. So, it is kind of saying that we start doing analysis this is what I believe or this is what we think, it will happen if this investigation is conducted why it is important because stating the hypothesis is the import is the first important step in understanding how to investigate the problem ok. So, the how to investigate the problem is given by the bike fairly setting the hypothesis.

So, let us take an example an example of a tyre manufacturing company tyre manufacturer assume that this companies manufacturing tyres for tyres for cars not automobiles car. So, car required for trials 4 tyre tyers plus 1 stepney or 1 back up this company manufactures this tyres and the aim is that this company is trying to say that the tire by this lets call this as company a the tyre by this company a tyre by this a company is better than other tyres on the market, if this is the claim made by the company and if it is a problem that is what something that need to be investigated.

So, then what is the hypothesis at this point or how do we say the hypothesis at this point hypothesis you can loosely state that ah. So, if this is what you want to prove or this is what you want investigate that whether the tyre is better than the accessing tyres in the market if not then what will happen the answer is that the tyre is as bad as are the tyres that are available in the market.

So, you can say that the new tyre is better than others that is one option at the other way to think about it is if not the new tyre is only as good as others in the market this is one way to look into it. So, here you are basically say trying to prove that the tyre is as good as for better if you say that is a new tire is bad then the other price in the market then the investigation that you are going to look for is other way. So, this the same way the judiciary system makes the promises that person is innocent this is the perception or the assumption or the belief and if the person is not innocent if not then guilty ok.

So, this kind of an approach this kind of a statement in a pair wise belief you have a belief on if that belief is not happening then what is going to be the next one that is what we called as a hypothesis or setting this in some mathematical form or verbal form it does not matter; that is what your hypothesis is? So, then all your analysis in the second case is to the; what the police will do they will only trying to find evidences to prove guilt ok. So, the same thing if we do this, then we only look for evidences in this case to prove superiority of tyre.

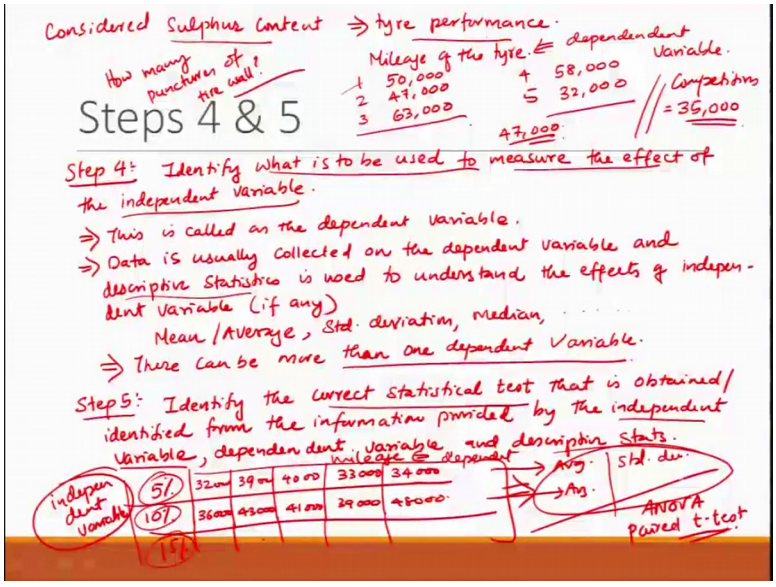
So, the analysis, how will you do the analysis is determined by how you state your hypothesis, then step 3 step 3 is simply put identify the cause ok. So, what does this mean here is that. So, in better words it is when the analyst states the hypothesis then what happens it results or forces them to identify and if I what they believe as the cause of the cause of it as the cause of it or in a better way what you think about it is what has forced it to occur ok.

So, when we think about the tyre manufacture when they say that the tyre is better the question is what makes it better maybe the answer is new process or new material something like this. So, this is where you identifying the cause why what has resulted in the occurrence of the problem the typically in statistics terms this cause this identified cause is usually known as the independent variable we call it as the independent variable.

So, the cause the one that resulted in this particular behavior or belief is what we call of independent variable because of it the phenomenon is being observed most of the time independent variable will half two or more levels this is a tricky one because what happens is in this case, you think about it the sulphur content let us call it as a sulphur content is the reason for the tyre quality is responsible for tyre quality if that is the case, then you can possibly say that the sulphur content can be 5 percent, 10 percent 15 percent something like this and one of the setting will along with some other values will give you the a good quality tyre. So, it is same content of the; what percentage of sulphur is content with the rubber is the one that actually determines the entire quality if you can think about it that way.

So, the sulphur content is the independent variable and these 5, 10, 15 percent are what we call as the levels of the independent variable hope you guys understand this.

(Refer Slide Time: 24:35)



And then we will know move to step 4 and 5 ah. So, the step 4 is ideally identify what is to be used what is to be used to measure to measure the effect of the effect of the independent variable independent variable. So, here to decide what is to be used; what is the phenomena that you are going to use the used to measure the effect of the independent variable and why is it cause.

So, this thing that is used to measure the effect this is called as the dependent variable and data is usually collected on the dependent variable on the dependent variable and

descriptive statistics descriptive statistics is used to understand if you to understand the effects of the effect of independent variable if any if there is an effect of the independent variable on the dependent variable then we use descriptive statistics to understand.

So, in the previous case we consider considered a sulphur content as a factor for the tyre performance now the question is how do you measure the tyre performance and one way to do it is mileage of the tyre. So, one way to do it is you manufacture maybe 5 tires or something. So, tyre 1, tyre 2, tyre 3, tyre 4, tyre 5, manufacture tyres and you put them on different vehicles and run them and the first tyres give you 50,000 kilometers, second will give you 47,000 kilometers, third is 63,000 kilometers, fourth is 58,000 kilometers, fifth is 32 kilometers something like this.

So, these value the mileage that you measure and which you are attributed to the performance of the tyre and you can saying that sulphur content influences the tyre performance which in turn influences the mileage. So, the mileage becomes the dependent variable through which we measure the impact of sulphur content. So, here you use descriptive statistics in a sense what we are trying to do is we are basically try to calculate what is the mean or average, ok, what is the standard deviation ok, what is the median then many other statistics range etcetera we calculate of this from there using that we try to find out whether the tyre is better.

So, if you say that the rest of the tyres in the market compete competitors are providing tires with let say 35,000 kilometer mileage that is ok. So, then if you find this average, let us say it comes to let us say 47,000 thousand kilometer, then you can possibly say that reasonably looks like this tyre is the new tyre is better than the other tyre and this new tyre is better barely because of the fact that the sulphur content is the one that is influencing the superior tyre performance, and also you should note that we should remember that there can be more than one dependent variable ok. So, one another example of this is identify how many punctures happened how many punctures of tyre wall this could be another dependent variable that you can see to measure the performance of the tyre.

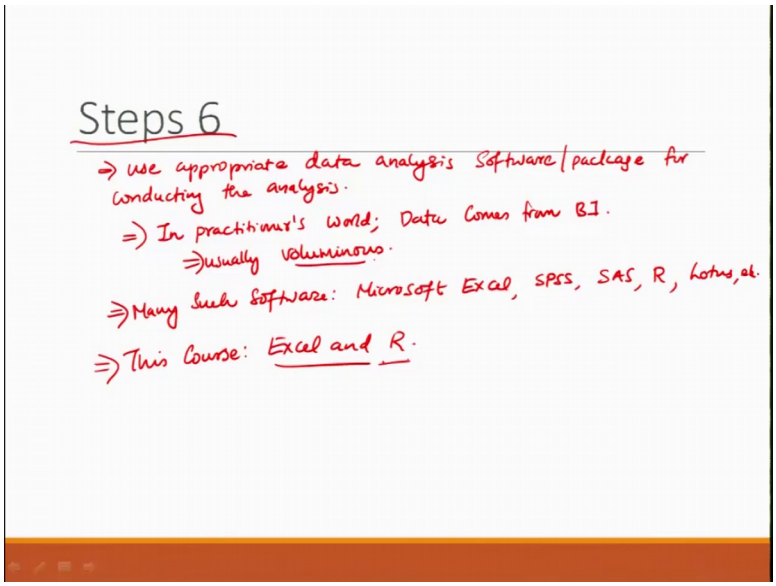
So, you can always there is no need of there is only one dependent variable that can be more than one dependent variable this is quite applicable and allowed then we talk about the step 5 fifth step which is identify the correct statistical test identify the correct

statistical test that is obtained or identified from the information provided information provided by the independent variable dependent variable and descriptive stats ok.

So, one example of this would be you are made tyres 5 tyres with 5 percentage content, then let us say 10 percentage content 5 tyres. So, you have like it this other 5 tyres and you put the mileage one let say 32,000; 39,000; 40,000; 33,000, something like this ah. So, you do your data on that let us say 34,000, these are the five data values you get and then 10 percent you got 36,000, 43,000, 41,000, 39,000, 48,000, something like this and then you can make tyres 5 tyres made out of 15 percentage like this and then get the data values and then using this data values, you can find what is the average of this row average of this row, etcetera standard deviations etcetera and then you can decide whether you want to do and analysis of variance for paired t test.

So, the data that data will tell you which is a probably better test it is for you to use to identify what percentage of the sulphur content gives you the best performance? So, the is the statistical test that is to be used to identify to make the decision of what sulphur content to use to get the best tyre performance should be done with the help of information that is provided by the independent variable, this is independent variable ok, this is the independent variable and this is the mileage this is your dependent variable and here is your descriptive statistics ok. So, these 3 put together tells you what test do and how to get the how to identify the correct statistical test to make the correct decision ok.

(Refer Slide Time: 33:30)



And the last step 6, ok, it is not step it is steps 6, then once this is done use appropriate data analysis software or package whatever you want to call it for conducting the analysis most of the time when you have a in business in practitioners world in practitioners world data comes from BI business intelligence department. So, usually voluminous I believe check the spelling if the data is voluminous than you require software probably doing manually might not help and it might also take too much of time. So, so there are many options many such software ok.

Starting with Microsoft excel, then you have SPSS, now made by IBM, then you have SAS; that is SAS institute R the open source alternative officers lotus spreadsheet etcetera, these are multiple options of the software that is available in this course this course we use excel and R other reset there will be tutorial of R in this course and you will learn what is R and how to use R further and for symbol cases we will also use excel in this regard with this we reach the conclusion of the process of how do we do analysis.

And how do we the stepwise process of conducting analytics and now from here after we will be starting to get into different tools of descriptive stats and graphical display tools because descriptive starch is one important parameters as p c identifying independent variable dependent variable and looking at the descriptive starch helps us to understand what is the right statistical test that is to be used it for the making the decision or doing the analysis ah. So, thank you for your patience hearing today and we will continue the

rest of the lecture rest of the new res of this course using the new topics in the coming classes.

Thank you