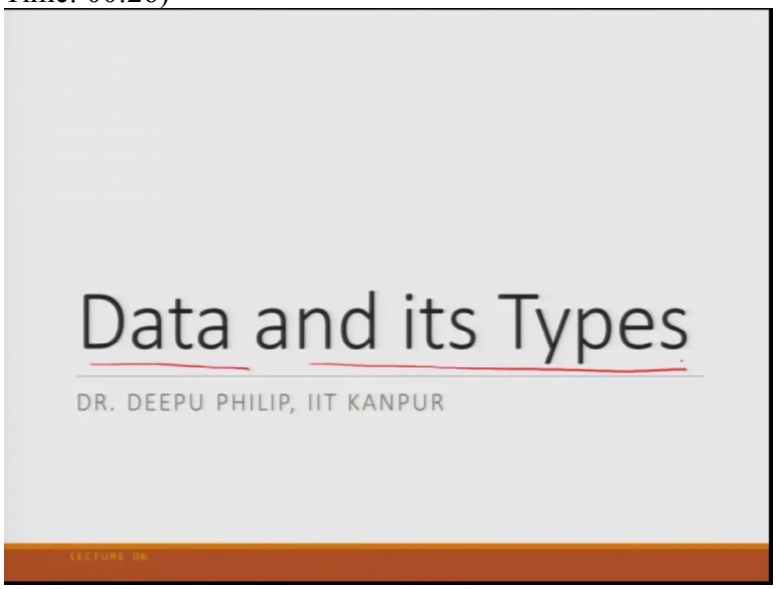
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

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

Lecture – 06
Data and its Types

Good evening students, we are in the second week now. We are getting into the 6th lecture on the data and its types and this course is a practitioners approach in analytics.

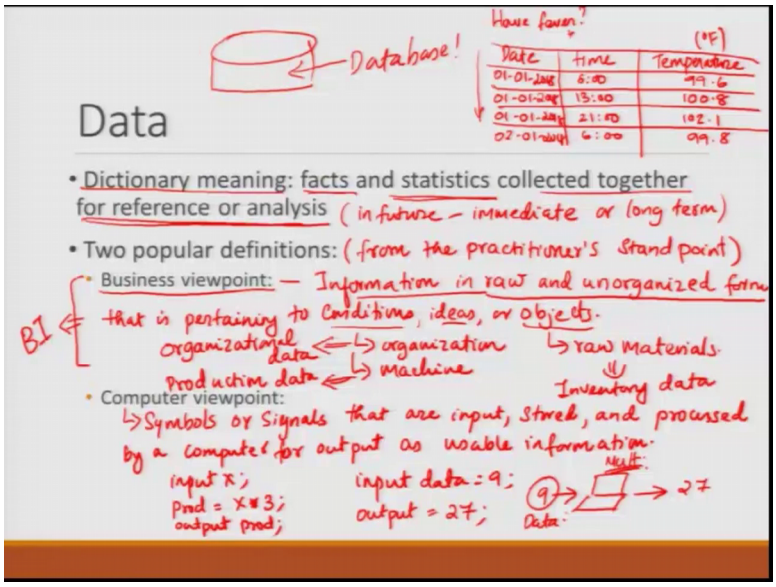
(Refer Slide Time: 00:26)



We are basically looking at the descriptive, predictive and prescriptive analytics components of it and today when we talk about the most important part of analytics about data and the different types of data and why we need to understand this types of data and I hope that you are so already been reading what is assigned for your reading from the syllabus and as well as you are following the text books that has prescribed as part of the course and from now onwards will be starting more focusing more towards the a applied side of the course, but also remember that we are also applying it from a practitioners view point a person who does it on a daily basis.

So, hence some other things that we would be interested in my not be too much into the theory, but more into how to add we apply the theory. With that we will start today's lecture and the topic today is data and its types.

(Refer Slide Time: 01:23)



Let us start with the first term data, the data if you ask somebody to talk about data sometimes people describe date and the form of something like this and say this is a database and whatever is stored in this is a data. Classical you know comical diagram that people would do on this and there is lot of other conceptions, misconceptions, confusions everything are associated with the data.

The dictionary meaning of data, when we talk about what does a dictionary says these are facts and statistics could be facts or could be refracts or it could be statistics collected together for reference for analysis. So, for example, if you are measuring then I say you have fever. So, let us say have fever is a true. So, how do we ensure that? So, let us take an example that we are recording the temperature of you in multiple times in a day. So, here we have is a date, then you have a set time and temperature. So, day we will say 11 2018.

So, these are the day at 6 am, say you temperature is 99.6 degrees and that 1 1 2018 at 13 hours your temperature is 100.8, 1 1 2018 at 21, the temperature is 102.1. So, the question is what is this temperature? Obviously, the temperature is in Fahrenheit ok. Then you have 2 1 2018 at 6, the temperature is 99.8 something like this. So, if

you think about this, this is fact on the particular date time your body temperature is measured using a thermometer and if this is given to a doctor, this collection this temperature in given to a doctor he would probably analysis and say you have more fever in the night. So, you might be having malaria something like that.

So, it is meant for reference it is also meant for your. So, everybody can know what is temperature was. So, this data is collected and as well as a doctor can analyse the data and decide what, what, what disease you are suffering from ok. So, again facts and statistics collected together for reference and analysis. When in future, the future can be immediate future or long term ok, could be far away into the time. The true popular definitions that is from the practitioners standpoint ok, these two popular definitions are from the practitioners standpoints, practitioners standpoint apologize for my spellings you how to check my spelling square feet only.

So, for the practitioners there is a business view point and what is the business view point the most important part, it is say is it is information, information in a raw, r a w raw and unorganised un organised. Information draw and unorganised form, form that is pertaining to pertaining to conditions ideas, or objects ok. So, these are raw and unorganised form, information in the raw and unorganised form about condition can be conditions about organisation, it can be about a machine, it can be a human being worker; etcetera or it could be about ideas, new ideas. Like a new business idea, new product idea or it could be about objects, it could be information about a particular raw material that have you going to use it in the manufacturing etcetera like that.

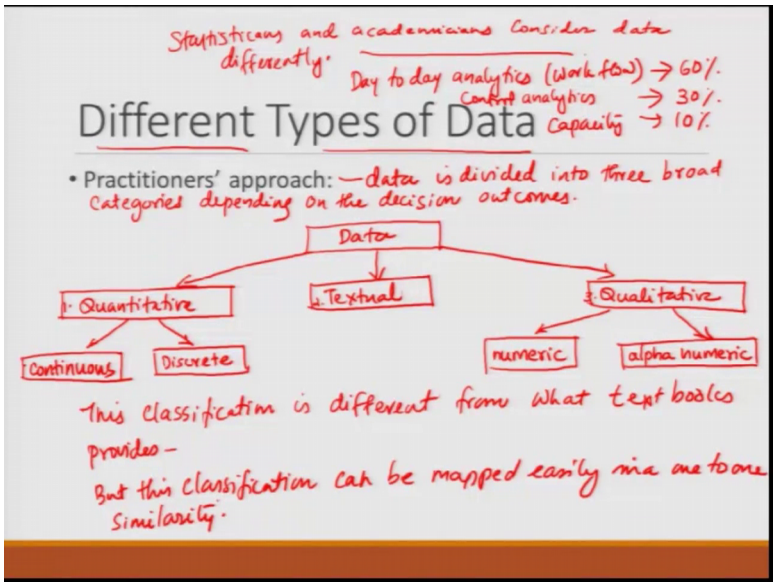
So, here is an example will be raw materials etcetera. So, whatever the raw and unorganised information collected on this behalf; that the business view point. So, when you say you are collecting data about the organisation you call it as organisational data, if you are collecting data about the machine sometimes people will call it as production data or machine data you can call which are way it is when you have things about raw materials you will say inventory data, stuff like this ok. So, depending upon what you doing in business you will can get lot of different type of data. So, when you collect all these data and put together in some format. So, that you can refer later that is where the concept of b; I comes into picture what we discussed earlier the classes business intelligence ok.

So, the business view point data is specifically as per the pertaining to conditions ideas for objects what about the raw and unorganised information. Where as in the computer science view point or the computer view point it is slightly different, it looks at the review point is here is that we assume here in the computer viewpoint that these are symbols or signals, symbols or signals that are inputted, that are input stored and processed, processed by a computer, computer for output ask usable information.

So, in an example if you write a program let us say input x let us see product equal to x times 3, output product let us see this is a computer program and you have provide the input data input data is provided as input data is equal to 9, then it does is it takes 9 multiplied by 3 and this is whatever the result 9 times 3 27. So, the input data and output what we will get out of this is output will be 27, 9 times 3 27. So, the input of 9 was given to a computer in the computer process the output to be 27 by doing the process of multiplication here ok.

So, the computer looks at this as a data that is provided to it, similarly can think about data that is being stored in the database and those kind of aspect.

(Refer Slide Time: 09:21)



With this what we will do is we will jump into the next thing the different type of data ok. So, we can always say that there is a statistician and academicians, academicians consider data differently ok. So, if you are looking in doing data from

my research done point then this is not the approach, here we are not looking we are just basically looking at how to do this for in the industry or the business industry as day to day (Refer Time: 10:06). So, our aim is to do day today analytics which we were called of the workflow which is the most, let us say, we cover about 60 percent of that as part of this; then, we also talk about control analytics ok. Which we were probably talk about 30 percent and then capacity we will talk about 10 percent. So, how will control will go through the courts apparently?

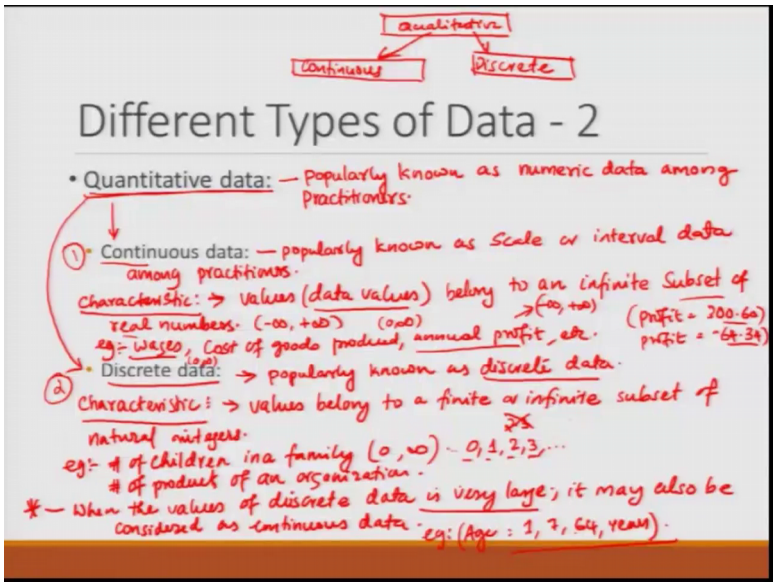
So, from that view point what we say it is a practitioners approach says this is that the data is divided into 3 buckets or 3 let us called as 3 broad categories, categories 3 broad categories depending on the, depending on the decision outcomes ok. So, since it is 3, it is better to draw it a diagram. So, let us make boxes, let us assume that this is data. So, the first classification of data is let us call it as the quantitative classification, this is the quantitative data that we talking about, second one we are talking about let us talk about is the textual data text later then the third classification let us talk about is the qualitative data. So, the data is broadly classified into 3 quantitative number 1, textual number 2 and qualitative number 3.

So, we will see what each one of them and this quantitative is further divided in to 2 as far as we are concerned and the first one is the continuous data ok, second one is the discrete data. So, quantitative data is divided into continuous and decretive then textual data is textual. So, there is no classification for that and qualitative data similar to quantitative data can be divided into 2 again and this division, one part is the numeric data and other part is the alphanumeric data right.

So, thus numeric data and alphanumeric, so now, you will see what each one of them, but in the broad sense this is it different compared to. So, this classification, situation is different from what a textbooks provides like nominal ordinal etcetera, but this classification can be mapped classification can be mapped easily in a one to one ah, one to one similarity.

So, let us see what are the individual aspects of this and this is our practitioner's classification of the data.

(Refer Slide Time: 14:28)



So, the first thing is we going to talk about is the quantitative data ok; obviously, quantitative data popularly known as, known as numeric data popularly known as numeric data among petitioners, practitioner. So, when practitioners typically talk about numeric data they are literally referring to quantitative data ok, there is some fine line in between, but that thing that we will clear out later ok. So, in this numeric data if you remember the diagram the qualitative data was divided like this, qualitative data was divided into the 2 things one was continuous the other one was discrete. So, let us see what these two are ok, the continuous data popularly known as, popularly known as known as scale or interval data among practitioner.

So, when the practitioner, says I am dealing with a scale data or interval data the; their talking about the continuous data which is within the quantitative data still ok. So, the most important thing is about this data, the characteristics that let us talk about the characteristic, characteristic of this is values this data value ok. When I say values V these are data values, data values is belong to and infinite subset infinite subset of real numbers ok.

So, the data values belong to an infinite set of real numbers, real numbers we know starts from minus infinity to plus infinity ok. So, let us take some examples and understand what it means, wages is an interval data, cost of goods cost of goods produced, annual profit etcetera. So, if you think about the wages, the wages can

technically vary between 0 to infinity a large number, cost of goods again 0 to infinity, annual profit can be a negative minus infinity to plus infinity something like this ok. So, but it is not really infinity there is some large number we think about it that way. So, if you think about it this is a subset of the real numbers. So, they are continuous data that is a subset of the real numbers definitely.

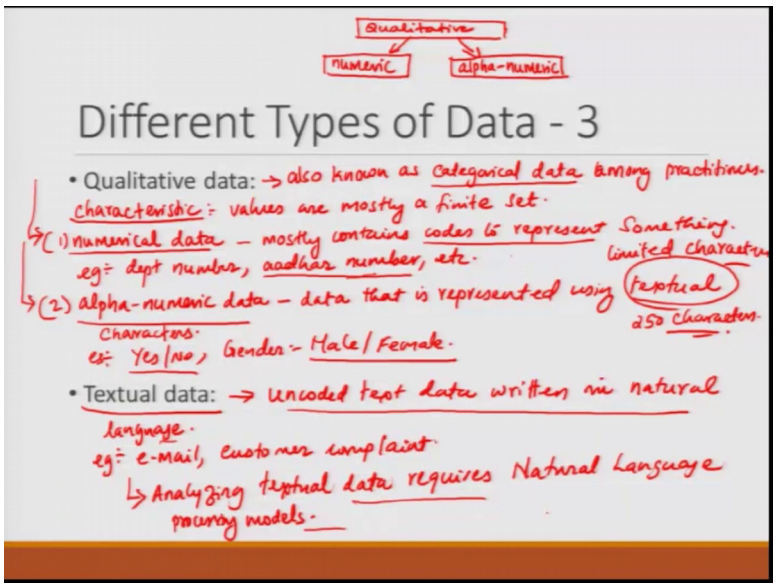
Now, the second one the discrete data which is also when the; it is popularly known as ok, known as the discrete data ah. So, when somebody says discrete data by practitioners; that means, they are talking about quantitative discrete data and the most important thing is the characteristic of this. Characteristic of this is the values, values belong to a finite or infinite finite or infinite subset of natural integers ok. So, here you will probably get a value of the annual profit could be something like for example, profit you can take a value of 300 rupees in 60 paise something like this or a profit could be another value would be minus 64 pi 64 rupees, 34 paise something like this.

So, these are real numbers because you are it is a continuous value, where as in the case of natural integers what we are trying to do is it is basically its countable finite. So, an example is example of this is number of children in a family. So, the values can vary between 0 to infinity only not infinity a large number. So, you will have numbers like this as there is could be 0 child, there could be 1 child, there could be 2 child, there could be 3 child, etcetera.

You will never get a 2.3 child because this is a wrong inform. So, these type of values where it is say countable natural integer value that kind of values are called as the discrete data ok. Another example is the number of products, number of products of an organisation that is another case like. So, the other part is this when the, this is an important point in this you should remember, when the value of discrete data is very large, very large it may also be considered, considered as continuous data this is where some people get confused and example I will give a very good example for this let us consider the age of a person ok. So, a person can have an age of the 1 year, 70 year, 64 years etcetera. So, these values ideally speaking ah, if you have a large set very very large, if you count the they if you document the data of the age of everybody in India then; obviously, you will get all sort of values multiple times.

So, then sometimes people will say that this value is similar to that of a continuous data. So, when people large very last set of discrete data as sequent to continuous data that is where the sometimes the confusion happens, but for the practical purposes, let us consider that what that the qualitative sorry, quantitative data is divided into 2. The continuous data and the discrete data and I told you also what condition where the discrete data when there is large set of it sometimes tend to be treated like a continuous data alright. So, we hope that I hope that you understand the concepts of the quantitative data aspect of this now let us move to the next type of data which is called as a qualitative data and remember qualitative data in our classification diagram.

(Refer Slide Time: 22:32)



Qualitative, qualitative data it was divided into 2 we call it as numeric data as 1 and other one we called it as alphanumeric ok. So, the qualitative data typically in the industry known so also known as, also known as categorical data, categorical data among practitioners when somebody say that I am working on categorical data, with the talking about their working on qualitative data ok. So, the major characteristics of a qualitative data, of a qualitative data is values are mostly a finite set, is unlike the other case it is not an infinite set up value it is there is a finite set a value for this ok.

So, the first thing about this is us as we said the numerical data when we talked about here which is the numerical qualitative data, that we talked about this it me mostly

contains codes to represent something. An example of this is a department number or like your Aadhaar number, etcetera. So, what up when is the Aadhaar number, when use Aadhaar number, you can identify that this Aadhaar number is used to represent an individual ok. The individual here is that whoever that person is and otherwise it does not really make much sense to anyone, similarly the other party is the what we talked about it as the alphanumeric or numeric data and in this case, it is data that is represented, that is represented using textual characters, textual characters.

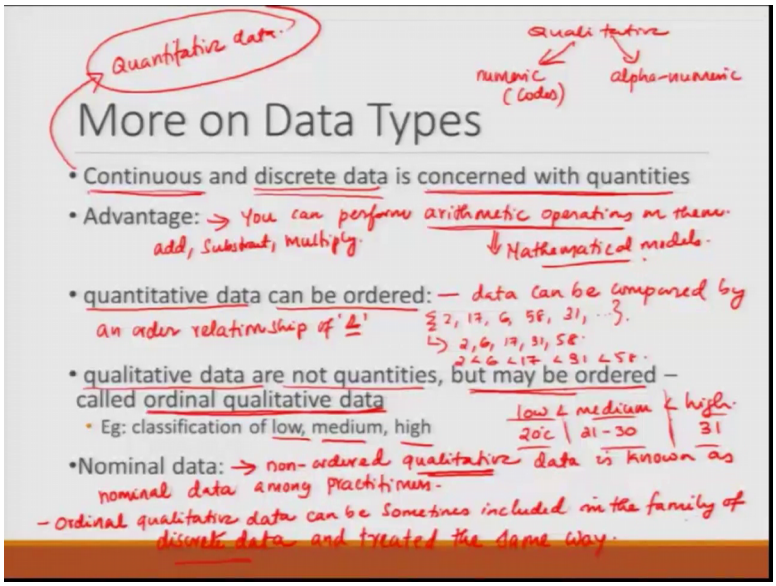
So, for example, of this is answering a question yes no or sometimes the gender where you reply does male, female etcetera ok. So, this is the important aspects of the qualitative data, now when we talked about it the other one which is the textual data and textual data in practitioners turns it is encoded, text data written in natural language, written in natural language. So, like for example, of this would be an email or a customer complaint etcetera the major difference between textual data and the alphanumeric qualitative data is that alphanumeric data typically contains limited amount of text or characters to describe something whereas, in the textual data the this is large one. So, analysing these data, analysing textual data required natural language processing models ok.

So, the qualitative data when people talk about it the qualitative data has a numerical data in this, but this numerical represented data is typically code of something ok. So, this like for example, postal code is an example of this which is a postal pin number and numerical values to assign to a particular post office. Similarly, in this case you have also called as alphanumeric data which is represented using textual characters where male and female or yes no something, something like that is not a textual character it is typically instead of the textual, I would say is that limited characters ok.

You do not have too many characters, sometimes people typically when say that this is at the maximum to 250 characters people make different classifications out of a. Whereas, here this is a much large a large amount of textual data type into the system or which denoting or describing a major incident ok, like an email or a complaint of a customer action and to do this you require multiple different models and if you get time in this class we will try to look some of these models (Refer Time: 28:32).

Before, I get into the classification of the other data aspects, which is used in most of the textbook statistical text books one of the aspects, I would like to bring to your guys notice is you should understand that this continuous and discrete data is always concerned with quantities ok.

(Refer Slide Time: 28:49)



So, what is the advantage of having a continuous and discrete data, where is continuous and discrete this continuous and discrete data is coming from your, this is a quantitative data and this data has a major advantage, this disadvantage is because it is concerned with quantities and the advantages that you can perform mathematical or arithmetic, arithmetic operations on them. So, you can do things you can do something like add, subtract, multiply, etcetera. So, if you have continuous and discrete data which is quantitative data you can do arithmetic operations on the data. So, that is the big advantage. So, the lot of your. So, people say when I am doing mathematical model and I am using data then when you are specifically mentioned it to either a continuous or discrete data which is quantitative in nature.

So, as we said that you can perform arithmetic operations on the quantitative continuous and discrete data, also another party is another major advantage is that quantitative data can be ordered, what do you mean?

When, data can be ordered which means data can be compared by an order relationship, relationship of less than or equal to. So, an example you see somebody

gives a data of 2 17, 6, 58, 31, etcetera, like this and you can always order this basically saying that 2, 6, 17, 31, 58, the relationship is that 2 is less than 6 which is less than 17, is less than 31 less than 58 like this. So, you can order them that ok. So, this is the major advantages of the quantitative data, you can you say less than or equal to relationship to order the date and this, then qualitative data or not quantities. So, that us we said we are seen either codes or they. So, remember in when you talk about qualitative data you have either. So, qualitative we had either numeric or which are basically codes or you had alphanumeric short text ok.

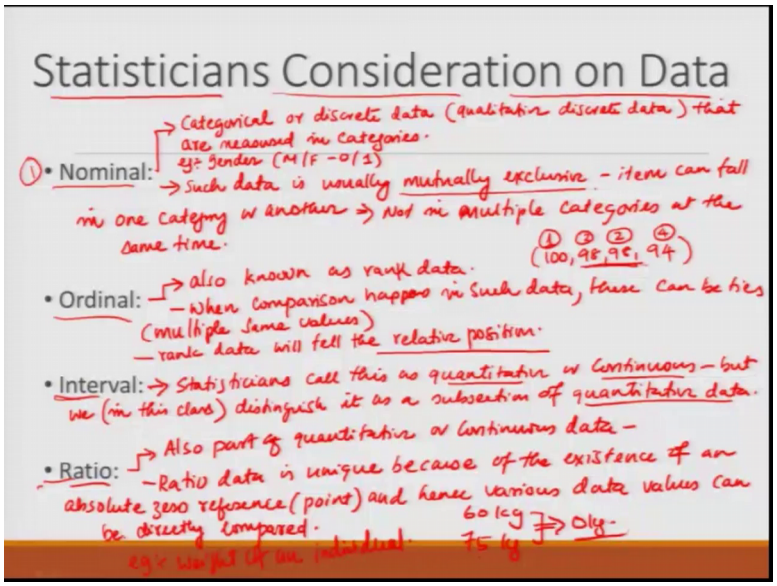
So, but in some cases it can be ordered and when you can do that is called order ordinal qualitative data. So, for example, if you are classified something as low medium and high then; obviously, low is less than medium, is less than high. So, a non-temperature could be temperature which is like maybe less than twenty degree Celsius or something like that and medium temperature could be something between 21 to 30 degree Celsius and greater than 31 degree Celsius could be high.

This you could think about it as a classified with a form of a low medium and high and then you can order this as well as that way so, but that type of data is called as the ordinal qualitative data, which is the ordered qualitative data alright. Then there is another day which is called as nominal data these are non-ordered qualitative data qualitative ok.

Non ordered qualitative data is known as nominal data, data among practitioners, the reason I am mentioning among practitioners because there is another nominal data that will come up which is typically what statisticians use and then there is slight difference on this. Also remember this ordinal qualitative data, ordinal qualitative data can be sometimes included in the family of, in the family of discrete data discrete data, this discrete data is your quantitative discrete data discrete data and treated the same way [noise, treated the same way.

So, what we are saying is a certain times ordinal qualitative data what we talked about it can be sometimes include in the family of discrete data, like for example, instead of the law medium and high if you are going with such kind of (Refer Time: 34:42) then that can be used in the ah, you can be treated in the way say same data as quantitative discrete data not every time sometimes.

(Refer Slide Time: 34:54)



Now, let us think about how the statisticians or the academic research considers the data ok. So, the statisticians considered they use the same price is like nominal ordinal interval ratio, but we can always make a one to one comparison with them. So, let us talk about the first one which is called as the nominal data ok. So, nominal data is categorical, categorical or discrete data which is, this is qualitative discrete data that are measured, that are measured in categories example of this is gender.

So, it can be a male, female or could be represented as a 0 1 something like this, alright. So, the most important aspect of this is ah, such data which means is usually mutually exclusive, which means item can for in one category or another not at the same time, not in or not in multiple categories at the same time ok. So, the nominal is typically category of discrete data that are measured in categories and usually search data mutually exclusive.

When statistician talk about nominal data we are talking mostly about the qualitative discrete data for the categorical data, then there is another data that is called as the ordinal data, an ordinal data we will kind of look at it this way is also known as, known as a rank data is called as rank data. Because when comparison happens, when comparison happens in such data there can be ties, ties means multiple same values ah. So, what this will tell is rank data will tell the relative position.

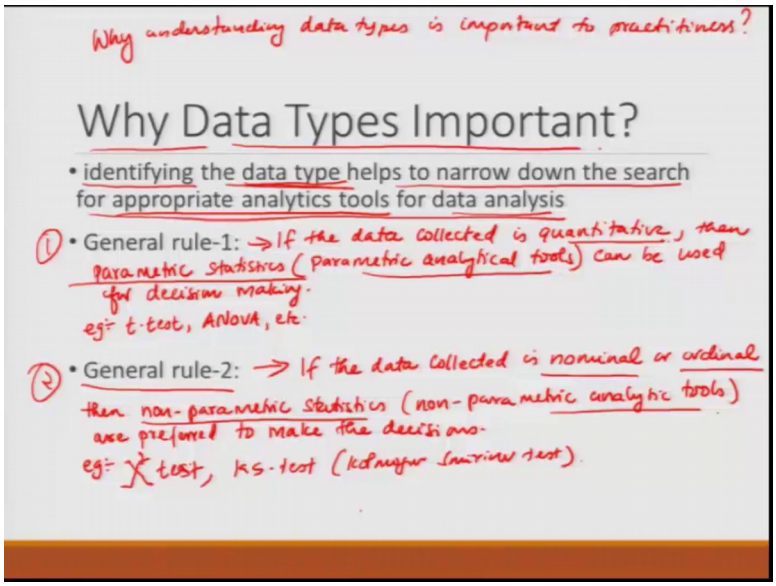
So, let us say when you guys write the final exam of this class and the scores are like out of 100 let us say, there is a 100, then 98, 98 and then 94 then the rank of this is this is the person who said rank 1. This both are rank 2 and as we as 2 rank 2 and this is rank 4 ok, is not the rank 3 at this because they rank 2 and 3 share by the 2 people who are exactly of the same marks. So, when you say that the 94 is the 4th rank and you are seen in 1 and 2.

So, obviously, the question where is 3 then you will; obviously, know that is to 2 ranks as one of the reason we got the fourth rank can (Refer Time: 39:16) ok. So, rank data will; obviously, tell you the relative position of things ok, then comes the next day which we called as the interval data. Statisticians call this as call this interval data as quantitative or continuous, quantitative or continuous ah, but we distinguish, we in this class we distinguish it as a sub section or subsection of quantitative data ok. So, when statistician says interval data they typically meant about us quantitative data continuous data continuous data, but for us continuous data is a subset of the quantitative data, alright.

So, remember if you remember the diagram when we had the surface of section of quantitative data, then comes the ratio data it is also this also a part of quantitative or continuous data which is again a subset for us in the qualitative data, the major thing is ratio data is unique, is unique why it is unique because of the existence of because of the existence of an absolute 0, existence of an absolute 0 point, 0 reference, reference means 0 point for us and hence various data values can be directly compared ok. So, if you for example, if you take the, weight of an individual if you think about it. So, somebody who comes with the weight of 60 kilograms and somebody comes with the weight of 75 kilograms, these 2 are comparable because of the absolute of this 0 kilogram weight. So, this is the person who is 75 kilograms is fifteen kilograms more than 60 kilogram percent ok.

So, because of the absolute 0 you can compare it that way ah. So, I hope that you guys now understand that the usage of statisticians, the nominal data the ordinal data interval data ratio data how does it maps to our classification of quantitative data and qualitative data textual data ok.

(Refer Slide Time: 42:57)



Now, the obvious question is why data types are important for us why do we need to learn all these different type of data types or why understanding data types is important, important to practitioners this is one of the most important question that we have to answer. Why do we need to do this, what is the most important aspect of it because reason is identifying the data types if you identify what is the data type it helps us to narrow down the search for appropriate analytic tools for data analyst. So, if you know the type of the data which data type it belongs to then it will help you to choose appropriate analytics tools rather than trying to use all analytical tools which are large in number, it is better than you can choose appropriate analytics tools depending upon the type of the data.

So that you can do analysis out of this. So, how what are the general rules, I mean how do you choose which one it is there to general rules which is typically advised for practitioner and either many other things also, but for the time being for practitioners just look at the rule. The first rule is if the data collected, if the data collected is quantitative, you have quantitative data then parametric, statistics or what we can call it as parametric analytical tools can be used for can be used for making the decisions, used for decision making.

So, when we say that you the data collected is quantitative the data is quantitative then parametric tools or parametric analytical tools can be used, what are some of the

parametric analytical tools that you might have heard across obvious example is t test is one example then analysis of variance for what we typically called anova, these are all examples of parametric data now another rule, rule number 2.

If the data collected, if the data collected is nominal or ordinal you know what it is nominal ordinal in the previous slide you gone through that then then nonparametric, nonparametric statistics or nonparametric analytic tools, tools is preferred are preferred, not is are preferred are preferred to make the decisions ok. So, of you said data is nominal ordinal, if there is if there an order data nominal data then we would require recommend that the use of nonparametric statistics ok, typically many people does not have heard too much about nonparametric statistics. But one very popular one is a chi square test this is not x square chi square stands a roman sorry Greek letter chi, chi square test and other one is k s test, which stands for Kolmogorov Simonov test I have no idea what the spelling Russian spellings are sometime far into me. So, these are the 2 general rules we follow rule number 1 and rule number 2, rule number 1 is about the radius quantitative use parametric statistics for parametric analytical tools and if the data is nominal of or ordinal data then we use nonparametric analytic tools.

So, with this we are able to understand that depending upon the data what we use or depending upon the type of data choosing appropriate tools is important and as well as having a data in one form and if you think that there is a better tool that is available in another form of data then you can think of ways to transform one data to another. Let us say for example, if you have the data the temperature of a one city for a particular time period continuously then you can basically classify them or bin them into different categories and we will see how that to be done and then we can say that this can be hot less hot called you made that kind of conditions can be put into this.

So, this allows you to transform data from one format to another as well, where you can choose appropriate data transformation tools. So, ask analytics or people who are practicing analytics it is more important for you guys to understand these aspects of the data and so that you can select appropriate tools that it will help you to make your decisions which is predominantly in the form of a work flow decisions on the form of a control decision.

Thank you for your patience listening and we will continue with the rest of the lecture in the next class.

Thank you.