Introduction of Data Analytics

Prof. Nandan Sudarsanam and Prof. B. Ravindran

Department of Management Studies and

Department of Computer Science and Engineering

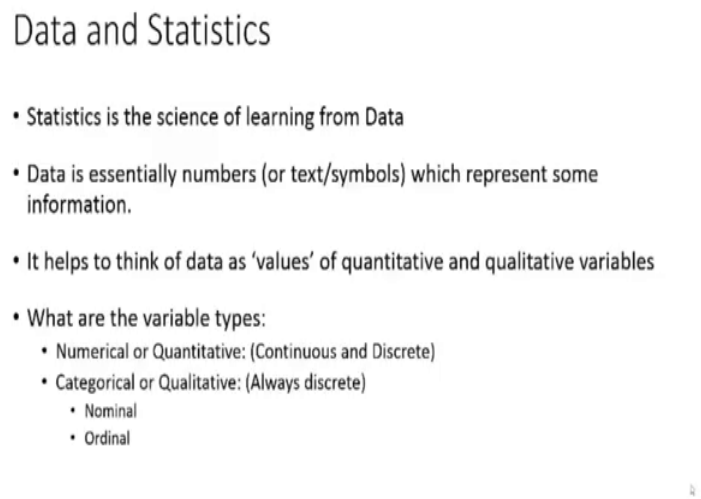
Indian Institution of Technology, Madras

Module - 01

Lecture - 03

Descriptive Statistics: Graphical Approaches

(Refer Slide Time: 00:31)



Hello, and welcome to the third lecture of the course Introduction to Data Analytics. My name is Professor Nandan Sudarsanam and today, we are going to be talking about Descriptive Statistics and more specifically about Graphical Approaches used in descriptive statistics. Now, before we jump into the content into descriptive statistics in the different types, it make sense for us to take a step back and talk more generally about data, what is data and most critically, what are the different types of data that you will encounter.

And it is important to do this, because the descriptive statistics that you use in the approaches that you use vary according to the different types of variables or different types of data that you will encounter and this is recurring theme in many other aspects of this course. So, make sense for us to talk about that for a few minutes right now. So, data

is essentially numbers, now it can also be texts, symbols, but more often the not, you will encounter numbers, which represents some kind of information.

And therefore, it is a kind of helps to think of data as values, because values is broad enough to cover numbers, text, symbols. So, you can think of data as values of quantitative and qualitative variables. Now, the variables themselves can be of different types and that is, what we are going to talk about right now. In statistics, you have various classifications of variables and so different depending on, who you ask, you find different classifications in different text books as well.

But, one broad classification that make sense and that is very useful for us is to differentiate variables as quantitative and qualitative variables. So, quantitative variables are also called as numerical variables and these variables are essentially, you know the best way to think of them is that they have meaning as a measurement, such as a person's height or weight or IQ or they can be some kind of count, such as the number of something number of days it is rained and so on and so forth.

But, a very intuitive way for me to, that is always been useful for me is to think of quantitative variables as variables, where some form of basic arithmetic like either adding or averaging or subtracting kind of make sense. So, I mean a definite requirement is that the quantitative variable is numerical, but some time you can also have numbers being used as symbols not as the actual numbers. So, a really simple rule that kind of helps me identify quantitative variables is to say that it has to be a numerical variable, that the values that the variables takes on are numerical and that, some basic forms of arithmetic kind of make sense on such a variables.

Now, within the quantitative variables you have continuous and discrete variables. Continuous variables are essentially one square within a certain interval and this is an interval that the variable could, where the variables could take on values. Within this interval any value is possible, if any value within this interval is possible, then this variable is said to be a continuous variable. So, let me give one example. Let us say that the variable we are interested in is the height of students, who are registered for this course.

And so, let us say I go take a sample of a 1000 students and write down and their heights and so this is a data set that I have. The data set have 1000 values and let say an interval

and the interval can, you can may be get the interval from taking the highest value to the lowest value or you can just, as long as it is a real interval that covers this data. The question is, is any value possible within this interval and the answer is yes.

So, let say my interval was 120 centimeter to 140 centimeters, is it possible that someone, who taken this course could have a height of 135 centimeters. Absolutely, there is nothing about heights that inherently stops people from having that particular height. Now, I can go even further, I can say is it possible for someone to have the height of a 135.156 centimeters and the answer is still yes. I mean I might not have a measuring scale that goes to a certain accuracy, but that is the measurement problem. There is nothing about heights that prohibits this value from existing in this data set.

So, continuous data is essentially one, where any value between certain interval and the interval is sometimes formally defined as the highest value in the data set to the lowest value. So, any value within this interval is potentially possible, then you have a continuous data set. The second kind is the discrete, so the discrete quantitative data is one, where this condition is not true essentially and it again helps to think of, what kind of a data set would be such that a value not all possible values are there and I will give you another example for that.

So, let say that I was interested in looking at the number of people, who enter IIT Madras every day, so the number of people, who… Let us make it interesting, the number of unique people, who enter IIT Madras every day. So, if you come in and go out, come in go out and times, I do not care, you are still one person. So, number of unique people, who enter IIT Madras on a given day is my variable and the actual values of this variable I get from doing this kind of a survey or a study for 1 year.

So, I have 365 data points, one data point for each day, which says the number of people who enter IIT Madras. Now, clearly this variable is discrete, because let say there is a lower bond, which is may be 0 people, nobody enter the IIT Madras highly unlikely, but on a given days. So, zero is the lower bond and the upper bond is some, you know 10,0000, 50,000 something. Now, within this range, can I have is every value possible, no. You could for instance never have a day, where two and half people entered or let say, you know 133.2 people entered.

So, here is discrete, because it is discrete in the sense that only the integer values are

possible, all the more; no, where you can never have values that are non integers. So, that is an example of a discrete value and in this particular case, it happens to be one of being discrete at the integers, but that is just, because of the example I came up. You can come up with the other examples, where the variable is numerical and it is I mean its quantitative, but the values you can take up are discrete.

We move on to the other class of variables, which are qualitative variables, these are also known as categorical variables and categorical variables essentially represents some characteristics, some characteristics, which can be categorized, which can be grouped. So, examples of that are things like a person gender, so that is the variable and the values of variable can take a male and female.

And then, you might have something like marital status or more interesting, one might be home town of or state within India. Let say, let us take all the people who registered for this course from within India, which should be bulk of them. And the variable we are interested in is, which state are you from, so that is the variable the state that you are from and the values that this variable can take up are the different states of India.

Now, the categorical variables again, because of their definition and their nature are always discrete, so that should be obvious. But, within these categorical discrete variables there are two classes again, there is nominal and the ordinal. With nominal, there is essentially… The big difference is that, with nominal there is no order. So, the great example of that would be this home state, which state are you from, variable. There is no order, which says that Madhya Pradesh is greater than or lesser than another state and so on and so forth.

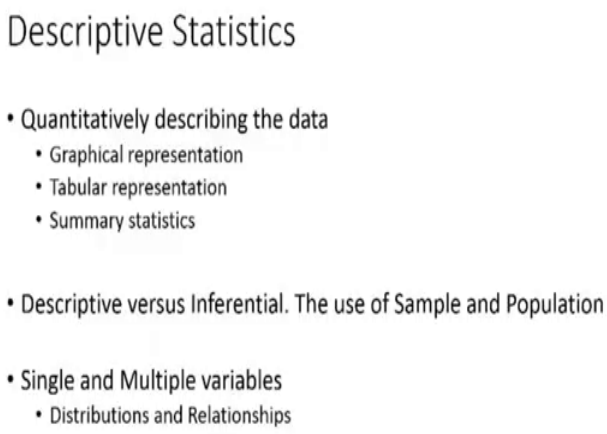
So, these are all, the values that these variables can take up cannot be ordered in a sensible way you know, whereas with ordinal data by definition of the variable, there is an order. Let me give an example of that, let say that we created a variable for, which is the color for terror alert, so some countries have this, the terror alert color signify something. And let say the possible values this can take is green, yellow, orange and red.

So, the variable stated terror alert color green, yellow, orange and red are the four values that this particular variable can take, where green represents low risk and red represents very high risk. See, so there again it is a categorical variables, because it is not like you can do arithmetic operation on green, yellow, red. The variable itself is a qualitative

variable, but yet there is some order.

Because, you can say things like if orange is worst than yellow and yellow is worst than green, that must mean orange is worst than green. So, you can get an idea also, that is an ordered categorical variables, whereas with a nominal variable you could not say if Madhya Pradesh is greater… First of all, you could never say greater or less than, so creating more complex relationships becomes impossible. So, that is just to give you a very quick idea about the different variable types.

(Refer Slide Time: 11:57)



So, now, let us jump into descriptive statistics. So, descriptive statistics is the idea of quantitatively describing data and you can do that through various means, you can do that through visualization techniques like graphical representation, tabular representation, but you can also do that through summary statistics. The idea here is that, you crunch the data, you work with the data and come up with 1 or 2 or 3 or 4 different numbers that summarized the data for you.

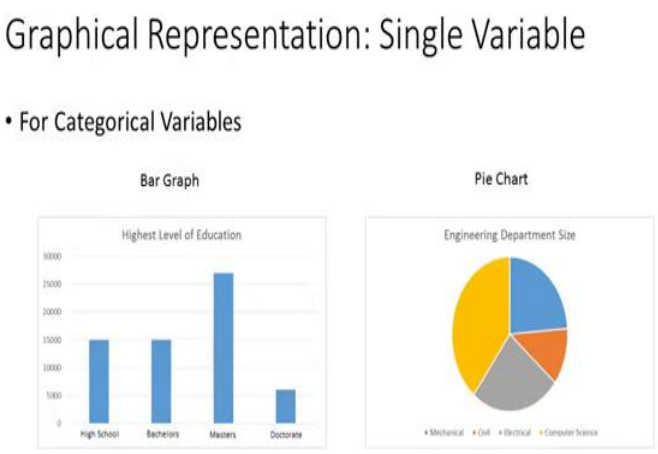
In this class we are going to be focusing more on the graphical and the tabular representation and the next module is going to be on the summary statistics, so that is the idea. Now, this is a very good time for us to just quickly review, you know in our overview classes we spoke about descriptive versus inferential statistics and this is the good point to just bring that up again and to kind of have a very quick idea, what descriptive statistics are really means.

The core idea in this dichotomy is that descriptive statistics focus or is the way to say something meaningful for the data that you have at hand. So, you have some data at hand, whether you call it sample of population or whatever, if you are making the statement based of that data about that data derived from that data, they are dealing with descriptive statistics. Descriptive statistics do not; however, allow us to make conclusions beyond the data we have.

So, you cannot look at the data, do something with the data and make and based of that make the generalization about potentially the source the data that, the data came from, you would need inferential statistics for that. Now, having said this; however, descriptive statistics is still very important, because you cannot just simply present raw data, it would be very hard to visualize, especially when the data is a lot. When you have a lot of data, you cannot just show the data, you need to present the data in a more meaningful way, which allows for some kind of simpler interpretation and that can be through the graph or through numbers, great.

So, and a final point I just want to make is that, descriptive statistics is not just confined to a single variable, it can be about multiple variables and when you are dealing with multiple variables, our topic of interest is relationships. So, how does one variables change with respect to another variable. So, in essence you will be doing two things which is summarizing each variable or describing each variable, but you are also interested in showing interrelationship between variables.

(Refer Slide Time: 15:09)



So, let us go ahead and now, that we have an understanding of different variable types, let us talk about some graphical representation techniques. If one is dealing with a single variable and let us say it is a categorical variable, a great way of representing data could be through a bar graph. So, that is the graph that you see on your left hand side here. So, here for instance let us say the example is one, where we sent out a survey and ask people, what their highest level of education is and highest level of education be the variable, the possible values that takes up our high school, bachelor's, master's and doctorate.

Hence, therefore, this is a categorical variable, there are only four possible qualitative states, that this particular variable can take up. And, what we plotting is the number that we, number of responses or the number of observations we counted in having this values. So, you sent the survey out let say it about 50000 people, may be 60000 from the local way. So, and about 15000 of them said that their highest level of education was high school.

So, the height represents the frequency about of occurrences of this particular value of this variable. So, this kind of a representation can quickly summarize, which is more which is less and so on. But, an interesting thing to note out here is that this categorical variable is actually ordinal, meaning there is an order of going high school, bachelors, masters, doctorate. You could have flip the whole graph around, but you would still have the order that is a sense that a doctorate is a more years, I guess than masters and which is more than a bachelors, which is more than a high school or whatever.

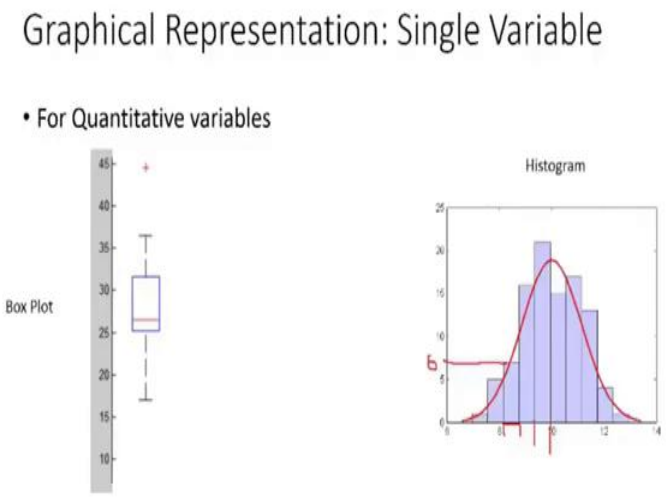
So, in some sense the variable itself has an intensive ((Refer Time: 17:19)) and the good thing is something like a bar graph, I loves for that. Just, because of the fact that there is this concept of a x axis, makes it very convenient to represent ordinal variables, which are categorical. Another way of representing categorical variables could be something like a pie chart, this is an example, where let say we looked at the number of students, who were in different engineering departments.

And your different engineering departments here are mechanical, civil, electrical and computer science. These are just some random departments I chose and again the frequency of occurrence is more represented as a percentage of the whole. So, this percentage of this full circle is computer science students and thus the idea behind using

the pie chart. And clearly a pie chart is not very suited for ordinal variables, which is more suited for nominal variables.

Because, there is no order, one that the fact that computer science shares the wall with mechanical and civil is just coincidental, that is not what a pie chart seeks to capture. Sometimes people will keep similar things together, but that is not a requirement. One important thing about pie charts is that, usually you want to represent all the values. So, if there are some engineering departments that are not being represented, usually a pie chart need not be the best way, because there is an impression that this is all the departments together. So, if there are more engineering departments, but you wanted to only show 3 or 4, may be you could use a bar graph rather than a pie chart.

(Refer Slide Time: 19:18)



Now, we move on to quantitative variables and with quantitative variables, you have a couple of different ways of representing a variable. One example is a box plot. So, with quantitative variables, remember that is numerical data and, so you might be interested in representing things like, what is the average, what is the variance and in our class on summary statistics, we are going to go to a great detail about it.

But, for purposes of this, a box plot is essentially something that captures central tendency, which is this red line that is there in, typically that tends to be the median of the data set. And you have the two bounce of the data set, so the top and the bottom of the box itself and that tense to capture in some way the variability in the data and the

way a box plot does that by representing the lower quartile and the upper quartile.

Now, the lower quartile and the upper quartile in really simple words is just 25th percentile and the 75th percentile and, so that kind of that range a gets captured there and the whiskers themselves take on different meanings depending on the, which version of box plot is using, but more often they are not it tends to be lowest value and the highest value in the data set. And typically red dots like this represent outliers in the data.

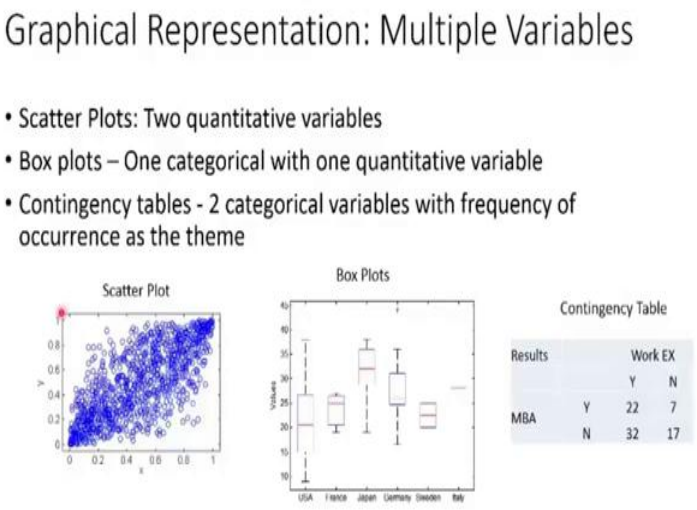
The box plot is itself something that will make more sense to you and we will talk about summary statistics, because you will understand, what exactly a median means you will understand, what a different way of representing spread an outliers and so on. But, it helps you at this stage to kind of say that this is one simple way to take a data set, which is full of numbers. So, let say this data set had you know 500 or 1000 numbers and it looks like this numbers are pretty much within the range of like 25 to 33 or, so and to represent all of these numbers in a single graphical representation, so great. Another way of representing quantitative data is through a histogram the histogram is what you see on the right hand side and histogram his arguably ah the richest representation of numerical quantitative data, because histogram essentially says how many data points do you have with in this range.

So, the x axis out here represents the different possible values that this data set has. So, if you take this as 8 to 10 this should be something like 8 to 8.66 and this should be till like 9.33 and this should possibly be 10. So, the question is in your data set how many data points do you have that have values greater than 8 and values less out here. So, another way of showing that I am going to try highlight it is, so here interested in this range right here this range. So, this is 8 and this is 8, let say 8.66 is from reading the graph right, how many data points do you have.

Because, this is a numerical quantitative data set that have values greater than 8 and less than 8.66 and the answer to that question is it looks like 6 data points right approximately may be 7 may be if I am reading graph correct. So, you answer that question for each of this bins this are all called bins each of this columns are called bins for each of this column you answer that questions and what you get is histogram and the histogram is the first step towards empirically constructing, what you will we will later learn is it distribution.

So, once you capture this, this entire picture out here you have a very clear representation of the entire data set. So, again just keep in mind that we are going to be talking about distribution we are going to talking about medians means and variances, but keep in mind also that this is the graphical way of representing these things.

(Refer Slide Time: 24:17)



So, now we move on to the multiples variables and the last section in graphical representation and there are three major of forms of representing ha this data and they are the following. The first is scatter plots this are very useful when you have two quantitative variables that is you know. So, two variables, which are both numerical you can you can very easily represent using scatter plots and, so in the key thing you should notice in this scatter plots is that it really helps you understand the relationship it does not do a very good job of understanding each individuals variable.

So, may be if you done distribution of x and distribution of y separately you would understood those two variables, but what it does a good job is of capturing the relationship between x and y. And this particular case the fact that in general when x increases it look like y also increased or y also high or you know you can always flip it the other way around in general when y is high x is a high when y is low x is also low.

So, that relationship gets captured for that reason this is the great graphical representation of two variables usually you can extend box plots if you feel like one of your variables is categorical and the other is quantitative. So, you are not just interested

you are interested, let say one variable, which is country and the other variable, which is some indicator let us say of the economy or crime or whatever I have just called it values here because it does not matter.

But, this variable is continuous right its mean I should not say quantitative I do not know if it is continuous it could be continuous or discreet, but this variable is quantitative, where as this variable is qualitative. So, one great way of look comparing different qualitative variables, which have data set that are on the quantitative scale is to perhaps use multiples box plots on the same graph and that gives you not just an idea of how on average his country is different, but there are also different in terms of their variability and their out so on, so forth.

Finally, we come to the use of contingency table out here on the extremely right and the idea here is that when you have two categorical variables and what you are interested in representing is the frequency of occurrence. So, the frequency of occurrence is the theme of the data set. Then, contingency tables are great ways to do that, so an example that I have come up with out here is how many people let say you go to a company and you take a survey of all the managers working in the company may be interested in asking how many of them have MBA.

So, y represents yes they have an MBA and represents no clearly this is categorical variables right it has only two states. Similarly, you could say before this people join the company did they have work experiences before they joined as managers and answer could be again be yes or no, so that is also categorical variables. So, here is an example why you have two categorical variables and what you are interested in his how many people belong to each combination.

So, how many people have MBA's and had work experience before joining how many people had MBA's did not have work experience before joining. So, this can be complex data set where it very neatly summarized in the contingency tables and that is something that could be quiet useful. So, I think that is about it for graphical approaches to ha representing data in the modules descriptive statistics and in the next class we will be looking more at ha summary statistics as a means of as the sub modules in descriptive statistics.

And then, we will move on to inferential statistics great, thank you for joining me and look forward to seeing you in the next lecture.