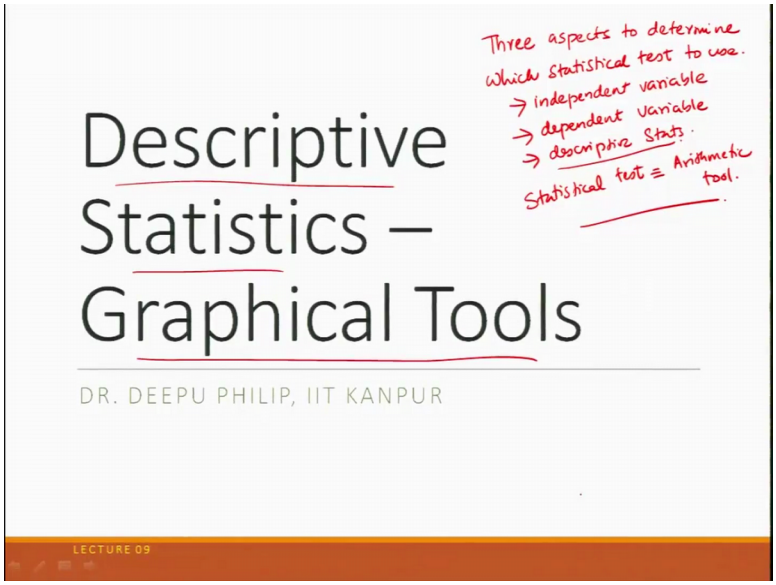
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

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

Lecture – 09
Descriptive Statistics--Graphical Tools (Private)

Good afternoon ladies and gentlemen. Today, again we are getting into the new lecture concept of the descriptive statistics and specifically the title is descriptive statistics and we are looking at the graphical tools ok.

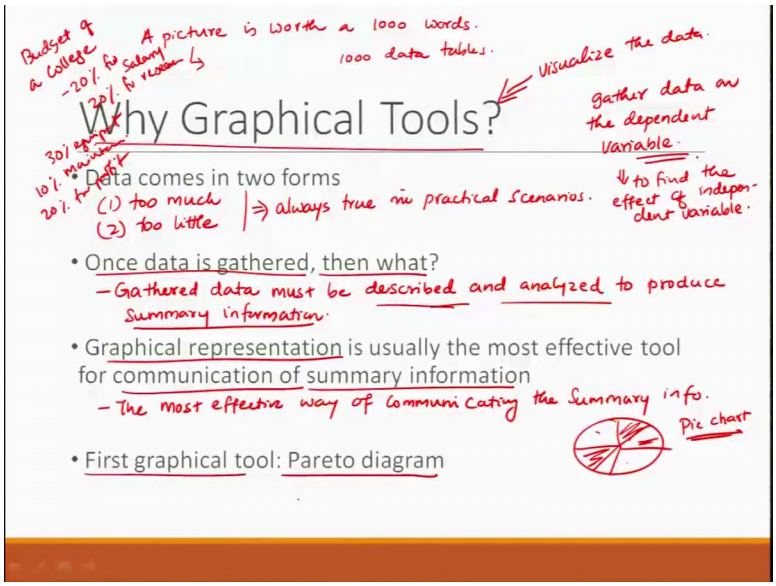
(Refer Slide Time: 00:33)



And remember, in the previous lecture, we have already talked about three aspects to determine which statistical test to use. So, three one number one was your independent variable ok, number two was your dependent variable and number three was your descriptive statistics and when we say statistical test think about it as an arithmetic tool you can think about it does not arithmetic tool for the time being ok.

So, what we are trying to do today in this class is? We will try to see a little bit of descriptive stats and one of the simplest tools that we are going to see in this case.

(Refer Slide Time: 01:45)



So, this is the ninth lecture of your this course and we will start by looking at why graphical tools; obviously, someone wise person once said a picture is worth a 1000 words, this is very true, especially when it comes in the statement of statistics, I could probably say that I can modify this and say that a picture is worth 1000 data tables or something like that or data values ok. So, graphical tools really help you to know help you to visualize the data and I said earlier data comes in two forms; number one too much of data, number two too little of data. This is very true when it comes in the practitioners case ok, this is always true in practical scenarios, you will never get the right set of data you have either too much of it or you are too little of it.

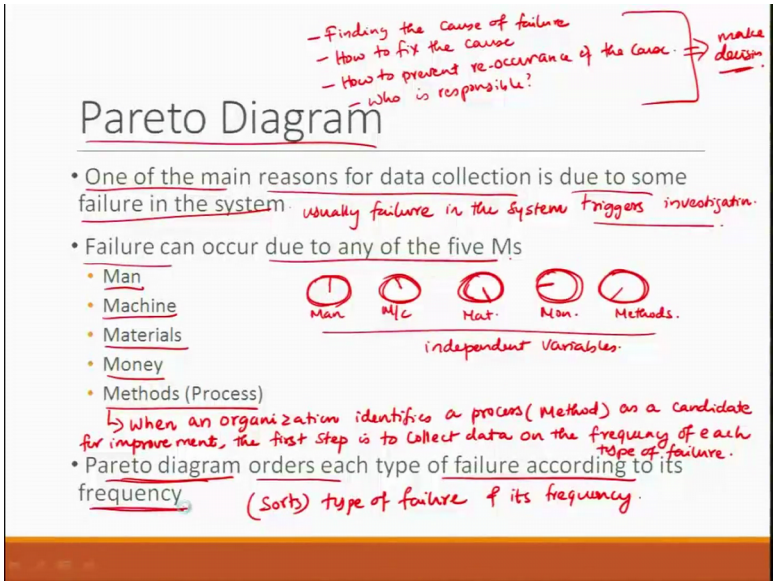
So, then you have to play around with it ok. So, once you gather the data and as I said earlier you can you gather data on the dependent variable why because by collecting this data you are trying to find the to find the effect of independent variable ok. So, due to that; so once you collect the data, then what then what do you do ok. So, gathered data must be described and analyzed gathered data must be described and analyzed to produce summary information. So, once you collect the data and then what you have to do it you have to must be described by an annualized, you should describe the data and analyze the data through which you will under producing what I call as the summary information.

So, once your summary information, then what then the best way to display the summary information or best way to communicate the summary information is using a graphical

representation ok. So, the most effective way of communicating the summary information, this is what we call as a graphical representation. So, we know if you say that the total budget of a. So, let us consider that budget of college ok, let us say this is if somebody says the total budget is divided into let us say 20 percent for salary 20 percent for research 30 percent for equipment ah. So, the and then 10 percent for maintenance, something like this and then the rest of the 20 percent for profit.

Let us say you think about a scenario like this, then displaying this in a pie chart would actually this is like 20 and then you have like 30 and then 10 and 20 something like this. So, you basically say that fine, these are the equal portions of the budget and these are the two large area. So, you can think about this as a pie chart which is an easy way of displaying this information to somebody which tells you the relative budget spending on each other thing rather than looking at a extensive list like this. So, in this class, what we are going to do is the first graphical tool we are going to study today here is called as the Pareto diagram. So, the Pareto diagram is an important tool and the Pareto diagram has quite a lot of aspects are related to it. So, we will talk about what it is ok.

(Refer Slide Time: 07:01)



Ah, so, before getting into the Pareto diagram let us think about why organizations collecting data one of the main reasons for data collection is due to some failure in the system usually failure in the system triggers investigation, what type of investigation

finding the cause of the failure finding the cause of failure ok, then how to fix the cause, then how to prevent reoccurrence of the cause who is responsible.

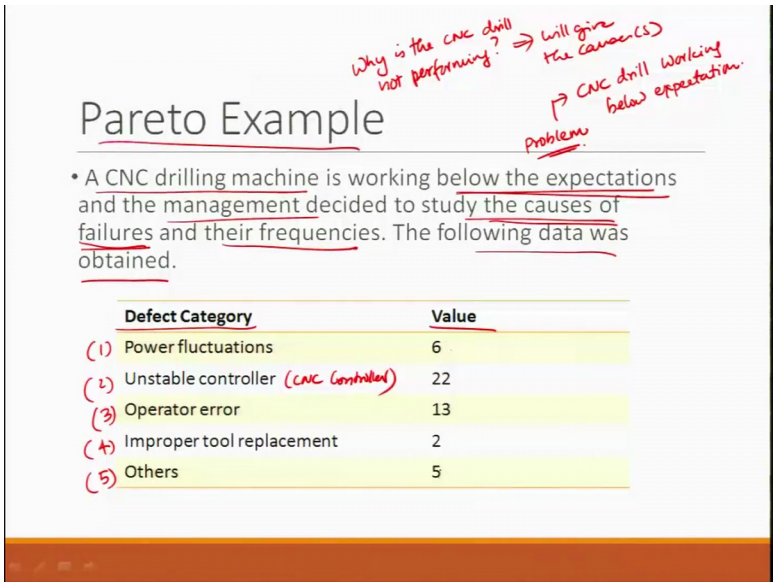
So, these kind of things these aspects are the ones that triggers the investigation this is kind of the triggered investigation and the outcomes of the investigations from there you use this to make decisions decision on how to sort out the problem most of the time failure in the organization occurred due to any of the 5 ms the first one is man or human beings machine your resources materials raw materials other things money or a body calls finance and method is what we call as the process ok.

You can think about this and any organization as a set of 5 dials ok, think about this was the man dial, then is the machine dial or a like kind of a dial, what am I say it is a regulator kind of a thing, then there is a materials dial then you have a money dial and then you have a methods or a process there something like this.

Yeah, so, this is your materials money and methods, if you have 5 dials, like these you can think about changing these dials changing the dials in the system to change the performance of it. So, these could be your independent variables and that results in that would cause some failure in the system which you measure it in some other case. So, when an organization when an organization identifies a process or a method as a candidate for improvement.

Why is it a candidate for improvement because it has caused the failure, then the first step first step is to collect data on the frequency of failures frequency of each type of failures or each type of failure. So, what does Pareto diagram do once you collect the data on each type of failure then the Pareto diagram orders each type of the failure according to its frequency orders in a sense it sorts into each type of the frequency and the each type of the failure and its frequency? So, the type of failure and its frequency is being considered as part of when the Pareto diagram is the one that orders each type of failure according to its frequency and then you can visualize how the what are the major causes of failure and how is it impacting that ok.

(Refer Slide Time: 11:47)



So, there is a principle behind Pareto, but before looking into that let us look into this Pareto example ok.

So, here is a trivial example a simple example that we can work in the class itself we do not need excel or anything to do this, but you can use excel if you want to, but the a pen and paper would do assume that your factory has a CNC drilling machine and it is working below the expectations ok. So, this is what the problem is what is a problem CNC drill working below expectation this is a failure in a sense it is a failure because you are not able to produce the amount of product that you want to produce ok. So, it is working below the expectations and the management wants to study the cause of it why is it why is the question is why is the CNC drill not performing this is the question ok.

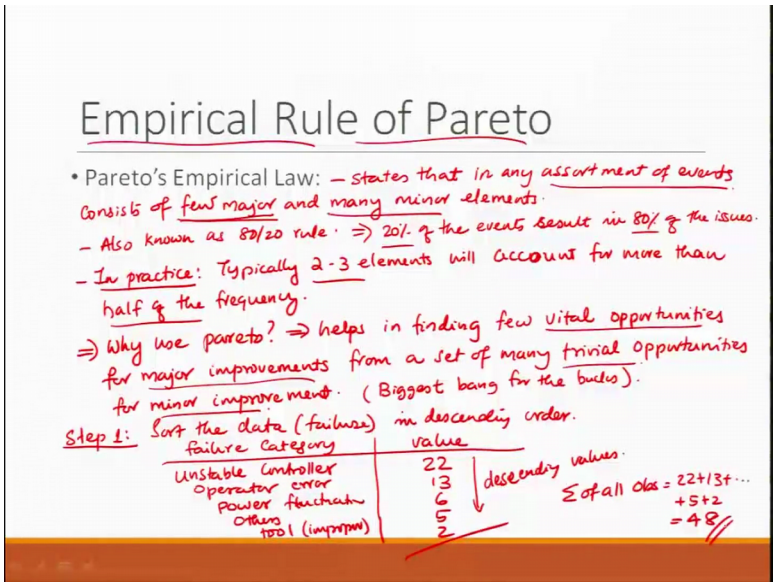
So, that is the cause this gives the answer to this question will give the cause or the causes there could be multiple cause ok. So, what the management did is it wanted to look into all set of failures and their frequencies and they interested the workers and the workers collected the data the data is like this there is a defect category and then the value. So, number one of the defect category is power fluctuations in the power to the machine and it resulted in 6 failures then the number two is the unstable controller the controller of the CNC.

So, this controller means the CNC controller the controller is unstable it just resulted in 22 failures then the operator the person who is operating the a machine because of the

operator error the mistakes committed by the operator it resulted in 13 failures then not being able to replace the tool in the proper time has resulted in 2 failures and other things which could be like failure of coolant not sufficient coolant break cage of tools some unexpected things or which are bunched resulted in 5 failures.

So, this is the data that is given to us and using this we will learn how to make a Pareto diagram.

(Refer Slide Time: 14:18)



So, one of the first things is the empirical rule of Pareto and there is an importance with this and the parameters empirical law that what does it stage, this law states that in any assortment in any assortment of events consists of few major and many minor elements.

So, if you start you take any assortment of events ok, take a lot of failure events or whatever, it is it will consist of few major events and many minor elements ok. So, the major will be few and the minor will be large in number it is also known as 80-20 rule which means 20 percent of or 20 percent of the events result in 80 percent of the issues or majority of the issues are created by minority of the few major events that is what the Pareto rule is ok.

So, in practice in practice because we have practitioners in practice typically 2 to 3 elements, 2 to 3 elements will account for account for more than half of the frequency more than half of the frequency ok. So, what we are saying here is we are not really

looking at the 80-20 perfectly as such we will say that typically 2 to 3 elements in the assortment of the events will account for more than half of the frequency. So, let us see how this actually works out.

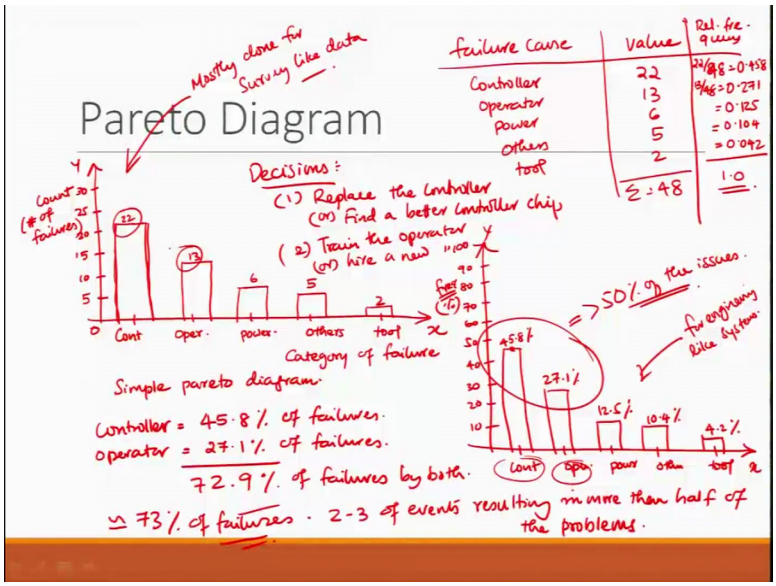
So, why use Pareto? What is the importance of using this Pareto? Because it helps in finding helps in finding few vital opportunities; for major improvements from set of many trivial opportunities trivial opportunities for minor improvement. So, what we are saying here is you how it will help us to identify few vital opportunities that results in major improvement than from a many trivial opportunities that will result only in minor improvement

So, what we are looking for is the biggest bang for the buck. So, what does that mean is; what is the best for maximum returns, that what we are trying to do. So, in this case what we are supposed to do is the first thing we need to do we have to look into the data that we talked about. So, we had 5 of these data values and the first thing for us to do it is the step want to create the Pareto diagram is to sort the data in the descending order ok. So, what we do is step one sort the data what data failure data; data in descending order ok. So, what we do here is. So, let us make a table out of this ok. So, here is the what we call as the failure category and we have the value. So, if we go back we will actually see that the largest descending order means the largest value first.

So, this is the largest ok. So, then what we do is we go back and we say unstable controller it is 22 that is the largest value then the next largest value was the operator error which is 13. So, we will say now it does operator which is 13; similarly we have power failure which is our power fluctuations operator it was operator error power fluctuations it resulted in 6 issues then others resulted in 5 and then tool improper tool it resulted in two failures. So, this is the descending sort. So, we can say descending value if you sum the total thing it is 22 plus 13 plus 6 plus 5 plus 2. So, summation sum of all observations is equal to 22 plus 13 plus all the way up to 5 plus 2 that is equal to 48 ok.

So, there are 48 total defects that were identified now with this the next thing is how do you do the Pareto diagram.

(Refer Slide Time: 20:58)



So, the Pareto diagram is a graph is x axis and the y axis and the x axis you have each of those defect categories being put here. So, we have 5 defect categories 3, 4 and 5. So, the first one is the controller second one is the operator the third one was the power fourth one was others, then the fifth one was tool. So, each one of the category. So, the category of failure goes into the x axis and the y axis is you can think about it does the count or number of failures that is that goes in your y axis.

So, it starts at the 0 and let us think about it does the maximum value is 25. So, we will say 5, 10, 15, 20, 25, 30, it goes like this. So, then the first observation of the controller failures was up to 22. So, we assume this is up to 22. So, we draw a rectangle here we say this is this value is 22 ok. So, the 22 failures are accounted right here in this one, then the next one is the operator which resulted in 13. So, somewhere here we draw a line and then we draw a rectangle and this is the 13 up of the operator failure then we had a power failure which were such sakes. So, 5, 6 is slightly bigger.

So, this is 6 then others was 5. So, slightly lower than this and the last one was the tool which was 2 of them. So, it is a small rectangle. So, this diagram typically will say that this is the this is the simple Pareto diagram ok, now in this case you remember we had a table and the table had failure cause value. So, we had a controller has 22, then operator has 13 power has 6 others has 5 and tool has 2 and we summed all of this sigma was equal to 48. So, if you look into this ok, then the relative frequency. So, here it will be 22

divided by 48 which will give you 0.45 eight and this will be 13 by 48 which will give you, 0.271 you can do the math by yourself the this value 6 by 48 will give you what you call as 0.125 the 5 by will give you 0.104 and this will give you 0.042 if you sum all of these the values will come to 1.

So, the idea here is that if you look at this then you can think about this particular fashion you can this is one way to do there is the count the other way to do the Pareto is also this fashion if you assume it as the accumulated value you can think about does the count 1, 2, 3, 4 and 5 same thing your controller operator power others and tool as the 5 factors and you can think about it as your the individual observations going right here 10, 20, 30, 40, 50, 60, 70, 80, 90 and maybe somewhere here is the 100, let us put it this way and so, the controller is 45 or 0.45 weight which means 45.8. So, somewhere here will be the controllers ok.

So, this is like the 45.8 percent, then the operator is 27.1 percent. So, that is someplace here is the operator which is 27.1 percentages then power is 12.5 percentages. So, somewhere here is the 12.5 percentage the others is the 10.4 percentage. So, it is like little bit less 10.4 percentage and the tool is the 4.2 percentage ok. So, something like this 4.2 percentage. So, now, if you look in this case ok. So, the this is the 45.8 percent, but if you look at the second one operator ok. So, the controller contributes to 45.8 percentage of failures and then operator contributes to 27.1 percent of failures.

So, if you do both of these, it will be 9 12; 72.9 percentage of failures by both ok. So, as we studied or as we looked in the Pareto principle what did we say 2 to 3 of events resulting in more than half of the problems ok? So, same way if you look into this you can say that these two guys put together results in greater than 50 percent of the issues. So, for you or for a decision maker, then it will say that the decision will be from here ways. So, what are what will be the decisions the decisions will be number one replace the controller the controller chip or whatever it is or find a better controller better controller chip number two trained operator that could be one decision or hire a new operator ok.

So, these kind of things by looking into this you can say that by solving these two cases by dealing with the controller and by dealing with the operator about 72 point or 73 percent approximately 73 percent of the failures or failures can be addressed by dealing

with this ok. So, this kind of a approach where we can either look in the count. So, here you are doing the count and here you are looking at the frequency or the percentage of the percentage contribution to the failure ok, I can think about as a frequency or you are called as relative frequency or you want to call it as a percentage of failures whatever it is. So, that y axis will have that and the x axis will have the category of failures. So, these two these are the two ways you can do the Pareto diagram for most of the this is done for what you call us typically mostly done for survey like data where you use the count and mostly done for engineering like systems ok.

So, in a hotel you might want to find out how many people actually do come. So, this count might matter to them whereas, in an engineering like system the percentages might matter the absolute count might not be of much importance. So, depending upon how the systems turn out you end up making the appropriate choice of the making the Pareto diagram and such a Pareto diagram can be easily made with the help of excel, you do not need any complicated tools to make these kind of Pareto diagram even in that case in the worst case scenario simple pen and a paper and a calculator will do the trick also. So, I hope that you guys understand the concept of Pareto diagram which helps us to identify the vital the few vital causes that actually results in majority of the problem and addressing them rather than focusing on the minor ones that actually give very minimal returns. So, this is where the focuses is on the biggest bang on the buck and then identify how to deal with it.

Thank you.