

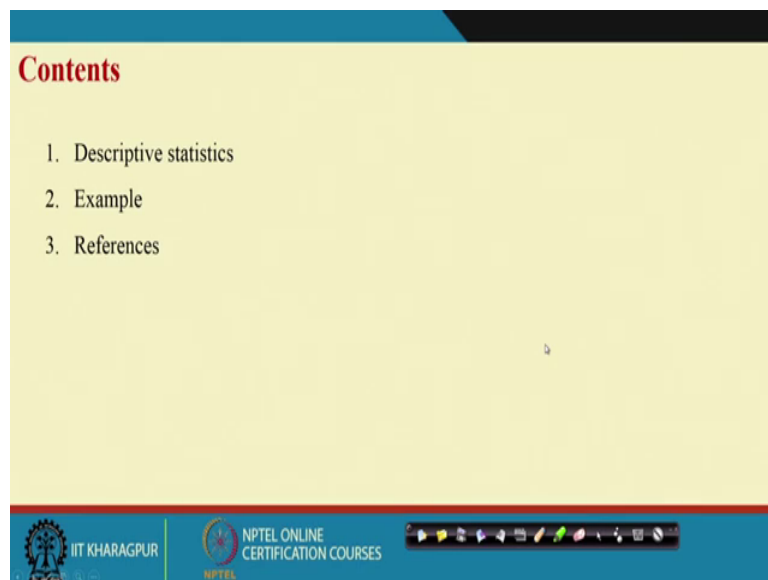
Industrial Safety Engineering
Prof. Jhareswar Maiti
Department of Industrial and Systems Engineering
Indian Institute of Technology, Kharagpur

Lecture – 47
Accident Investigation & Analysis: Descriptive Analytics

Hello everybody welcome to this lecture. Last lecture you have seen Accident Investigation and also what kind of analysis is require. We also shown you that, what are the different information that will be should be collected and in today's lecture what I will do just I will describe the Descriptive Analytics part assuming that you have some kind of data available with you and those data are basically come from the accident in or incident investigation reports or records ok.

So, in this lecture we will see descriptive analytics or other way you can say descriptive analysis also, but today's word is analytics driven whenever you have data you must try to do analytics so, that you will get the best benefit out of it ok.

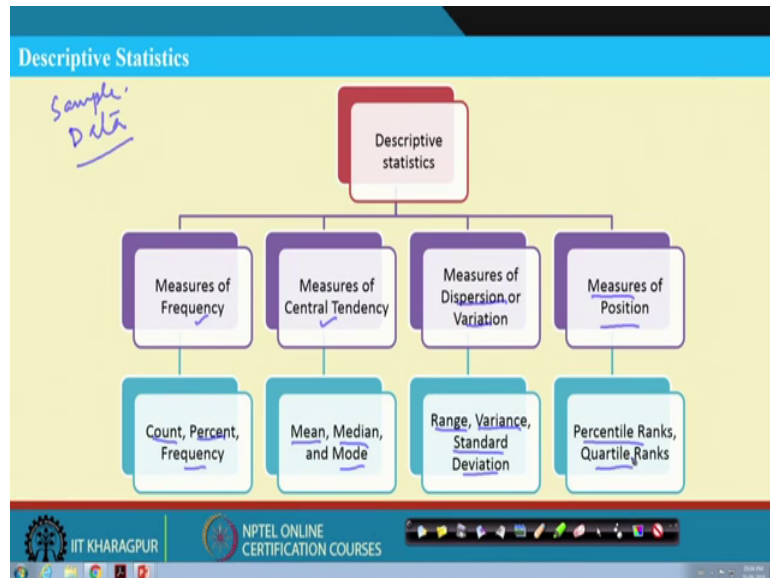
(Refer Slide Time: 01:17)



So, what are the things we will see today? We will see first the descriptive statics and then one example and finally, all those descriptive statistics, we will try to relate to the data and give some of the insights to you and which will ultimately help you in

describing the accidents or incident or safety status of a plant for which you are doing this analysis.

(Refer Slide Time: 01:52)



So, when you talk about a statistics, basically statistics means when you have data you have collected lot of data is there. So, this data are usually we say that sample data, which is collected from a population or from a process or a system, then this data can be used to infer or to quantify or to measure some of the basic properties of the population or the process of the system so; that means, what happened that descriptive statistics include that the summarizing the data, then your chatting the data, and then getting location and dispersions measures from the data, and this basically actually describe the what is happening in the system or the process or a from statistics point of view in the population.

So, there will be different kinds of measures; one is measures of frequency then measures of central tendency, measures of dispersion or variation, measure of measures of position under measure of frequency you measure, count, percent, frequency under central tendency you measure mean, median and mode and under measure of dispersion you measure range, variance and standard deviation and under measure of position you basically measure the percentile ranks, quartile ranks and something like this.

So, I will not be discussing all those things, but for your benefits some of the things will be discussed today and rest of the things you please do yourself study.

(Refer Slide Time: 03:50)

Example - 1

A company is experiencing serious accidents over the last five years. They have been collecting records of each of the accidents occurred in the plant. The data consisting of 500 records are retrieved from the company for analysis with an aim to provide descriptive statistics of the data. The attributes used in the dataset are provided below.

Attributes	Types of attributes	Category	Ranges
Month	Categorical	12	-
Day	Categorical	7	-
Location	Categorical	5	-
Incident Events	Categorical	5	-
Working conditions	Categorical	2	-
Machine conditions	Categorical	3	-
Observation types	Categorical	4	-
Incident types	Categorical	3	-
Employee types	Categorical	2	-
Time-shift	Categorical	3	-
Gender	Categorical	2	-
Working temperature	Numerical	-	30.1-44.99
Heart rate	Numerical	-	60.27-109.95
Blood pressure (Systolic)	Numerical	-	90.01-189.49
Health rating	Numerical	-	1.01-9.99
Incident outcomes	Categorical	3	-
Risk	Categorical	3	-

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us consider a data set, a company is experiencing serious accident over the last five years. They have been collecting records of each of the accidents occurred in the plant. The data consists of 500 records are retrieved from the company for analysis with the aim of to provide descriptive statistics of the data. The attributes used in the discrete below.

So, please remember that means, investigations have been made whenever incident has taken place and led in this fashion the report is available, report can content many other things, but for the time being we have created hypothetical datasheet and that hypothetical data sheet is like this. So, that the incident every incident report has attributes like month, day, location, incident event, working condition, machine condition, observation type, incident type, employee shift, gender, working temperature, heart rate, blood pressure, health rating, incident outcomes risks, so, many things are there.

So, you see some of the are basically timestamp, some of them are location stamp, some of them are event related detailing, some of them are basically related to the particular big team, some of them are related to the physical process say physical environment like temperature and also there are some outcomes related to incidental, incident may be fatality property damage and something like this. So, most of the things are categorically

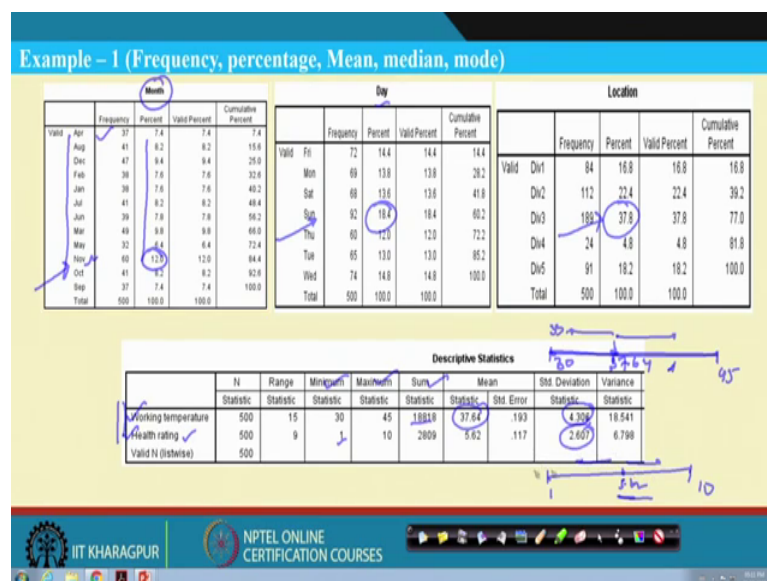
nature. Categorical means they are basically just giving you some kind of identity like day; day has 7 days Sunday, Monday, Tuesday to Saturday.

So, they will from Sunday and Monday they gives you the identity of the day only nothing more nothing less ok. Similarly location they will tell you which location whether it is in the plant on the road, in the subfloor or within the subfloor whether in section 1, section 2, section 3 like this; so, that since they are all categorical in nature. Now temperature is numerical, heart rate is also numerical. blood pressure numerical.

So, working temperature related to the physical environment, heart rate and blood pressure related to the victim and health rating also related to the victim that is also numerical, and incident outcomes basically what happened to the victim and risk means basically its basically the that event we have seen the probability of happening that event and times the severity. In some scale it is measure, this is also categorical in nature.

So, how many categories are there? In month 12 categories day 7 categories like different categories are there and for the numerical variables you will get range values that from which value to minimum to maximum this value will get ok. Suppose you have collected that mean what happen? These are the attributes you have collected from the accident reports, there are 500 such records so, you have good amount of data set. Now, we want to do some of descriptive analytics here or descriptive statistics here.

(Refer Slide Time: 07:15)



So, here first you see that month wise, that in different months, what is the frequency, the percentage? Now if you see the frequency and percentage you find out that this percentage wise this is the 12 percent is in November, otherwise more or less they are similar ok. Now if you go by day I think the day wise also percentage if you say so, from these data 18.4 is little more other why others are, so, almost equal. Similarly when we go by division, division wise you see that division 3 is having percentage this much percentage 37.8 percentage, but otherwise others are almost similar. Now these are basically from month, day, location this categorical data where different categories and their frequency and percentage is calculated you can calculated in this manner also. So, you can use a software and get it, but when we are talking about some like working temperature, health rating then they are numerical variables.

So, in this case you are able to find out the minimum value, maximum value they are sum and total 500 sum, then their mean value, the standard deviation value ok. So, working temperature if I want to analyze what is happening? Working temperature minimum value is 30. So, this is 30 and maximum value is 45 and then your mean is 37 point somewhere here 37.64 standard deviation is 4.63. So, that mean if you if you see the data is basically running from 30 to 45 with mean value here and standard deviation here 4.36.

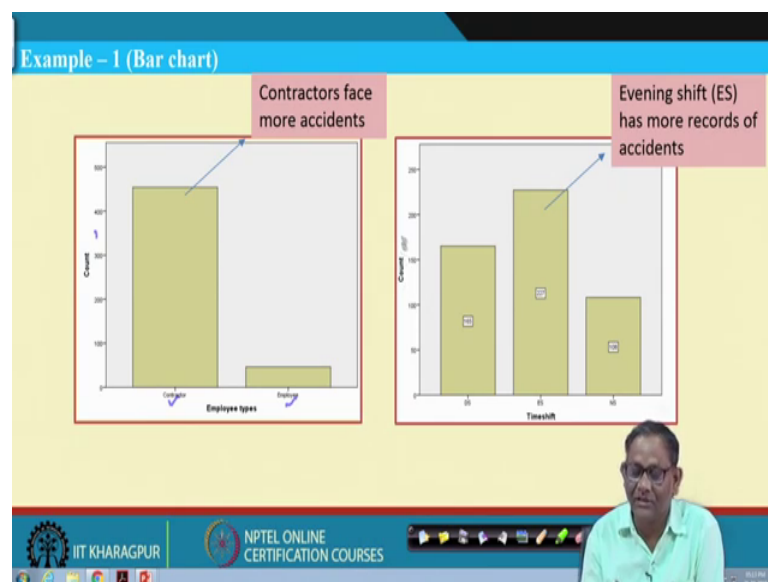
So, it is this side 4.36. So 37.43 if you go this side 4.36, this is 33. So, this is the range variation over the mean, but if you see health rating its minimum value is 1 health rating if I go minimum value 1 maximum value 10 and it is mean value is almost middle point 5.62 and it variants for this 2.6. So, this say 2.6 this say 2.6 this same (Refer Time: 09:41).

So, the mean will give you the central tendency and variants (Refer Time: 09:45) the standard deviation The dispersion what is happening and this is possible, which you will get when you have the data in the numerical scale or numerics scale. When you have data in the categorical scale you will get the frequency values and some probability values or the percentage occurrences you will get and accordingly you can take the decisions. For example, what will be the decision here? Decision for the month wise you may say that why November mean value percentage is more you want to look into this, similarly Sunday the value is more or division 3 per frequency is more.

So, immediate attention will be why, what happened to division 3, what happened is happening on Sunday and what is happened in November? That may be the quest for from the safety management point of view and working temperature if you see that mean value is 37.64 you may it may be ok and health rating point of view suppose it is basically 5.62, but the rating should be when it is again in the middle value, but it is not good because it should be at the higher level. So, that since you have to understand.

So, that mean now what happened that mean the people I can say the health rating is not good, so, their health to be improved, may be health ultimately lead to safety problems ok. So, this is what is the logic of doing this kind of descriptive analysis.

(Refer Slide Time: 11:25)

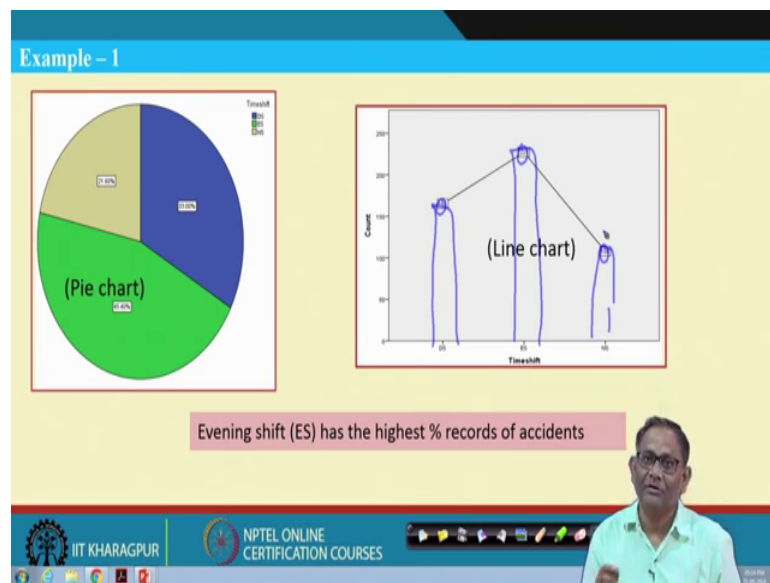


Now, there are many charts that can be used for example, when you are doing the which kind of employee is are more prone to accident Then you can use simple bar chart and then bar chart you find out that contractor are more than the employees. Now if you see the percentage also if the contractor workers and company employees they are having equal in number then this count is easily contractors are more prone to accident and it may be because contractors are there are many contractors coming from diverse background, then they are may not be having proper training for those workers they may the contractor workers might be changing time to time.

So, they are familiarity about the working condition is also not that good. Similarly if you do bar chart for the shift wise analysis you see that evening shifted the accident is

more. So, it may be you have to see that why evening shift accident is more, you do further analysis on the evening shift and then find out the reasons behind it, but never the less we I want to showing the use of bar chart here ok. So, bar chart what happened? In the X axis there will the categories and Y axis there will be the frequency or count. So, here employee type two categories contractor and employee they said count and here in case of that shift there are three shift and count ok.

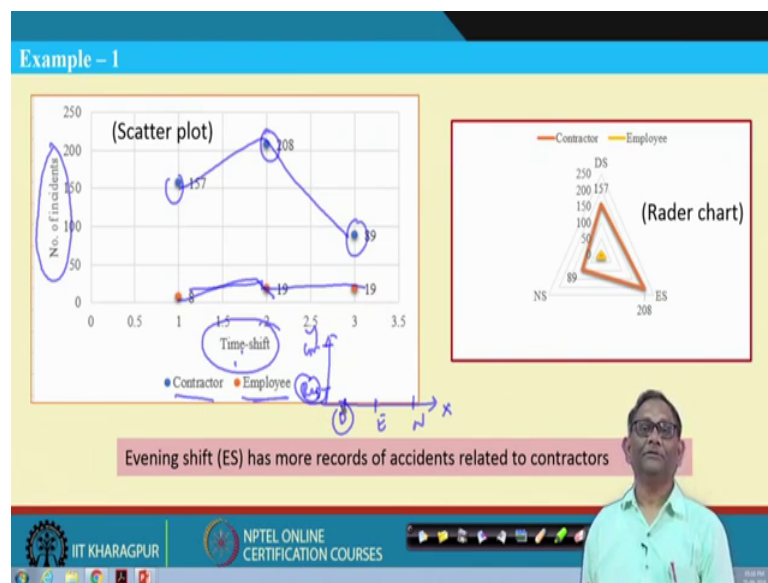
(Refer Slide Time: 13:05)



So, you can use Pie chart. Pie chart in last class I have shown you pie chart we will give you basically that when you capture the totality it will give you the area of the chart talks about the percentage of contribution for that particular category. For example, time shift day, evening and night here you have seen that the your major contribution is evening shift, so that is why the area here is more share is more and that same type if you want to do in line chart, line chart here is similar to that bar chart here, but except putting bar like this you are what you are doing? You are putting one point here another point here another point in the drawing and joining them.

So, that is line chart giving you the similar analogy or similar interpretation of the accident or incident occurrences for the hypothetical plant studied.

(Refer Slide Time: 14:13)

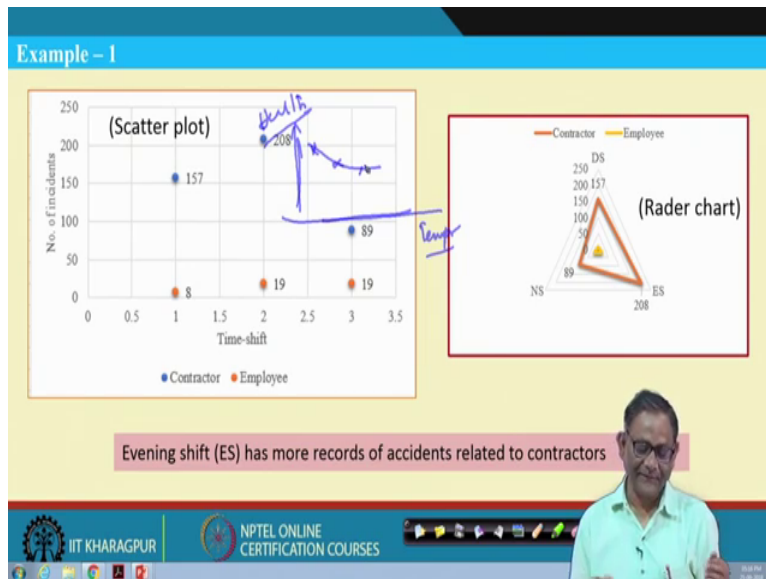


Then there is Scatter plots. Scatter plot when you do scatter plot is nothing but basically you have to consider two variables at a time or two attributes, in this side you are saying that number of incident or incident count and this side that shift. So and here what a time shift and then here Contractor and Employee two kinds of that them workers are considered.

So, then the contractor and contractor everywhere they are more it is more ok. So, although we have basically shown you that time versus time versus the type of employee with the number of incident, but scatter plot usually the general meaning of scatter plot is there will be two variable this side x and this side y now what is the relation between X and Y ok.

So; that means, what will happen if we if we do suppose here that your shift is your day shift, evening shift and night shift and this side your regular and contractor then if you if you put the value so, day shift day shift and day shift and regular.

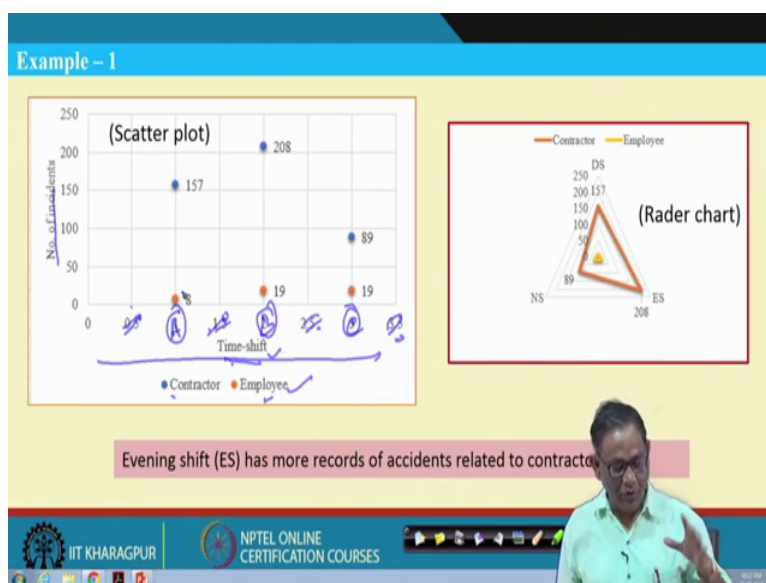
(Refer Slide Time: 15:46)



So, and this kind of this some kind of plot you will get, but better will be when we have suppose this side this side let it be temperature and this side let it be your health rating health rating for the worker.

So, if temperature increasing health rating decreasing or something like this it may so happen you will find out something like this curve, this kind of things are described in scatter plots, scatter plot means; it is a bivariate one, two variable case and this is a special type of scatter plot what we have done we have considered to employee type, as well as time and also we have considered the counts ok.

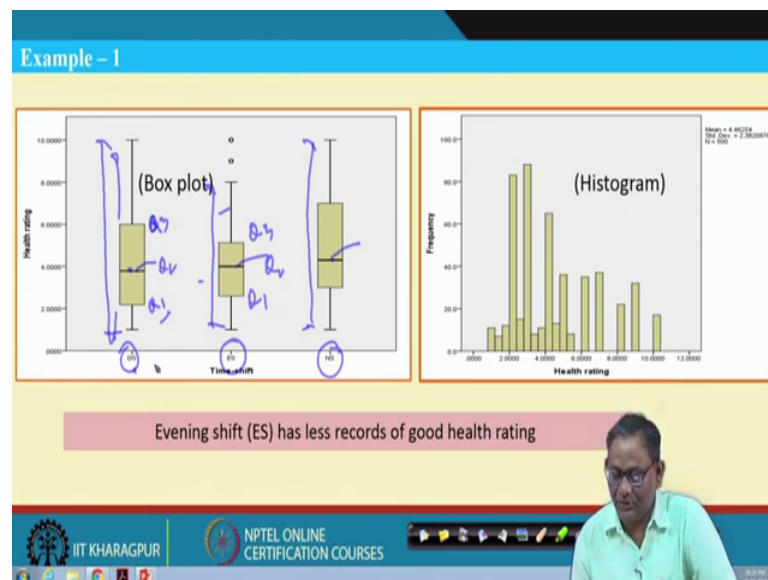
(Refer Slide Time: 16:22)



So, if I consider number of counts versus time shift only then what happens shift means basically that shift 1 and shift 2 and shift 3. In fact, time in between means middle of the shift here middle of the shift here middle of the shift here middle of the shift here. So, that 18 here, but usually we will not do in this fashion, we will write here suppose shift A or shift day shift B or shift C this middle shift if you want to write shift then A B C otherwise you write time 24 hours of time from 0 to 8, 8 to 16, 16 to 24 this side then contractor and regular employee like this ok.

So, these are the different ways of plotting the data, but it is clear from this data that, regular employee regular employee they are having less number of accident or incident compared to contractor employee and if you compare shift wise then definitely winning shift is having more number of accident compared to other two shifts. And then that kind of things you can plot Rader chart, what happened here that contractor employee type and then night shift, evening shift, day shift some Rader is formed and just to show you that you know what is happening in terms of contractor and employee with reference to the 3 different shifts this is another kind of plot.

(Refer Slide Time: 18:20)



So, when if you have your suppose a health rating or some kind of continuous variable and you want to see with reference to categorical variable, suppose he had shift day shift sorry here day shift evening shift and night shift. So, then you have you know that how many accident incident has taken place in day shift evening shift and night shift also.

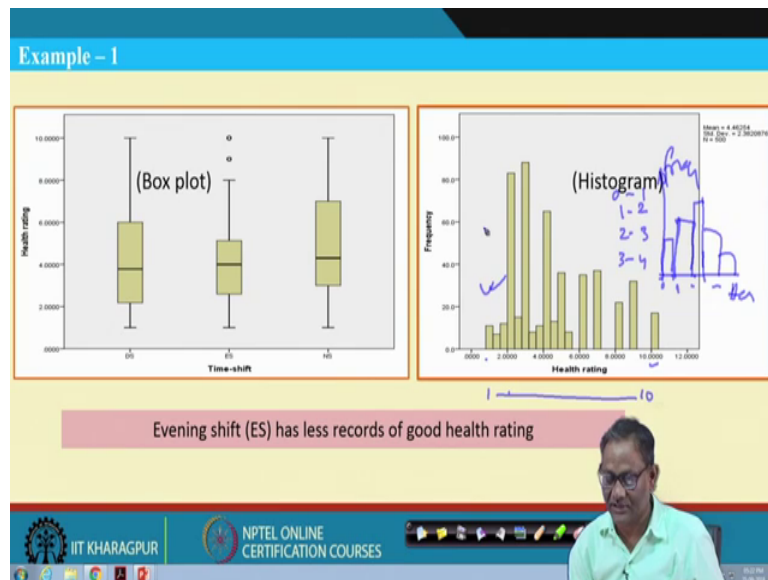
Now you know the victims and their health rating also you know. So, you want to see that what is the variability and central tendency for people in health rating for people who are injured or who are victim in day shift in the evening shift and this shift.

So, then this kind of box plot will give you a beautiful comparison. What is box plot box? Plot basically talks about median value the middle one is a median value and this is your first quartile 1 and this is your third quartile 3. So, median is quartile 2 this is quartile 2. So, quartile 3 and quartile 1 and then this is the whiskers. So, this ultimately this lower than quartile 1 more than quartile 3.

Some whiskers means, what is the spread possible from here to here for these here to here and for these here to here. So, if you see the mean value in terms of health rating then in day shift this median value is lower or almost they are similar, but if you see the spread then day shift workers spread is much more compared to the evening shift people. It may be so, that the people who are victim in the evening shift their health rating is spread is less. So, now doctor if you ask doctor then they can tell you why the this pattern of health rate rating is observed ok.

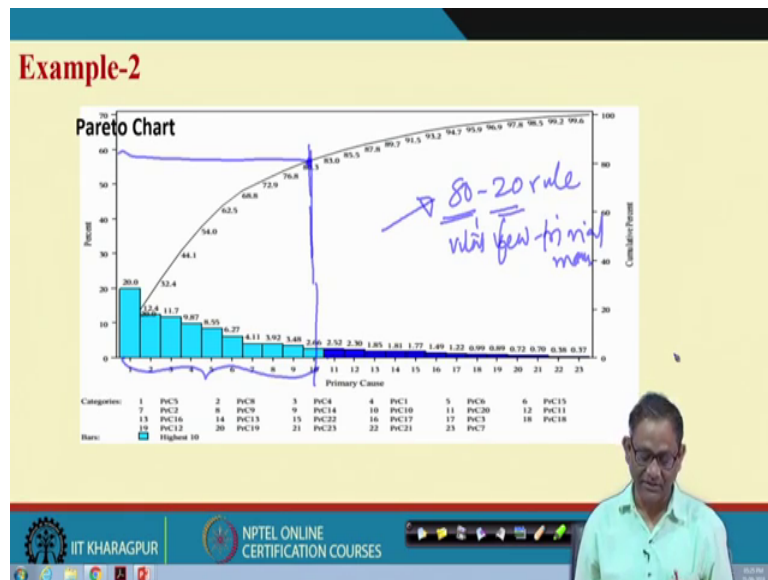
So, doctor will help you. So, another way of representing this is this through Histogram. Histogram when you talk about histogram talk about only one particular variable, which is basically numeric in scale ok, so that mean here health ratings. So, it what it does basically histogram from the from the minimum to maximum with certain range small range number of counts number of frequencies or frequency is counted for example, I can say that health rating it is 1 to 10, so, then up to 10 suppose you want to 2, 1 to 2, 2 to 3 may be 0 to 1 to 2, 2 to 3, 3 to 4 like this.

(Refer Slide Time: 21:10)



So, you create 0 to 1, 1 to 2 and then find out this x is the frequency 0 to 1 suppose this much then 1 to 2 let it be this much then 2 to 3 let it be this much, this much like these. So, this kind of chart you get here this side is frequency and this side is the particular variable numeric variable health rating which is plotted here. What it will give from the shape of the shape of the histogram? You will know many of the important characteristics for example, where the mean lies, where the whether it is rightly skewed or left skewed; that means, which kind of occurrences are more frequent, which kind of occurrence are less frequent and more importantly from histogram you can go to some probability distribution and then you can generalize the population with reference to that particular variable.

(Refer Slide Time: 22:28)



Another important chart which you can use in say that descriptive analytics of the incident data is the Pareto chart. Pareto chart is a chart basically, which was developed in the probably in the 18th century I think 18th or 9th I am not clear, but when you long back it was developed and the Vilfredo Pareto is the Italian economist, he observed that that time that 80 percent of the world wealth was consumed by 20 percent of the world population.


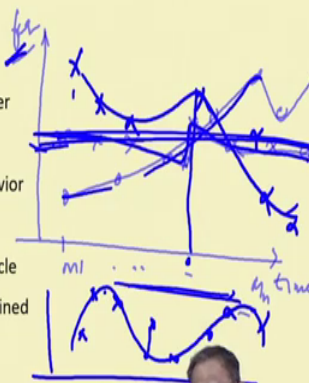
So, that is why it was also known as 80, 20 rule 80, 20 rule because 8 percent of world wealth was consumed by 20 percent of the world population; however, now the scenario is different I think now, it is 98 percent of the world wealth are consumed by 2 percent of the world population similar such kind of statistics are available I am not 100 percent sure, but that is the case, but Vilfredo Pareto, he observed this. So, that mean this one that is why this is also known as vital few and trivial vital few and trivial menu trivial menu. What is vital few? If you see this suppose you want that 80 percent where is 80 percent here is 80 percent ok.

So, 80 percent 80 percent of the 80 percent of this values problems consumed by a here there are many accident causes, but thin causes are contributing to 80 percent of the occurrences ok. So, this access is basically related to percent and this curve this is basically cumulative percent. So, in the cumulative value is 80 here. So, when we come down here. So, that mean 10 different causes that pc these are the causes primary causes

So, that mean, there are vital few causes which are primarily responsible for most of the incidents that is occurring in the plant. So, your objective will be to overcome those causes and reduce the accident or incident occurring ok.

Pattern in Time Series Data

- Average: the mean of the observations over time
- Trend: a gradual increase or decrease in the average over time
- Seasonal influence: predictable short-term cycling behavior
- Cyclical movement: unpredictable long-term cycling behavior due to business cycle or product/service life cycle
- Random errors: remaining variation that cannot be explained by the other four components



NPTEL ONLINE CERTIFICATION COURSES

So, this is this is basically a pattern in the data you may find out the situation is like this ok. So, or you may find out the situation is like this so, many things ok. Now if you want to this from the plant level, suppose you want to see organization level over past 30 years. So, that time you may find year 1 year 2 year 3 like this, you may find something like this also ok. So, what happened this data when you it is measured overtime is known as time series data. It may be frequency of incidents occurring or in the operations management production management it is mostly the demand per month or per year.

So, here frequency incident per month or per unit time. So, then if you may find out that this data will be one is average for example, that the second one here almost there is no change because more or less number of incident every months is same this is average data then trained there is increase or decrease if you see this data it is increase in trend. So, that mean the safety of the system is deteriorating there is seasonal influences.

For example, there can be there can be suppose this is the 6 month seasonality may be like this then again after 6 month it will or weekly seasonality or quarterly seasonality will be there it may be there then, cyclic moment means when you go for long period long time you consider and then you will may find out that some kind of cycle is created means mostly these are all long term cycling behavior which we object in the business cycle.

But for safety data incident data you may observe if you go over the time, but usually it is not in overtime as everything is improving it is it actually usually decreases, but cycling moment is like this cycle this second cycle it is there then otherwise random remaining what happened either all those 3 4 things will be there average trend seasonal influence and cyclical moment these are pattern otherwise data is random, random over average random over trend random over cycle some kind of it is not that always you will be when you plot the data you will plot like this.

So, some this is not falling on the on this smooth curve. So, this is the these are there will be some kind of now, if you if you say this is the cycle then you subtract the cyclic value from every actual observation whatever value live that is that will be a zigzag random type that will be the random year ok. So, time series data is also very important for safety studies ok.

So, when you do a record accident and those accidents or incidents are recorded overtime then either daily monthly weekly quarterly yearly some plot can be made and this will talk about the progress of the performance of the department of the plant of the organization with reference to accidents or incident occurring. So, we have given you some idea of descriptive part descriptive analytics part and I hope that you have understood this.

Thank you very much.